UNIVERSITÄT ZU LÜBECK
INSTITUTE OF MEDICAL INFORMATICS

From the Institute of Medical Informatics
of the University of Lübeck
Director: Prof. Dr. rer. nat. habil. Heinz Handels

Generative Deep Learning Models for the Automatic Analysis and Synthesis of Medical Image Data Featuring Pathological Structures

Dissertation
for Fulfillment of
Requirements
for the Doctoral Degree
of the University of Lübeck

from the Department of Computer Sciences and Technical Engineering

Submitted by
Hristina Uzunova
from Stara Zagora, Bulgaria

Lübeck, 2021

First referee: Prof. Dr. rer. nat. habil. Heinz Handels
Second referee: Prof. Dr.-Ing. Alfred Mertins

Date of oral examination: 02.12.2021

Approved for printing. Lübeck, 09.12.2021

Abstract

Medical image processing methods, especially approaches based on artificial intelligence, have gained popularity in recent years since they enable a facilitation of the daily clinical routine by automatizing time-consuming and error-prone processes. Deep neural networks have shown to be particularly suitable for many medical image processing tasks due to their non-linear and flexible nature which allows them to learn the complex relationships and large variability of medical images. The main challenge related to employing neural networks is, however, the immense amount of annotated images required for their training. Typically, the data annotation process is expensive and time intensive and most commonly can only be fulfilled by experts. With growing variety and complexity of the data, the amount of minimum required training samples also increases, since the entire variability needs to be covered by the training dataset. For this reason, it is exceptionally challenging to train neural networks for images containing pathologies. This is due to the vast variety of pathological structures compared to healthy anatomical tissue. Yet, a large amount of images containing pathological anomalies emerges in the daily clinical routine, thus, their automatic processing is of great relevance for the field.

In this work, different deep learning-based generative models are explored and developed in order to cope with some of the most common problems related to the pathology occurrence in medical images. The development of strategies to reduce the required amount of annotated data for the training of neural networks is one of the main objectives throughout the work. On the one hand, variational autoencoders are utilized for the unsupervised detection of pathologies in medical images, where a rough detection of the pathological structures can be established without using any annotated data for training. On the other hand, a method based on generative adversarial networks is developed in order to synthesize realistic annotated images with pathological structures and use them for the training or evaluation of neural networks. Furthermore, the presented approaches are applied for the improvement of existing registration and segmentation algorithms, emphasizing their potential to optimize the automatic processing of images containing pathological abnormalities.

Overall, the developed deep learning generative models enable the realistic synthesis of images with annotated anatomical and pathological structures and facilitate the automatic analysis of medical image data featuring pathologies.

Zusammenfassung

Automatische Bildanalysealgorithmen, insbesondere Lösungen basierend auf künstliche Intelligenz, sind von immenser Bedeutung für die medizinische Bildverarbeitung und haben das Ziel die klinische Routine deutlich zu erleichtern. Deep-learning-basierte Algorithmen haben sich im Bereich der medizinischen Bildverarbeitung als besonders geeignet erwiesen, da sie durch ihre flexible nicht-lineare Art die große Variabilität medizinischer Daten erlernen können. Die Schwierigkeit solche Algorithmen zu trainieren, besteht allerdings darin, dass eine große Datenmenge mit gegebenen Expertenannotation benötigt wird. Je komplexer und unterschiedlicher die Daten, desto größer sollte der Trainingdatensatz sein, um die gesamte natürliche Variabilität abzudecken. Aus diesem Grund ist die Erstellung solcher annotierter Datensätze für Bilddaten mit vorhandenen pathologischen Strukturen besonders herausfordernd, da die Variabilität der Pathologien verglichen mit normalen anatomischen Strukturen enorm ist. Dies erschwert die Anwendung von deep-learning-basierten Lösungen auf Bilddaten mit Pathologien. Da im medizinischen Alltag jedoch viele Bilder mit pathologischen Anomalien entstehen, ist die Entwicklung von Algorithmen für deren automatische Verarbeitung von großer Relevanz.

In dieser Arbeit wurden deep-learning-basierte generative Modelle eingesetzt und weiterentwickelt, um mehrere Herausforderungen verbunden mit dem Auftreten von Pathologien in medizinischen Bilddaten zu bewältigen. Vorwiegend wurden Strategien entwickelt, um die benötigte Datenmenge für das Training von neuronalen Netzen deutlich zu reduzieren. Einerseits wurde ein variationeller Autoenocoder für die unüberwachte Detektion von Pathologien entwickelt, um die Lokalisation von pathologischen Strukturen ohne den Einsatz von annotieren Daten während des Trainings zu ermöglichen. Auf der anderen Seite wurden Ansätze basierend auf GANs (generative adversarial networks) entwickelt, um realistische künstliche annotierte Bilder mit integrierten pathologischen Strukturen zu generieren und sie für das Training und Testen von neuronalen Netzen zu verwenden. Weiterhin wurden die vorgestellten Ansätze für die Verbesserung der Bildregistrierung und Segmentierung von Bildern mit Pathologien eingesetzt. Somit wurde deren Potenzial für die optimierte automatische Verarbeitung von Bildern mit pathologischen Veränderungen unter Beweis gestellt.

Zusammenfassend ermöglichen die in dieser Arbeit entwickelten intelligenten generativen Modelle die realistische Synthese von annotierten medizinischen Bilddaten und die Verbesserung der automatischen Analyse von medizinischen Bildern mit vorhandenen pathologischen Strukturen.

Contents

Chapter 1

Introduction

1.1 Motivation and Medical Background

Medical image processing methods, especially AI-based approaches, have gained popularity in recent years, since they aim to facilitate the daily clinical routine by automatizing the otherwise time-consuming and error-prone processes. Various image analysis methods have established for the automatic image processing of medical applications, including approaches handling the two most prominent tasks of medical image analysis: semantic segmentation and spatial registration. The semantic segmentation of medical images is the process of delimiting anatomical or pathological structures from the rest of the image by labeling each image pixel as a part of the particular structure or the background. This is useful for e.g. tracking the development of diseases, measuring the size of organs or pathological structures, postoperative progress control, 3D visualization for surgery planning and many more. The pixel-wise labeling of structures in this manner is very time consuming, especially when it comes to 3D medical data, and typically medical experts with knowledge of the specific organs or diseases are required for fulfilling the task. Thus, automatizing this process would significantly improve the clinical workflow. Another common medical image processing area is the spatial registration of images, meaning that the corresponding regions of two images of the same scene are aligned to each other. Registration is particularly important for the comparison of follow-up images, adjustment of images acquired with different imaging modalities or to enable atlas-based segmentation. Of course, next to those examples, many more image analysis methods are relevant for the clinical routine.

One of the major challenges for automatic medical image analysis methods is the vast variety of medical images. Starting from the different imaging devices, acquisition techniques and the choice of body area, all the way to the uniqueness of the anatomy and disease appearance of each individual patient and the intra-patient variations appearing over time, the range of possible image combinations and appearances is enormous. Through the rapid development of deep learning approaches in the last years, this variability becomes more and more manageable due to the extremely flexible and non-linear nature of deep learning approaches, that typically learn the data variability from a given labeled dataset [Lu et al., 2019, Henry et al., 2021].

However, the processing of medical images containing pathological structures remains challenging even when considering the most recent developments. One of the reasons for this lies in the extremely diffuse nature of pathological structures, thus, in order to capture their variability, neural network-based approaches would require a large amount of annotated images. Next to the fact that such annotations are rarely available and hard to obtain, neural networks can only be trained for the processing of a very specific pathological structure, e.g. the segmentation of glioblastoma multiforme in multi-modal brain MRIs [Kamnitsas et al., 2016]. This specific tumor type is, as its name already suggests, a representative example for the huge variability in which pathological structures manifest. Moreover, pathological structures are not only different in size or shape, their variability is also expressed in their differing textures and inner structures which are far less homogeneous than the tissue of the normal anatomy. Hence, for each specific pathology type a new representative and fully-annotated dataset is required for training, which given the wide variability of possible pathology appearances, is not feasible. For this reason, approaches that generalize for a multitude of pathologies and methods that do not require annotated data for training, i. e. weakly supervised or unsupervised methods, are of high interest. Such approaches are for example [Schlegl et al., 2019], where the segmentation of pathological structures in retinal images is established by modeling the natural variability of healthy images. Another example is [Baur et al., 2020], where a multi-scale autoencoder-based approach is used for the unsupervised detection of brain MRI anomalies.

Furthermore, the modeling of pathological structures is not only important for their direct detection and segmentation. Pathological structures often change the appearance of an image significantly, impairing algorithms that target normal anatomical structures. Fig. 1.1 visualizes some of the problems that might be caused by the presence of pathological structures. For example, pathological data are specifically challenging for image registration algorithms since pathologies cause missing correspondences, and thus, no reliable image alignment can be established. Often, prior knowledge of the pathological structure and its position is necessary in order to be integrated into the registration method, like in [Chitphakdithai and Duncan, 2010], where the authors consider the masks of brain tumors during registration to prevent the missing correspondences from influencing the registration results.

Another issue related to pathological structures, is the occlusion of anatomical regions. E.g. a large tumor or a lesion in the brain might overlay the ventricles, thus an algorithm designed for the segmentation of brain ventricles would most likely fail, since the shape of the ventricles strongly deviates from the learned variability. Still, the segmentation of brain ventricles is a crucial step for assessing medical conditions like Alzheimer disease [Karaca et al., 2020] or ischemic strokes [Qian et al., 2017]. Not only do pathological structures overlay healthy tissue, they also often deform the surrounding anatomical structures that are not directly affected by the pathology. For example, pathological fluids in the retina displace the retinal layers significantly, thus,

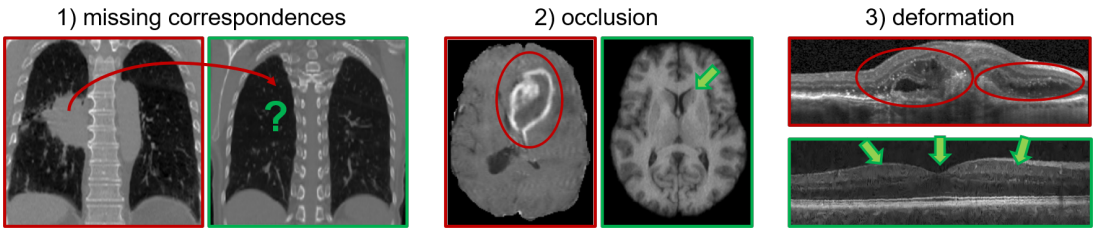


Fig. 1.1: Visualization of some problems occurring due to pathological structures featured in medical images. Red border indicates the presence of visible pathologies and green border indicates normal appearance. The shown problems are: 1) Missing correspondences from pathological and healthy images, here a large lung tumor visible in a thorax CT does not have any correspondence to the structures of a normal lung. 2) Pathological structures occluding healthy structures, here the ventricles visible in healthy individual's brain MRI are overlayed by tumor tissue. 3) Normal structures are strongly deformed by pathologies, here retinal fluids deform the retinal layers visible in an OCT image. Image sources: lung CTs: [Castillo et al., 2013]; brain MRIs: [Menze et al., 2015, Shattuck et al., 2008]; retinal OCTs: [Bogunović et al., 2019, Farsiu et al., 2014].

their segmentation with an algorithm trained on healthy subjects is highly infeasible (Fig. 1.1). Another example is the so-called mass effect where tumors gradually push away and displace the surrounding tissue. This effect has deep bio-physiological foundations and is typically modeled by algorithms considering complex dependencies like finite element methods [Mohamed and Davatzikos, 2005].

To cope with such problems by using deep learning, the training of neural networks on data containing pathological structures with given annotations of the anatomical regions is crucial. However, such data are rarely available in the medical image field. Typically, openly available datasets of images containing pathological structures only feature the annotation of the pathologies [Menze et al., 2015, Maier et al., 2017], while datasets of healthy subjects typically contain the labels of anatomical regions [Hammers et al., 2003, Shattuck et al., 2008]. The lack of anatomical labels also impairs the quantitative evaluation of standard algorithms applied on pathological images. For example, if a brain ventricle segmentation network trained on healthy subjects is applied on images containing brain tumors, it would be crucial to estimate the network's performance on the pathological data. This is important since the images acquired in clinical practice are often different from the used training dataset, and thus, the algorithm's accuracy for the different image types should be known in order to help the clinician interpret the results. A commonly proposed approach to cope with the lack of annotated data is data augmentation and specifically, data augmentation by generating realistic images for the training of neural networks. In [Uzunova et al., 2017], a model-based approach is used for the generation of annotated brain and heart MRIs for the training of a registration network on the synthetic data yielding significantly

improved results compared to using only a few available real images. The authors of [Frid-Adar et al., 2018] propose a neural network-based method for the generation of pathological liver images, that strongly improves the training of a liver lesion classification network. Similarly to that, [Shin et al., 2018] propose a sophisticated approach for the generation of brain tumor MRIs and successfully train a tumor segmentation network on the generated data. These methods do significantly improve the data situation connected to training algorithms targeting pathological structures, however, they do not feature the generation of annotations for normal anatomical structures. Furthermore, the existing methods do not explicitly consider occlusions or pathology-induced deformations of the surrounding healthy tissue. Thus, many of the mentioned problems connected to the presence of pathologies in medical images remain unsolved.

1.2 Objectives of the Work

This work aims to cope with several issues related to the occurrence of pathologies in medical images and the problems connected to the automatic processing of such image data.

Due to the enormous amount of required annotated data for the training of neural networks, their application for pathological images is not feasible in any case where the variability of pathologies is particularly large or the pathology types are rare, thus, a large dataset cannot be collected and their annotation requires special knowledge typically only provided by clinical experts. In this work, strategies for reducing the needed dataset size are considered. For this aim, two different concepts will be pursued. First, unsupervised methods not requiring any annotated training data for the detection of pathologies will be developed and investigated. Second, an approach for the direct generation of realistic pathological images with ground truth annotations will be designed in order to enable a training of neural networks on synthetic data and significantly reduce the number of needed real annotated images.

Algorithms targeting normal anatomical structures in pathological images are a further focus of this work since they might encounter difficulties due to effects like pathology-induced tissue deformations and missing correspondences. Here, these problems should be addressed in a multitude of ways. On the one hand, explicit knowledge in form of segmentation masks achieved by unsupervised segmentation of pathological structures can be integrated into processes like e.g. registration and, thus, prevent bad registration accuracy due to missing correspondences. On the other hand, synthetically generated images with injected pathological structures should be used to improve the training of neural networks, e.g. aiming facilitated image segmentation. Furthermore, pathology-induced tissue deformations should be explicitly modeled and integrated into the synthetically generated training images to help cope with the tissue distortion of the surrounding anatomy.

The common lack of anatomical annotations in pathological images, however, does not only impair the training of networks, it moreover impedes the quantitative evaluation of algorithms targeting the normal anatomy of images containing pathologies. In order to cope with this problem, a generation of images containing ground truth annotations of the anatomical and pathological regions is intended, where the synthetic pathological images are of such realistic appearance that a reliable estimation of the algorithm's accuracy for real images can be established.

In summary, the following objectives of the work can be derived from the above problem definitions:

- **Reduce the need of annotated images containing pathological structures for the training of neural networks.**
- **Facilitate automatic image processing algorithms targeting normal structures in images featuring pathological structures.**
- **Enable quantitative evaluation of neural networks on images featuring pathologies.**

The methodological focus of the work for the achievement of the above-mentioned goals lies in the utilization of diverse deep learning-based generative models. Generative models aim to describe the data distribution of a given training set, e.g. using a probabilistic model, so sampling points from this model would result in new realistic observations that do not exist in the training dataset. Since this work focuses on medical images, a generative model can, for example, be trained on a dataset containing brain MRIs from many different patients. Sampling from the trained model should then result in brain MRIs that look realistic but are not available in the used training set. Unlike the wide-spread discriminative deep learning models, that typically predict a certain label from a given observation, generative models are able to predict the observation itself (possibly considering a given label). Furthermore, discriminative models are most commonly deterministic, while generative approaches are probabilistic, hence, they require a given stochastic element that affects the generation of diverse samples. In this work, next to the ability to generate realistic new samples, the representation ability of deep generative models is considered and their ability to reconstruct real samples unseen during training is investigated.

This work also provides solution approaches to major practical challenges connected to the development of such methods. Generally, the properties of the different deep learning generative models are investigated and quantified. There is a wide variety of generative models available, thus, a systematic and quantitative comparison of methods and an investigation of their concrete advantages and drawbacks is required for their advanced development and utilization. Furthermore, many deep learning generative

models are known for their lack of training stability and overall hard training, so such issues are tackled in this work.

A further major problem, associated with neural networks in general and amplified by generative models in particular is the fact that they require an enormous amount of computational resources and are, thus, mostly intractable for large 3D medical image volumes. However, the plausibility of such models for 3D images is a key for their utilization in medical imaging. Thus, developing techniques to reduce the required computational resources for large volumes including patch-based and multi-scale approaches is a further objective of the work.

Finally, a quantitative assessment is crucial for the interpretation of the achievements presented here. A quantitative evaluation of the results of generative models is not as trivial as the evaluation of e.g. segmentation results. Especially, assessing the realism and plausibility of generated images is not standardized, and thus, suitable evaluation techniques are required in order to enable a reliable quantitative evaluation.

1.3 Organization and Contributions

To simplify the organization of this work, each of the methodical chapters 2, 3, 4 and 5 begin with a short summary and description of the own contributions to the chapter (grey-colored box). After the explanation of the basics of the work in Chapter 2, three methodological chapters follow, their structure is visualized in Fig. 1.2.

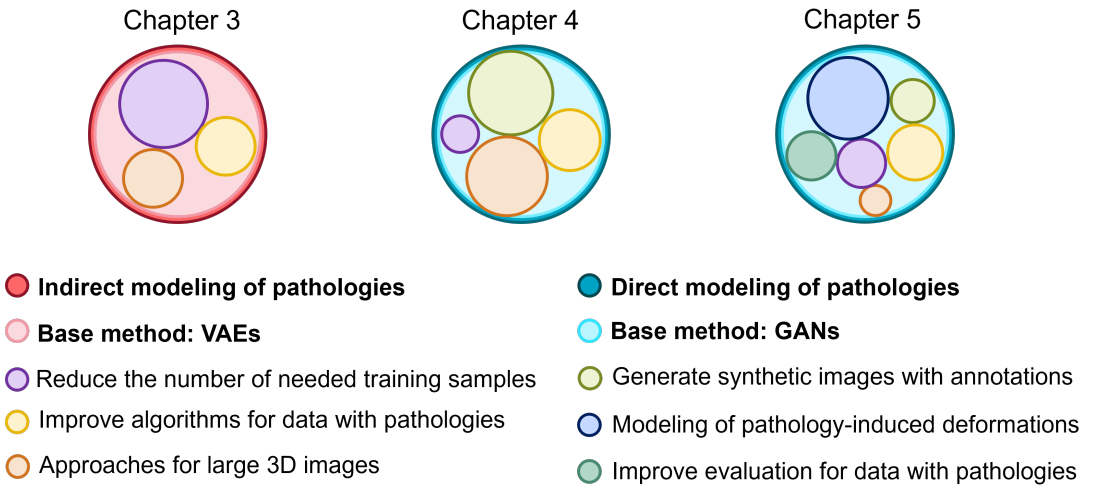


Fig. 1.2: Schematic presentation of color-coded contents of the methodological chapters 3-5. The size of the circles represents the subjective perception about the relative contributions of the individual topics to the chapter.

More specifically, this work is organized as follows:

- In the first part of Chapter 2, the methodological foundations of some prominent deep learning generative models are presented. Furthermore, a systematic comparison of those methods points out their advantages and disadvantages and justifies the choice of the methods observed during the work. The basics for this comparison were laid in the works [Uzunova et al., 2021b, Uzunova et al., 2020c].

- Chapter 3 utilizes variational autoencoders (VAEs) for the unsupervised modeling of pathological structures, based on the main assumption that pathologies can be presented as deviations from a learned healthy norm, thus, pathologies are modeled in an indirect manner. The VAE-based unsupervised pathology detection is first published in [Uzunova et al., 2018]. An extension of this work, applicable to 3D images and considering the explicit pathology knowledge integration into image registration algorithms was later published in [Uzunova et al., 2019d]. Furthermore, this method was utilized for the explanation of black-box pathology-classification approaches, implicitly yielding an unsupervised pathology localization in [Uzunova et al., 2019b].

- A different generative model is chosen for Chapter 4, where pathological structures are explicitly generated using generative adversarial networks (GANs). The main emphasis lies in the design of an approach that is able to generate fully-annotated pathological images, published in [Uzunova et al., 2019c] and the development of an architecture and a training procedure for GANs such that the generation of full-resolution medical 3D volumes is enabled. The main findings of the developed approaches were described in the works [Uzunova et al., 2019a, Uzunova et al., 2020b].

- Chapter 5 adds to the realistic appearance of the images generated by the previous approach by modeling the pathology-induced tissue deformations of the surrounding healthy tissue. The generated images are used for the training and evaluation of neural network and lead to significant improvements. The findings of this chapter were mainly published in [Uzunova et al., 2020a].

- In the last Chapter 6, a summary of the described approaches and results is presented and final main conclusions about the findings of the work are drawn.

Chapter 2

Deep Generative Models for Medical Image Processing

In this chapter, the foundations of three deep learning generative models are presented: autoencoders, variational autoencoders and generative adversarial networks. These approaches are used throughout the work and are the basis of the presented developed methodologies. In the second part of the chapter, a systematic comparison of the methods among each other and to conventional statistical generative models highlights their main properties and gives an insight into the advantages and disadvantages of the different methods.

2.1 Introduction and Motivation

Generative models aim to capture the underlying data distribution of a given training dataset, such that sampling from the trained model would result in generating new realistic samples not contained in the training set, while simultaneously a reliable representation of the training data is learned. In medical imaging, generative models are most commonly applied to learn the shape and appearance variations of anatomical or pathological structures observed in a given population of subjects. Such models are characterized by two main properties – reconstruct samples unseen during training and generate new realistic samples – that open up many possibilities in the field of medical image processing. For example, generative models can provide prior information about plausible shape and appearance of certain anatomical structures facilitating standard image processing tasks like segmentation and registration [Kirschner et al., 2011, Hu et al., 2015]. Furthermore, in an era where medical image analysis is primarily based on deep learning, generative models can be utilized for the generation of realistic training data and data augmentation for neural networks [Uzunova et al., 2017, Karimi et al., 2018]. Also, deep learning-based generative models provide many new possibilities like representation learning [Fleitmann et al., 2021, Chen et al., 2017] or unsupervised learning [Schlegl et al., 2019, Schlegl et al., 2017].

Classical examples of generative models are statistical shape models (SSMs) that use principal component analysis (PCA) on a training set consisting of point-wise shape representations to describe the underlying shape variability in a lower dimension [Cootes et al., 1995]. In addition to shape modeling, SSMs can be extended by intensity value information resulting into simultaneous statistical shape and appearance models (SSAMs) [Cootes et al., 1998]. This step improves the representation of anatomical structures and enables the generation of realistic images for e.g. data augmentation [Uzunova et al., 2017]. Even though such statistical models have had great success in the past, especially in the field of medical image segmentation [Heimann and Meinzer, 2009], they come with some major disadvantages. Due to their linear nature, SS(A)Ms are not particularly flexible and cannot capture complex non-linear relations. Furthermore, they require explicit shape representations for training, e.g. corresponding landmarks, thus, a time-consuming and error-prone data preprocessing is needed. Some of those issues have been addressed by previous works considering specific extensions of the main methodology. For example, [Krüger et al., 2015] propose a probabilistic approach to avoid using corresponding landmarks. Further extensions aim to increase the general flexibility of the models and, thus, increase their representative abilities [Wilms et al., 2017, Kirschner et al., 2011].

With the research focus shifting towards deep learning approaches, many of the above mentioned shortcomings can be addressed by deep learning generative models. The most prominent examples for such methods are autoencoders [Hinton and Salakhutdinov, 2006], variational autoencoders [Kingma and Welling, 2014] and generative adversarial networks [Goodfellow et al., 2014] that share some similarities with classical generative models like dimensionality reduction, data generation and data representation. However, generative deep learning approaches do not require explicit shape information, thus, tedious preprocessing and extraction of corresponding points can be omitted. Such approaches are indeed able to learn the shape and appearance of the training images simultaneously. Furthermore, deep learning neural networks are well-known for their ability to capture complex data due to their flexible non-linear nature. Yet, a major drawback of deep learning generative models is the fact that they require larger training datasets and are often perceived as black-box functions that are hard to interpret.

Due to the large multitude of classical and deep learning generative models with various extensions, this chapter presents a systematic comparison of their specific advantages and drawbacks in a scenario based on brain MRIs. First, the foundations of the most prominent deep learning generative models and their extensions relevant for the further course of this work are presented, followed by a comparison of the models to classical SSMs and their locality-based extension. The investigated properties of the comparison feature the ability to generate new realistic samples and reconstruct unseen samples, compactness of the models, interpretability and training set size. Fi-

nally, the findings of the comparison are used to substantiate the choice of models and their specific extensions used throughout the work.

2.2 Autoencoders

Autoencoders (AEs) are neural networks that typically aim to learn a low-dimensional representation from high-dimensional input data [Hinton and Salakhutdinov, 2006]. They consist of an encoder $f_\phi : \mathbb{R}^d \to \mathcal{Z}, \mathcal{Z} \subseteq \mathbb{R}^r$ that maps the input $\mathbf{x} \in \mathbb{R}^d$ to a latent space variable $\mathbf{z} \in \mathcal{Z}$; and a decoder $g_\theta : \mathcal{Z} \to \mathbb{R}^d$ that aims to reconstruct the input image as flawlessly as possible by only considering its latent representation. To ensure a good reconstruction of the input image, usually a loss $\mathcal{L}_{rec}$ between the real input image and its reconstruction is calculated. Thus, the network's objective is to minimize the following loss:

$$\mathcal{L}_{AE}(\phi, \theta) = \sum_{i}^{N} \mathcal{L}_{rec}\Big(\mathbf{x}_i, g_\theta\Big(f_\phi(\mathbf{x}_i)\Big)\Big) \tag{2.1}$$

for the given training samples $\{\mathbf{x}_i\}_{i=1}^N$. Most commonly, $\mathcal{L}_{rec}$ is set to an L1 or an L2 loss, however, depending on the data structure, more sophisticated loss functions like the structural similarity index measure (SSIM) for intensity images [Zhou Wang et al., 2004] or generalized Dice loss for labels [Sudre et al., 2017] can be used. During training, the parameters ϕ and θ are jointly optimized using e.g. stochastic gradient descent techniques.

Using the simple basis of reconstructing an image by propagating it through a bottleneck, autoencoders open many perspectives in medical image analysis. A significant property of autoencoders is that the latent vector of an input image contains all the important information to reconstruct the images. So, the latent representation can be observed as a very descriptive low-dimensional feature. Thus, AEs are commonly used for feature extraction [Chen et al., 2017]. Furthermore, due to the limited size of the latent representation, the reconstruction of an input image only features the most important structures, while e.g. noise or anomalies are eliminated [Vincent et al., 2008]. Those applications emphasize that the size of the latent space is crucial for autoencoders. A too large latent space might enable a very good image reconstruction, however, it would contain irrelevant information or noise. Too small latent vectors may not allow for a sufficiently accurate reconstruction of the input, since they do not capture all of the important image information.

Overall, a trained AE can be observed as a generative model, where an unseen sample $\mathbf{x}_u$ can be reconstructed as $\mathbf{x}_u \approx g_\theta(f_\phi(\mathbf{x}_u))$ and new samples can be generated by sampling a random latent vector $\mathbf{z}_n$ and computing $\hat{\mathbf{x}}_n = g_\theta(\mathbf{z}_n)$.

However, using AEs for the generation of new samples in this straightforward manner might be infeasible due to the unknown underlying distribution of the latent space.

Thus, a random sample can lie far away from the learned distribution and produce unrealistic samples.

2.3 Variational Autoencoders

Variational autoencoders (VAEs) are often viewed as an extension of conventional autoencoders that prevent two main problems [Kingma and Welling, 2014]. Firstly, conventional autoencoders could simply learn an identity function (e.g. subsequent down- and upsampling of the input image). Secondly, drawing random samples from the latent space is impaired due to the unknown underlying distribution.

VAEs cope with these problems by proposing to directly describe the observations in a probabilistic manner. More formally, VAEs aim to approximate the unknown data distribution given the latent representation $\mathbf{z}$. Thus, the objective is:

$$\max p_\theta(\mathbf{x}) = \max_\theta \int p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z}, \tag{2.2}$$

where θ are the network's parameters, $p_\theta(\mathbf{x}|\mathbf{z})$ is the decoder's probability to reconstruct the training data $\mathbf{x}$ given the latent variable $\mathbf{z}$, while $p(\mathbf{z})$ is the prior distribution of the latent space. A typical choice for $p_\theta(\mathbf{x}|\mathbf{z})$ would be a Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ with a mean $\boldsymbol{\mu}$ and a standard deviation $\boldsymbol{\sigma}$. The probability $p(\mathbf{z})$ is usually set to a multivariate normal distribution $\mathcal{N}(0, \mathbf{I})$. Thus now, Eq. 2.2 needs to be maximized using these assumptions. A standard approach to estimate a distribution like $p_\theta(\mathbf{x})$ would be to randomly sample a large amount of $\mathbf{z}$'s and calculate $p_\theta(\mathbf{x})$. However, in this case this is intractable since most of the $\mathbf{z}$'s would result into $p_\theta(\mathbf{x})$ being zero or very close to zero, thus, an immense amount of samples would be required. For this reason, values of $\mathbf{z}$ that will most probably lead to the generation of $\mathbf{x}$ should be sampled. To accomplish this, a second distribution function $q_\phi(\mathbf{z}|\mathbf{x})$ is presented, that aims towards constraining the distribution of $\mathbf{z}$ to a smaller space than its a-priori assumption $p(\mathbf{z})$. This function is designed as conditional on the given observations and parameterized by ϕ. To measure how well $q_\phi(\mathbf{z}|\mathbf{x})$ approximates $p_\theta(\mathbf{z}|\mathbf{x})$, the Kullback-Leibler (KL) divergence can be calculated:

$$\begin{aligned} \mathcal{D}_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})] &= \sum q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})} \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})} \right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log q_\phi(\mathbf{z}|\mathbf{x}) - \log p_\theta(\mathbf{z}|\mathbf{x}) \right] \end{aligned} \tag{2.3}$$

with $\mathbb{E}$ denoting the expectation value of $\mathbf{z} \sim \mathbf{x}$. Using Bayes theorem the following equation can be formulated:

$$\begin{aligned}\mathcal{D}_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})] &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log q_\phi(\mathbf{z}|\mathbf{x}) - \log \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}\right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log q_\phi(\mathbf{z}|\mathbf{x}) - \log p_\theta(\mathbf{x}|\mathbf{z}) - \log p(\mathbf{z}) + \log p(\mathbf{x})\right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log q_\phi(\mathbf{z}|\mathbf{x}) - \log p_\theta(\mathbf{x}|\mathbf{z}) - \log p(\mathbf{z})\right] + \log p(\mathbf{x}).\end{aligned} \tag{2.4}$$

Since $p(\mathbf{x})$ is independent of q, only the first term of Eq. 2.4 needs to be considered for optimization. The negative of this term is called the evidence lower bound (ELBO) and can be calculated as:

$$\begin{aligned}\mathcal{L}_{VAE}(\phi,\theta) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}) + \log p(\mathbf{z})\right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log q_\phi(\mathbf{z}|\mathbf{x}) + \log p(\mathbf{z})\right] \\ &= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{decoder}} - \underbrace{\mathcal{D}_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))}_{\text{encoder}}.\end{aligned} \tag{2.5}$$

In Eq. 2.5 the main structure of VAEs becomes obvious: an encoder tries to generate a latent representation $\mathbf{z}$ that adheres to a particular prior distribution $p(\mathbf{z})$; and a decoder aims for the generation of the data $\mathbf{x}$ given the particular vector $\mathbf{z}$ (Fig. 2.1).

2.3.1 Implementation and Training

In most cases the log-likelihood from Eq. 2.5 for gray-value images is assumed to correspond to the reconstruction loss of classical autoencoders, e.g. L1 or L2 loss. So the last obstacle is the minimization of $\mathcal{D}_{KL}$. Common choices for the distributions of the encoder are a normally distributed prior $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{z}|0,\mathbf{I})$ and a factorized Gaussian $q_\phi(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x}))$. These choices allow the KL divergence to be calculated in closed form as follows:

$$\mathcal{D}_{KL}\left[\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})||\mathcal{N}(0,\mathbf{I})\right] = \frac{1}{2}\left(\boldsymbol{\mu}^T\boldsymbol{\mu} + \mathrm{tr}(\boldsymbol{\Sigma}) - k - \log|\boldsymbol{\Sigma}|\right), \tag{2.6}$$

where k denotes the dimensionality of the distribution, $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma}^2)$ is a diagonal matrix and $\mathrm{tr}(\cdot)$ is the trace function. Since $\boldsymbol{\Sigma}$ is a diagonal matrix, its determinant can be computed as a product of its diagonal. So in reality, $\boldsymbol{\Sigma}$ can be computed as a vector and Eq. 2.6 can be simplified to:

$$\begin{aligned}\mathcal{D}_{KL}\left[\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})||\mathcal{N}(0,\mathbf{I})\right] &= \frac{1}{2}\left(\sum_k \boldsymbol{\Sigma}_k + \sum_k \boldsymbol{\mu}_k^2 - \sum_k 1 - \log\prod_k \boldsymbol{\Sigma}_k\right) \\ &= \frac{1}{2}\left(\sum_k \boldsymbol{\Sigma}_k + \sum_k \boldsymbol{\mu}_k^2 - \sum_k 1 - \sum_k \log \boldsymbol{\Sigma}_k\right) \\ &= \frac{1}{2}\sum_k (\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k^2 - 1 - \log \boldsymbol{\Sigma}_k).\end{aligned} \tag{2.7}$$

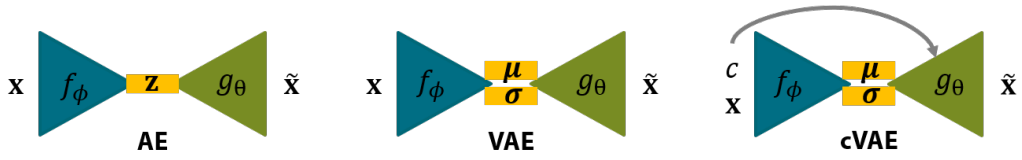


Fig. 2.1: Schematic visualization of the structure of AEs (left), VAEs (middle) and cVAEs (right). Encoder (■); decoder (■); latent space (■).

The functions $\boldsymbol{\mu}_\phi(\mathbf{x})$ and $\boldsymbol{\sigma}^2_\phi(\mathbf{x})$ are generated by the encoder function $f_\phi(\mathbf{x}) = \left(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x})\right)$. In order to increase numerical stability, the output of the network is designed as $\log \boldsymbol{\sigma}^2$ to avoid calculating the logarithm of a negative number. The variance $\boldsymbol{\sigma}^2$ can be computed by simply taking the exponential function. Then a $\mathbf{z}$ is sampled from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ and the decoder function $g_\theta(\mathbf{z}) = \tilde{\mathbf{x}}$ outputs a reconstruction of $\mathbf{x} \approx \tilde{\mathbf{x}}$ (Fig. 2.1).

In order to optimize the network's parameters, the gradient $\nabla_{\theta,\phi}\mathcal{L}_{VAE}(\theta, \phi)$ needs to be calculated. However, calculating the gradients of the sampling function for $\mathbf{z}$ is not possible. Thus, the so-called "reparametrization trick" [Kingma and Welling, 2014] is applied, in order to express $\mathbf{z}$ as a differentiable function:

$$\mathbf{z}(\mathbf{x}) = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \boldsymbol{\omega}, \boldsymbol{\omega} \sim \mathcal{N}(0, \mathbf{I}). \tag{2.8}$$

Here, $\odot$ denotes the element-wise multiplication. Thus, the random sampling operation is implemented outside of the back-propagation graph and does not impede the gradient calculation (Fig. 2.2).

When observing VAEs as generative models, the reconstruction of unseen samples can be done analogously to conventional AEs, nonetheless, the main difference lies in the generation of new samples by sampling a $\mathbf{z}_n \sim \mathcal{N}(0, \mathbf{I})$ and computing $\hat{\mathbf{x}}_n = g_\theta(\mathbf{z}_n)$. Since the latent space is constrained to a normal distribution, the sampling from a known distribution is guaranteed and the generated samples have a higher probability to be realistic.

2.3.2 Conditional VAEs

Conditional VAEs (cVAEs) are an extension of VAEs which also considers a prior given information about the input data in the form of so-called condition [Sohn et al., 2015]. Typically, global image labels are used as conditions, however, many possibilities such as coordinates or pixel-wise annotations are worth consideration. The condition c is usually propagated through the encoder together with $\mathbf{x}$ and is also used as input to the decoder together with $\mathbf{z}$, so, Eq. 2.5 changes to:

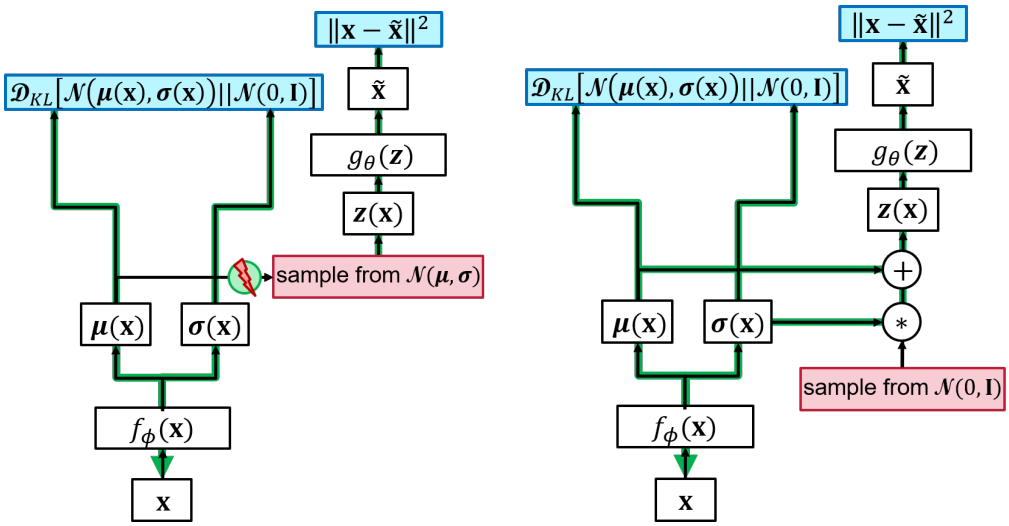


Fig. 2.2: Schematic visualization of the reparametrization trick. Left: no reparametrization trick; right: with reparametrization trick. Loss functions (■); sampling operation (■); black arrow: feed-forward; green arrow: back-propagation flow. Inspired by [Doersch, 2021].

$$\mathcal{L}_{cVAE}(\phi,\theta) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},c)}\left[\log p_\theta(\mathbf{x}|\mathbf{z},c)\right]}_{\text{decoder}} - \underbrace{\mathcal{D}_{KL}\Big(q_\phi(\mathbf{z}|\mathbf{x},c)||p(\mathbf{z})\Big)}_{\text{encoder}}, \tag{2.9}$$

where both the encoder and the decoder are also conditioned on c. This results into the structure shown in Fig. 2.1. The main advantage of cVAEs is that they generate a normally distributed latent space per condition. Meaning that in the inference phase, new samples can be generated given a certain condition, e.g. only images of a given class can be generated. Formally, for a fixed c a new sample $\hat{\mathbf{x}}_{nc}$ can be generated by sampling $\mathbf{z}_n \sim \mathcal{N}(0,\mathbf{I})$ and decoding $\hat{\mathbf{x}}_{nc} = g_\theta(\mathbf{z}_n, c)$.

2.4 Generative Adversarial Networks

Generative adversarial nets (GANs) are known for their exceptional ability to generate realistic images [Goodfellow et al., 2014]. GANs map a random noise vector $\mathbf{z} \sim p(\mathbf{z})$ to an image $\mathbf{x}_f \sim q(\mathbf{x}|\mathbf{z})$ by using a generator function $g_\theta : \mathcal{Z} \to \mathbb{R}^d$. To ensure that the generated images are as realistic as possible, GANs enclose a so-called discriminator $d_\xi : \mathbb{R}^d \to \{0,1\}$ in an adversarial training process, that is trained to distinguish between real image samples $\mathbf{x}_r \sim p(\mathbf{x})$ and the generator's fakes $\mathbf{x}_f \sim q(\mathbf{x}|\mathbf{z})$. In an alternating manner, the generator aims to produce realistic images and, hence, to fool the discriminator that the generated images are real. Subsequently, the discriminator

learns to distinguish between fake and real images better in each iteration, such that the generator needs to improve the synthesized image quality in order to fool the discriminator. This results in a minimax game where g_θ minimizes the log probability for d_ξ to determine the generated image as fake and d_ξ maximizes the log probability to correctly determine whether the samples are real or fake. The following value function of the minimax game can be formulated:

$$\min_\theta \max_\xi V(g_\theta, d_\xi) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log d_\xi(\mathbf{x})\right] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(1 - d_\xi\Big(g_\theta(\mathbf{z})\Big)\right)\right] .scal \quad (2.10)$$

Since practically, the generator's and discriminator's parameters are optimized in an alternating manner, two separate value function can be formulated:

$$\begin{aligned} \max_\xi V(d_\xi) &= \underbrace{\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log d_\xi(\mathbf{x})\right]}_{\text{recognize real images}} + \underbrace{\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(1 - d_\xi\Big(g_\theta(\mathbf{z})\Big)\right)\right]}_{\text{recognize fake images}} \\ \min_\theta V(g_\theta) &= \underbrace{\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(1 - d_\xi\Big(g_\theta(\mathbf{z})\Big)\right)\right]}_{\text{fool the discriminator}} . \end{aligned} \quad (2.11)$$

This minimax game lays the main foundations of GANs: a generator aims to generate images so good that the discriminator is not able to recognize them as fake, while the discriminator is trained to distinguish between real and generated samples. For given real training samples $\mathbf{x}_i$ and randomly sampled vectors $\mathbf{z}_i$ with $i \in [1, N]$ and N denoting the number of training images, the objectives for the training of a GAN can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{GAN(d)}(\xi) &= -\sum_i^N \left(\log\Big(d_\xi(\mathbf{x}_i)\Big) + \log\left(1 - d_\xi\Big(g_\theta(\mathbf{z}_i)\Big)\right)\right) \\ \mathcal{L}_{GAN(g)}(\theta) &= -\sum_i^N \left(\log\left(d_\xi\Big(g_\theta(\mathbf{z}_i)\Big)\right)\right). \end{aligned} \quad (2.12)$$

Typically the network parameters are updated in an alternating manner: in each iteration, ξ is updated first by back-propagating through the discriminator's loss and next θ is updated by back-propagating through the generator's loss (Alg. 1).

GANs have shown to be very suitable for image generation by sampling a random $\mathbf{z}_n$ (most commonly $\mathbf{z}_n \sim \mathcal{N}(0, \mathbf{I})$) and computing $\hat{\mathbf{x}}_n = g_\theta(\mathbf{z}_n)$ (Fig. 2.3). Yet, GANs are not suitable for straightforward reconstruction of input data, since the mapping of an input sample to the latent space is not known. To cope with this problem, the authors of [Schlegl et al., 2019] propose to set up an optimization problem for finding the latent vector $\mathbf{z}_n$ that leads to a best possible reconstruction of the input sample $\mathbf{x}_n \approx g_\theta(\mathbf{z}_n)$. However, this optimization is time-consuming and slows down the network's inference significantly [Uzunova et al., 2019d].

Algorithm 1: Training procedure of GANs

for *number of training iterations* **do**

//update discriminator

- sample minibatch of m noise samples $\mathbf{z}_1,\ldots,\mathbf{z}_m$ from noise prior $p(\mathbf{z})$
- sample minibatch of m real samples $\mathbf{x}_1,\ldots,\mathbf{x}_m$ from the real data distribution $q(\mathbf{x})$
- feed forward: 1) $\hat{\mathbf{x}}_i = g_\theta(\mathbf{z}_i)$ 2) $d_\xi(\hat{\mathbf{x}}_i)$ 3) $d_\xi(\mathbf{x}_i)$
- update the discriminator by ascending the stochastic gradient of its loss:

$$\xi \underbrace{\leftarrow}_{\text{update}} \nabla_\xi \mathcal{L}_{GAN(d)}(\xi)$$

//update generator

- sample minibatch of m noise samples $\mathbf{z}_1,\ldots,\mathbf{z}_m$ from noise prior $p(\mathbf{z})$
- feed forward: 1) $\hat{\mathbf{x}}_i = g_\theta(\mathbf{z}_i)$ 2) $d_\xi(\hat{\mathbf{x}}_i)$
- update the generator by ascending the stochastic gradient of its loss:

$$\theta \underbrace{\leftarrow}_{\text{update}} \nabla_\theta \mathcal{L}_{GAN(g)}(\theta)$$

end

2.4.1 Implementation and Training Stability Issues

Due to the complex minimax game underlying the training process of GANs, their optimization suffers several common problems:

- Vanishing gradient: The training of GANs often suffers from vanishing gradients. One reason for this is the design of the generator's loss function $\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\log(1 - d_\xi(g_\theta(\mathbf{z})))$ since it yields small gradients for small values of $d_\xi(g_\theta(\mathbf{z}))$ and large gradients for large values of $d_\xi(g_\theta(\mathbf{z}))$. This means that when the generator delivers bad results in the beginning of the training process and $d_\xi(g_\theta(\mathbf{z})) \approx 0$, the loss function is rather flat, so it yields small gradients enabling only a minor weight correction. However, when the generator delivers realistic results and $d_\xi(g_\theta(\mathbf{z})) \approx 1$, the loss function has high gradients, even though the weights of g_θ do not need to be strongly adapted. A common solution for this problem is to minimize $\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[-\log(d_\xi(g_\theta(\mathbf{z})))]$ instead, since this function has the desired properties for training.

Using a proper loss function helps facilitate convergence, nonetheless, the gradient vanishing problem remains a rather common issue. The reason is that the discriminator typically has an "easier" task to learn, thus, it can distinguish between real and fake samples fairly well only after a small number of training iterations. Contrary to that, the generator usually takes a large amount of iterations to enable the generation of rudimentarily realistic images, hence, the task to distinguish between real and generated samples is very simple. Once the discriminator has learned to (nearly) perfectly distinguish between real and fake images, only very small gradients are back-propagated to the generator. This issue is usually handled by using additional losses to boost the performance of the generator or training schemes where the generator is primarily trained in the beginning and the discriminator does not get the chance to gain its performance.

- Mode collapse: A typical requirement for a GAN is to produce a large variety of outputs. This is, however, not always given since the generator might only produce a particularly plausible output that the discriminator cannot distinguish from real samples. Thus, this output gets especially preferred by the generator and at some point, only this or a few very similar outputs are produced. This is known as mode collapse. The discriminator should learn to classify this image as fake, but if it is stuck in a local minimum, the generator would continue to optimize against this single mode. This issue can be prevented by applying more sophisticated losses [Arjovsky et al., 2017] or using conditional GANs [Isola et al., 2017] as explained in the further course of the work.

- Non-convergence: Theoretically, a converged GAN would find the so-called Nash equilibrium, where $\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\log(1 - d_\xi(g_\theta(\mathbf{z})))] = 0.5$ and $\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[\log d_\xi(\mathbf{x})] = 0.5$ which means that the discriminator can only randomly guess whether a sample is real or fake. Yet, this is hard to achieve in practice. One of the reasons is the alternating training nature that enables the generator to synthesize the same samples in alternating manner. Assume that the generator is only able to produce the sufficiently different samples a and b in some advanced iteration of the training. First a is generated, the discriminator accepts the generated sample as realistic in the beginning and then learns to identify it as fake in the next few iterations. So the generator is pushed to generate against the sample b, which, since it is very different from a the discriminator assumes to be realistic and then learns to identify as fake. Hence, the generator goes to producing a again, causing a non-convergence of the training process. Furthermore, since the best possible discriminator accuracy is 50% (random guess), towards the end of the training, the discriminator's feedback gets less meaningful. So the generator adapts its weights against a meaningless error function. Most commonly, this issue can be coped with by using standard neural network regularization functions like drop-out or batch norm, applying data augmentation strategies or normalizing the network's weights.

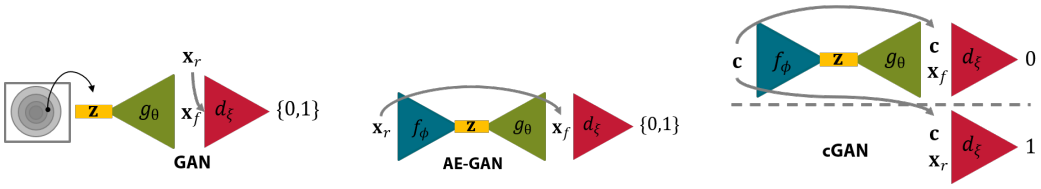


Fig. 2.3: Schematic visualization the architectures of a GAN (left), an AE-GAN (middle) and a cGAN (right). Encoder (■); generator (■); latent space (■); discriminator (■).

2.4.2 Autoencoder GANs

An idea to enable GAN-based image reconstruction are the so-called AE-GANs [Larsen et al., 2016] that combine an autoencoder architecture with an adversarial discriminator. As Fig. 2.1 and Fig. 2.3 show, AEs and GANs share some similarities: both consider a low-dimensional latent space and a generator function (decoder for AEs) that synthesizes images. These properties enable an easy fusion of both architectures. An AE-GAN consists of an encoder $f_\phi : \mathbb{R}^d \to \mathcal{Z}$, a decoder $g_\theta : \mathcal{Z} \to \mathbb{R}^d$ and enclosed adversarial discriminator $d_\xi : \mathbb{R}^d \to \{0, 1\}$. Typically, the reconstruction loss of AEs is combined with the GAN loss, yielding the following objectives:

$$\begin{aligned} \mathcal{L}_{AEGAN(d)}(\xi) &= -\sum_i^N \Big(\log d_\xi(\mathbf{x}_i) + \log \Big(1 - d_\xi\Big(g_\theta(f_\phi(\mathbf{x}_i))\Big)\Big)\Big) \\ \mathcal{L}_{AEGAN(g)}(\phi, \theta) &= -\sum_i^N \log d_\xi\Big(g_\theta(f_\phi(\mathbf{x}_i))\Big) + \sum_i^N \mathcal{L}_{rec}\Big(\mathbf{x}_i, g_\theta\Big(f_\phi(\mathbf{x}_i)\Big)\Big), \end{aligned} \tag{2.13}$$

that get minimized similarly to the alternating optimization of GANs, while the parameters ϕ and θ are jointly optimized. An advantage of using an additional reconstruction loss for the generator is the fact that it stabilizes the training by additionally pushing the generator to synthesize realistic images faster. A scheme of this approach is shown in Fig. 2.3. Of course even further extensions of this idea are possible, e.g. considering an additional KL-loss for the latent space in order to ensure normal distribution. Using AE-GAN enables an easy, fast and reliable reconstruction of unseen samples $\mathbf{x}_u$ by calculating $\mathbf{x}_u \approx g_\theta(f_\phi(\mathbf{x}_u))$.

2.4.3 Conditional GANs

As mentioned above, (V)AEs and GANs share some similarities, so similar techniques can be applied. Like VAEs, GANs can also be conditioned on additional information about the input images. However, conditional GANs (cGANs) are usually not conditioned on a single global label, moreover they are typically conditioned on another image, thus, the condition $\mathbf{c} \in \mathbb{R}^d$ [Isola et al., 2017]. A prominent example for the

usage of cGANs is the input of pixel-wise segmentations and output of images corresponding to the segmentations. To achieve this, cGANs consider pairs of conditions and real images $(\mathbf{c}, \mathbf{x}_r)$ and use a structure where a generator $g_\theta : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$ maps the condition image $\mathbf{c}$ and a noise vector to an image $\mathbf{x}_f$ that is an approximate reconstruction of the corresponding real image $\mathbf{x}_r$. An important difference compared to conventional GANs is, that the discriminator $d_\xi : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$ gets a pair of images as its input: $(\mathbf{c}, \mathbf{x}_f)$ or $(\mathbf{c}, \mathbf{x}_r)$ and learns to distinguish whether its input is a real or a fake image pair (Fig. 2.3). The loss from Eq. 2.12 can be adapted to:

$$\begin{aligned} \mathcal{L}_{cGAN(d)}(\xi) &= -\sum_i^N \Big(\log d_\xi(\mathbf{x}_i, \mathbf{c}_i) + \log \Big(1 - d_\xi\Big(g_\theta(\mathbf{c}_i, \mathbf{z}_i)\Big)\Big)\Big) \\ \mathcal{L}_{cGAN(g)}(\theta) &= -\sum_i^N \log d_\xi\Big(g_\theta(\mathbf{c}_i, \mathbf{z}_i), \mathbf{c}_i\Big) + \sum_i^N \mathcal{L}_{rec}\Big(\mathbf{x}_i, g_\theta(\mathbf{c}_i, \mathbf{z}_i)\Big), \end{aligned} \tag{2.14}$$

where it is also typical to consider a reconstruction loss for the training of the generator. A common modification of this cGAN definition is to omit the random noise vectors $\mathbf{z}_i$ during training and testing, since the generator would most commonly learn to ignore them [Isola et al., 2017]. This yields deterministic results, that might not be required by all applications, so the non-deterministic network behavior can be enforced by additional drop-out layers in the network architecture. Using cGANs typically greatly improves training stability, since the generator is lead into a particular direction and the different conditions enforce variability of the output data. Using conditional GANs as generative models is, however, only possible when a condition is given or can be easily generated for new samples. Using cGANs, unseen samples can only be implicitly reconstructed by using their conditional images $\mathbf{c}_u$ as inputs $\mathbf{x}_u \approx g_\theta(\mathbf{c}_u)$, where new samples can be generated by forwarding a conditional image $\mathbf{c}_n$ through the network $\hat{\mathbf{x}}_n = g_\theta(\mathbf{c}_n)$. Note that, here, the random input vectors are omitted.

2.5 A Comparison of Generative Models

In order to concretely point out the modeling abilities of the presented approaches and to highlight their advantages against conventional statistical approaches, a comparison between different models similarly to [Uzunova et al., 2021b, Uzunova et al., 2020c] is established in this section. Concretely, AEs, VAEs, AE-GANs and two statistical modeling approaches are compared to each other. Here, no conventional GAN models are used, since the reconstruction abilities of the models play an important role. In the experimental setup, firstly, the modeling abilities in terms of reconstructing unseen samples, generating new samples and their realism are considered, and secondly, the latent space properties of the methods regarding their compactness, normal distribution and ambiguity are inspected.

2.5.1 Baseline Methods

Classical generative modeling approaches seek to capture the information needed to describe the distribution of certain anatomical structures in a given population. Typically, statistical generative models consider the shape and appearance variations between samples independently and build separate shape and appearance models which are combined at the end. Contrary to that, deep learning approaches are able to capture appearance and shape changes simultaneously. Most statistical shape modeling approaches are based on PCA [Cootes et al., 1995] and, thus, have the disadvantage to only represent linear dependencies. A further drawback are the required point-based correspondences for training. However, many developments and extensions cope efficiently with those problems [Wilms et al., 2017, Davatzikos et al., 2003, Krüger et al., 2017]. Furthermore, the additional consideration of image intensities can be established by statistical shape and appearance models (SSAMs). In this comparison scenario a classical SSAM and its locality-based extension (LSSAM) are used as baseline methods for the opposed deep learning generative models.

Statistical Shape and Appearance Models: SSMs are built using a training set of n discrete, vectorized shape representations $\mathbf{x}_1 \dots \mathbf{x}_n$ [Cootes et al., 1995]. Here, each $\mathbf{x}_i \in \mathbb{R}^{dm}$ is composed of the d sub-coordinates of all m landmarks representing the object's shape in a d-dimensional space. PCA of the training set is used to create a mean shape $\mathbf{x}_\mu$ and an orthonormal basis $\mathbf{U} \in \mathbb{R}^{dm \times p}$ for projecting shape representations into a low-dimensional latent space $\mathbf{z} \in \mathbb{R}^p$ via $\mathbf{x}_n = \mathbf{x}_\mu + \mathbf{U}\mathbf{z}$ or to generate new shapes by varying $\mathbf{z}$.

SSMs can be extended to SSAMs by considering appearance images $\mathbf{a}_i$ and warping their shape representations $\mathbf{x}_i$ to $\mathbf{x}_\mu$ with the corresponding deformation field φ_i. The resulting "shape-normalized" images can be used in a similar manner for a PCA-based modeling of the intensities sampled on multiple points. After the sampling of the appearance parameters, the resulting images need to be warped back with φ^{-1} [Cootes et al., 1998].

Locality-based Statistical Shape and Appearance Models: In classical SSMs, the number of training samples influences the flexibility of the model, since the size of the latent space is limited by the size of the training set. Usually a small dataset and a large dimensionality of the samples result in a model that mainly captures global variations. However, additional flexibility can be achieved by assuming that local shape variations have a limited effect in distant areas and breaking global relationships. This idea can be integrated in classical SSMs by manipulating the covariance matrix in a way that distant parts of the samples are not co-dependent. This is typically done in a multi-resolution manner considering multiple levels of locality using different distance thresholds to determine which parts are independent. Such models are called locality-

based SSMs (LSSMs) [Wilms et al., 2017], or LSSAMs when additionally to the shape information appearance is also considered analogously to classical SSAMs, and they have shown to perform on par with other extensions of classical SSMs like Gaussian process models [Wilms et al., 2020].

2.5.2 Evaluation Metrics

The necessary metrics for the evaluation of the experiments conducted in the further part of this section are introduced here. The evaluation features two main parts: the quality of the generative models in terms of reconstruction and sampling abilities, and the properties and plausibility of the latent space.

2.5.2.1 Assessing the Quality of Generative Models

Generalization Ability: Generalization ability (GA) is the ability of a generative model to reconstruct samples unseen during training [Davies et al., 2002]. Formally, given a set of real test images $\mathcal{R}_{test}$, GA is measured as $\frac{1}{N_{\mathcal{R}}} \sum_i^{N_{\mathcal{R}}} dist(\mathbf{x}_i, \widetilde{\mathbf{x}}_i)$, where $dist(\cdot, \cdot)$ is a suitable image-wise distance metric and $\widetilde{\mathbf{x}}_i$ is the reconstructed input $\mathbf{x}_i$. The reconstruction of unseen samples is established as previously explained depending on the used approach.

Specificity: The specificity is the ability of a generative model to generate new samples that are similar to the real samples of the training dataset, i.e. realistic samples can be generated. Specificity [Davies et al., 2002] is measured by generating a dataset $\mathcal{G}$ of $N_{\mathcal{G}}$ synthetic images and calculating $\frac{1}{N_{\mathcal{G}}} \sum_i^{N_{\mathcal{G}}} \min\{dist(\mathbf{x}_i, \hat{\mathbf{x}}_j) | \mathbf{x}_i \in \mathcal{R}; \hat{\mathbf{x}}_j \in \mathcal{G}\}$ where $\mathcal{R}$ is a set of real images. Specificity measures the distance of a generated image to its best fitting real image, however the diversity of the generated samples is not considered. This means that a good specificity may indicate that the model only generates one image, that suits a particular real image very well. Thus, the diversity of the generated images is also considered by the following evaluation method.

Likeness: The likeness score proposed by [Guan and Loew, 2020] evaluates the realism of generated images by considering the aspects creativity, inheritance and diversity. The authors propose using a distance-based separability index (DSI). Given the set of real images $\mathcal{R}$ and the set of generated images $\mathcal{G}$ with $N_{\mathcal{R}}$ and $N_{\mathcal{G}}$ images correspondingly, the intra-class distance sets (ICD) are defined as:

$$\begin{aligned} \{d_{\mathcal{R}}\} &= \{dist(\mathbf{x}_i, \mathbf{x}_j) | \mathbf{x}_i, \mathbf{x}_j \in \mathcal{R}; i \neq j\} \\ \{d_{\mathcal{G}}\} &= \{dist(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) | \hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j \in \mathcal{G}; i \neq j\}, \end{aligned} \tag{2.15}$$

where $dist(\cdot,\cdot)$ is a distance function between two images. Similarly, a between-class distance set (BCD) for the distances between both sets can be determined as:

$$\{d_{\mathcal{R},\mathcal{G}}\} = \{dist(\mathbf{x}_i, \hat{\mathbf{x}}_j) | \mathbf{x}_i \in \mathcal{R}; \hat{\mathbf{x}}_j \in \mathcal{G}\}. \tag{2.16}$$

The similarity of the distributions of the ICD and BCD sets can then be examined using a Kolmogorow-Smirnov (KS) statistic

$$\begin{aligned} s_{\mathcal{R}} &= KS(\{d_{\mathcal{R}}\}, \{d_{\mathcal{R},\mathcal{G}}\}) \\ s_{\mathcal{G}} &= KS(\{d_{\mathcal{G}}\}, \{d_{\mathcal{R},\mathcal{G}}\}). \end{aligned} \tag{2.17}$$

The two-sample KS statistic calculates the maximum distance between two cumulative distribution functions (CDFs):

$$KS(p,q) = \sup_x |F_p(x) - F_q(x)|, \tag{2.18}$$

where $F_p(x)$ and $F_q(x)$ are the CDFs of the distributions p and q respectively and $\sup_x$ is the supremum. Given the ordered observations X_i, CDF can be defined as:

$$F(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty,x]}(X_i) \tag{2.19}$$

with $I_{[-\infty,x]}$ being the indicator function:

$$I_{[-\infty,x]} = \begin{cases} 1 \text{ if } X_i < x \\ 0 \text{ otherwise.} \end{cases} \tag{2.20}$$

Although there are multiple statistical measures to determine the similarity of two distributions like Kullback-Leibler divergence or Jensen-Shannon divergence, they typically require equally large input sample sets. Since the sets $\{d_{\mathcal{R}}\}, \{d_{\mathcal{G}}\}, \{d_{\mathcal{R},\mathcal{G}}\}$ have a different number of samples, the KS measure is well suited for this purpose.

To now calculate the DSI, the average of the KS statistics $DSI(\mathcal{R},\mathcal{G}) = 0.5 \cdot (s_{\mathcal{R}} + s_{\mathcal{G}})$ is considered. This metric is further referred to as *likeness* since it describes the similarity between the distributions of the real data and the generated data. The likeness ranges from 0 to 1 where smaller values indicate better synthetic images.

Distance Measurements: To calculate the presented metrics, distances between images need to be measured. Thus, here the used distance metrics are defined between two given images $\mathbf{x}$ and $\mathbf{y}$.

MSE and MAE: The mean squared error (MSE) and mean absolute error (MAE also called L1) are common metrics for the evaluation of the difference between two images in a pixel-/voxel-wise manner as follows:

$$\begin{aligned} MSE(\mathbf{x}, \mathbf{y}) &= \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{x}^{(i)} - \mathbf{y}^{(i)} \right)^2 \\ MAE(\mathbf{x}, \mathbf{y}) &= \frac{1}{N} \sum_{i=1}^{N} \left| \mathbf{x}^{(i)} - \mathbf{y}^{(i)} \right|, \end{aligned} \tag{2.21}$$

where the superscript (i) indicates the spatial position of a pixel and N is the overall number of pixels in the images. The smaller the MSE and MAE values are, the more similar the input images.

SSIM: Contrary to the pixel-wise metrics, the structural similarity index measure (SSIM) [Zhou Wang et al., 2004] takes the structural information of the images into account by considering a spatially close neighborhood region of the input pixels. In this way, the particular intensity of a single pixel is less significant to the metrics output than the structural similarity of the images. So, the SSIM measure is calculated on small corresponding image patches $\mathbf{p} \in \mathbf{x}$ and $\mathbf{q} \in \mathbf{y}$ in a sliding window manner as follows:

$$SSIM_{patch}(\mathbf{p}, \mathbf{q}) = \frac{(2\mu_{\mathbf{p}}\mu_{\mathbf{q}} + C_1)(2\sigma_{\mathbf{pq}} + C_2)}{\left(\mu_{\mathbf{p}}^2 + \mu_{\mathbf{q}}^2 + C_1\right)\left(\sigma_{\mathbf{p}}^2 + \sigma_{\mathbf{q}}^2 + C_2\right)}. \tag{2.22}$$

Here C_1 and C_2 are stabilizer terms in case that either $(\mu_{\mathbf{p}}^2 + \mu_{\mathbf{q}}^2)$ or $(\sigma_{\mathbf{p}}^2 + \sigma_{\mathbf{q}}^2)$ is close to zero. The variables $\mu_{\mathbf{p}}, \mu_{\mathbf{q}}$ denote the mean value, $\sigma_{\mathbf{p}}, \sigma_{\mathbf{q}}$ the standard deviation of the patches $\mathbf{p}$ and $\mathbf{q}$, respectively, and $\sigma_{\mathbf{pq}}$ is the covariance calculated as follows:

$$\sigma_{\mathbf{pq}} = \frac{1}{N-1} \sum_{i}^{N} \left(\mathbf{p}^{(i)} - \mu_{\mathbf{p}} \right) \left(\mathbf{q}^{(i)} - \mu_{\mathbf{q}} \right). \tag{2.23}$$

To achieve a global image metric, the mean over all M patches is calculated:

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{1}{M} \sum_{j=1}^{M} SSIM_{patch}\left(\mathbf{p}_j, \mathbf{q}_j\right), \tag{2.24}$$

where $\mathbf{p}_j \in \mathbf{x}$ and $\mathbf{q}_j \in \mathbf{y}$ are image patches of the corresponding images. Based on the original work [Zhou Wang et al., 2004], the mean and standard deviation values are calculated within an 8×8 pixel window that slides pixel-by-pixel over the entire image. Here, SSIM is also extended to the use on 3D data by considering a 8^3 voxel window. The balancing parameters are set to: $C_1 = 0.01$ and $C_2 = 0.03$. Strictly, by this definition SSIM is not a distance, since a high SSIM value (maximum 1, minimum 0) corresponds to high image similarity. Thus, $(1 - SSIM)$ should be used as a distance measure when calculating generalization ability and specificity.

2.5.2.2 Quantifying the Latent Space

Compactness: The main property of the latent space is its dimension r. In SSAMs the latent space dimension is determined by the number of modes describing the desired percentage of variability (95%) and can be automatically determined. For deep learning models, the process of determining the right latent size is rather empiric. Here, a grid-search strategy ($r \in [64, 2048]$) is used to determine the optimal dimension while minimizing the reconstruction error. The size of latent vectors is crucial for the network's performance, since a good trade-off between noisy representations from too large latent vectors and too blurry reconstructions resulting from too small vectors is required. Overall, a smaller number of latent dimensions contributes to the interpretability of the model.

Normality: An important assumption of all approaches is that new samples can be generated by sampling a vector $\mathbf{z}$ from a normal distribution. This requires a normal distribution of the learned latent space. To assess this property, a set of real samples is encoded and the distribution along each dimension using a Shapiro-Wilk test [Royston, 1982] is examined. In this manner, the percentage of not-normally distributed components of the latent space can be determined. A further visual evaluation of the smooth distribution of the latent space can be established by interpolating between the projections of two random images and visualizing the decodings of the intermediate latent vectors.

Latent Ambiguity Score: A desired property of generative models is, for a given image $\mathbf{x}$ mapped to its latent variable $\mathbf{z}$ and reconstructed to $\tilde{\mathbf{x}}$, the reconstruction $\tilde{\mathbf{x}}$ should also get mapped to the same latent vector. This is given by default for statistical models, however, not mandatorily fulfilled by deep learning approaches. To asses the extent of the problem, the latent ambiguity score (LAS) is proposed. Given a set of real images $\mathcal{R}$ and a trained model with encoder f_ϕ and decoder g_θ, the mean distance

$$\mathcal{D}_{mean} = \frac{1}{N_\mathcal{R}} \sum_{i}^{N_\mathcal{R}} \left\| f_\phi(\mathbf{x}_i) - f_\phi\Big(g_\theta\Big(f_\phi(\mathbf{x}_i)\Big)\Big) \right\|_2, \text{ with } \mathbf{x}_i \in \mathcal{R} \tag{2.25}$$

and the baseline

$$\mathcal{D}_{BL} = \frac{1}{N_\mathcal{R}(N_\mathcal{R}-1)} \sum_{i}^{N_\mathcal{R}} \sum_{j \neq i}^{N_\mathcal{R}} \left\| f_\phi(\mathbf{x}_i) - f_\phi(\mathbf{x}_j) \right\|_2 \tag{2.26}$$

are computed. Since the latent space of each model is distributed in a different range, the normalized LAS is calculated

$$\mathcal{D}_{LAS} = \mathcal{D}_{mean} / \mathcal{D}_{BL}. \tag{2.27}$$

A score close to zero corresponds to an unambiguous latent space, whereas a score close to one indicates that the latent mappings of the real and the reconstructed images are nearly randomly located. Hence, models with high LAS eventually sample the entire latent space by simply subsequently inputting the reconstructions of the previous input rather than sampling the desired latent vectors.

2.5.3 Experiments and Results

Data and Experimental Setup: The used dataset contains 600 3D T1-weighted brain MRIs from the publicly available IXI dataset[1]. A dataset split of 300/290/10 test/train/validation is used. Since deep learning methods are very resource demanding, the images and labels are cut to a central area around the ventricles of size $64 \times 96 \times 64$ voxels.

The intensity image volumes can be directly used by the neural networks, while SSAMs consider shape distortions (given by registration displacement fields) and appearances separately. Each model is trained on varying training set sizes sampled from the 290 training images and the generalization ability, specificity and likeness are assessed in accordance to that. This allows an observation of the quality of the generative models depending on the training set size N, so their robustness can be assessed. To avoid choosing unsuitable training data for smaller training set sizes, a multi-fold training is performed (four folds for $N < 100$, two folds for $N \geq 100$), averaging the results over the folds. The latent space evaluation is carried out for the largest possible training sets for each experiment.

To facilitate comparability, the architectures for the presented neural networks have a simple design. Each encoder (or discriminator) contains three 3D convolutional layers each increasing the number of channels in the order $[1, 20, 40, 80]$ and simultaneously decreasing the image size by half using strides in each layer. Finally, a fully-connected layer is used to obtain the latent vector. The decoders contain a fully-connected layer for the latent space, followed by three subsequent upsampling and convolutional layers that decrease the input channels and increase the image size in an inverse order as the encoders. Since the image intensities are scaled in the range $[-1, 1]$ during training, a tangens hyperbolicus activation is used after each convolutional layer. An SSIM loss [Zhou Wang et al., 2004] is chosen for training since it yields sharper images compared to L1 and L2 losses. For the deep learning approaches a z-space size of 1024 delivered best results, while the latent space of the statistical models automatically results from the chosen variability percentage (here: 95%) and the training set size as demonstrated in further experiments. The networks are trained for a maximum of 300 epochs, however, an early-stopping strategy based on the validation dataset might interrupt the training in a more suitable epoch.

[1] `https://brain-development.org/ixi-dataset/`

Modeling Ability Results: The results for the modeling abilities in terms of generalization, specificity and likeness are shown in Fig. 2.4. Clearly, the deep learning approaches perform better in terms of specificity and generalization for all training set sizes and measures. In terms of likeness, the VAE yields the worst results, while it delivers the best specificity for large training set sizes. The intuition behind it, is the fact that VAEs tend to generate more "average" images in order to satisfy the normal distribution requirements for the latent space.

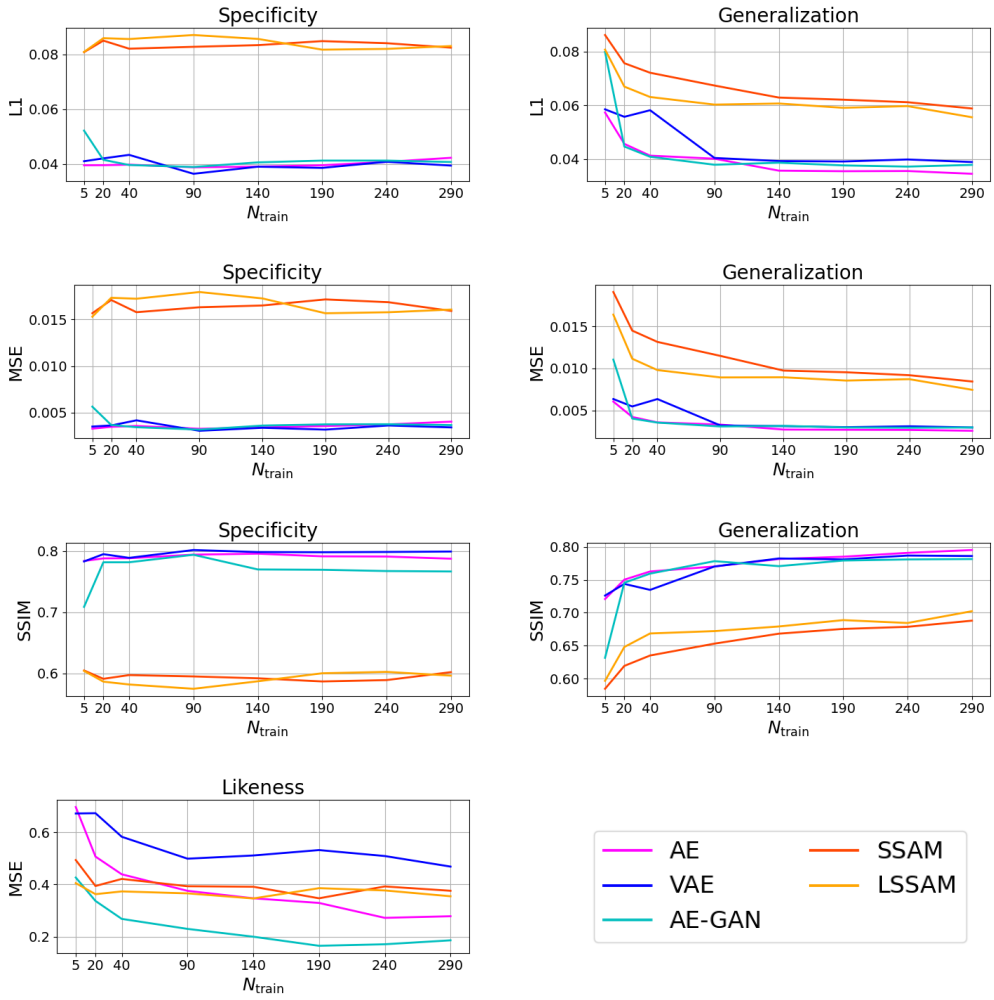


Fig. 2.4: Generalization ability, specificity and likeness for the modeling ability of all models (see legend) measured in L1(↓), MSE(↓) and SSIM(↑). Likeness is only measured in terms of MSE(↓) as proposed by [Guan and Loew, 2020]. For colors see legend (botton right).

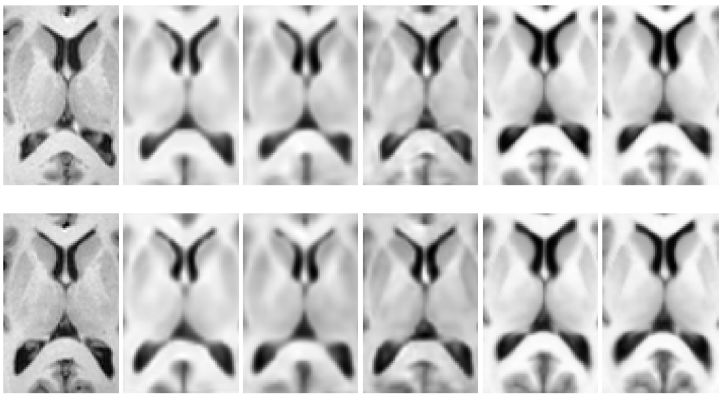


Fig. 2.5: Example reconstructions for two real unseen samples (left) using all methods. Shown are the center axial slices of the 3D volumes. Best viewed digitally.

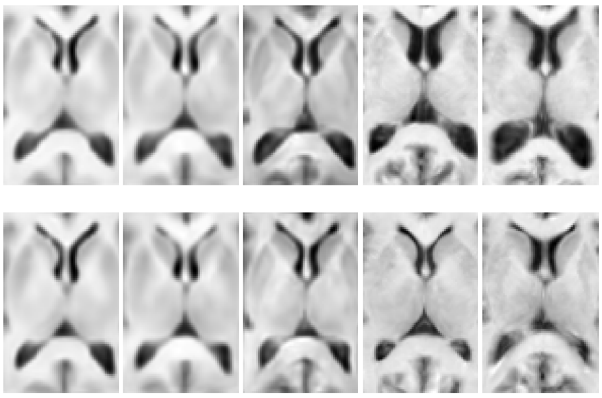


Fig. 2.6: Example random sample generation (top and bottom) using all methods. Shown are the center axial slices of the 3D volumes. Best viewed digitally.

This also becomes obvious in the shown example images for the reconstruction (Fig. 2.5) and generation of new samples (Fig. 2.6). Hence, the VAE model also tends to yield very similar samples and the lack of diversity leads to a bad likeness score. Due to the lack of this constrain, the AE models are able to better reproduce unseen samples and, thus, reach slightly improved generalization ability. The best likeness score by far is achieved by the AE-GAN approach. The example images also show a realistic appearance of the synthetically sampled images and good reconstructions of the unseen real samples, confirmed by the quantitative results for specificity and generalization ability.

Latent Space Properties: The quantitative evaluation for the latent spaces of the various methods is presented in Tab. 2.1. Typically, statistical shape models tend to be rather compact compared to deep learning approaches [Uzunova et al., 2021b, Uzunova et al., 2020c], yet, since they model shape and appearance separately and the complexity of the data is high, they only show a small improvement in terms of compactness in the presented experiments. A significant drawback of the deep learning approaches is, however, the fact that they have an ambiguous latent space (LAS $\gg 0$). This indicates that the whole latent space could be sampled by simply inputting the reconstruction of the previous image to the encoder. Also, different reconstructions results would be produced, while this is impossible for statistical models per definition. Yet, it can be observed that the latent space regularization in VAEs leads to smaller unambiguities.

A crucial property of generative models is the normality assumption of the latent space that enables the generation of new samples by sampling latent vectors from a normal distribution. Tab. 2.1 shows that the statistical methods have up to 40% of non-normally distributed dimensions of their latent space. This is due to the fact that statistical models are of linear nature and cannot map non-normally distributed input data to a normal distribution. This leads to bad specificity and poor likeness of the generated images (confirmed by the visual results in Fig. 2.6). The deep learning methods are able to map the input data in a non-linear manner and, thus, the resulting latent space distribution is closer to a normal distribution.

To enable visual assessment of the latent space, a visualization of the interpolation between the latent vectors of two images is shown in Fig. 2.7. Visually and also indicated by the colorbars, all methods deliver feasible results and generally smooth interpolations. However, the AE and VAE yield less similar reconstructions of the two real input images. Also the AE and statistical models show less smoothness in the transitions. The smoothest transitions and overall best results are observed in the AE-GAN scenario, which is consistent with the quantitative results. Also, the sampled images are realistic and detailed, underlining the plausibility of the model.

Table 2.1: Latent space metrics. 1) Latent ambiguity score (LAS): small values indicate an unambiguous latent space. 2) Non-normality: percentage of non-normally distributed dimensions of the latent space. 3) Compactness: for the statistical methods, the compactness is specified in a range depending on the training set size. Also, since they observe shape and appearance separately, the values are noted as $dim_s + dim_a$.

data	SSAM	LSSAM	AE	VAE	AE-GAN
ambiguity (LAS) ↓	**0**	**0**	0.28	0.04	0.14
non-normality ↓	40%	33%	12%	**7%**	**7%**
compactness ↓	**[4+4,257+189]**	[37+41,383+351]	1024	1024	1024

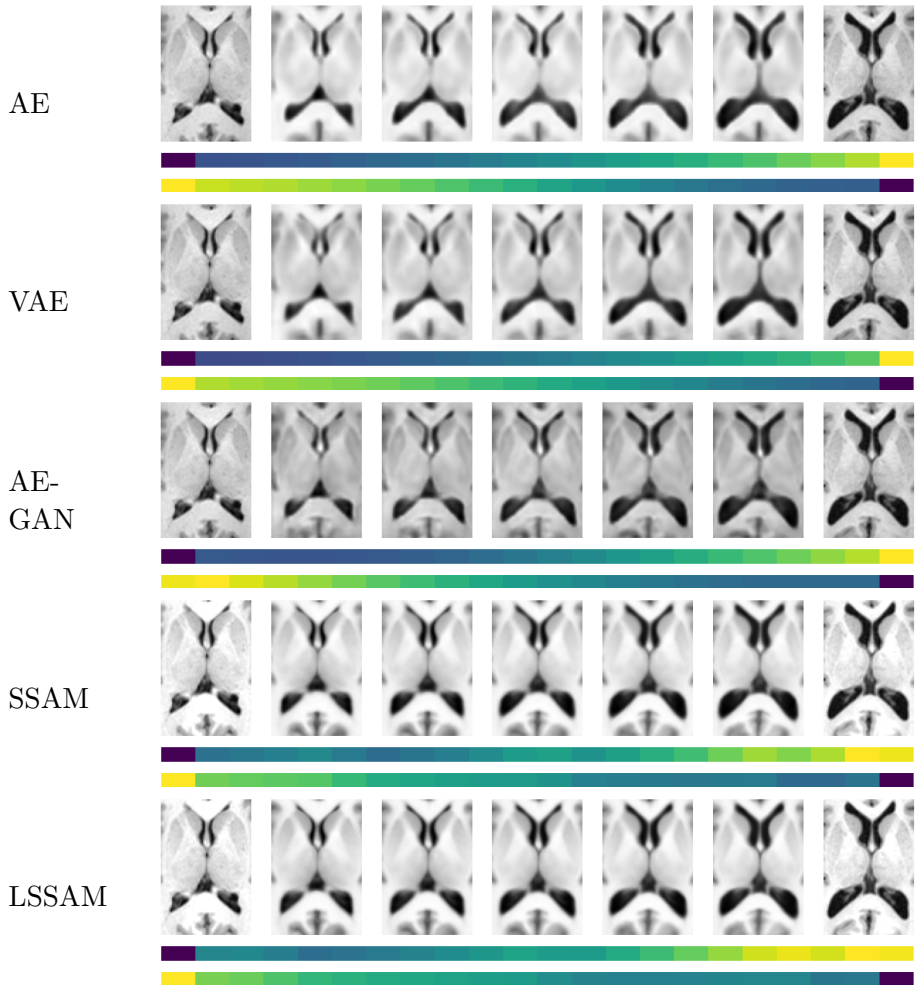


Fig. 2.7: Visualization of the linear interpolation between latent vectors of two images (first and last in a row). The bars underneath indicate the MSE between a reconstruction (20 steps) and the first shape (top bar) and a reconstruction and the last shape (bottom bar): max ( ) → min ( ).

2.6 Discussion and Conclusion

In this chapter, the generative models used as basis throughout this work are described. The considered autoencoders (AEs), variational autoencoders (VAEs) and generative adversarial networks (GANs) are the main generative deep learning models considered here. In general, all presented methods use a low-dimensional latent space for the representation of the learned images, yet, they have different architectures at their disposal. On the one hand, AEs and VAEs have an encoder-decoder architecture, where the encoder maps an input image to a low-dimensional latent space and

the decoder aims to reconstruct the input image given only a latent vector. VAEs additionally consider enforcing a normal distribution on the latent space to facilitate image generation. GANs, on the other hand, map low-dimensional random vectors to realistic images using a generator (functionally equivalent to a decoder) and an adversarial discriminator improves the realistic appearance of the generated images by introducing a training in the manner of a minimax game. Furthermore, various combinations and extensions of the methods are possible. In the first part of the chapter, the mathematical foundations and intuition behind those methods are presented.

In the second part, a systematic comparison among the methods next to a comparison to classical statistical models gives insight into the properties of the different models justifying the choice of deep learning models for the further developments in the work. Experiments on 3D brain MRI images show, that all deep learning approaches are indeed able to generate more realistic new images and reconstruct unseen images with a higher accuracy than the classical statistical models. This is underlined by visual as well as quantitative results. Especially in terms of specificity (ability to generate new realistic samples), but also in terms of generalization ability (ability to reconstruct unseen samples) the statistical methods perform significantly worse. One of the reasons for that becomes obvious through a close observation of the latent space. Other than expected, it is not normally distributed, thus, sampling new images from the assumed normal distribution yield unrealistic results. The reason for that lies in the linear nature of statistical models making it generally impossible to map non-normally distributed input data to a normal distribution in the latent space with a limited number of parameters and functions. Due to the non-linear nature of deep learning approaches, this problem is not as present. Further observing the latent spaces of the deep learning approaches shows the advantages of the latent space normalization implemented by VAEs, namely delivering a more normally distributed and a less ambiguous latent space, which is a common problem for deep generative methods. Overall, the reconstructed images and new samples generated by the deep learning models are of realistic appearance and behave as expected. The most crisp and realistic images are, however, generated by the GAN-based approach, which is emphasized by the visual as well as the quantitative results. Still, VAEs show some interesting properties considering their latent space. Based on the investigated properties, VAE- and GAN-based approaches are further pursued in this work.

Chapter 3

Autoencoder-based Unsupervised Modeling of Pathological Structures as Deviations from the Healthy Norm

This chapter introduces the developed methodology for the unsupervised modeling of pathological structures by assuming that pathologies can be recognized as deviations from the normal anatomical appearance. This is an advantageous representation since the variability of images of healthy subjects from a certain domain is significantly smaller than the variability of pathologies and can be learned using deep learning approaches. For the modeling of the healthy tissue, here, a VAE-based approach is designed, where pathologies can be recognized as deviations in the image space as well as the latent space. First, the main methodology is developed for 2D data and later, it is extended for 3D images. Three different medical applications of the developed method are shown and thoroughly evaluated: 1) detection and rough segmentation of pathologies; 2) pathology masking for the registration of images with the presence of pathologies; 3) pathology perturbation for the interpretability of black-box pathology classification methods.

3.1 Introduction and Motivation

The occurrence of pathologies in medical images is common, but still a major challenge for most automatic image processing methods. The reasons for this are various, but mostly lie in the irregular nature and enormous variability of pathological structures. Machine learning methods are able to grasp large varieties of complicated structures and issues. Therefore, many approaches are developed for the straightforward modeling of a certain pathology type. In [McKinley et al., 2016] random decision forests are applied for ischemic stroke segmentation based on various manually crafted descriptors. Other typical supervised learning approaches are convolutional neural networks. One

example for their successful use is [Kamnitsas et al., 2016], where brain glioblastomas are segmented in a supervised manner. However, for methods that directly model the entire variability of a certain pathology type, a tremendous amount of annotated training data is required. In the medical domain, large datasets are hard to access, due to e.g. privacy reasons. Moreover, their annotations are very time-consuming and usually require an experienced clinician. And while there are a few openly available datasets with certain annotated pathology types, e.g. brain lesions [Maier et al., 2017] or brain glioblastomas [Menze et al., 2015], they only cover a small variety of pathologies, so only specialized methods can be trained. Such methods are good at modeling a particular pathology group, but once trained, they fail on test data containing other pathological structures. For example, a neural network trained for the segmentation of brain stroke lesions is not suitable for the segmentation of brain tumors.

Inspired by the way how experienced clinicians who have seen many healthy subjects images would immediately recognize an abnormality in a given image, a novel pathology modeling approach is obtained here. Namely, an unsupervised approach based on the intuition that pathological structures can be recognized as abnormalities in medical images, i.e. structures that strongly differ from the normal healthy tissue. The main idea of the proposed method is, thus, to learn the healthy tissue variability of a certain domain and detect pathologies as a deviation from the learned norm. This allows to consider any possible pathology types, even if they are not known at the time of the training. Also, no pixel-wise annotations of the pathologies are needed for training [Uzunova et al., 2018, Uzunova et al., 2019d].

A similar idea has been explored in [Krüger et al., 2015] where the healthy variability of brain tissue is modeled with a statistical approach, such that brain lesions are detected in an unsupervised manner. In [Schlegl et al., 2017, Schlegl et al., 2019], GANs have been applied for the segmentation of pathological fluids in retinal OCTs. While GANs have a very good data generation ability, they are not able to directly map data to their latent representation and need an extra step to do so. AEs, on the other hand, have poorer generalization abilities, nonetheless, their encoder-decoder structure allows for direct mapping of an input to a latent vector and enable more precise reconstruction. In [Pawlowski et al., 2018], a work developed parallelly, AEs have been applied to model the healthy appearance of brain tissue for the unsupervised segmentation of brain tumors. In those approaches, the models are trained on a dataset containing only healthy subjects and are, thus, able to only reconstruct healthy images. Therefore, when an image containing a pathology is forwarded through a trained model, its reconstruction resembles healthy tissue, hence, pathological structures can be detected as large differences between the input image and its reconstruction.

This idea is also the base of the proposed approach, however, it is one of the first methods to utilize VAEs and conditional VAEs (cVAEs) for this purpose [Uzunova et al., 2018, Uzunova et al., 2019d]. More recent works confirm the plausibility of the chosen methodology by also employing different types of VAEs for anomaly de-

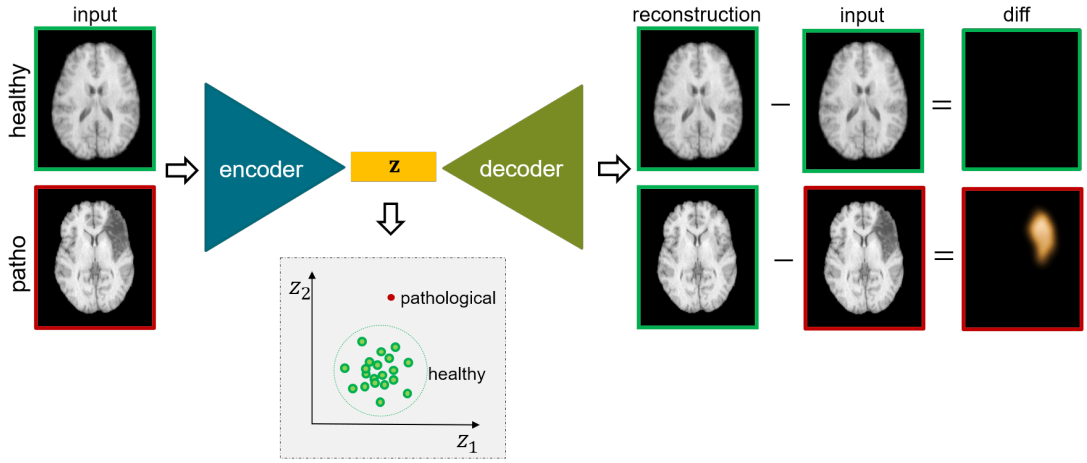


Fig. 3.1: Overview of the VAE-based pathology detection method. Red border indicates pathological images (here a stroke lesion); green border indicates a healthy subject's brain.

tection [Yao et al., 2019, Banerjee and Ghose, 2020, Baur et al., 2021]. Next to the reconstruction abilities of the model, a further property connected to the latent space of the variational autoencoders is also considered: since it is constrained to a normal distribution during the training on healthy images, pathological images are expected to be mapped far away from the learned normal distribution.

3.2 Unsupervised Modeling of Pathologies with VAEs

In Sec. 2.5 it has been shown that (V)AEs can reliably be used for representation learning making them suitable to learn the variability of the healthy data of a certain domain, e.g. in [Xue et al., 2017] where they manage to learn the representation of cardiac images for cardiac index prediction. Furthermore, VAEs assume a prior distribution of the latent space (typically a normal distribution) motivated by the logical distribution of the input data on some particular domain. This is indeed an advantageous feature, since a reasonable distribution of the healthy tissue representation of some particular domain in a lower dimensional space (z-space) can be obtained. For this purpose, a VAE needs to be trained on healthy data only. Fortunately, large datasets of unannotated images of healthy subjects are available more broadly in the medical image domain.

The pathology detection approach developed for this work [Uzunova et al., 2018, Uzunova et al., 2019d] is schematically shown in Fig. 3.1. Two assumptions, further confirmed by the experiments, lay the foundations of this method: 1) pathologies can be described as a reconstruction error in the image space; 2) the latent space encodings

of pathological structures do not conform to the learned normal distribution. Those two assumptions are not strictly disjoint and, as experiments show, should be used together, combined to a third assumption: pathologies are regions that show large reconstruction as well as z-space errors.

3.2.1 Pathologies as a Reconstruction Error

An advantage of VAEs is that due to their latent space constrains, they are not able to simply learn the identity function, e.g. downscaling in the encoder and upscaling in the decoder. They are moreover able to learn the underlying distribution of the data. This is why, when trained to reconstruct healthy images only, VAEs are not able to reconstruct pathological regions in the test images and rather interpret them as healthy tissue. Hence, big distances between the input image and its reconstruction would occur in the area of the pathology.

Concretely, using the notation from Sec. 2.3, given a VAE trained on healthy images, the underlying distribution of healthy data should be learned. Then, during the inference phase, a given input test image $\mathbf{x}$ can be encoded as $\mathbf{z} = f_\phi(\mathbf{x})$ and decoded as $\tilde{\mathbf{x}} = g_\theta(\mathbf{z})$. Since the healthy tissue variability has been learned, normal anatomical structures would get reconstructed fairly well. However, pathological structures would not get good reconstructions, thus, they can be recognized by their large pixel-wise distances $dist_{pixel}(\tilde{\mathbf{x}}, \mathbf{x})$. The function $dist_{pixel}(\cdot, \cdot)$ can be chosen accordingly as a pixel-wise distance, here MSE as defined in Sec. 2.5.2 is chosen.

3.2.2 Latent Space Pathology Detection using Patch-based Conditional VAEs

As the VAE learns to map healthy tissue to a normal distribution while training, in the test phase healthy tissue should conform to the learned distribution. Pathological tissue, however, differs from healthy tissue drastically, which leads to the assumption that its latent representations lie far outside from the normal distribution. So, the distance from a z-vector of a pathological image to the mean of the normal distribution of the healthy training data is large.

Using the above annotations, $\mathbf{x}$ can be classified as pathological if it lies far away from the learned mean vector determined by the formula $dist_z(\boldsymbol{\mu}_{\boldsymbol{train}}, f_\phi(\mathbf{x})) \gg \epsilon$, where ϵ is a threshold value, $\boldsymbol{\mu}_{train}$ is the mean vector of the z-space and $dist_z(\cdot, \cdot)$ is the Euclidean distance between vectors. Here, ϵ is set to $3\boldsymbol{\sigma}$ as typical for e.g. statistical shape models in order to capture ca. 99.7% of the data complying with a normal distribution. In this manner, the z-space can only be used to classify a whole image as pathological or healthy. In order to enable pixel-wise classification, a patch-based approach using conditional VAEs is proposed in the further coarse of this work.

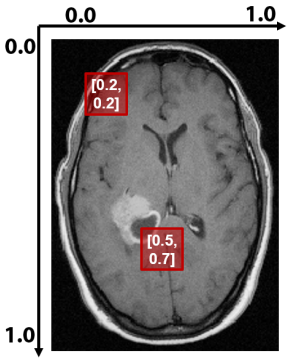


Fig. 3.2: An example for patch positional conditions c. Note that the two patches contain a similar structure with a high intensity. However, one of the patches contains tumor tissue and the other contains the normal anatomy of the head border.

As explained in Sec. 2.3, conditional VAEs (cVAEs) are an extension of VAEs that allow for an additional prior semantic information about the data, e.g. a label. The condition c is concatenated to the image as input to the encoder and additionally to the latent vector as an input to the decoder (see objective in Eq. 2.9).

In this way, a normal distribution pro condition is learned, thus, new data can be generated in a controlled manner given a certain condition. For the most part, cVAEs have been used as generative models for computer vision applications, e.g. [Yan et al., 2016, Zhang et al., 2020], however, here, they are applied for patch-based descriptor learning. It can be assumed that the latent vectors are very descriptive and can actually resemble a good image descriptor. Based on the success of patch descriptors [Faktor and Irani, 2012, Lotan and Irani, 2016, Hanif, 2019] and due to the benefits connected to training/testing computational demand, the cVAE is trained on image patches. It is, furthermore, important to integrate location context of the patches, since a healthy looking patch on a certain position might be what pathologies look like at another position, e.g. a patch of high-intensity voxels in a brain MRI T1 sequence resembles healthy tissue around the head border, but is most certainly abnormal if found in the center of the brain (Fig. 3.2). Locality information about the patches is integrated by using the continuous positions of their centers $c \in [0, 1] \times [0, 1]$ (analogous for 3D) relative to the image size as condition of a cVAE (Fig. 3.2). Thus, the healthy tissue variability is mapped to a normal distribution in z-space for each positions. In order to ensure that patches of same positions contain similar structures over all images of the training set, an assumption about the rough alignment of the images has to be made, e.g. the images are affinely or rigidly preregistered. Here, even a simpler alignment approach is applied: the images are cropped around the structure of interest using a bounding box, ensuring that the scaling and translation approximately match among the different images with respect to their size. This method enables the calculation of distances per patch and not the whole image at once. Accordingly, when overlapping patches are sampled, a distance per voxel can be determined as the average value of all overlapping patch voxels. Especially for medical images, patch-based approaches

also have further advantages. Medical images are often large 3D volumes and their straightforward processing with neural networks features extraordinarily high memory requirements (usually GPU RAM). Splitting the input images in patches results in facilitated memory usage and enables the processing of large medical volumes of high resolutions.

3.2.3 Concept Analysis

The reconstruction ability of VAEs has been already shown in Sec. 2.5. However, the spatial conditioning and the assumed latent space properties are crucial for the proposed approach. For this reason, a preliminary experiment on non-skull-stripped healthy T1-weighted brain MRI slices from the IXI dataset [Hammers et al., 2003] [2] is conducted. A patch-based VAE is trained on 5 000 randomly sampled patches from 40 images. Fig. 3.3 shows an example training image and several generated patches for given positions. It can be observed that patches on particular positions contain structures typical for the corresponding locations. E.g. the position $c_1 = [0.1, 0.9]$ yields patches containing structures typical for the border of the head, while patches sampled on the position $c_2 = [0.5, 0.5]$ contain structures like parts of the ventricles from the middle of the brain. Those first results show that using the location of the patches as a condition yields plausible results. The variety of different patches per position indicates that the natural variability of the healthy looking brain can be captured using the proposed approach.

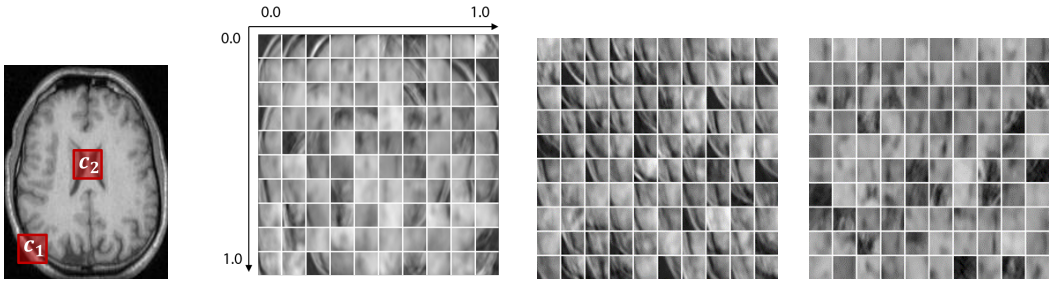
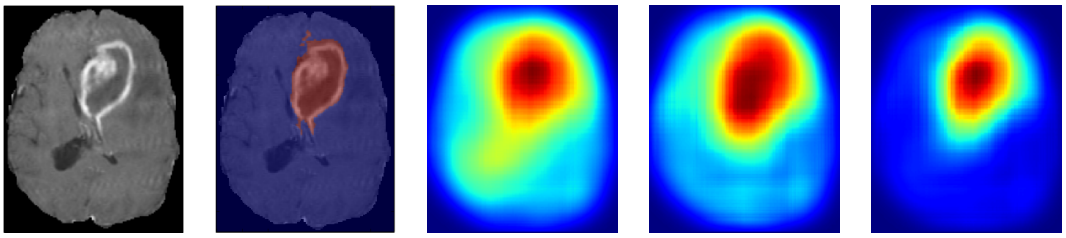


Fig. 3.3: An example training image and examples of sampled patches on different locations. The locations c_1 and c_2 are illustratively shown on the example image.

In a further experiment, the validity of the pathology detection assumptions is investigated. For this purpose, the patch-based VAE is trained on 10 000 patches from 220 brain MRIs from the BRATS dataset [Menze et al., 2015], where slices without visible pathologies are chosen for training. In Fig. 3.4 an example test image containing a large tumor and its ground truth segmentation is displayed. Next to them, the

[2] `https://brain-development.org/ixi-dataset/`



test image | ground truth | latent space $dist_z$ | reconstruction $dist_{pixel}$ | combined $dist_{pixel} * dist_z$

Fig. 3.4: Examples of the detection heat maps for a brain tumor with both strategies: reconstruction error in the image space and distance to the mean vector of the latent space. A multiplicative combination of both delivers best results.

results from both detection assumptions are shown in the form of color-coded distance maps. It can be observed that the tumor can be detected in the latent space as well as the image space. A further interesting observation is that, the pathology detection benefits from the multiplicative combination of both distance maps, corresponding to a logical conjunction of the two assumptions. This is also consistent with more recent works like [Zimmerer et al., 2019].

3.3 Unsupervised Pathology Detection and Segmentation using VAEs

Pathology detection and segmentation are crucial but complex tasks of medical image computing and analysis. Typically, the challenge of pathology detection can be managed by supervised machine learning methods that learn the variability of the underlying pathology by using a large training dataset where the pathological structures are annotated in a voxel-wise manner. Especially supervised deep learning approaches have shown great success in terms of segmentation in recent years. For example, in [Lu et al., 2019], the authors enable the segmentation of retinal fluid in optical coherence tomography images using a convolutional neural network. In [Myronenko, 2019], brain tumors in 3D brain MRIs are segmented with a supervised neural network approach combined with an autoencoder regularization. Even though very precise segmentation results are achieved, huge annotated training datasets are required and the quality of labels is crucial. Since such datasets are a rarity in the medical image domain, semi-supervised and unsupervised approaches gain importance. Unsupervised methods are often less accurate in terms of segmentation ability [Schlegl et al., 2017, Schlegl et al., 2019]. While some applications (e.g. tumor measurement or treatment monitoring) require accurate segmentations of the pathologies, other would only need a detection of the pathology and a rough approximation of its outline, for example to assist the

physician in a visual inspection or for the retrieval of images with pathologies at specific locations from a database.

Here, the unsupervised pathology modeling approach from Sec. 3 is directly applied for rough segmentation and detection of pathological structures like in [Uzunova et al., 2018]. The patch-based cVAE method is utilized for this purpose to, first, make use of the latent space in a patch-wise manner and, second, apply the approach on 3D medical volumes without significant GPU RAM constrains. By simple thresholding techniques the resulting distance heat maps can be transformed to segmentation masks. Thorough experiments for 2D and 3D single-modal and multi-modal images are carried out. Qualitative and quantitative evaluation of the results show that a rough segmentation is possible and, thus, the method is overall suitable for pathology detection. Also, a comparison to a GAN-based approach shows the advantages of using a VAE-based architecture for the given use case.

3.3.1 Architectures and Implementation Details

Different cVAE architectures for 2D [Uzunova et al., 2018] and 3D image data [Uzunova et al., 2019d] are developed. The best performing ones are shown below and further employed for various experiments. In both cases the patch-based cVAE approaches are applied, and thus, the latent space can be utilized for pathology detection. Therefore, patches extracted from the whole images are used as inputs of the networks. The cVAE architecture for 2D images is shown in Fig. 3.5. A simple fully-connected architecture containing three hidden layers in the encoder and decoder each is chosen.

The more sophisticated 3D architecture is shown in Fig. 3.6. Here 3D convolutions are used in the encoder and 3D deconvolutions (transposed convolutions) in the decoder. The architecture is shown for 3-channel input images but can be analogously used for one-channel images.

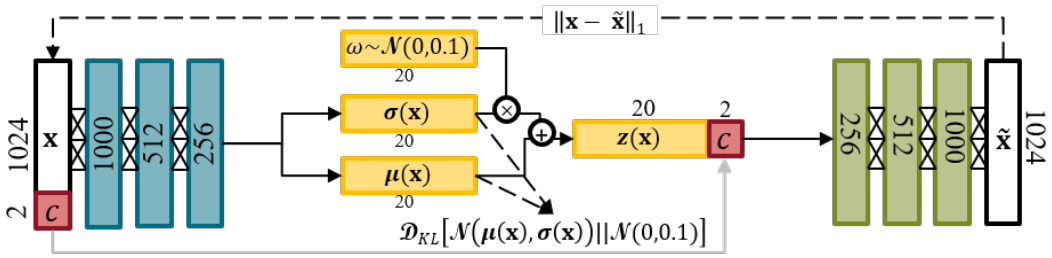


Fig. 3.5: cVAE architecture for 2D images. Vectorized patch $\mathbf{x}$ and reconstruction $\tilde{\mathbf{x}}$; fully-connected encoder (■) and decoder (■) layers; condition (■); z-space (■) where $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \times \boldsymbol{\omega}$, with random $\boldsymbol{\omega} \sim \mathcal{N}(0, 0.1)$. Dashed lines correspond to loss functions: $\mathcal{D}_{KL}$ – KL divergence and $||\cdot||_1$ – L1 norm. Numbers denote size of layers.

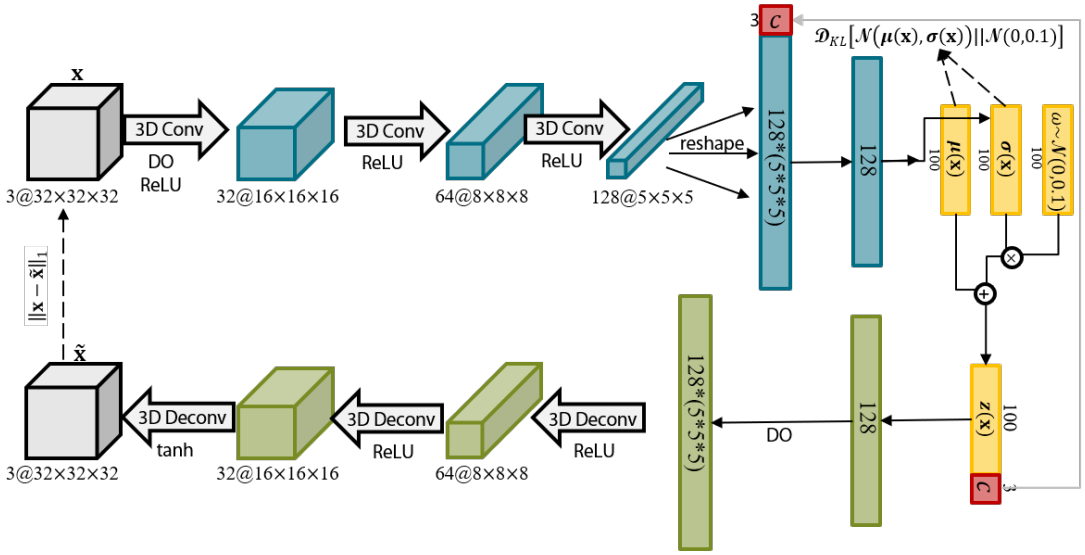


Fig. 3.6: cVAE architecture for 3D images. 3D patch $\mathbf{x}$ and reconstruction $\widetilde{\mathbf{x}}$; convolutional encoder with fully-connected latent space encoder (■) and decoder (■) layers; condition (■); z-space (■) where $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \times \boldsymbol{\omega}$, with random $\boldsymbol{\omega} \sim \mathcal{N}(0, 0.1)$. Numbers denote size of layers (channels). Solid black lines correspond to fully-connected operations. DO stays for Dropout; ReLUs are used everywhere if not mentioned otherwise. Loss functions (dashed lines) as in Fig. 3.5.

Note that the dimension of $\mathbf{z}$ is crucial – too large z-vectors contain too much insignificant information and noise; too small $\mathbf{z}$'s might be insufficient to store all the important information about the input. The latent dimension is empirically chosen, with 20 for patches of size 32×32 and about 100 for 3D patches sized $32 \times 32 \times 32$ voxels.

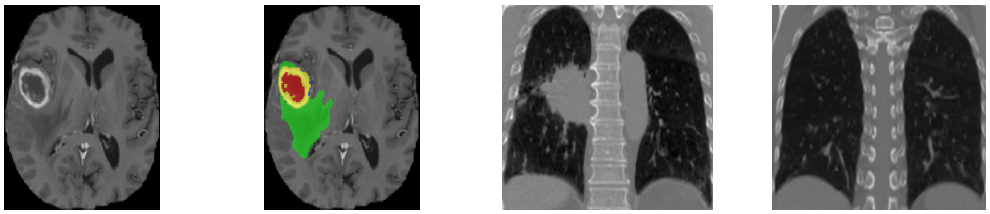
Another important detail is the choice of the loss functions. While it is rather common to choose the Kullback-Leibler divergence $\mathcal{D}_{KL}$ as a latent space loss, the reconstruction loss may vary. In literature, it is usual to take the sigmoid cross entropy loss when processing binary data [Kingma and Welling, 2014], but the most natural choice for grey-valued data is the mean squared error (MSE). Yet, MSE leads to learning very smooth reconstructions with small variety between images. Choosing the L1 loss (sum of absolute differences) leads to more sharp and diverse results and is, thus, integrated in the presented architecture.

In the usual cVAE architecture, $\mathbf{z}$ is calculated using the "reparametrization trick" as $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \times \boldsymbol{\omega}$, where $\boldsymbol{\omega} \sim \mathcal{N}(0, 1)$. However, because $\mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\sigma}(\mathbf{x})) \sim \mathcal{N}(0, 1)$ is enforced, the same amount of learned information and random noise coming from $\boldsymbol{\omega}$ are used during the training. To enforce the stronger dependence of the latent vectors on the input data, here, the randomness is reduced by sampling $\boldsymbol{\omega}$ from distributions

with smaller standard deviations (e.g. $\mathcal{N}(0, 0.1)$), which eventually results in having less noisy reconstruction images and distributions in z-space.

3.3.2 Data and Experimental Setup

For the conducted experiments two datasets showing different modalities, anatomical regions and pathologies were used (Fig. 3.7).



brain MRI image slice | brain MRI with segmentation | thorax CT test image | thorax CT train image

Fig. 3.7: Examples of the training and testing data. From left to right: A slice of the T1c sequence of a *Brain MRI* image; Ground truth segmentation of the tumor tissue: enhancing tumor (■), necrotic tissue (■), edema (■). An example test image of the *Thorax CT* dataset; An example train image of the *Thorax CT* dataset.

Brain MRI: 220 brain MRI volumes with ground truth segmentations of high-grade glioblastomas from the BRATS 2015 challenge dataset [Menze et al., 2015, Bakas et al., 2017] are used. The volumes contain four modalities per subject (T1, T1c, T2, Flair). Two types of experiments are conducted with the data: 2D experiments where only center axial slices of the T1c volumes that offer the best tumor core visibility are extracted; and 3D experiments where a one-channel approach with T1c modalities and a multi-modal version with 3-channel 3D volumes (T1c, T2 and Flair) are considered. Here, the T1 sequences are not used since the T1c sequences are already considered, and from experience, the T1 sequences do not add to the information value.

A major advantage of this dataset is that the available segmentations allow a selection of non-pathological patches for training and a quantitative evaluation while testing. For the experiments on this dataset, a 4-fold cross-validation is implemented. For training 150 000 patches not containing pathologies are randomly selected. The test dataset only contains subjects with visible pathologies. For the 3D dataset, images that suffer from bad resolution in z-direction (e.g. follow-up images) are excluded from the experiments. Since the skull-stripping of the brain MRIs yields very rough edges causing high distances in the border areas, all distance maps are multiplied by the Gaussian smoothed masks of the brain.

Thorax CT: This dataset contains 2D coronal thorax CT slices from 46 patients with various pathologies, e.g. lung cancer or fibrosis. There is no suitable ground truth available which makes picking healthy patches more challenging and only enables a qualitative visual evaluation. Thus, for training, 35 image slices containing no obvious pathologies are chosen and 5 000 patches are extracted from them, another 11 slices from the remaining images are used for testing.

As previously mentioned, the images are cropped using a bounding box around the structures (brain or lungs), ensuring that the structures of interest have the same relative positions to the image size. Furthermore, all extracted training patches are sampled on random positions to avoid learning only discrete positions. A patch size of 32×32 pixels (or $32 \times 32 \times 32$ for 3D) was used in all experiments, since it empirically showed to deliver the best results. For the test phase, same-sized patches are sampled densely with large overlaps ($> 50\%$ of the patch size).

For those experiments, the 2D (Fig. 3.5) and 3D (Fig. 3.6) cVAE-based architectures were applied. All three detection assumptions were investigated: reconstruction distances $dist_{pixel}$, z-space distances $dist_z$ and combined distances $dist_{pixel} * dist_z$. The distances were calculated in a patch-wise manner, resulting in a single distance value per patch. The reconstruction distances correspond to the L2-norm of the pixel/voxel-wise difference of the original and cVAE-reconstructed patches; for the z-space Euclidean distances between the latent representation of an input patch and the mean of the learned latent vectors of the healthy training data are calculated; and the combined distances are the element-wise multiplication of both. To ensure a smooth appearance of the results, patch distances are averaged over the patch overlaps.

To evaluate the experiments, visual assessment is used for the *Thorax CT* dataset due to the lack of a suitable ground truth. For the *Brain MRI* experiments, a quantitative evaluation with the metrics presented in the following is possible since fortunately ground truth segmentation labels of the different tumor tissues are given. However, only the labels of the tumor tissue types that are visible in the particular modalities are assessed: the tumor core for the T1c modalities and additionally the tumor edema for T2 and Flair. For the 3-channel 3D brain MRI experiments, every modality is evaluated separately and a combined multi-modal distance (MM) is introduced $\mathcal{D}_{MM} = \alpha\mathcal{D}_{T1c} + \beta\mathcal{D}_{T2} + \gamma\mathcal{D}_{Flair}$ where $\alpha + \beta + \gamma = 1$ and $\mathcal{D}_{T1c}, \mathcal{D}_{T2}, \mathcal{D}_{Flair}$ are the distance maps for each modality respectively. The values $\alpha = 0.5, \beta = 0.2, \gamma = 0.3$ showed the best results.

3.3.3 Evaluation Metrics

Several evaluation methods have been used to determine whether the methods presented here are suitable for the detection and segmentation of pathologies.

Histograms: The aim to determine the distribution of distances between pathological and healthy pixels. The distances per voxel are displayed as frequency distributions, e.g. by fitting a Gaussian curve on the histograms of the distances. In the case of a bimodal frequency distribution, the pathological and non-pathological pixels can be distinguished on the basis of their distances. In Fig. 3.8 possible distributions are shown. The case of perfect separability is rather uncommon (left), so, a separability with a certain error tolerance (right) is typically expected. In the case of overlapping Gaussian curve peaks, no distinction based on distances is possible.

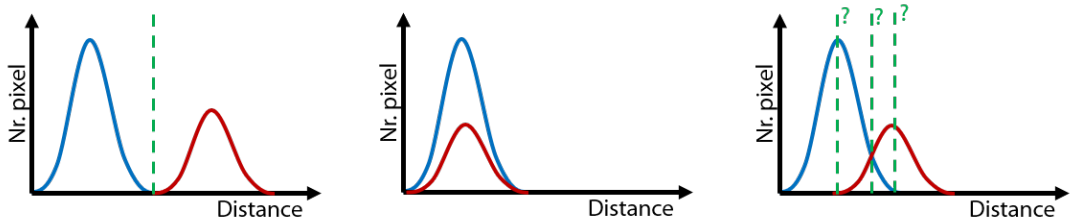


Fig. 3.8: Gaussian curves adapted to histograms. Left: two perfectly separable distributions; middle: two distributions that cannot be distinguished; right: two distributions that conditionally separable. Possible threshold to distinguish two structures (■); non-pathological pixels (■); pathological pixels (■).

ROC Analysis: The classes in bimodal frequency distributions are usually not perfectly separable. So, a threshold should be chosen such that the probability of correct assignment for both classes is maximized simultaneously. For example, in Fig. 3.8 (right) several possible threshold values are feasible. To determine the best threshold value, a method called *Receiver Operating Characteristic* (ROC) analysis can be applied. For this purpose several thresholds are tested and for each the classification quality is evaluated in terms of *sensitivity* and *specificity* (for definition see below). In this way, a diagram – with an x-axis corresponding to $(1 - specificity)$ and a y-axis to *sensitivity* – depicting the values for each threshold value can be obtained. In this way a curve is formed (Fig. 3.9). A curve that is close to the diagonal corresponds to a nearly random process. Accurate ROC curves are as close as possible to the upper left corner of the diagram. In ROC analysis, the optimal threshold value is reached when sensitivity and specificity are optimal. This can be calculated e.g. using the Youden index $(sensitivity - specificity - 1)$. Furthermore, in order to quantify the separability of both distributions, the area under the curve (AUC) can be calculated. The closer to one the AUC is, the better the classes can be separated. AUC values close to 0.5, indicate differentiation ability close to a random process (see Fig. 3.9).

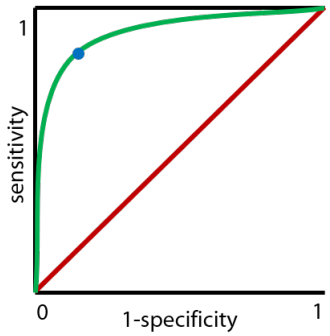


Fig. 3.9: ROC curve. Random probability that a pixel is pathological or healthy (■, AUC=0.5); good separability between pathological and healthy pixels (■, AUC≫ 0.5); best threshold for maximum differentiation between both structures (■).

Sensitivity and Specificity: The sensitivity and specificity values of a predicted segmentation compared to the ground truth are defined as follows:

$$\begin{aligned} sensitivity &= \frac{TP}{TP+FN} \\ specificity &= \frac{TN}{TN+FP}, \end{aligned} \tag{3.1}$$

where TP, TN, FP and FN are the predicted segmentation pixels compared to the ground truth as described in the confusion matrix Tab. 3.1. The two measures quantitatively describe how well the predicted segmentation matches the ground truth: values close to one indicate good segmentation. It is also crucial to consider both measures simultaneously since individually their informative value is limited, e.g. sensitivity can reach its maximum if all pixels of the image are featured in the predicted segmentation, but in this case, the specificity will be low.

Table 3.1: Confusion matrix.

		predicted	
		object	**background**
ground truth	**object**	True Positive (TP)	False Negative (FN)
	background	False Positive (FP)	True Negative (TN)

Dice Coefficient: The Dice coefficient is commonly used to evaluate the quality of segmentations. Similar to sensitivity and specificity, it measures the overlap of two segmentations, where the highest value of one corresponds to a perfect agreement between the segmentations and a Dice of zero corresponds to no overlap between them. The Dice coefficient is calculated as follows:

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|} \Leftrightarrow$$
$$Dice = \frac{2TP}{2TP + FP + FN}, \tag{3.2}$$

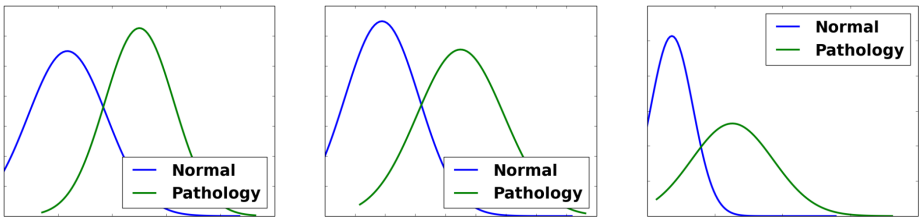
where X and Y are the pixel sets of the two segmentations and the operations applied on them correspond to set operations. This can also be expressed by using TP, FP and FN. Note that no TN pixels are considered by the Dice coefficient, so that correctly segmented background pixels are ignored.

3.3.4 Experiments and Results

Pathology Detection on 2D Images: The results computed on the 2D *Brain MRI* data are shown in Tab. 3.2 and Fig. 3.11. The ROC analysis clearly shows that the method is able to establish distinction between healthy and pathological tissue (AUC = 0.95). This is also visible in the distribution plots in Fig. 3.10, where pathologies clearly tend to yield larger values than normal tissue. Best results in terms of all

Table 3.2: 2D detection results for the *Brain MRI* dataset. From left to right: cVAE reconstruction $dist_{pixel}$, z-space $dist_z$ and combined distances $dist_{pixel} * dist_z$; anomaly score of anoGAN from [Schlegl et al., 2017]. Evaluation metrics are averaged over all images and folds: AUC over all subjects, sensitivity and specificity at Youden index and Dice coefficient of segmentation with ROC-picked threshold. Superscript $\star$ indicates statistical significance ($p < 0.001$) in a one-sided t-test compared to combined.

	$dist_{pixel}$	$dist_z$	$dist_{pixel}$* $dist_z$	anoGAN
Dice ↑	$0.49 \pm 0.26^\star$	$0.51 \pm 0.26^\star$	$\mathbf{0.55 \pm 0.27}$	$0.51 \pm 0.28^\star$
sens ↑	0.89	0.88	**0.91**	0.90
spec ↑	0.81	0.85	**0.86**	0.81
AUC ↑	0.94	0.92	**0.95**	0.93



z-space distances reconstruction distances combined distances

Fig. 3.10: Distribution plots of the distances of normal and pathological test patches.

evaluation metrics are achieved when the combination of reconstruction and z-space distances is used, which corresponds to the intuition that pathological tissue has both properties: it gets reconstructed poorly and its latent vectors get mapped far away from the learned distribution. This can also be seen in the example in Fig. 3.11 where the oversegmentation of the latent space and the undersegmentation of the reconstruction are combined to a nearly optimal segmentation of the brain tumor.

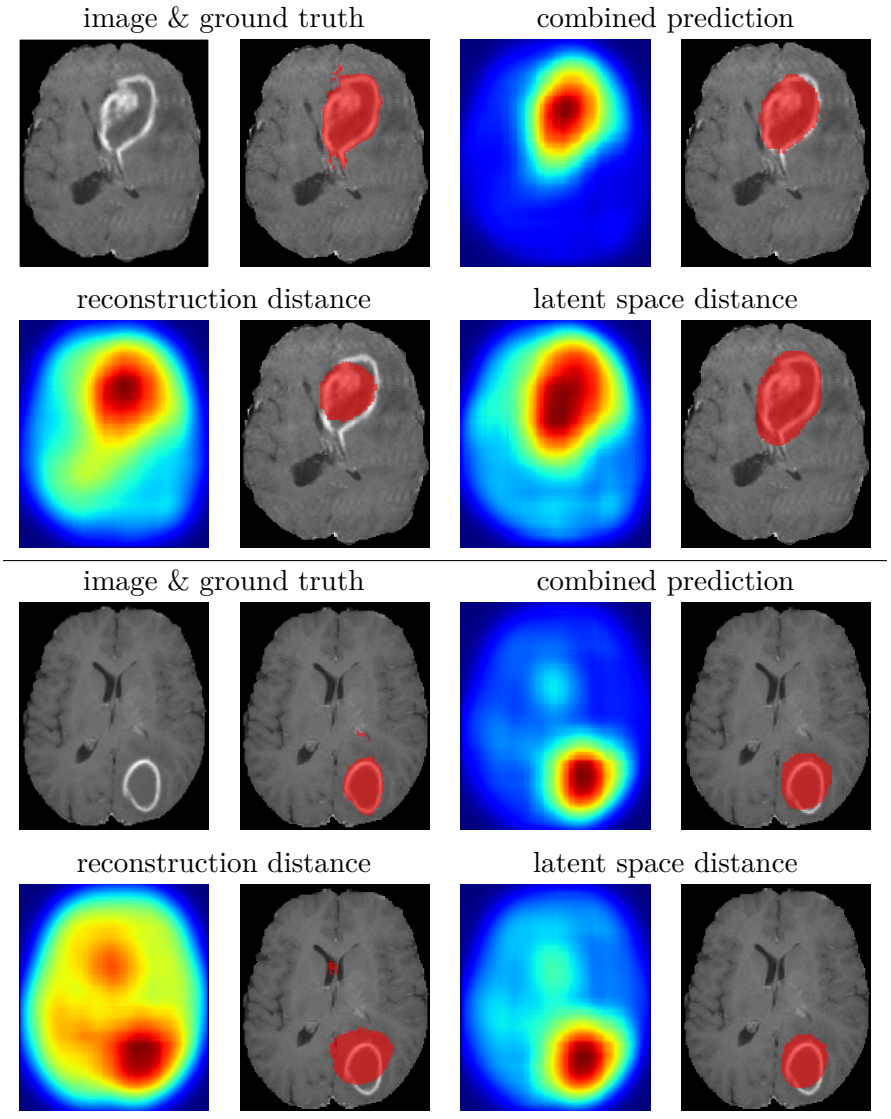


Fig. 3.11: 2D *Brain MRI* segmentation results. Shown are two example images with their ground truth segmentations, the predicted distance heat maps and resulting segmentations for the reconstruction distance, latent space distance and their multiplicative combination.

The achieved mean Dice of 0.55 is a segmentation result comparable to the results from the 2015 BRATS challenge [Menze et al., 2015, Bakas et al., 2017] where the proposed approach would reach the 6th place, but not to advanced methods like [Kamnitsas et al., 2016, Isensee et al., 2021]. Still, those approaches are supervised and trained for the explicit segmentation of glioblastomas, while the proposed method establishes an unsupervised detection of different pathologies that are not necessarily considered by the used ground truth (Fig. 3.12). Furthermore, the 2D layers often contain partially overflowing 3D segmentations, so the detection results can be distorted.

To establish a comparison to other unsupervised methods, a GAN-based method is also implemented here. The so-called anoGAN [Schlegl et al., 2017] roughly follows the same idea of learning the representation of healthy images. The authors also combine a latent space distance and an image reconstruction distance for detection, defining an anomaly score to distinguish between pathologies and healthy tissue. However, GANs learn a direct mapping $g_\theta(\mathbf{z}) = \tilde{\mathbf{x}}$ from the latent $\mathbf{z}$ with g_θ being the generator, but present no possibility to map an input image $\mathbf{x}$ to a latent variable $\mathbf{z}$, s.t. $\tilde{\mathbf{x}} \approx \mathbf{x}$. So, to find the latent mapping of a test image $\mathbf{x}$, an additional optimization problem needs to be solved by finding an optimal $\mathbf{z}$ s.t. $g_\theta(\mathbf{z}) \approx \mathbf{x}$. Naturally, this step significantly slows down the inference of the model and its accuracy is limited.

Since the anoGAN is applied patch-wise on densely sampled overlapping patches analogously to the cVAE, the per-patch anomaly score is used to generate distance images. This process leads to considerably longer computational time per patch (ca. 15 times longer in the performed experiments) yielding infeasibility for a whole image slice or even a 3D volume.

Despite this fact, the exact same experimental setup for the 2D brain MRI images is recreated using anoGAN and the results are also shown in Tab. 3.2. Note that the approach was originally developed for OCT data, therefore, the architecture was adapted for the specific use case. It could be observed that anoGAN is able to reconstruct the input images with a higher quality, thus, the pathological as well as the healthy tissue get reconstructed well. This can be intuitively explained by the lack of positional conditions for the patches, the missing constraint of the latent-space distribution and

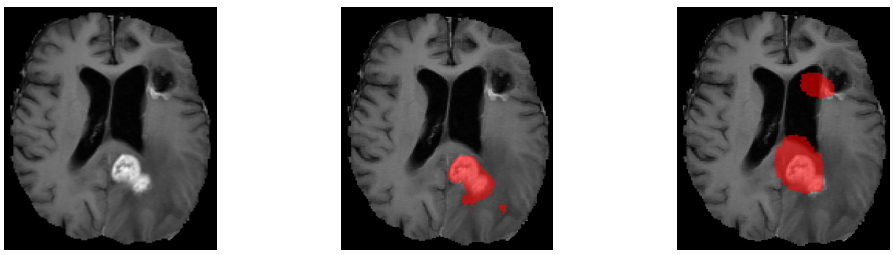


Fig. 3.12: An example, where the proposed method segments an additional anomaly in the brain, which is not annotated as tumor tissue by the experts.

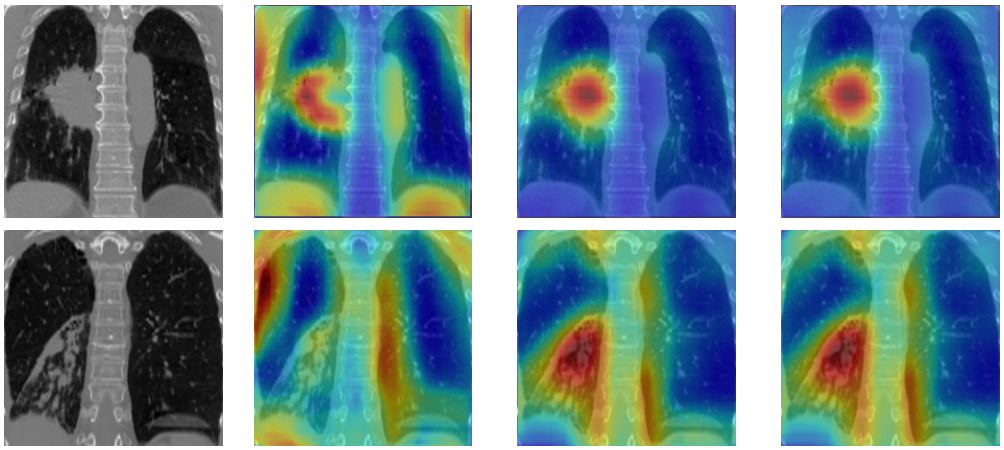


Fig. 3.13: 2D *Thorax CT* distance images. Test pathological images (left column) with their three types of corresponding distance images shown as overlayed heat maps.

the indirect optimization-based z-space mapping. Still, anoGAN demonstrates a good detection ability, nonetheless, it delivers significantly lower Dice scores compared to the presented approach and seems to be less suitable for the specific task.

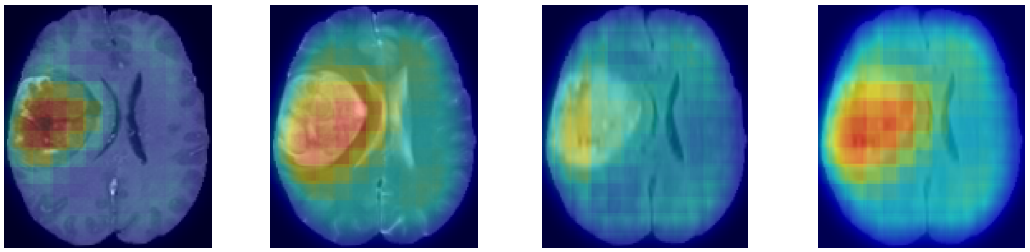
To show the generalization abilities of the method, an experiment with the fundamentally different *Thorax CT* data is carried out. Despite the missing ground truth segmentations, a visual evaluation shows the feasibility of the method for this data as well (Fig. 3.13 big distances correspond to pathological structures). Interestingly, z-space distances here seem to rather correspond to edges than pathological structures since pathologies typically have the same intensity values as healthy tissue. Still, combining them with the reconstruction distances performs well.

Pathology Detection on 3D Multi-modal Images: The same experimental setup is followed on the 3D *Brain MRI* dataset for pathology detection. The results are shown in Tab. 3.3. Clearly, 3D detection volumes are more challenging for this detection approach, even though a more sophisticated network architecture is used. The model is trained on one-channel volumes, containing the T1c modalities only, and on three-channel volumes consisting of the T1c, T2 and Flair modalities. The additional benefit of multi-modal data can be observed in the quantitative results, since the detection abilities not only improve by considering a multi-modal distance, but also only assessing the T1c distance images.

However, when training on only one modality, the values are not convincing. Contrary to that, best results regarding the segmentation quality based on Dice are achieved when using the multi-modal distance, where a Dice value of 0.5 shows to be significantly better in a one-sided t-test. An interesting observation is that higher

Table 3.3: 3D detection results for the *Brain MRI* dataset. From left to right: the combined distances for one-channel training; multi-channel training and assessing the modalities T1c, T2, Flair separately and the multi-modal (MM) distance. Evaluation metrics are averaged over all images and folds: AUC over all subjects, sensitivity and specificity at Youden index and Dice coefficient of segmentation with ROC-chosen threshold. Evaluated labels: tumor core and tumor edema (for T1c only the tumor core is evaluated due to visibility). Superscript $\star$ indicates to statistical significance ($p < 0.05$) in a one-sided t-test compared to the other experiments.

	train 1-channel	**train 3-channel**			
	T1c only	T1c	T2	Flair	MM
Dice ↑	0.32 ± 0.24	0.47 ± 0.17	0.29 ± 0.22	0.40 ± 0.17	$\mathbf{0.50 \pm 0.18}^{\star}$
sens ↑	0.89	0.92	**0.93**	0.92	0.91
spec ↑	0.86	**0.87**	0.80	0.80	0.81
AUC ↑	0.95	**0.96**	0.93	0.93	0.94



T1c (Dice = 0.74) T2 (Dice = 0.27) Flair (Dice = 0.42) MM (Dice = 0.80)

Fig. 3.14: 3D *Brain MRI* multi-modal distance images (axial slices). Shown are different modalities with corresponding overlayed distance heat maps. From left to right: T1c; T2; Flair; Multi-modal distance (overlayed on Flair)

Dice values not necessarily correspond to higher AUC values, e.g. the 3-channel T1c approach has a higher AUC value than the MM approach. This can be explained by the considerably higher specificity based on the large amount of TN predicted pixels typical for an undersegmentation. Contrary to that, TN values are not considered by the Dice coefficient. In Fig. 3.14 the distance images for an example slice of a 3D volume are shown. All of them tend to have large values around the tumor location and their combination ensures best differentiation between the healthy and pathological tissue. Since the 3D volumes contain an immense amount of voxels, the patches are only sampled on every 16th voxel in each direction, leading to the clear patch artifacts in the distance maps. This naturally also impairs the segmentation results since the obtained segmentations are rough.

To further prove the importance of the positional conditioning of the patches, an ablation experiment is also conducted without the used conditions. In this case, a signifi-

cant performance drop can be noticed: the Dice coefficient decreases from $\mathbf{0.50 \pm 0.18}$ to $\mathbf{0.42 \pm 0.13}$ and the AUC from **0.94** to **0.88** using the exact same setup. This demonstrates the intuition, that positional conditioning is a meaningful additional prior information about the data. When the positional information is missing, it can partly be learned and stored in larger $\mathbf{z}$ vectors. Yet, oftentimes, patches on different positions look alike but contain completely different structures. The hypothesis that healthy tissue is only detectable in the context of its localization (e.g. very dark patches are detected as healthy in the ventricles, but pathological in the brain tissue on T1 sequences) is confirmed by the better performance achieved by the method with integrated positional conditions.

3.4 Unsupervised Pathology Detection for 3D Pathological Image Registration

After showing the general ability of the proposed method for rough detection of anomalies, in this section, the application of the detection approach as a preprocessing step for further automatic image analysis methods is aimed. Here, the emphasis lies on medical image registration of pathological data [Uzunova et al., 2019d]. Image registration has become an essential part of medical image processing. The purpose of image registration is to align two or more images featuring the same object, so that significant points correspond optimally. This can be used e.g. for atlas-based segmentation or image fusion. The usual definition of image registration uses a template image, which is aligned to a chosen reference image by finding proper transformations such that corresponding points match. However, it is still a significant challenge to register images containing major pathological structures, since they typically cause missing correspondences. This problem is addressed by several works e.g. [Chitphakdithai and Duncan, 2010], where a map of the healthy tissue with available correspondences is build to boost the registration of the healthy regions. In [Yang et al., 2016] the authors use a variational autoencoder to map a pathological image to a semi-healthy image and use it for registration. An even more sophisticated approach to generate a semi-normal image from a brain tumor MRI is developed in [Han et al., 2020], where a deep neural network simultaneously reconstructs the healthy appearance of an image containing a tumor and warps it to a given atlas.

Here, a different approach is chosen pursuing the aim to use the rough pathology detection as a weight map integrated in the registration approach, s.t. regions that cause missing correspondences get less involved into the registration objective [Uzunova et al., 2019d]. In this manner, a proof of concept is established where rough segmentations can be indeed helpful as pre-processing steps for further automatic image processing methods.

In the experiments presented here, 3D lesion brain MRIs are used in an atlas-to-patient registration scenario, where pathology weight masks significantly improve the registration results.

3.4.1 Unsupervised Pathology Weight Masking

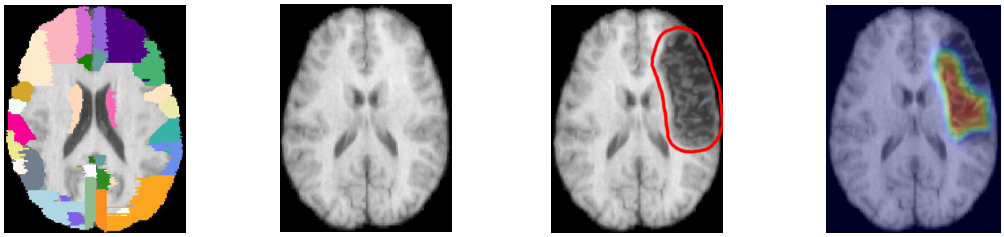
The proposed pathology detection method is examined in atlas-to-patient registration scenarios assuming that a healthy atlas image is registered to a patient image containing pathological structures. It is investigated whether masking the pathological tissue in the images leads to better registration results. The variational non-linear registration method from [Werner et al., 2014] is chosen, since it performs along with the best algorithms and is freely available [Ehrhardt et al., 2015]. For the presented experiments, a normalized local cross correlation distance metric (NCC) with curvature regularization is used during the registration; all parameters are chosen according to [Ehrhardt et al., 2015].

To integrate a-priori information about possible pathological tissue, the pathology detection approach described in Sec. 3.3 is used to generate a pathology map for weighting the registration objective. Hence, the locations that most likely cause missing correspondences are not essential in the registration process, contrary to locations of the images that represent healthy tissue and feature good correspondences.

To compute the weight mask, the combined reconstruction and z-space distances $\mathbf{d}$ from the cVAE (see Fig. 3.6) are thresholded with an AUC-chosen threshold values resulting into a binary segmentation $\mathbf{s}$. Then, $\mathbf{s}$ is smoothed using a Gaussian filter with a standard deviation $\sigma = 3$. The weight mask $\mathbf{w} \in [0, 1]$ is the inverted version of $\bar{\mathbf{w}} = \max(resc(\mathbf{d}), gauss(\mathbf{s}))$, where $resc(\cdot)$ denotes rescaling the images to the interval $[0, 1]$, $gauss(\cdot)$ is the Gaussian blurring function and $\max(\cdot)$ is a voxel-wise maximum function. The final inverting of the values ensures that $\mathbf{w} \in [0, 1]$ with $\mathbf{w} = 0$ for regions detected as pathology with a high certainty. The resulting mask locally weights the distance measure and, thus, suppresses the influence of pathological regions.

3.4.2 Data and Experimental Setup

This experimental setup features synthetically generated phantom brain lesion MRIs to enable direct comparison between registration of healthy and pathological images and ensure the availability of ground truth segmentations for evaluation purposes. As a basis, the LPBA40 dataset [Shattuck et al., 2008] is chosen, which is a publicly available dataset of 40 healthy subjects whole-head MRI volumes with 56 labeled anatomical regions each. This dataset has previously been used for evaluation studies of image registration algorithms [Klein et al., 2009]. For experimental purposes, phantom pathologies are simulated in the images using manually segmented stroke lesions



atlas image | image w/o lesion | image with lesion | w. weight mask

Fig. 3.15: Example images from the LPBA40 dataset. From left to right: atlas image with overlayed ground truth labels; image without lesions; phantom image with simulated lesion (red border); image with the pathology-detection weight mask.

extracted from the ISLES challenge data [Maier et al., 2017]. The lesion simulation method as proposed in [Krüger et al., 2020] is formulated as follows:

$$\mathbf{x}^p = HE(T(\mathbf{x}^l)) \cdot \mathbf{m}^l + \mathbf{x}^h \cdot (1 - \mathbf{m}^l) \tag{3.3}$$

where $\mathbf{x}^l$ denotes an image from the ISLES dataset containing a lesion, $\mathbf{x}^h$ is a healthy subject's image from the LPBA40 dataset and $\mathbf{m}^l$ is a smoothed segmentation mask of the lesion. The intensities of both images are matched using histogram equalization $HE(\cdot)$ and the healthy image is registered to the lesion image using an affine transformation $T(\cdot)$. Here, one example lesion is chosen and simulated onto the LPBA40 images resulting in 40 healthy and 40 corresponding pathological images with known ground truth segmentations of the pathologies (Fig. 3.15).

The following atlas-to-patients registration experiments are performed: 1) using the original (healthy) LPBA40 images, 2) using the phantom pathological images, and 3) using the phantom pathological images combined with a weight mask. The subject which has the greatest average similarity to all remaining subjects is determined as an atlas and registered to the 39 remaining subjects. Analogously to the previous experiments, a 4-fold cross-validation over the patients is applied using 30 healthy subjects images for training (6 000 patches extracted) and the remaining nine for testing, making sure that the train and test datasets are disjoint.

3.4.3 Experiments and Results

First, the pathology detection accuracy is evaluated. Similarly to the previous experiments, an ROC-analysis is performed and Dice coefficients are calculated. Averaged over all images and folds, a Dice coefficient of $\mathbf{0.49 \pm 0.11}$ and an AUC of **0.93** are achieved. These results are comparable to the previous experiments (Tab. 3.3) and indicate a good generalization ability of the method.

The registration setups are evaluated using the given ground truth segmentations and Dice coefficients to determine the overlap of the aligned images. The results are

shown in Tab. 3.4. The Dice values, averaged over all labels and subjects, indicate that pathological structures cause a decrease in registration performance. Contrary to that, using the proposed weighting approach, the registration results improve significantly ($p < 0.001$ in a two-sided t-test) since the influence of the non-corresponding areas is suppressed.

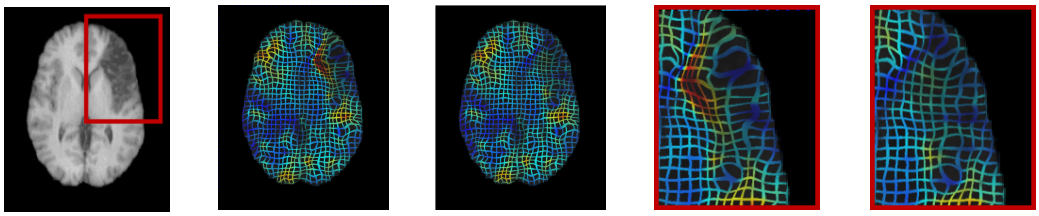
Typically, it is desirable to apply smooth deformation fields to medical images in the context of non-rigid registration. However, pathologies influence the smoothness of the resulting transformation, since missing correspondences might cause implausible distortions of the displacement field which result in higher displacement gradients. This can be observed in Fig. 3.16 where, in the area of the lesion, the displacement field is significantly distorted when no weight mask is used. When the weight mask is incorporated into the registration process, a much smoother displacement field can be achieved. For quantitative evaluation, the standard deviation of the Jacobian is used to measure the smoothness of the displacement field (as proposed in [Werner et al., 2014]). A significantly reduced Jacobian standard deviation ($p < 0.001$ in a t-test) is achieved with the application of pathology masking (Tab. 3.4). Overall, the results of the experiments show that integrated knowledge about the presence and location of pathologies can improve registration performance.

3.5 VAE-based Interpretability of Black-Box Pathology Classifiers

In the previous sections, it has been successfully shown that VAEs can be used to model the variability of healthy tissue appearance of a particular domain. In this section, this property is explored in the context of explanation of black-box classification methods for pathological images. The approach is motivated by the fact that machine learning approaches, especially neural networks, play an essential role in the medical image processing domain, since they show exceptional accuracy and generalization ability.

Table 3.4: Image registration results: Dice coefficients and standard deviation of the Jacobian for the three registration experiments. Dice coefficients are averaged over all 56 labels and all subjects. The standard deviation of the Jacobian indicates smoothness of the resulting displacement field [Werner et al., 2014] and is averaged over all subjects. The superscript (*) indicates statistical significance ($p < 0.001$) in a two-sided t-test compared to using no weights.

	no lesion	with phantom lesion	
		no weights	with weights (proposed)
Dice ↑	0.74 ± 0.01	0.72 ± 0.01	$\mathbf{0.73 \pm 0.02}^\star$
std. Jac. ↓	0.47	0.55	$\mathbf{0.37}^\star$



lesion image | no weights std. Jac. 0.46 | with weights std. Jac. 0.21 | no weights zoomed | with weights zoomed

Fig. 3.16: Registration displacement fields (atlas to patient with pathological structures). The magnitude of the displacement is color-coded as a heat map. From left to right (zoomed to the red border): a reference image with a lesion (red border); overlayed registration displacement fields without pathology masking; and with pathology masking.

However, a major problem with those approaches remains their so-called *black-box* character. Due to their data-driven nature, once neural networks are trained, it is hard to retrace what information – and how strongly weighted – affects the decision-making process in test time.

An example application of neural networks for the medical image domain is the accurate classification of healthy images and images containing pathological structures. Nonetheless, black-box classifiers are not feasible for clinical use since their decision-making process is not comprehensible by the clinician and, thus, they are harder to trust. For this reason, the need for explanation methods for black boxes such as trained neural networks is considerable.

There are several established methods that aim to gain an insight into the network's training process and learned weights like guided backpropagation (backProp) [Springenberg et al., 2015] and gradCAMs [Selvaraju et al., 2017]. A major drawback of such methods is the fact that they typically depend on the network's architecture and are mostly heuristic. A more explicit explanation approach is to find the region of an image that influences the classification result, e.g., by using perturbations [Fong and Vedaldi, 2017]. The intuition behind such approaches is that, if a region is removed from an image and the classifier's decision changes, then this region is substantial for the decision-making process of the black-box classifier (Fig. 3.17). Next to their explicity, such methods have the advantage to be model agnostic and can be applied to any classifier. However, the perturbation type is crucial to the success of the method. Replacing image values with constant values (e.g. zeros) to delete certain regions might seem intuitive, but does not always lead to the desired results.

For example, replacing the values of a brain lesion in a T1 sequence of a brain MRI by zeros, yields an appearance similar to the pathological image, hence the classifier could fail to categorize the perturbed image as healthy. Thus, perturbations need to

be a natural choice of what would replace a pathological region (e.g. healthy brain tissue).

In [Fong and Vedaldi, 2017], the topic of meaningful perturbations is addressed by proposing perturbations with naturalistic imaging effects composed from constant values, noise and image blurring. Such effects are suitable for natural images, but rather inappropriate for the medical domain, since medical images are often noisy and blurry by nature or due to artifacts. A meaningful perturbation in the context of pathology classification is the replacement of pathological tissue by its healthy equivalent (Fig. 3.17). For example, if a classifier is trained to distinguish between brain MRIs containing a lesion and healthy subjects brain MRIs, a perturbation changing the class from "pathological" to "healthy" is the replacement of the lesion tissue with healthy brain tissue.

To cope with the problem of meaningful perturbations, a VAE trained on healthy images is used, such that in the test phase the autoencoder generates the "nearest" healthy image. Using VAE-generated healthy tissue as a perturbation method [Uzunova et al., 2019b] in combination with the explanation algorithm from [Fong and Vedaldi, 2017] delivers plausible explanations for the classifier's decision.

3.5.1 Explanation of Black Boxes with VAE-based Perturbations

Explanation of Black Boxes by Perturbations: The explanation of black-box classifiers is described in [Fong and Vedaldi, 2017] as the following optimization problem:

$$\mathbf{m}^\star = \text{argmin}_{\mathbf{m}\in[0,1]^\Lambda} \lambda_1 ||1-\mathbf{m}||_1 + \lambda_2 \sum_{u\in\Lambda} ||\nabla \mathbf{m}(u)||_\beta^\beta + f_c\Big(\Phi(\mathbf{x}_0;\mathbf{m})\Big). \tag{3.4}$$

Here, $\mathbf{x}_0 \in \mathcal{X}$ is an input image, $\Phi(\mathbf{x}_0;\mathbf{m})$ is a perturbation operator that causes the "deletion" of particular regions of $\mathbf{x}_0$ by applying a multiplicative mask $\mathbf{m}$. The function $f_c : \mathcal{X} \to \mathcal{Y}$ is a black-box function, e.g., a neural network, that expects an

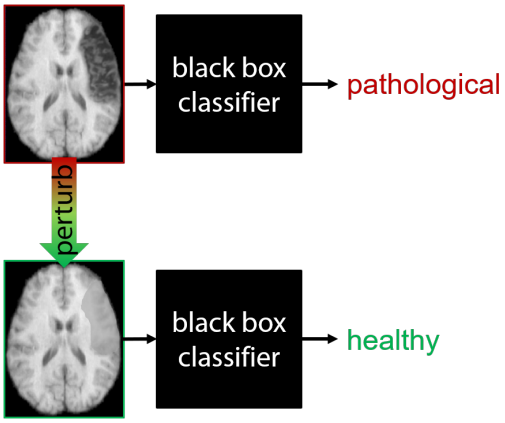


Fig. 3.17: Overview of the main idea of the perturbation approach: Fool a black-box classifier by replacing the pathological tissue (■) by healthy tissue (■).

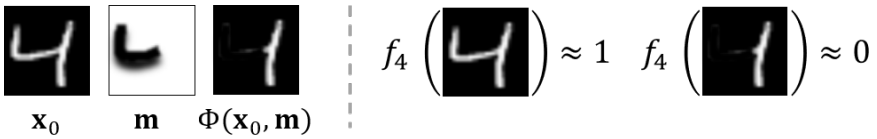


Fig. 3.18: Example of the perturbation method for a digit from the MNIST dataset [Lecun et al., 1998]. The original digit is classified as "4", when a significant region is blacked out, the classifier's result changes since the digit rather represents a "1".

image $\mathbf{x}_0$ as an input and outputs the probability $f_c(\mathbf{x}_0) \in [0, 1]$ for class $c \in \mathbb{R}^C$ of C possible classes. Intuitively, this corresponds to finding an appropriate variation $\mathbf{x} = \Phi(\mathbf{x}_0; \mathbf{m})$ of the original input $\mathbf{x}_0$, such that $f_c(\mathbf{x}) \ll f_c(\mathbf{x}_0)$.

In other words, the probability that a particular class is recognized by the classifier is minimized by perturbing the regions of the image crucial for this decision, e.g. in Fig. 3.18 a region of the digit "4" is varied such that the number looks more like a "1", thus, the probability for the class "4" would significantly decrease. To ensure that only a specific small region of $\mathbf{x}_0$ is deleted rather than perturbing the whole image, λ_1 encourages $\mathbf{m}$ to mainly contain zero values. The last part of this problem is to ensure that $\mathbf{m}$ is smooth and regular. This is done by minimizing the total variation β-norm $\sum_{u \in \Lambda} ||\nabla \mathbf{m}(u)||_\beta$ of $\mathbf{m}$ – balanced with the weight λ_2. An advantage of this minimization problem is that it can be solved with an iterative gradient descent method, e.g. Adam, and, thus, efficiently calculated on the GPU. Here, all parameters are chosen according to [Fong and Vedaldi, 2017].

The proposed deletion game concept is quite comprehensible, though, its practical realization stumbles upon the definition of the term "deletion". Formally, the deletion concept can be defined as follows. Given the deletion mask $\mathbf{m} : \Lambda \to [0, 1]$ that associates each pixel $u \in \Lambda$ with a value $\mathbf{m}(u)$, the perturbation is defined as:

$$[\Phi(\mathbf{x}_0; \mathbf{m})](u) = \mathbf{m}(u)\mathbf{x}_0(u) + \Big(1 - \mathbf{m}(u)\Big)p(u), \tag{3.5}$$

where $p(u)$ is a function delivering the replacement values of the masked pixels. An intuitive concept for deletion would be to simply replace image values by zero (or any other constant value) with $p(u) = 0$. This is not very plausible for pathologies since it would typically result in structures that still look pathological (see Fig. 3.19, left). More plausible naturalistic perturbations like blurring and noise are considered in [Fong and Vedaldi, 2017] where $p(u) = [gauss(\mathbf{x}_0)](u) + \eta(u)$ with $gauss(\cdot)$ being a Gaussian smoothing function and $\eta(u)$ being Gaussian noise. However, such perturbations are also not suitable for medical images since they are typically naturally noisy and/or blurry. Also, blurring large homogeneous regions (e.g. whole organs or large pathologies) yields unsatisfactory results (Fig. 3.19, right).

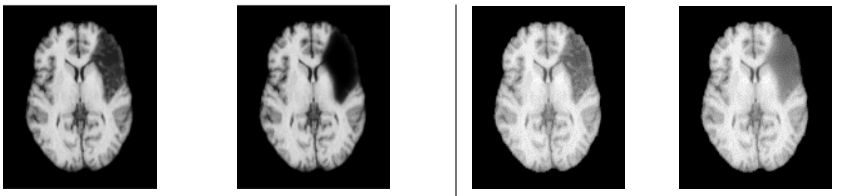


Fig. 3.19: An example of lesion perturbation by blacking and blurring. Both lead to images very similar to the original one and can still be mistakenly classified as pathological. Remark: lesion 1 and 2 are contrast-varied versions of the same lesion.

Building on that, a novel pathology "deletion" approach is proposed. Based on the previous sections, the main assumption is that pathologies can be replaced by their healthy tissue equivalent using variational autoencoders.

Perturbing with Variational Autoencoders: The proposed VAE-based perturbation can be straightforward integrated into Eq. 3.4 [Uzunova et al., 2019d], by modifying Eq. 3.5 as follows:

$$\left[\Phi(\mathbf{x}_0;\mathbf{m})\right](u) = \mathbf{m}(u)\mathbf{x}_0(u) + \Big(1-\mathbf{m}(u)\Big)\left[p_{VAE}(\mathbf{x}_0)\right](u), \tag{3.6}$$

where p_{VAE} is the trained VAE model (Fig. 3.20).

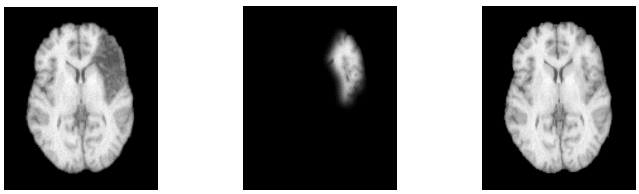


Fig. 3.20: An example of lesion perturbation by replacing it with its healthy tissue equivalent.

Analogously to the previous applications, the VAE is trained on healthy images only, hence pathological structures are interpreted as healthy tissue in the inference phase. Since a reconstruction of the entire input images is soughed, the patch-based approach is omitted. Instead, a new convolutional 2D VAE takes whole image slices as input and reconstructs the images in full size. The concrete architecture is shown in Fig. 3.21. The application of dropout layers and the size of the z-space were chosen experimentally, since they yield best reconstruction results. Similarly to Sec. 3.3.1, a standard deviation of 0.1 is used for the distribution of the latent space. The VAE is trained as usual in a batch-wise manner for a maximum of a 1 000 epochs. Early stopping is applied using 5% of the training data set for validation. The input images

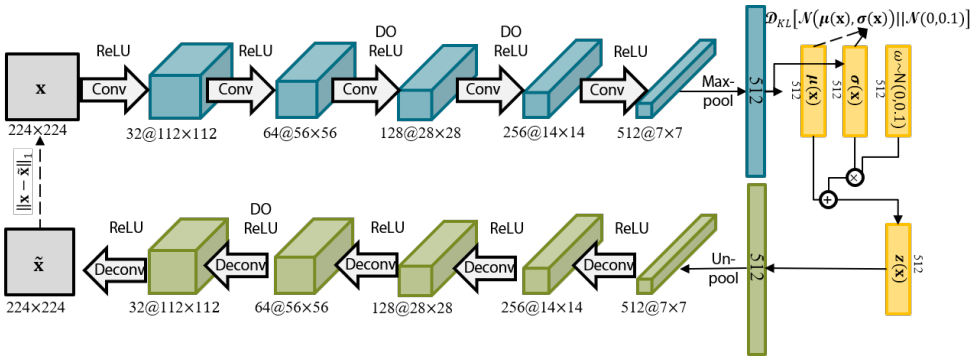


Fig. 3.21: Convolutional VAE architecture. Input image $\mathbf{x}$ and reconstruction $\widetilde{\mathbf{x}}$; convolutional encoder with fully-connected latent space (■) and decoder (■) layers; z-space (■) where $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \times \boldsymbol{\omega}$, with random $\boldsymbol{\omega} \sim \mathcal{N}(0, 0.1)$. Numbers denote size of layers (channels). Solid black lines correspond to fully-connected operations. DO stays for Dropout; ReLUs are used everywhere if not mentioned otherwise. Loss functions (dashed lines) are same as in Fig. 3.5. First two convolutions and last two deconvolutions use kernel size 5×5, the rest uses kernels of size 3×3.

are resized to 224×224 pixels and slight data augmentation using random affine transformations is applied: maximum 5° rotation, 10% translation in x and y directions and scaling in the range [90%, 110%].

3.5.2 Data and Experimental Setup

For the conducted experiments, both healthy subjects images (for training) and pathological images (for testing) of the same domain are required. Also, ground truth segmentations of the pathologies of the test images are needed to enable quantitative evaluation. The following two datasets fulfilling those requirements were chosen:

Phantom Lesion Brain MRIs: The brain MRIs with synthetically simulated lesions from Sec. 3.4.2 are also applied for the experiments of this section. Here, four different lesions are chosen which results in *40 healthy* and *160 pathological images* (examples in Fig. 3.22).

Retinal OCT: *Healthy training images* are chosen from the DUKE dataset, since it contains 115 healthy subjects retinal OCT volumes [Farsiu et al., 2014]. Selecting 60 2D slices around the center of each volume yields 6 900 non-pathological images. In addition, the images are flattened and denoised using a BM3D filter [Dabov et al., 2007] in order to match their appearance to the following pathological images. The *pathological retinal OCT* counterparts are chosen from the RETOUCH challenge dataset [3].

[3] `https://retouch.grand-challenge.org/`

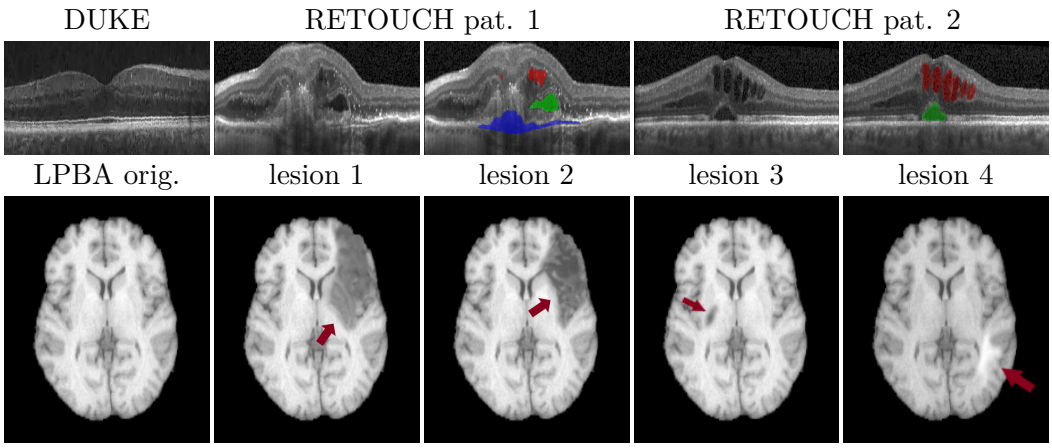


Fig. 3.22: The used datasets. First row from left to right: Retinal OCT – healthy image and two patients (pat. 1 and pat. 2) pathological images with ground truth labels IRF (■), SRF (■) and PED (■). Second row from left to right: LPBA40 – original healthy image and four simulated lesions. The hardly visible lesion 3 and lesion 4 are marked by a red arrow.

Those images contain three different pathology types with their ground truth expert segmentations: intraretinal fluid (IRF), subretinal fluid (SRF) and pigment epithelium detachments (PED). Since the images are roughly aligned and only slightly noisy, no pre-processing steps are undertaken. Furthermore, 49 2D slices are extracted per volume (examples in Fig. 3.22).

The main idea of the performed experiments is to train a black-box classifier that distinguishes between different pathology types and healthy images, and apply the proposed explanation method to judge the classifier's reliability. For this purpose global classes are extracted for the training of the classifier. For the brain MRIs, a binary classification between healthy (0) or pathological (1) is performed. For the retinal OCTs, a further differentiation between the different pathology classes is established, e.g. [0,0,0] for healthy or $[1(IRF), 1(SRF), 1(PED)]$ for all three pathologies available simultaneously. For the classification approach, a neural network with the DenseNet121 [Huang et al., 2017] architecture is selected since it shows reliable classification results for the datasets at hand. The last fully-connected classification layers are changed to fit the number of classes and a binary cross entropy loss is applied in the case of multi-label classification. The assumption that a good explanation of a pathological structure roughly corresponds to the outline of the structure is used as a basis for the quantitative evaluation. Thus, segmentation and detection evaluation strategies can be applied to assess the overlap of the explanation region and actual pathology segmentation.

For the experiments, the proposed perturbation approach [Uzunova et al., 2019b] is compared to two other perturbation strategies: using three different constant values for perturbation (0, 127 and 255) – *const* and using the method proposed in [Fong and Vedaldi, 2017] combining blurring and noise – *blur*. Furthermore, two popular heuristic explanation methods are also used as baselines: gradCAM [Selvaraju et al., 2017] and backProp [Springenberg et al., 2015].

3.5.3 Experiments and Results

The Need for Explanation Methods: The aim of the following experiment is to emphasize the general importance of explanation methods. For this purpose, a classifier is trained to distinguish between healthy and pathological retinal OCTs. However, the images are intentionally chosen from two different studies: the healthy images are acquired in the DUKE [Farsiu et al., 2014] study and the pathological images are collected from the RETOUCH challenge [Bogunović et al., 2019]. On the test dataset, the classifier reaches 100% accuracy indicating its reliability for this classification task.

The applied explanation methods reveal that, in fact, the classifier did not learn to generally distinguish healthy from pathological OCTs, but moreover between the two datasets. Since the DUKE images contain more noise in the background, the explanations of nearly all methods, are concentrated on the background, obviously used by the classifier as a distinguishing feature (see Fig. 3.23). This example underlines the necessity for explanation methods since, oftentimes, overlooked distortions and bias of the training dataset can lead to an unforeseen learning process.

Explaining Multi-label Retinal OCT Classifier: In this experiment, the classifier is trained for multi-label classification of the three conditions in the RETOUCH data. The results are assessed in a 4-fold cross-validation manner using a ROC-analysis and calculating the AUC value. On the leave-out test datasets, the classification achieves promising performance with the following AUC values: **0.96 (IRF)**, **0.98**

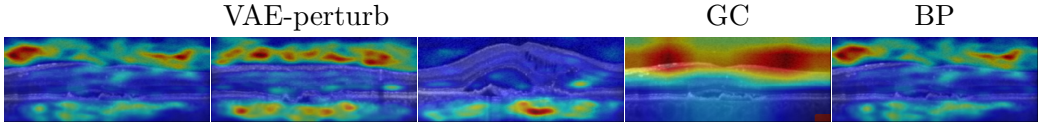


Fig. 3.23: Example for the need of explanation methods. Explanations for a classifier trained to differentiate between healthy DUKE images and pathological RETOUCH images visualized as overlayed heat maps. Shown are the proposed method (VAE-perturb) for three patients and gradCAM (GC) and guided backProp (BP) for the first patient. All explanations focus on the background and not the pathological structures.

(SRF) and **0.97 (PED)**, indicating reliability. Further, the different perturbation approaches and the baseline methods are applied to evaluate the plausibility of the learned features. Note that for the proposed VAE-based perturbation, the VAE is trained on the healthy DUKE data.

Some results of the perturbations can be seen in Fig. 3.24. The regions highlighted by the proposed approach roughly correspond to the segmentations of the respective pathological structures. Also, in the perturbed images, those pathologies are less visible. The blur perturbation also shows good results for smaller structures, but the explanations of larger pathologies are rather infeasible. Using constant perturbation values (here: 127) also yields implausible results. Intuitively, this can be explained by the need of additional knowledge of the healthy tissue's gray values on a certain position, e.g. the IRF structures of pat. 1 can reliably be perturbed by gray constant values but not by black. It also becomes obvious that gradCAM delivers particularly blurry results, connected to the fact, that gradCAM relies on the resolution of the last convolutional layer in the classification network. The least satisfying results are achieved by the backProp method, where very noisy heat maps are generated even after post-processing with a Gaussian smoothing filter.

These qualitative observations can further be confirmed by the quantitative evaluation presented in Tab. 3.5. The proposed method delivers the best results for SRF and second best results for the structures IRF and PED. Although the blurring per-

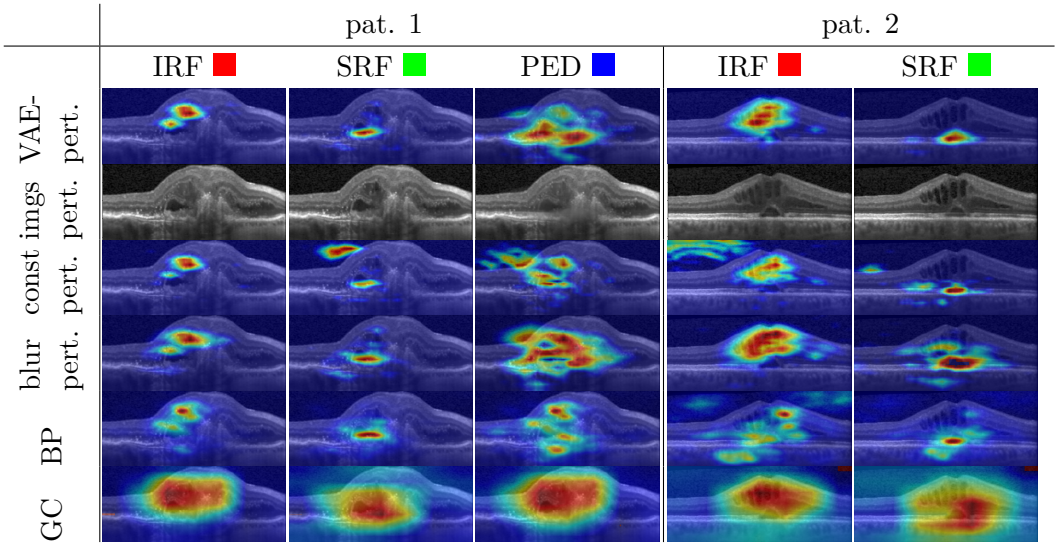


Fig. 3.24: Explanation techniques for the multi-label OCT classifier on two example patients visualized as overlayed heat maps. Row-wise: VAE-pert. – our proposed method; const pert. – perturbations with constant values; blur pert. – perturbation by blurring; BP – guided backpropagation; GC – gradCAM.

Table 3.5: AUCs (↑) of the explanation methods for each label for the multi-label OCT classifier (first three rows) and the single-label brain MRI classifier (last row). From left to right: perturbation methods: proposed VAE-perturbation, constant value perturbation (127) and blur perturbation; BP: Backpropagation; GC: gradCAM.

		explanation method				
dataset	label	VAE	const	blur	BP	GC
RETOUCH	IRF	0.89	0.85	**0.90**	0.89	0.86
RETOUCH	SRF	**0.90**	0.83	0.67	0.87	0.89
RETOUCH	PED	0.85	0.77	**0.86**	0.77	0.78
LPBA40	lesion	**0.95**	0.60	0.78	0.91	0.91

turbation approach yields the best results for two structures, it performs significantly worse on the third label, indicating a lack of generalization ability. On the contrary, the proposed method constantly performs well for the different structures and shows to be robust against different sizes and intensity values of the pathological structures.

Explaining Single-label Brain MRI Classifier: Analogously to the previous experiment, the classifier is trained on the brain MRIs to distinguish between healthy and pathological images. The results are also evaluated in a 4-fold cross-validation manner over all subjects. On the test data, the classification achieves AUC of 1, implicating a perfect classifier. However, the qualitative evaluation (Fig. 3.25) reveals, that none of the explanation methods, is able to reliably detect lesion 4, leading to the conclusion that the learned classification of those images is not necessarily correlated to the pathology presence. An intuitive explanation is that images with stronger gray value variability are classified as pathological. Similarly to the previous observations, backProp yields noisy and not class-discriminative results. Also, that the same few regions are highlighted for the different pathological structures, meaning that the network learns the possible location of the structures. This is naturally based on the small variability of lesions and images for this experiment. Although gradCAM generally detects the rough location of the pathologies, its explanation maps suffer from bad resolution.

An interesting observation is, that the proposed method only detects the structures responsible for the pathological appearance in lesion 1 and lesion 2, e.g. brain furrows have typically lower intensities in healthy brains and, thus, those areas do not affect the explanation. Tab. 3.5 (last row) also shows the quantitative results for this experiment. Since, generally, none of the methods explain the classification of lesion 4 reliably, it is not included in the calculations. It can be observed that the presented method delivers the largest conformity of explanations and actual pathology segmentations.

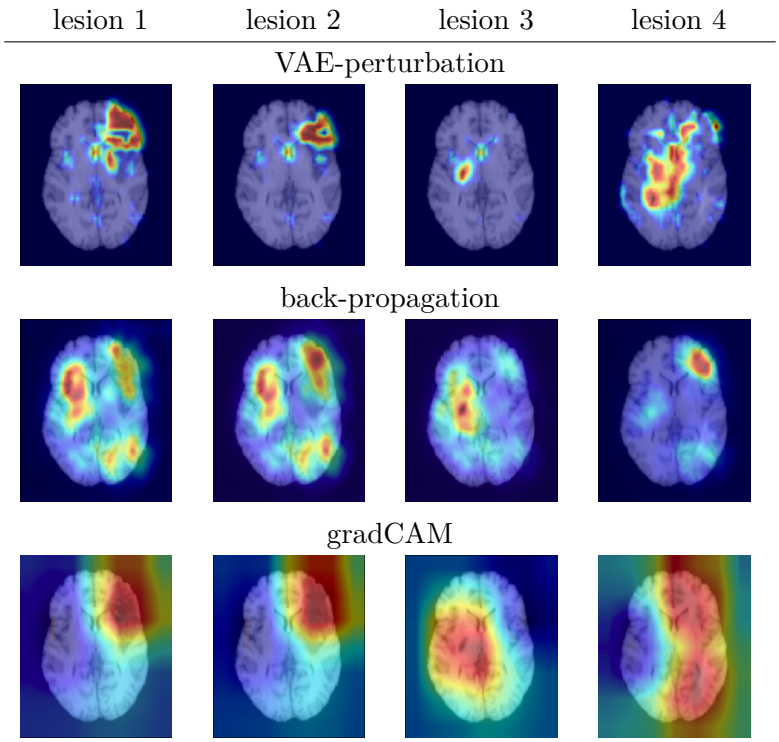


Fig. 3.25: Different explanation techniques for the lesion classifier. First row: proposed method for all four lesions; second row: back-propagation; third row: gradCAM.

The proposed VAE-based perturbation once again shows to be reliable and robust compared to other methods.

Based on these results, it can be noticed that such explanation methods can also be considered semi-supervised pathology segmentation approaches given a reliably trained classification neural network. However, a noteworthy drawback of the proposed perturbation method, is the fact that only different pathology classes can be explained. Thus, the lack of pathological structures, in other words, the classification as healthy, cannot be explained. This is typical for all presented explanation methods, since an image can only be classified as healthy by observing the whole image, while a pathological structure is typically limited to a small region.

3.6 Discussion and Conclusion

Pathological structures in medical images tend to have a large variety of different shapes and appearances, thus, it is hard to straightforward learn their natural distribution. A good approximation can be achieved by training supervised deep learning methods on large annotated datasets. Since these are hardly available in the medical

image domain, here, a method that is able to approximately learn the healthy tissue variability of images of a certain domain is developed. Following an exclusion logic, a pathological structure can then be recognized as tissue that does not fit into the learned healthy norm. To learn the healthy variability from a dataset, the proposed approach specifically concentrates on the usage of VAEs since they are not only able to reconstruct healthy images, they also learn their underlying distribution in a low-dimensional space. So, pathological structures can be detected as structures that lie outside the learned latent distribution and also get poorly reconstructed by a trained VAE. In fact, experiments show that these assumptions are not only reasonable by themselves, their combination also has a high potential. To be able to calculate the latent distances in a voxel-wise manner and also process large 3D medical volumes, a novel patch-based cVAE approach considering positional information is proposed.

The presented unsupervised pathology modeling approach is used in three very different setups to underline its various applications. The first and most intuitive application is to use the method for the segmentation and detection of pathologies in medical images. Generally, a rough segmentation of the pathologies is possible, as shown in the experiments performed on a multitude of datasets. The proposed method performs better than a GAN-based unsupervised approach and the importance of the positional conditions and the patch-based design are clearly shown in the experiments. However, typical for an unsupervised approach, the predicted segmentations are not exceptionally accurate making the method rather unsuitable for pixel-level precise segmentation. Nonetheless, other applications like pre-processing for further automatic image analysis methods for pathological images are conceivable. Therefore, in a second experiment, the application of the rough pathology detection as a-priori information in a pathological image registration scenario is explored. Since pathological structures typically cause missing correspondences, the registration of pathological images to a given healthy atlas yields bad results. These results can be significantly improved by applying a weight mask obtained from the proposed approach to scale the objective of the registration method. This leads to the conclusion that even a rough pathology detection is beneficial to the registration of images with missing correspondences.

In a third application setup, the capability of the proposed approach to model the healthy variability of images is further exploited. Specifically, a VAE-based perturbation approach is developed for the interpretability of pathology classification black-box methods. Among many established methods, the proposed approach delivers the most plausible, robust and class-discriminative explanations and shows its versatile applications for a wide range of medically relevant applications. Also, the explanations of a reliable classifier can be observed as a weakly supervised pathology detection approach, since a good segmentation of the pathological structures is established.

The unsupervised pathology modeling approach shows overall convincing results, still it can be observed that the generated images are of rather blurry nature, the heat maps are partly unspecific and pathological structures are roughly localized. This

might be sufficient for many applications, however, the rapid development of deep learning approaches urges for more accurate modeling approaches and direct involvement of pathological structures into the training process of neural networks. These requirements are therefore addressed in the approach proposed in the next chapter of the work, where a direct modeling of pathological structures is obtained.

Chapter 4

Generative Adversarial Networks for the Synthesis of Full-resolution Medical Volumes with Pathological Structures

In order to reduce the amount of needed training images for neural networks and especially to minimize the manual annotation effort of relevant anatomical and pathological regions, a GAN-based approach for the generation of synthetic annotated images featuring pathologies is developed here. The presented approach consists of two main methods that overcome significant practical hurdles. 1) A method for the translation from healthy to pathological images by keeping the shape of the healthy structures and, thus, be able to use the available annotations. A pathological appearance is achieved by explicitly integrating tumor tissue and learning the appearance of the pathological image domain. 2) Overcoming the immense computational resource requirements of GANs by developing an approach for the efficient generation of realistic high-resolution 3D medical volumes using only a fraction of the previously required memory. At the end of this chapter, thorough experiments prove the general capability of the method for the generation of realistic healthy images and especially underline the possibility to generate annotated data containing pathologies for e.g. augmentation purposes.

4.1 Introduction and Motivation

The general motivation of the previously presented methods is based on the assumption, that medical images with ground truth segmentations of pathological structures are rarely available in large quantities. For this reason, an indirect modeling of pathological structures deduced from the learned variability of normal tissue was proposed. However, many applications require direct modeling of pathological structures, e.g. for

the training of neural networks, a training dataset containing exact pathology segmentations is crucial. In this chapter, an explicit pathology modeling approach is presented for the generation of pathological images containing annotations of both the pathological and the healthy anatomical structures. Analogously to the previous approach, the image generation is based on the availability of datasets of healthy images, which also frequently contain segmentations of normal anatomical structures. The main idea is to keep the topology of the healthy subject images preserved in order to directly apply the given segmentations. However, the appearance of the images is modified to fit the appearance of pathological data and pathological structures such as tumors are modeled and injected explicitly. Thus, ground truth annotations of the healthy tissue can be preserved next to the labels of the modeled pathological structures obtaining a synthetic dataset with high quality ground truth annotations.

A prominent example for the use of synthetically generated pathological images in medical image processing is data augmentation and generation of datasets for the training of machine learning approaches. For example, in [Frid-Adar et al., 2018, Xing et al., 2019] images containing pathologies are synthesized to boost the performance of the segmentation and classification of specific pathological structures. In [Shin et al., 2018], tumors are simulated on healthy-appearance brain MRIs, leading to an improved performance of a tumor segmentation network. In [Waheed et al., 2020], a GAN is used to synthesize chest X-rays and use them additionally to a dataset of real images in order to boost the detection of Covid-19 biomarkers in chest X-rays. Those methods address certain pathology types and require ground truth annotations of the pathologies, yet image analysis tasks targeting normal anatomical structures can also be strongly influenced by the presence of pathological structures [Kwon et al., 2014]. For example, the evaluation of automatic image processing methods on images with abnormal structures could strongly profit from pathological data with annotated healthy structures. Typically, many image processing approaches are developed for the normal anatomy, e.g. segmentation of healthy anatomical structures or atlas-to-image registration. However, images containing pathological structures are common in the clinical context, therefore it is crucial to estimate the accuracy of the existing algorithms for such data. Yet, without suitable annotations, a quantitative evaluation is infeasible.

Furthermore, machine learning approaches targeting normal anatomical structures on pathological data (e.g. brain ventricle segmentation of brain MRIs with tumors) need suitable training data. Namely, images containing ground truth annotations of both the normal and the abnormal structures are required for the training of approaches engaging with this task. Unfortunately, most of the large publicly available datasets containing some type of pathologies are commonly designed for the segmentation (detection/ localization) of the particular pathological structure and only contain expert segmentations of the latter, e.g. [Menze et al., 2015]. Contrary to that, datasets con-

taining ground truth annotations of normal anatomy (as used for e.g. atlas generation) are usually generated from healthy populations [Shattuck et al., 2008].

Since the generation of realistic medical images is pursued for the fulfillment of those requirements, the approach presented in this chapter is based on the state-of-the-art image generation neural networks GANs. GANs so far have shown exceptional achievements in the field of photo-realistic images [Isola et al., 2017, Karras et al., 2019, Emami et al., 2021]. First applications to medical images [Shin et al., 2018, Frid-Adar et al., 2018] also show promising results. In this chapter, an approach for the generation of medical images with ground truth annotations of pathological and normal structures is proposed. In the first step, a method for the topology-preserving healthy-to-pathological image translation is outlined [Uzunova et al., 2019c]. Based on the fact that GANs have high computational memory requirements, a second step features the development of a novel memory-efficient approach which is more suitable for 3D medical volumes of large sizes [Uzunova et al., 2019a, Uzunova et al., 2020b].

4.2 Healthy-to-Pathological Image Generation using GANs

The basis of the following method is the assumption that two datasets of different appearances are given: 1) a dataset of healthy images containing ground truth annotations of certain anatomical regions; 2) a dataset of pathological images, possibly with given segmentation masks of the pathological structures. Given that, the aim is to evaluate an algorithm targeting the normal structures in the pathological images, or, furthermore, to train a supervised approach targeting the normal structures on pathological images. To create a suitable dataset addressing this problem, an algorithm that keeps the anatomical labels of the healthy images, but recreates an appearance similar to the pathological dataset is necessary. Or, in other words, a topology-preserving domain translation approach needs to be obtained. Since the availability of paired datasets is not given, the preserving of the source topology is challenging.

In the following methods, conditional GANs (cGANs) as presented in Sec. 2.4 are considered since they are often applied for various domain translation tasks especially in the field of computer vision.

4.2.1 Unpaired Unsupervised Domain Translation

When it comes to unpaired domain translation, the most prominent architecture is the so-called CycleGAN [Zhu et al., 2017] schematically shown in Fig. 4.1. Given the healthy domain $\mathcal{H} \subseteq \mathbb{R}^d$ and the pathological image domain $\mathcal{P} \subseteq \mathbb{R}^d$, the CycleGAN architecture has two generators ($g_{\mathcal{H}\to\mathcal{P}} : \mathcal{H} \to \mathcal{P}$ and $g_{\mathcal{P}\to\mathcal{H}} : \mathcal{P} \to \mathcal{H}$) translating respectively from healthy to pathological and vice versa. The discriminators ($d_{\mathcal{H}}$ and

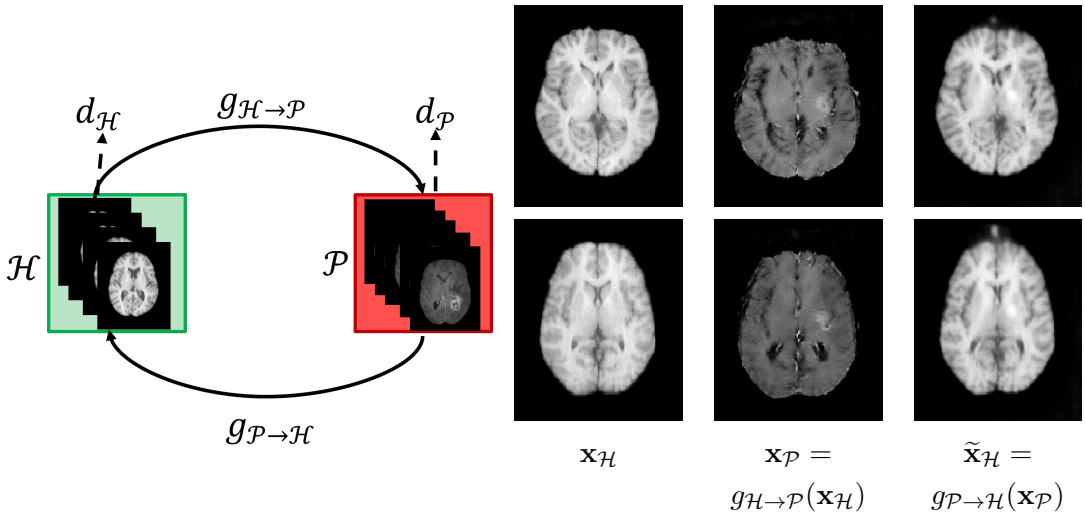


Fig. 4.1: Left: schematic display of the CycleGAN architecture. Two generators $g_{\mathcal{H}\to\mathcal{P}}$ and $g_{\mathcal{P}\to\mathcal{H}}$ translate from a healthy (■) to a pathological (■) domain and vice versa. The discriminators $d_{\mathcal{H}}$ and $d_{\mathcal{P}}$ are responsible for the realistic appearance of healthy and pathological images respectively. Right: results of CycleGAN. Original healthy images; images translated to the pathological domain; synthetic pathological images translated back to the healthy domain.

$d_{\mathcal{P}}$) additionally encourage the generation of images that are indistinguishable from the given target domain. Furthermore, the domain mappings are regularized by a cycle-consistency loss, that objectifies the intuition that if an image is translated to another domain and back, the resulting image should be similar to the starting image. Based on this intuition, the cycle-consistency loss enforces that $g_{\mathcal{P}\to\mathcal{H}}(g_{\mathcal{H}\to\mathcal{P}}(\mathbf{x}_{\mathcal{H}})) \approx \mathbf{x}_{\mathcal{H}}$ and $g_{\mathcal{H}\to\mathcal{P}}(g_{\mathcal{P}\to\mathcal{H}}(\mathbf{x}_{\mathcal{P}})) \approx \mathbf{x}_{\mathcal{P}}$, where $\mathbf{x}_{\mathcal{H}} \in \mathcal{H}$ and $\mathbf{x}_{\mathcal{P}} \in \mathcal{P}$. Note that the network's trainable parameters are omitted here to facilitate readability. Due to the successful application of CycleGAN for medical images in previous works [Armanious et al., 2019, Zhou et al., 2020], it is an intuitive method for the given unpaired domain translation task.

In the following, some preliminary experiments are carried out to determine whether this approach is suitable for the task of topology-preserving unpaired domain translation. Two datasets of brain MRIs are representative for the healthy (LPBA [Shattuck et al., 2008]) and pathological (BRATS [Menze et al., 2015]) domains. The data have been thoroughly described in Sec. 3.3.2 and Sec. 3.4.2. Here, only 2D slices are used for the purposes of a toy example. Some results of these experiments are displayed in Fig. 4.1. Overall, CycleGAN produces realistic images, however, several disadvantages for the desired application become apparent. A major disadvantage is, that, since the shape of the input images is not explicitly modeled and there are no topology constrains, the shape of the images is not entirely preserved (e.g. ventricles and head

outline in Fig. 4.1). Furthermore, it can be observed that the variability of the generated pathologies is fairly small, since they are also generated heuristically. This also leads to minor pathology hallucinations in the reconstructed healthy images, typical for CycleGANs [Cohen et al., 2018]. Since this method does not guarantee a preserving of the source topology, further approaches are examined.

4.2.2 Unpaired Topology-preserving Domain Translation

Another cGAN-based domain translation approach is the so-called pix2pix [Isola et al., 2017]. This method learns the translation from one domain into the other in a paired manner. As typical for cGANs, the pix2pix discriminator also takes a pair of images rather than a single image and determines whether the presented pair is real or synthetic (see Sec. 2). Thus, the objective of pix2pix can be formulated as:

$$\min_{\theta} \max_{\xi} \mathbb{E}_{\mathbf{c},\mathbf{x}} \left[\log d_{\xi}(\mathbf{c}, \mathbf{x})\right] + \mathbb{E}_{\mathbf{c},\mathbf{z}} \left[\log(1 - d_{\xi}(\mathbf{c}, g_{\theta}(\mathbf{c}, \mathbf{z})))\right], \tag{4.1}$$

where $\mathbf{c} \in \mathbb{R}^d$ is a conditional image and $\mathbf{x} \in \mathbb{R}^d$ is a real image from the target domain.

An obvious disadvantage of this architecture is that it requires paired image data, e.g. different brain MRI sequences of the same subject as in [Armanious et al., 2020]. Contrary to that, no paired images are available in the considered scenario. For this reason, a trick for customizing the pix2pix approach for unpaired image translation is developed here [Uzunova et al., 2019c]. In the original pix2pix work, the authors show that it is possible to translate from intensity-independent shape information to a realistic appearance conforming to the given shape [Isola et al., 2017]. For example, a realistic building appearance can be reconstructed from simple segmentation masks of the building; or a realistic cat photography can be synthesized from a sketch of the cat. Since the shape information and the appearance image are strictly paired, the training procedure is considered supervised. Fortunately, shape information such as sketches can simply be extracted from images using e.g. a Canny filter. Such operations can be considered as "free of charge" due to their simplicity in contrast to pixel-wise expert annotations.

The main hypothesis here is that if an image-to-image translation network is trained for the generation of pathological images from given sketches, then in the test phase, the sketches of arbitrary images (of the same scene) can be translated to the learned pathological domain. Hence, healthy images can be translated into a pathological image domain. To explicitly model the pathological structure, its contours can be integrated into the test input sketch. An example of this pipeline based on tumor brain MRIs is shown in Fig. 4.2. During the training, the contours of real tumor MRIs are used as input and their corresponding appearance is learned. In the inference phase, the contours of healthy subject brains are extracted and combined with the contours of a tumor. Using such sketches as inputs to the trained GAN results in realistic

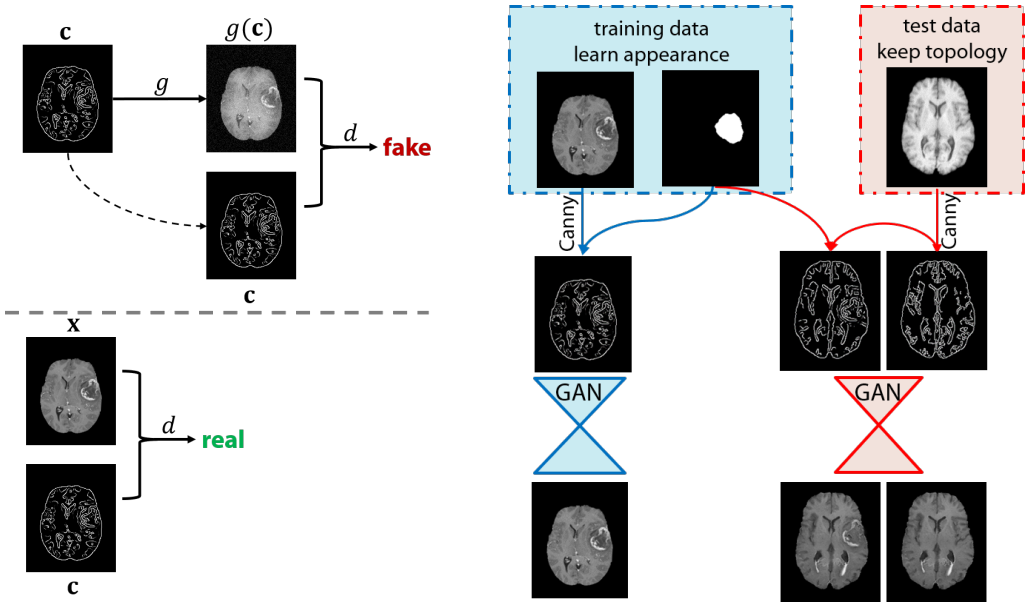


Fig. 4.2: Left: schematic display of the training of the pix2pix method, adapted for the application on medical images. Right: training (■) and testing (■) scheme for unpaired domain translation.

images containing tumors. Note that the intensity profile of the target domain is also mimicked.

4.2.3 Concept Analysis

In order to assess the general suitability of the proposed method for the unpaired topology-preserving domain translation, some preliminary experiments are conducted. For this purpose, a pix2pix-based network is trained on the brain tumor T1c-weighted MRIs from the BRATS dataset with Canny-extracted edges and the overlayed tumor outlines as inputs. Again, only 2D slices are considered here. In the original architecture [Isola et al., 2017], a U-Net [Ronneberger et al., 2015] generator is applied, while the chosen ResNet blocks [He et al., 2015] deliver better results here. A further observation is that the method profits from sketches weighted by the gradient magnitude of the input images, rather than binary edges. For this reason, all further experiments consider magnitude-weighted sketches. For testing, 40 LPBA healthy image slices with given ground truth anatomical annotations and 4 pathologies from the BRATS dataset are used. Five different synthetic images per real healthy image are generated by inserting each of the four different pathologies and also considering no pathological structures.

In these first experiments, this approach shows to be suitable for unsupervised domain translation as can be seen in Fig. 4.3 (middle). The healthy appearances are

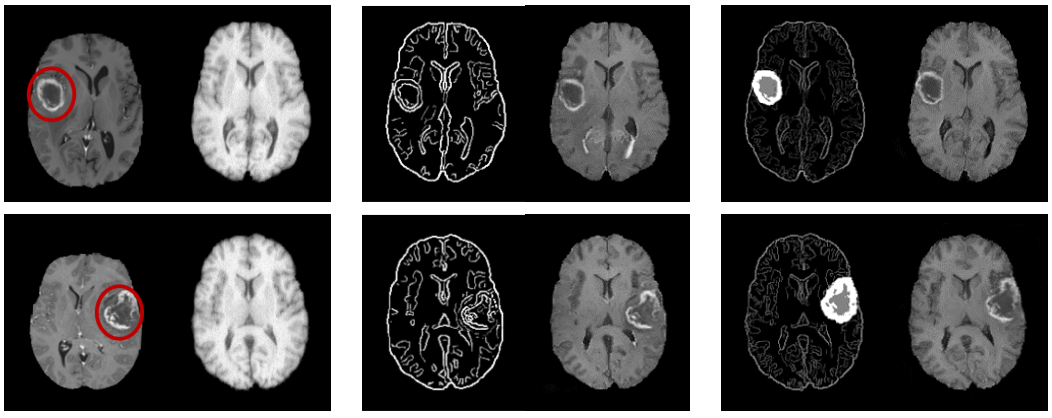


Fig. 4.3: Examples of the pix2pix unpaired domain translation approach. The tumors (red circle) are applied to the sketches of the healthy images as overlayed contours (implicit approach) or masks (explicit approach).

successfully translated into a pathological appearance and as far as visual evaluation can assess, the topology of the input image is preserved. However, a typical image translation problem can be observed in the generated images. Some structures, e.g. the ventricles have intensity values imitating contrasting tumor tissue. This so-called feature hallucination [Cohen et al., 2018] appears due to the non-explicit distinction between healthy and pathological tissue in the intensity-independent sketches. For this reason, an explicit separation of healthy and tumor tissue is required. Since the segmentation masks of the tumors of the BRATS datasets are available, those are used for explicit tumor modeling and overlayed over the contours of the healthy images. This leads to significantly less feature hallucination as seen in Fig. 4.3 (right-hand side).

An important assumption about the presented method is that the topology of the input healthy images is preserved during the generation process. While visual evaluation confirms this intuition, a quantitative evaluation is important in order to objectively assess this main assumption.

To prove whether the source topology can be preserved, allowing for a direct transfer of the ground truth anatomical labels of the LPBA dataset, the edges of the synthetic images are extracted using a Canny filter. This enables comparing the edges of an input healthy image and its corresponding translated image, naturally excluding the pathological area. The average symmetric contour distance (ASCD) between such contour pairs is calculated as the mean over all images. A mean ASCD of $\mathbf{0.58 \pm 0.007}$mm is achieved, which indicates sub-pixel accuracy and adequate topology preserving. The worst (ASCD 0.83mm) and the best result (ASCD 0.46mm) are shown in Fig. 4.4. On both images, the transferred labels seem to overlap with the corresponding anatomical

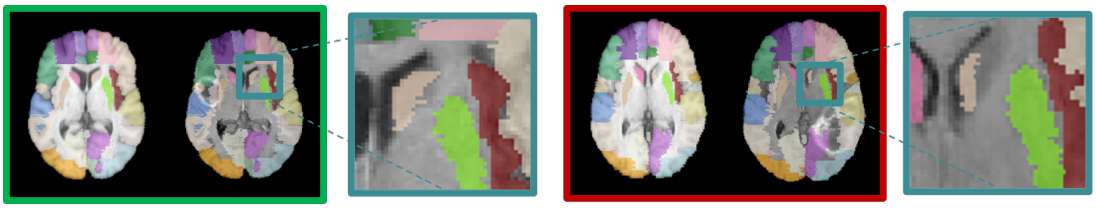


Fig. 4.4: An example of two translated images with overlayed ground truth segmentations from the healthy images. In the zoomed regions, the labels seem to fit the anatomical structures. Shown are the best (**0.46mm**) and worst (**0.83mm**) topology-preserving results.

regions. As baseline, the ASCD between each healthy image and its best corresponding image from the BRATS dataset is calculated yielding $\mathbf{1.75 \pm 0.13}$mm. Obviously, this baseline cannot serve as a direct comparison, but shows that the ASCD values resulting from the domain-translated images are clearly in favor of the topology preserving assumption. In summary, the established unpaired topology-preserving domain translation is suitable for the generation of realistic pathological images and the direct transfer of the normal anatomical annotations is ensured. Here the GAN architecture is applied only for 2D images, since GANs and especially cGANs require immense computational resources, typically GPU RAM. With images of size 181×217 pixels, the presented architecture is already on the verge of the possibilities of typical consumer hardware. So, 3D image generation is still infeasible using this approach and is, therefore, explored in the further sections of this chapter. Another issue leading to an unnatural appearance of the generated images is the fact that simply overlaying the pathological structures over the healthy tissue does not consider possible pathology-induced deformations of the surrounding anatomical structures. Tissue deformations around brain tumors are known as the mass effect and are a particularly common phenomenon. This problem is also coped with in the course of this work.

4.3 MEGAN: Memory-efficient GAN for the Generation of High-resolution 3D Medical Images

GANs have shown their ability to generate images of exceptional quality, especially in the computer vision field [Wang et al., 2018, Karras et al., 2019, Jam et al., 2021]. Next to the previously presented applications, GANs could boost many medical image processing applications, e.g. by generating data for augmentation purposes [Uzunova et al., 2017, Frid-Adar et al., 2018], image reconstruction [Yang et al., 2018], or domain translation for multi-modal data [Lei et al., 2019]. In recent years, the progress of GANs for the generation of high-resolution images has been developing rapidly. For example, in [Karras et al., 2018], the authors develop a training procedure that starts

off generating a small resolution of the image and successively adds more details until the highest resolution is reached. Highly detailed images of sizes up to 1024×1024 pixels are achieved by this approach. In [Wang et al., 2018], even larger images are generated by a succession of two networks: one for the generation of low-resolution images and one for the refinement to high-resolution images. However, those methods are applied to 2D images only and already require enormous GPU RAM capacities (16 respectively 24 GB) leading to the conclusion that larger images or image volumes would require special and expensive hardware. Medical images typically feature an enormous amount of voxels: LPBA [Shattuck et al., 2008] $181 \times 217 \times 217$; BRATS [Menze et al., 2015] $155 \times 240 \times 240$; COPDgene [Castillo et al., 2013] $512 \times 512 \times > 100$; VISCERAL [Jimenez-del-Toro et al., 2016] $> 800 \times 512 \times 512$. Thus, GAN-based methods aiming for the generation of medical images usually do not consider full-resolution volumes. In [Shin et al., 2018], the authors consider images downscaled to the half of their size ($128 \times 128 \times 54$) due to computational restrictions. In another attempt [Yu et al., 2018], an image size of 127^3 voxels is achieved, which still does not satisfy the requirements for most medical datasets. A common approach to overcome such computational restrains is to use patch- or slice-based techniques [Chen et al., 2018, Lei et al., 2019, Zhou et al., 2020], that unfortunately often lead to artifacts on the non-continuous transitions of the patches/slices. Most commonly, this issue is avoided by sampling overlapping patches and averaging the values in the overlapping regions like in the patch-based approach presented in Sec. 3, which typically causes blurry and less detailed images.

One possibility to prevent inconsistencies between patches is to additionally observe their neighborhood. In [Kamnitsas et al., 2016], the authors propose such an approach for the segmentation of brain MRI volumes. However, this approach cannot be applied for unpaired image generation and its effectiveness is limited for images of size radically exceeding the chosen patch size. In order to address these issues, a novel multi-scale patch-based GAN with constant memory usage regardless the image size is introduced here [Uzunova et al., 2019a, Uzunova et al., 2020b]. Due to a sophisticated multi-scale approach, the patch generation is context aware and does not introduce any artifacts between the patches. This approach is further referred to as *MEGAN* (Memory-Efficient GAN).

In order to underline the importance of a memory-efficient GAN training method for large medical volumes, multiple popular methods are compared in terms of their GPU RAM requirement for different image sizes. The results of this examination can be seen in Fig. 4.5. The three approaches DCGAN [Wu et al., 2016], Pix2Pix [Isola et al., 2017] and PGGAN [Karras et al., 2018], are state-of-the-art GAN-based approaches and have proven to produce good results for small image sizes. Here, the approaches are straightforward adapted for 3D images (replacing 2D convolutions by 3D convolutions etc.) using their openly available source code. Their memory requirement for one forward-backward pass and storing the image data on the GPU RAM for a batch size

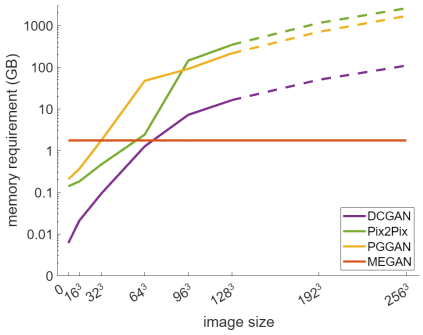


Fig. 4.5: GPU RAM requirements for 3D GANs. State-of-the-art approaches: DCGAN [Wu et al., 2016], Pix2Pix [Isola et al., 2017] and PGGAN [Karras et al., 2018]. The proposed MEGAN has constant memory requirements [Uzunova et al., 2019a]. Dashed lines: cubic regression. Log-scaled y-axis.

of one is calculated. In Fig. 4.5, it can be observed that their memory demand grows cubically with respect to an isotropic side size of a volume. Calculations for image sizes over 128^3 are already impossible on the available hardware (Titan XP 12GB GPU), yielding the high infeasibility for straightforward 3D GANs. Contrary to that, the proposed approach copes with this problem by only requiring constant GPU RAM regardless the image size due to the generation of image patches (Fig. 4.5).

4.3.1 Multi-scale Patch-based GANs

The concept of using multiple resolution scales to achieve better image quality have been successfully implemented in the history of image generation. In [Denton et al., 2015], the authors use a training approach inspired from the Laplacian pyramid, applying a separate GAN for each pyramid level and are able to generate detailed naturalistic images. Recently, [Karras et al., 2018] proposed a new progressively growing training procedure for GANs, where a GAN is trained on increasing image resolutions successively. This innovative method delivers impressive results and is further pursued in [Karras et al., 2019, Karras et al., 2020] where only two resolutions are needed to achieve an even better image resolution.

Using multiple resolution scales is intuitive since, in this manner, a complex task is decomposed in several simple tasks. The first step involves learning the global information in a very low resolution and each further step features resolution refinement by taking into account the available global information from the previous steps. However, the methods mentioned above require the propagation of the whole full-resolution input image trough the network at some point of the training which increases the memory demand significantly. Even storing a large 3D volume on a GPU might be a challenge, hence the forward and backward pass through a network further exacerbate the problem. Thus, a method which is able to generate full-resolution image volumes without the excessive memory requirement for storing and propagating them is proposed here [Uzunova et al., 2019a, Uzunova et al., 2020b].

The basis of the presented approach is to first generate the whole image in a low resolution on the first scale s_0 by using a so-called low-resolution GAN (LR GAN). To

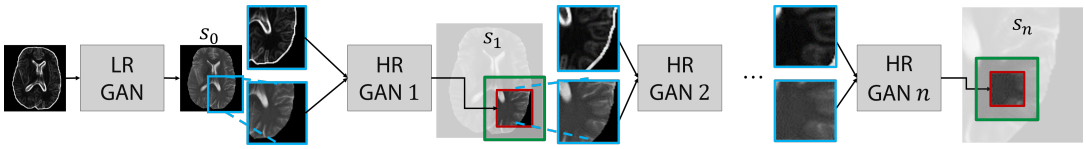


Fig. 4.6: Overview of the presented multi-scale patch-based approach. The whole image is first generated with a low resolution (LR) by an LR GAN. Then the resolution is increased subsequently across the following scales using high-resolution (HR) GANs. Blue border: an input patch at the current scale; red border: patch of the previous scale; green border: reception field of generated patch.

reach the last resolution scale s_n, a succession of n conditional high-resolution GANs (HR GANs) is applied. Each HR GAN $i, i > 0$ is trained in a patch-wise manner such that a patch from the previous scale s_{i-1} is used as an input. The HR GAN outputs a patch at scale s_i of the same size representing the center of the low-resolution patch. Thus the patch on scale s_i shows a smaller portion of the final image than the patch on scale s_{i-1} but has a higher resolution. For example, if an input patch of size 32^3 is chosen as input for HR GAN 1 (Fig. 4.6 blue patch at s_0), HR GAN 1 generates a patch of size 32^3 as well (Fig. 4.6 red patch at s_1) which only represents a sub-region of the input patch. On this scale the input patch would have the size 64^3 (Fig. 4.6 green patch at s_1). Using this scheme enables the propagation of the global information from s_0 up to the last resolution. Furthermore, the generation of a sub-region of the input patch at each scale ensures that the neighborhood information is considered and inconsistencies between the patches and the border regions can be prevented.

In order to involve the style transfer approach presented in Sec. 4.2.2, an additional conditioning on the image sketches is implemented. Hence, the LR GAN receives the lowest resolution image edges and generates the first image scale; the HR GANs receive the sketches of the image patches of the corresponding scales additionally to the low-resolution input.

In formal terms, the objective of Eq. 4.1 can be modified as follows: for the conditional sketch images $\mathbf{c}_0 \dots \mathbf{c}_n$ on resolution scales $s_0 \dots s_n$, the outputs $\mathbf{x}_0 \dots \mathbf{x}_n$ are generated using the generators $g_0 \dots g_n$ and discriminators $d_0 \dots d_n$ with the training objectives:

$$\begin{aligned}
\min_{g_0} \max_{d_0} \mathcal{L}_{cGAN}(g_0, d_0) =& \mathbb{E}_{\mathbf{c}_0, \mathbf{x}_0} \left[\log d_0(\mathbf{c}_0, \mathbf{x}_0)\right] + \mathbb{E}_{\mathbf{c}_0, \mathbf{z}} \left[\log\left(1 - d_0\Big(\mathbf{c}_0, g_0(\mathbf{c}_0, \mathbf{z})\Big)\right)\right] \\
\min_{g_i} \max_{d_i} \mathcal{L}_{cGAN}(g_i, d_i) =& \mathbb{E}_{\mathbf{c}_{p_i}, \mathbf{x}_{p_i}, \mathbf{x}_{p_i-1}} \left[\log d_i(\mathbf{c}_{p_i}, \mathbf{x}_{p_i-1}, \mathbf{x}_{p_i})\right] + \\
& \mathbb{E}_{\mathbf{c}_{p_i}, \mathbf{x}_{p_i-1}, \mathbf{z}} \left[\log\left(1 - d_i\Big(\mathbf{c}_{p_i}, \mathbf{x}_{p_i-1}, g_i(\mathbf{c}_{p_i}, \mathbf{x}_{p_i-1}, \mathbf{z})\Big)\right)\right].
\end{aligned} \tag{4.2}$$

Here, $\mathbf{c}_{p_i}$ and $\mathbf{x}_{p_i}$ are patches from $\mathbf{c}_i$ and $\mathbf{x}_i$, respectively, with $i \in [1, n]$ being the current scale number. Note that the training parameters θ and ξ are omitted for better readability.

4.3.2 Concept Analysis and Parameter Tuning

In order to verify the assumption, that the proposed MEGAN is suitable for the generation of large 3D medical image data and no patch-artifacts appear, a few preliminary experiments are performed. To prove the suitability for exceptionally large medical volumes, the GAN-cascade is trained on thorax CT images of size 512^3 voxels. For this illustrative example, the extracted edges of the thorax CTs are used as input sketches. The training is performed on overall four scales: the LR GAN delivers images of size 64^3 and each of the three following HR GANs doubles the image size until 512^3 voxels are reached. In Fig. 4.7 an example result that demonstrates the resolution enhancement on each scale is shown. Also, a comparison to a naive patch-based approach using averaging of the overlapping regions is established. This corresponds to using the HR GAN for the highest scale, but training it without a low-resolution conditioning. This yields blurry results and some patch artifacts are visible. However, the images generated by the proposed approach are of resolutions comparable to the original image. No visible patch artifacts are available in any of the successive scales. Furthermore, MEGAN requires a constant memory amount regardless the image size

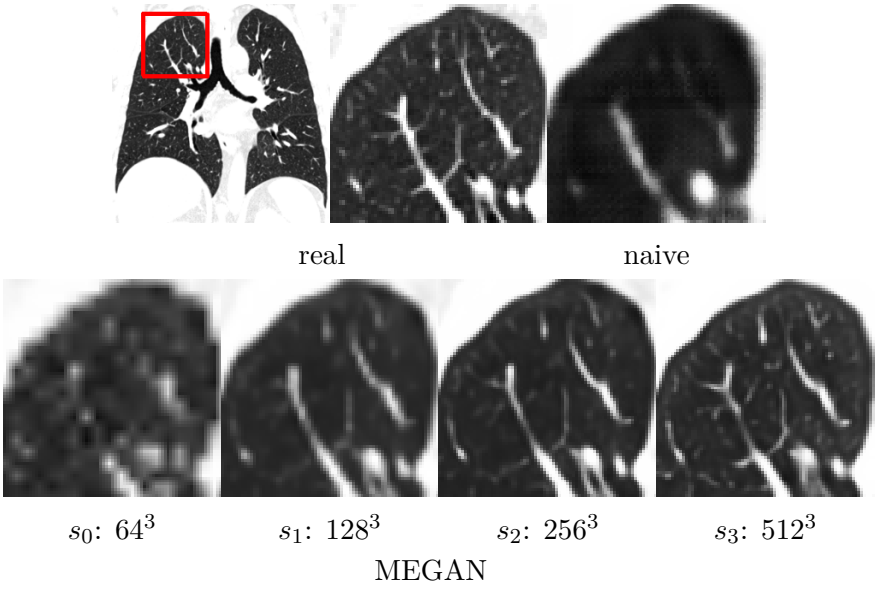


Fig. 4.7: Thorax CT image generation. First row: real image and a zoomed-in (red patch); a naive patch-based approach; Second row: the four scales of the proposed approach MEGAN – no patch artifacts visible.

since only patches of same sizes are considered. This is a major advantage and enables the generation of arbitrary large medical image volumes of high quality for the first time.

A general problem occurring with cascading approaches like this is the fact that they are prone to propagating errors from lower scales to the higher ones. This issue is addressed by applying data augmentation approaches to the low-resolution conditions of each HR GAN. Randomly applied Gaussian noise, Gaussian blurring and resolution variation ensure that the GANs do not over-adapt to their input and are able to generalize well. To demonstrate this improvement, the proposed method is applied on the previously mentioned brain tumor MRIs from the BRATS dataset, since artifacts tend to appear in abnormal tissue like brain tumors. In Fig. 4.8 an example of artifacts appearing in the tumor tissue on scale two and their prevention on scale three is shown.

The last investigation is designed to prove the assumption that both multi-scale and edge-information are necessary for each HR GAN. The intuition behind this assumption is that, similarly to a Laplacian pyramid, edge information provides fine information about the image, while the conditioning from the lower scales provides coarse information about the global gray value distribution. The impact of ablating each of the conditions during training or only during testing is investigated in this experiment. All of the setups result in a poor image quality (Fig. 4.9). As expected, when training on both conditions but not using them during the inference phase, the results are very poor. On the one hand, using only sketches for training leads to visible patch artifacts. On the other hand, leaving the edge information out leads to blurry and homogeneous images due to the lack of high-resolution information.

Overall the proposed multi-scale patch-wise approach is able to generate large medical volumes with a constant memory demand regardless the image size. It fulfills the requirements for high quality and no visible patch artifacts. Thus, this GAN training procedure can be used to enable the healthy-to-pathological image domain translation as in Sec. 4.2.2 for 3D medical image data.

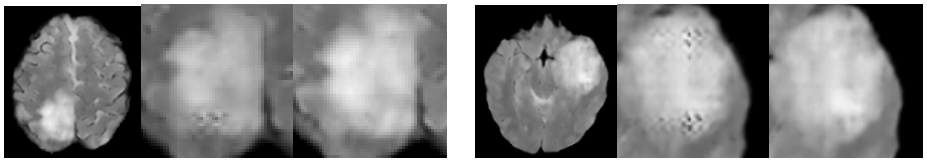
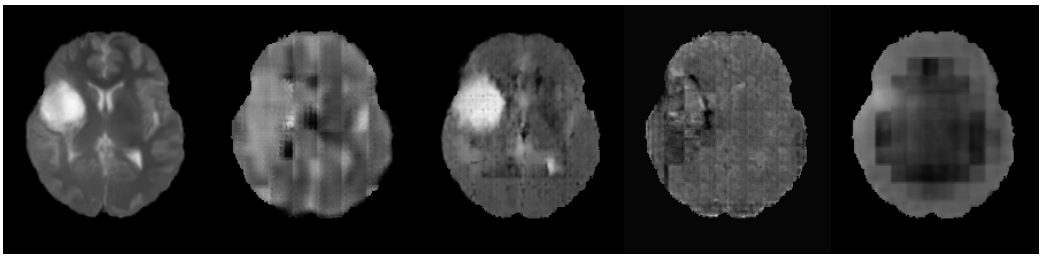


Fig. 4.8: Two generated 3D images (axial slices) and zoomed in tumor tissue from scale 2 and 3. Artifacts generated on lower scales are not propagated to the higher scales.



both	only edges	only multi-scale	only edges	only multi-scale
train&test	train	train	test	test

Fig. 4.9: Testing the necessity of both conditions: edges and multi-scale. Shown are examples of a generated image when one of the inputs is left out. From left to right: using both conditions while training and testing; lower-scale left out during training and testing; edges left out during training and testing; lower-scale left out during testing; edges left out during testing.

4.4 MEGAN for the Generation of Realistic Healthy and Pathological Image Volumes

The presented MEGAN approach enables the generation of large medical volumes for the first time. In the following experiments, the general suitability of the method for medical domain translation is investigated as a pre-step for the pathological image generation [Uzunova et al., 2019a, Uzunova et al., 2020b].

Medical image domain translation is a common image processing task since there is a large variety of different acquisition parameters in medical imaging [Kaji and Kida, 2019]. The different acquisition techniques ensure an optimal contrast and better visibility of the structures of interest, yet, they cause a heterogeneous data situation. The different MRI acquisition parameters are a common example for this issue, since different pulse sequences can be chosen (e.g. T1-, T2-weighted or FLAIR). Many algorithms require a multitude of sequences or a particular sequence, e.g. FreeSurfer [Fischl, 2012] requires T1 sequences for brain segmentation. For this reason, generating missing sequences is a crucial task considered by various works from the community. E.g. [Jog et al., 2017], is an established strictly paired random forest method for the translation between different brain MRI sequences. Furthermore, deep learning approaches like [Yang et al., 2020] enable the generation of more realistic synthetic brain MRI sequences. Here, the authors show the need for image translation as a pre-processing step for a registration algorithm, however, their approach is also trained in a paired manner. In [Armanious et al., 2020], an unpaired domain translation for PET-to-CT is proposed. Though this approach is effective, the topology maintenance between the translated images is not given. A further application is the translation between e.g. the different reconstruction kernels or different doses used in CT [Jin

et al., 2019, Yang et al., 2021]. Most of the existing approaches require paired data or do not ensure topology preservation of the input image and are only able to transfer from one particular image domain. Here, a method that has the ability to translate any domain into the target domain with a single training is proposed. Furthermore, contrary to the state-of-the-art, MEGAN is suitable for high-resolution 3D medical image data, demonstrated in the following experiments.

A further advantage of MEGAN is the ability to explicitly model pathologies in the generated data which, unlike implicit approaches like [Zhu et al., 2017], prevents pathology hallucination [Cohen et al., 2018]. Furthermore, this process facilitates data-balancing and pathological data augmentation which are important tools for improving the training data situation for deep learning. An important feature is the ability to preserve the topology of the source image, thus, a healthy-to-pathological domain translation is possible resulting in images containing both: ground truth annotations of the healthy anatomical structures and labels for the pathological structures.

4.4.1 Architecture and Implementation Details

For the approach presented in Sec. 4.3, commonly used architectures are selected in order to decouple the effectiveness of the proposed approach and the performance gain due to architectural search. The architectures are shown in Fig. 4.10. Since the tasks learned by the HR and LR GANs are substantially different, different architectures are chosen for each. On the one hand, the LR GAN should be able to generate the whole image in a low-resolution, in other words its generalization ability should be high, while no particularly high resolution is required for the generated image. For this purpose, a U-Net architecture is applied since its bottleneck is able to filter unimportant details that ensures high generalization ability but might lead to blurry appearance of the

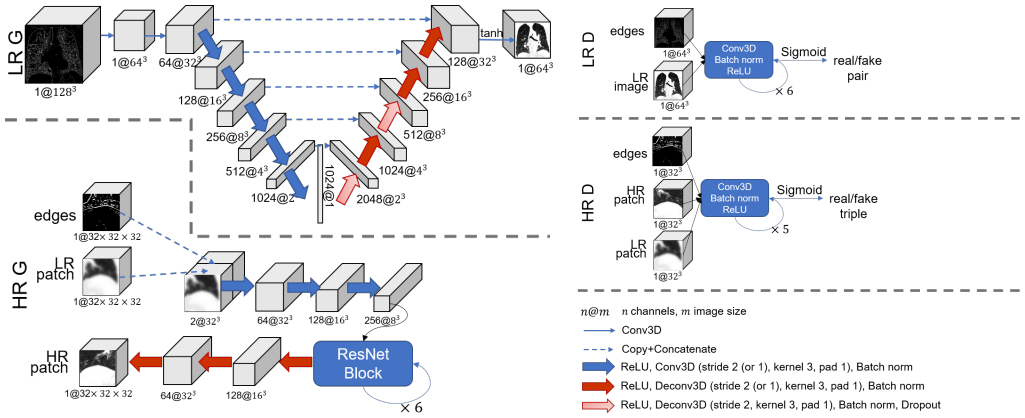


Fig. 4.10: Architectures of the generators (HRG and LRG) and discriminators (HRD and LRD) used in the training cascade.

images. For the patch generation using the HR GANs, on the other hand, high-resolution patches are required. Here, ResNet blocks [He et al., 2015] are chosen since they are able to generate sharp output images due to keeping the input image resolution unchanged. The higher overfitting of this architectural choice is negligible because of the stronger conditioning and the overall large number of training data resulting from extracting multiple patches per image. For all discriminators, a simple fully convolutional architecture is implemented (Fig. 4.10).

Furthermore, to stabilize the training, the loss function is combined with a pixel-wise L1 loss for the training of the generator. This prevents the discriminator from learning at a faster rate than the generator and leads to an improved training procedure which is consistent with [Isola et al., 2017]. The straightforward usage of a noise vector $\mathbf{z}$ is replaced by dropout layers and noise-based data augmentation also proposed in [Isola et al., 2017].

As it can be noticed in Fig. 4.10 the LR GAN and the HR GANs have different input and output sizes. Those can be varied depending on the specific application and overall image size. Experience showed that the LR image size should be a good trade-off between large enough to capture the complexity and the general structure of the image, but not too large in order to avoid noise generation and the allocation of unnecessary computational resources. An interesting observation is also that a larger sketch size (here: 128^3 vs. 64^3) leads to a significant improvement of the appearance of the LR image since the fine details of the contours might vanish for very small resolutions. Those considerations lead to the following image sizes for the experiments: input and output of the HR GAN 32^3; input of the LR GAN 128^3 and output of the LR GAN 64^3. As mentioned in Sec. 4.2, the sketch for the LR GAN input is generated by the element-wise multiplication of Canny-extracted edges and the gradient magnitudes of the source image. Furthermore, in order to avoid padding artifacts, only a region of the network's receptive field is extracted from the generated patches during the final image stitching. This way, no patch overlaps are required.

During the training of each HR generator, the low-resolution input patches are augmented with noise, Gaussian blurring of 30% of the patches and halving the resolution of 20% of the patches. These values are determined empirically, such that they roughly represent the distribution of corrupted patches generated during the inference phase. Further details and the preferred parameters can be seen in the publicly available code[4].

4.4.2 Data and Evaluation Metrics

Data: In the following experiments, three domain translation scenarios are presented and thoroughly evaluated using the following datasets.

[4] `https://github.com/hristina-uzunova/MEGAN`

Lungs COPD: A thorax CT dataset containing the thoracic CT volumes of 56 subjects with different degrees of chronic obstructive pulmonary disease (COPD) [Ehrhardt et al., 2016]. For each subject, the data are reconstructed with three different CT reconstruction kernels: soft (B20f), sharp (B50f) and very sharp (B80f). Each volume features up to 512^3 voxels.

Low-dose Lungs: 4D thoracic volumes of sizes around 300^3 voxels. The images are acquired with a lower dose of 120kVp, 40mAs during the free breathing of the patients, resulting in 166 3D images in total [Castillo et al., 2013].

Tumor Brain MRI: A subset from the BRATS challenge and commonly used throughout this work. Here, multiple MRI sequences are considered: Flair, T1- and T2-weighted. 220 whole volumes sized $155 \times 240 \times 240$ are used in the experiments.

Healthy Brain MRI: Images from the LPBA40 dataset, also commonly mentioned throughout this work. The original size $181 \times 217 \times 217$ of the 40 T1-weighted MRI volumes is used in the following experiments.

Evaluation Metrics: A direct quantitative evaluation of the generated image quality is enabled, when a real ground truth corresponding image is available. Formally, if the aim is to translate an image $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ from the domain $\mathcal{X}$ to an image $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^d$ from the target domain $\mathcal{Y}$ by using the function $f(\mathbf{x}) = \tilde{\mathbf{y}}$ (e.g. a trained neural network), then, the difference between the real $\mathbf{y}$ and the generated $\tilde{\mathbf{y}}$ can be quantitatively evaluated using a suitable metric $\mathcal{D}_{eval}(\mathbf{y}, \tilde{\mathbf{y}})$. In the presented experiments, the following commonly used metrics are chosen to enable objective quantitatification of the results and comparability among different works.

MSE/MAE: Mean squared error (MSE) and mean absolute error (MAE) as explained in Sec. 2.5.2.

PSNR: Peak signal-to-noise ratio (PSNR) is a widely used metric to measure similarity. In signal theory, it describes the ration between the maximum power of a signal and its corrupting noise. In the case of images, the maximum signal is described by the maximum image intensity I_{max} and the corruption noise is most frequently calculated using MSE. So, the following formula can be constructed:

$$PSNR(\mathbf{y}, \tilde{\mathbf{y}}) = 10 \log_{10} \frac{I_{max}^2}{MSE(\mathbf{y}, \tilde{\mathbf{y}})}. \tag{4.3}$$

Obviously, this metric is derivable from MSE, yet, it is presented here for enhanced comparability throughout different works.

SSIM: Structural similarity index measure (SSIM) [Zhou Wang et al., 2004] as explained in Sec. 2.5.2.

FID: The Fréchet Inception Distance (FID) is a metric that does not compare an image and its corresponding ground truth directly, moreover, it compares the distributions of the real and the generated images [Heusel et al., 2017]. The closer the distributions are, the better the quality of generated images. To enable a feature space

distribution of the domains, feature vectors representing the images are considered rather than the images themselves. The feature vectors are extracted using the pre-trained inception network as in [Salimans et al., 2016]. Assume that $\mathcal{R} \subseteq \mathbb{R}^m$ is the set of feature vectors with length m of the real images and $\mathcal{G} \subseteq \mathbb{R}^m$ is the set of feature vectors extracted from the generated images, then the mean feature vectors $\boldsymbol{\mu}_\mathcal{R}$ and $\boldsymbol{\mu}_\mathcal{G}$ and the covariance matrices $\Sigma_\mathcal{R}$ and $\Sigma_\mathcal{G}$ can be calculated respectively. The Fréchet distance is calculated as follows:

$$FID(\mathcal{R}, \mathcal{G}) = ||\boldsymbol{\mu}_\mathcal{R} - \boldsymbol{\mu}_\mathcal{G}||_2^2 + \mathrm{tr}(\Sigma_\mathcal{R} + \Sigma_\mathcal{G} - 2\sqrt{(\Sigma_\mathcal{R}\Sigma_\mathcal{G})}), \tag{4.4}$$

where $\mathrm{tr}(\cdot)$ denotes the trace function. According to the authors of [Salimans et al., 2016], this metric roughly imitates the human quality perception of images, where high scores correspond to bad image quality and vice versa.

4.4.3 Experiments and Results without Pathology Injection

Sharp to Soft Kernel Domain Translation: The influence of the different CT reconstruction kernels on the image appearance typically impairs the compatibility of automatic image processing methods. For example, the emphysema index is usually calculated on images reconstructed with a soft B20f kernel, thus, the computation of the emphysema index on images reconstructed with a sharp kernel would be significantly overestimated [Gierada et al., 2010, Boedeker et al., 2004]. Contrary to that, clinicians would typically prefer sharper kernels for visual assessment of the images. In order to address this problem, MEGAN is utilized for the domain translation of the sharper kernels (B80f and B50f) to the soft kernel (B20f). A main advantage of the proposed method is that a single training is required for the inference of both source domains to the target. The cascading GANs are trained on the B20f images in a 5-fold cross-validation manner. In the test phase, the corresponding sketches of the B50f and B80f reconstructions of the images not used for training are propagated trough the cascading nets. This results into images represented by the shape of the input sketches and an appearance typical for soft kernels. Examples can be seen in Fig. 4.11 and App. A. Since the B50f kernel delivers less noisy images, the translation from this kernel yields more detailed results. Nonetheless, the images resulting from the B80f kernel are also of high quality.

Since this dataset contains all images reconstructed simultaneously with each kernel, the corresponding real B20f for each B80f/B50f image is given and, thus, a ground truth for a quantitative evaluation is available. Note that the B20f images are only used for evaluation and not during training nor testing. The generated and real images are compared using the previously described metrics SSIM, PSNR, MSE and MAE.

In order to underline the effectiveness of MEGAN, three baseline methods are implemented for comparison. 1) A straightforward patch-wise GAN is developed and trained in the exact same manner as the HR GAN on the last scale, excluding the

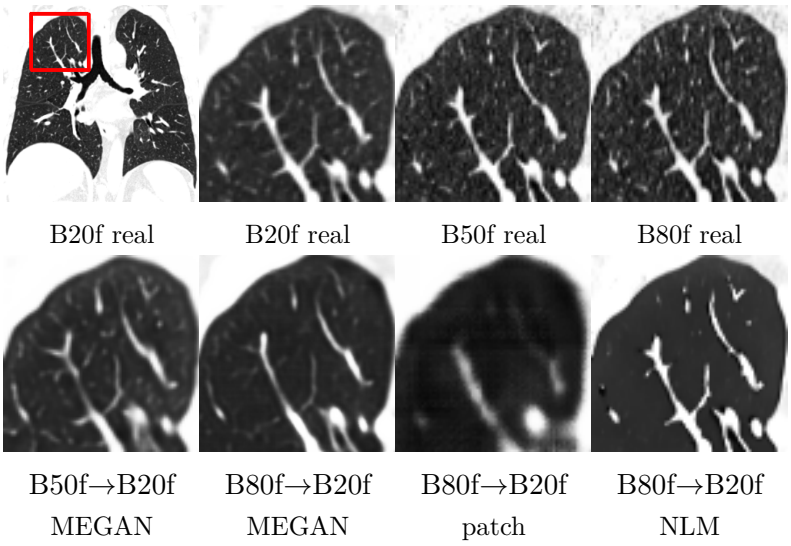


Fig. 4.11: Examples of the domain translation for sharp (B80f, B50f) to soft (B20f) kernels. Top row: real reconstruction kernels (zoomed in for better visibility). Bottom row: generated B20f images by the proposed method MEGAN from the B50f images; from the B80f images; a straightforward patch-wise approach from the B80f images; NLM smoothed B80f images. For larger images, see App. A.

low-resolution conditional information. To avoid patch artifacts, overlapping patches are sampled during the test phase and again only the network's reception field region is extracted from the generated patches. 2) A GAN generating a small image and then rescaling to the original image size using conventional linear resampling. This corresponds to the LR GAN, since this image size is about the computational limit of the hardware at hand. 3) A conventional method to achieve a soft kernel reconstructed image from a sharp kernel image is to use smoothing. This is also implemented here by using the state-of-the-art non-local means (NLM) smoothing filter that is known for its edge-preserving properties [Buades et al., 2011]. Examples of the qualitative results of 1) and 3) can also be seen in Fig. 4.11.

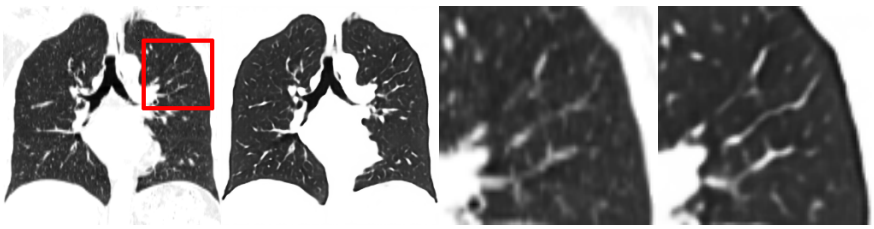
The quantitative results for both scenarios and all comparison methods are shown in Tab. 4.1. According to the shown values, the B50f and B20f kernels yield fairly similar images initially, thus, the domain adaption does not result to a substantial enhancement. However, the domain translation from the B80f to the B20f images results in a strongly improved image similarity compared to the initial results. A major drawback of the chosen metrics is that they do not fully reflect the desired image quality. The NLM smoothed images show similar quantitative results to MEGAN, yet qualitatively (Fig. 4.11) the images lack many important details. Furthermore, the filtering requires an around 100 times longer computational time than the network's

Table 4.1: Quantitative results for the sharp to soft kernel translation scenario. Superscripts correspond to statistical significance ($p < 0.0001$) in a paired two-sided t-test for the baseline methods compared to the proposed method MEGAN in terms of: (*) all measures; (†) only SSIM.

	method	**SSIM (↑)** mean	**MAE (↓)** mean	**MSE (↓)** mean	**PSNR (dB, ↑)** mean
B80f→B20f	MEGAN gen.	**0.773**	0.033	0.004	24.1
	patch gen.*	0.706	0.049	0.008	21.2
	small gen.*	0.633	0.058	0.011	19.5
	NLM B80f	**0.773**	0.031	0.004	19.5
	orig. B80f*	0.480	0.065	0.012	19.2
B50f→B20f	MEGAN gen.	**0.794**	0.033	0.003	24.8
	patch gen.*	0.698	0.050	0.008	21.1
	small gen.*	0.636	0.055	0.012	19.3
	NLM B50f*	0.478	0.283	0.213	19.3
	orig. B50f†	0.722	0.028	0.002	26.3

inference. The other two compared approaches deliver significantly worse results in terms of quantitative and qualitative evaluation.

Low-dose to High-dose Domain Translation: The CT dose used for image acquisition also significantly impairs their automatic comparability. To underline the generalization ability of the proposed domain translation method and emphasize its unpaired nature, the *low-dose lungs* dataset is translated to the higher dose B20f images using the already trained networks. The edges of the low-dose volumes are extracted analogously to the previous experiments and propagated through the cascading GANs. An example result of this experiment can be observed in Fig. 4.12 and App. A. Due to the lacking ground truth for this experiment, i.e. no corresponding high-dose CTs are available, no straightforward quantitative evaluation is possible. However, next to the qualitative evaluation showing that the generated images look less noisy and have



low-dose high-dose (fake) low-dose high-dose (fake)

Fig. 4.12: Low-dose to high-dose domain translation: whole image slices and zoomed in area (red) for better visibility.

higher quality, the feature space distributions of both domains are compared using FID. The FID is calculated between: the original low-dose and the original high-dose images (FID=150); the synthetic high-dose and the original high-dose images (FID=131); the synthetic high-dose and the original low-dose images (FID=157). The small FID value indicates the best correspondence between the translated high-dose images and the real high-dose images.

4.4.4 Experiments and Results with Pathology Injection

In the following experiments, the pathological MRI images are used for training. All images are cropped using a bounding box around the brain to ensure a rough alignment of the brain tissue. Due to the good tumor visibility in the T2 sequences of the volumes, they are used as the target domain. The training is established in a 4-fold-cross-validation manner using one LR and two HR GANs. Since the tumor segmentation masks are given, they are strictly integrated into the input sketch by overlaying them over the extracted edges. This explicit pathology injection prevents pathology hallucination and provides control over the pathological structures enabling easy pathology augmentation.

Pathological Brain MRI Modality Translation: Translating between different MRI modalities is a common medical image processing task since many algorithms are specialized for certain image sequences that might not have been acquired in the clinic. In this experiment, a T1/Flair-to-T2 translation is aimed, thus, for inference the masks of the left-out T1 and Flair images are used as inputs to the trained GAN-cascade. Here, one training is sufficient to establish both the Flair→T2 and the T1→T2 translation. Since all three sequences are co-aligned in an intrapatient fashion, the ground truth T2 sequences are available for quantitative evaluation, however, they are not used either during training nor during testing.

For comparison, two state-of-the-art paired image translation methods are considered: REPLICA [Jog et al., 2017] and MedSynth [Nie et al., 2018]. REPLICA is a random regression forest approach using manually crafted features in a multi-scale manner. MedSynth is a patch-based GAN approach that applies an auto-context model for patch stitching. For those models, the code provided from the authors was adapted for the performed experiments. The quantitative evaluation of the methods is established with the metrics MAE, MSE, PSNR and SSIM as presented in 4.4.2. The results can be seen in Tab. 4.2.

All compared methods produce similar results in terms of MSE and MAE and, as expected, paired methods generate intensities more similar to the real T2 images. Since MEGAN is of an unpaired nature and does not consider the intensities of the input images, the voxel-wise distances are higher. Nonetheless, in terms of structural similarity described by SSIM, the presented method performs significantly better than

Table 4.2: Quantitative results for the pathological brain MRI translation. Metrics are calculated between a generated image and its ground truth and averaged over all folds and training images. Top and bottom: T1 to T2 translation and Flair to T2 translation. Compared to the proposed method MEGAN are REPLICA and MedSynth. Superscripts (*) indicate statistical significance ($p < 0.0001$) in a paired two-sided t-test compared to MEGAN in terms of SSIM.

	method	**SSIM (↑)**	**MAE (↓)**	**MSE (↓)**	**PSNR (dB, ↑)**
		mean	mean	mean	mean
T1→T2	MEGAN*	**0.911**	0.017	0.003	26.0
	REPLICA	0.854	0.017	0.003	26.9
	MedSynth	0.613	0.025	0.003	27.0
Flair→T2	MEGAN*	**0.905**	0.021	0.004	24.6
	REPLICA	0.833	0.019	0.002	25.9
	MedSynth	0.734	0.020	0.002	27.5

the other approaches demonstrated by a two-sided t-test. The low SSIM values of the MedSynth approach are due to visible patch artifacts, especially in the tumor regions. Better results are achieved by REPLICA, but it still does not reach the SSIM values obtained by MEGAN, which can be explained by the blurry nature of the REPLICA images.

Some of the results of this experiment are visualized in Fig. 4.13. As already evident from the quantitative results, the qualitative evaluation shows that the Flair-to-T2 translation is a more challenging task. Since the input Flair sequences have a rather low contrast, the good edge extraction is impaired. Overall, the paired image translation methods REPLICA and MedSynth do not produce qualitatively improved images compared to MEGAN. REPLICA yields less noisy images while MedSynth suffers from patch artifacts (red arrows). The developed approach MEGAN is able to generate visually more realistic results and the Flair-to-T2 translated images feature more details. Furthermore, MEGAN is trained in an unpaired manner and only one training is sufficient for the translation from both source domains. Overall, the resulting synthetic images generated by the proposed approach are of high resolution and quality.

Healthy-to-pathological Brain MRI Translation: A considerable advantage of the presented method is its ability to generalize to nearly arbitrary images of the source domain. To highlight this, in this experiment, the GAN-cascade pre-trained on the T2 images is used to establish a healthy-to-pathological image translation. Here, the healthy subject images are collected from the entirely different LPBA40 dataset. The edges from the healthy images are extracted analogously to the previous experiments, and, since the LPBA40 images do not contain pathologies, tumor masks from the BRATS training set are synthetically overlayed. In this way images with pathological appearance can be obtained (Fig. 4.14). An advantage of the explicit tumor

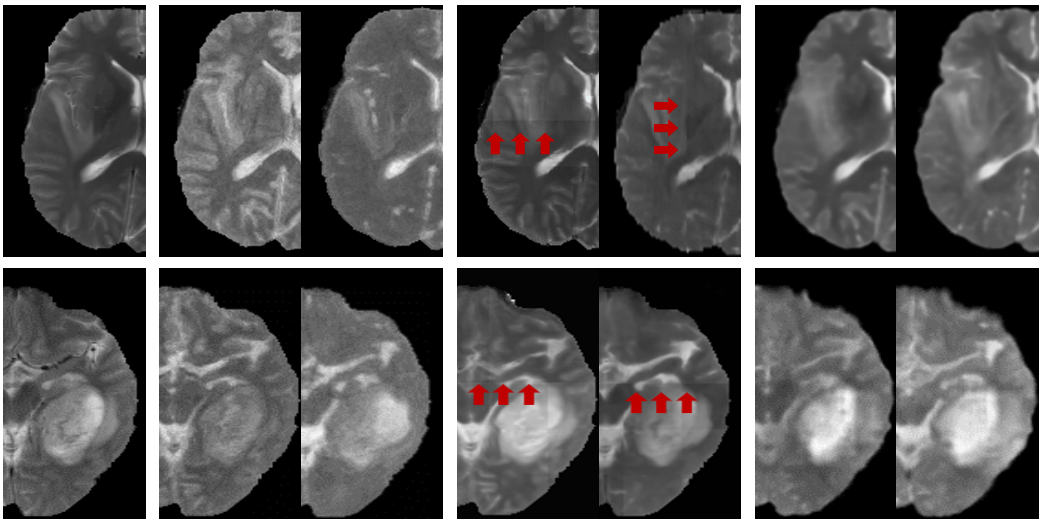


Fig. 4.13: Results of pathological MRI sequence translation T1→T2 and Flair→T2 (cropped center axial slices). Shown are the real T2 sequence and domain translated images generated by REPLICA, MedSynth and MEGAN. Red arrows show patch artifacts resulting from MedSynth. Best viewed digitally.

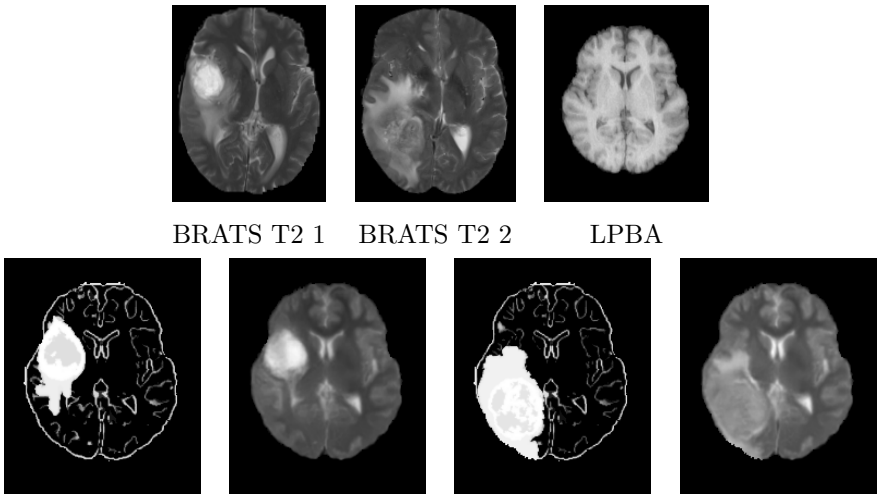


Fig. 4.14: Healthy-to-pathological domain translation examples. First row: Real images; Second row: Extracted edges and generated images.

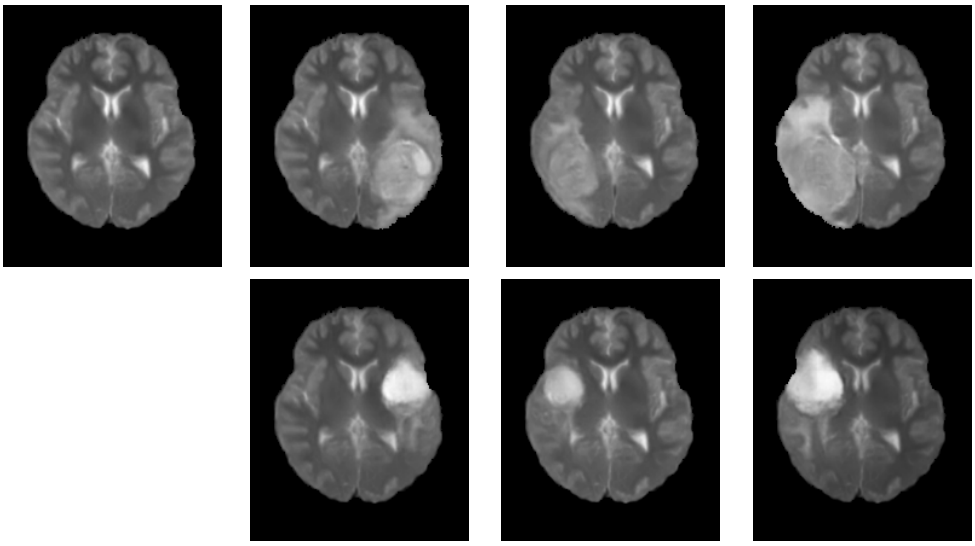


Fig. 4.15: Tumor augmentation examples. Row-wise: Two different tumors applied on the same image (tumors from Fig. 4.14) and their augmentation featuring mirroring, shrinking and zooming.

integration is also shown in Fig. 4.15: contrary to other unpaired image translation methods (e.g. [Zhu et al., 2017]), when the tumor mask is not used, an image of healthy appearance is generated without visible hallucinated pathologies.

Furthermore, the explicit control of the tumor enables more data augmentation possibilities by performing simple transformations on the input tumor mask. In the shown experimental setup, the tumor mask transformations feature mirroring and shrinking/zooming by 15%.

Although the generated images look highly realistic, they are not conceptualized for the clinical use. Moreover, they can be used to improve further automatic image analysis and processing approaches, e.g. data augmentation and data balancing for the training of neural networks, or the generation of annotated images for algorithm evaluation. In further experiments, synthetically generated images are indeed used in similar scenarios. At this stage, the synthetic images still lack an important feature of pathology presence in medical image data: pathologies, especially brain tumors, oftentimes push the surrounding tissue away resulting into deformed shapes of the neighboring anatomical structures. The simulation of such tumor-induced deformations is developed in the next chapter.

4.5 Discussion and Conclusion

In this chapter, fundamental methods for the realistic pathological image generation were developed. Under the assumption that healthy subjects images are typically available and the annotations of their anatomical regions are often given, an approach that translates healthy images to images of pathological appearance by keeping the initial healthy topology is developed. In the first step, a topology-preserving cGAN approach is investigated and an explicit pathology injection is proposed. It is shown that a cGAN conditioned on intensity-independent automatically extracted sketches is able to preserve the initial topology. Furthermore, an explicit pathology injection ensures that no hallucinated pathologies occur and the pathological label is preserved. Despite the success of this method for 2D images, a remaining hurdle is the fact, that neural networks and especially GANs are very memory-consuming, making their application on large 3D volumes highly infeasible. Yet, medical images are typically 3D volumes with an enormous amount of voxels. To address this issue, in a second step, the MEGAN method is developed: a GAN-based approach for the generation of high-resolution 3D medical image data with a constant memory demand regardless the image size. The main idea of MEGAN is to apply a combination of patch-based and multi-scale strategies in such a manner that no patch artifacts appear and a high image quality is achieved. This approach, combined with the previous pathology-injecting method, enables the feasible generation of 3D medical images with explicitly simulated pathological structures. A great advantage of the proposed method is the fact, that the images are directly generated with given anatomical and pathological annotations. Thus, the time-consuming and costly process of image annotation is avoided. Also, the anatomical annotations of the generated images can be used for training and evaluation of methods commonly applied on healthy subjects.

In the conducted experiments, various applications of the proposed pathological image generation method are investigated. The first experiment shows the feasibility of MEGAN for the unsupervised domain translation of full-resolution 3D medical image on the examples of different thorax CT acquisitions and brain MRI sequences. Further, the approach enables healthy-to-pathological image translation by explicit pathology injection. This enables the prevention of feature hallucination and facilitates pathology augmentation. Even though the generated images have a high quality underlined by an extensive quantitative analysis, a major drawback of the presented method is that simply overlaying pathological structures over healthy tissue is not sufficient for a realistic pathological appearance. Pathologies, oftentimes, cause distortions of the neighboring healthy tissue, e.g. tumors might cause the so-called mass effect. This problem is coped with by proposing an inverse probabilistic approach for the simulation of pathology-induced tissue deformations in the next chapter.

Chapter 5

Generation of Annotated Brain Tumor MRIs with Tumor-induced Tissue Deformations for the Training and Evaluation of Neural Networks

In this chapter, the previously developed method for the generation of realistic images containing pathologies is extended by a novel pathology-induced deformation simulation approach that approximates the healthy tissue distortion around a given pathology. Unlike existing methods, the presented approach is able to predict distortions caused by pathologies by deducing the healthy appearance from a single image. Using the proposed method, the realism of the synthetically generated pathological images can be additionally enhanced by the simulation of pathology-induced tissue deformations. To underline the necessity of simulating annotated images featuring pathologies and the important role of pathology-induced tissue deformations, here, brain tumor MRIs with tumor mass effect are synthesized and used in multiple experiments for the training and evaluation of neural networks.

5.1 Introduction and Motivation

Tissue deformations in the surroundings of pathological structures have been a focus of research for a long time. A common example is the so-called mass effect, most commonly caused by tumor tissue in the brain or other organs, which deforms the surrounding tissue with a radial-like force and compresses adjacent anatomical structures. For the modeling of the tumor mass effect, the biophysical properties of the tumor growth have been frequently used as a starting point [Hogea et al., 2007, Ezhov et al., 2019]. Unfortunately, such models are typically time-consuming and their accuracy is limited due to the multitude of required parameters. Furthermore, these

methods are build around a very specific pathology type and growth assumptions that, in particular cases, cannot be determined. These hurdles make the modeling of pathology-induced tissue deformations challenging, while their importance for medical image processing algorithms is crucial. In [Nie et al., 2018], the authors generate brain tumor MRIs by simply overlaying the pathological structure over the healthy brain tissue and demonstrate the efficiency of the method as an augmentation technique used for a tumor segmentation network. However, when normal anatomical regions around the tumor tissue are targeted by e.g. a segmentation network, the shape distortion of the structures is of great significance. For example, the segmentation of ventricles in brains containing brain tumors could be impaired due to the common deformation of the ventricles by the tumor mass effect.

A possible step to lower the complexity of the classical tumor growth models would be to solve the problem by applying deep learning approaches. Yet, a major difficulty is to design an approach that is able to determine the pathology-induced deformation from a single pathological image at a given state. In [Elazab et al., 2020], the authors use a GAN to predict the growth of a tumor by training on images containing three different time points. While their approach is accurate in predicting the next time step, they do not determine the entire mass effect. In order to be able to deduce the healthy tissue deformation caused by a pathological structure, a straightforward approach would require the healthy correspondence of the given pathological image. Unfortunately, longitudinal data are hardly ever available and most of the patients are not screened before the first appearance of a pathological structure. As far as is known, there are no established deep learning approaches for mass effect simulation or pathology-induced tissue deformations from a given pathological image to this end. For this reason, a novel deep learning approach is established for the prediction of pathology-induced tissue deformation from a single image by deducing its corresponding healthy shape and, hence, implicitly learning the inverse pathology-induced deformation [Uzunova et al., 2020a].

5.2 Simulation of Pathology-induced Tissue Deformations for the Synthesis of Realistic Images Containing Pathological Structures

5.2.1 Inverse Probabilistic Tissue Deformation Prediction

The key idea of the method developed here is to apply an inverse approach that deduces the healthy shape from a given pathologically deformed shape. In Sec. 3.3, it has been successfully shown that the healthy tissue variability can be modeled using deep learning approaches. Here, this strategy is used to learn the variability of shapes of healthy anatomical structures. Thus, an inverse pathology-induced deformation can

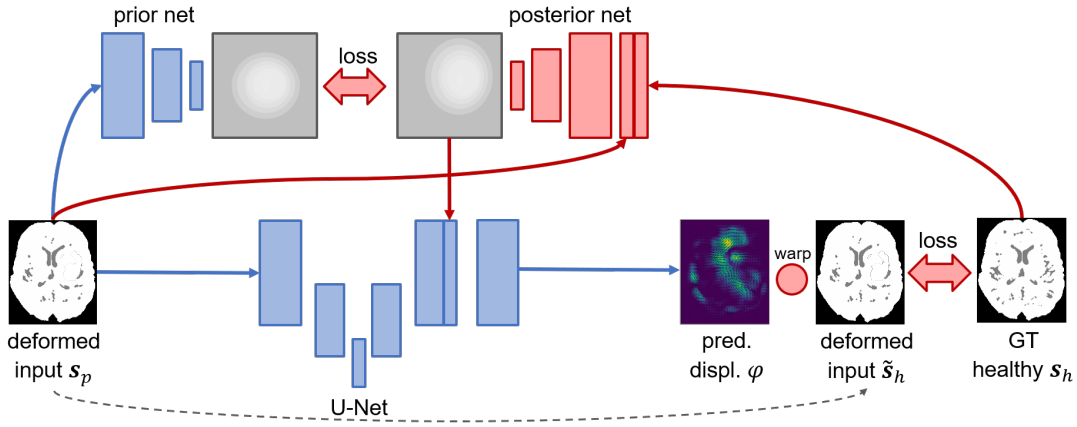


Fig. 5.1: Scheme of a probabilistic U-Net [Kohl et al., 2018]. The red and blue pathways are used in the training phase, while the red path is not used during the inference phase.

be deduced from a pathologically deformed tissue by predicting its healthy correspondence. In other words, a network that predicts the healthy shape from a deformed shape needs to be designed. Formally, let $\mathbf{s}_p \in \mathbb{R}^d$ be the shape of a pathologically deformed image and $\mathbf{s}_h \in \mathbb{R}^d$ its healthy shape equivalent. Then $f : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times m}$ is a probabilistic function (here: a neural network) with $f(\mathbf{s}_p) = \varphi$ and $\mathbf{s}_p \circ \varphi = \tilde{\mathbf{s}}_h \approx \mathbf{s}_h$ where φ is a dense displacement field and $m \in \{2, 3\}$ is the image dimension (Fig. 5.1).

Since multiple possible healthy shapes could correspond to a given pathological image, a probabilistic function f is required to statistically represent this variability in a large synthetic dataset. This work applies a probabilistic U-Net to estimate these unknown tissue deformation parameters [Kohl et al., 2018]. A scheme of the probabilistic U-Net is shown in Fig. 5.1. Basically, a probabilistic U-Net is a combination of a variational autoencoder and a U-Net [Ronneberger et al., 2015]. During training (blue and red paths), a meaningful embedding of the predictions in the latent space is ensured by introducing a so-called posterior net that maps the variant of the current ground truth to the latent space. The latent spaces of the prior and posterior nets are typically trained to be similarly distributed using a Kullback-Leibler loss. In the test phase (blue path), samples are directly predicted from the input image and a posterior net is not necessary. The main discrepancy between the proposed method and the original approach, presented in [Kohl et al., 2018], is that, here, a probabilistic U-Net is used to predict a dense displacement field that warps the deformed input shape to a healthy non-deformed shape. The warping process is implemented in a fully differentiable manner and can be integrated into the training procedure with backpropagation (similarly to [Jaderberg et al., 2015]).

Having established a method to determine the inverse pathology-induced displacement, a pipeline as in Fig. 5.2 can be set up. The network is trained on pathologically

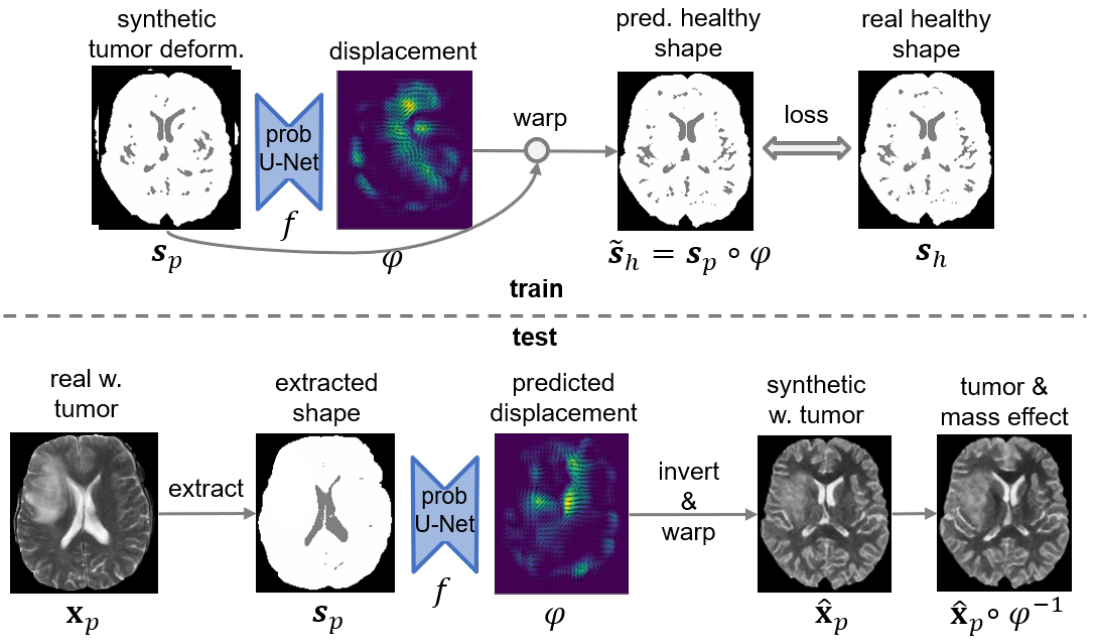


Fig. 5.2: Overview of the training and testing for the generation of tumor-induced displacements.

deformed shapes $\mathbf{s}_p$ and generates a displacement field φ that transforms the input to a healthy shape $\mathbf{s}_h \approx \mathbf{s}_p \circ \varphi$. In the test phase, the trained probabilistic U-Net predicts a displacement field from the shape extracted from a real pathological image. A synthetically generated image with an overlayed pathology $\hat{\mathbf{x}}_p$ (e.g. using MEGAN) can then be warped with an inverse generated field $\hat{\mathbf{x}}_p \circ \varphi^{-1}$ resulting into an image containing a pathological structure and a corresponding pathology-induced deformation. Having this pipeline as a base, several practical issues need to be addressed.

Generating Training Data: The underlying assumption of this method is that shape information of some kind is available. Here, a simple yet efficient method for extracting the intensity-independent shape information from images by employing a binary-thresholding approach is considered. The intensity images are segmented using thresholding into three classes: background, main brain tissue and fine details of the brain tissue, mostly featuring the ventricles and the brain sulci. Using such training shape images mirrors the intuition that the main distortions caused by a tumor can be captured by observing the deformation of the ventricles and the sulci.

Furthermore, pathologically distorted image shapes and their healthy correspondences are needed for training. However, such data are, as the assumption of the presented method states, not available and, hence, need to be synthetically generated.

Starting from healthy images, their shape representations are extracted as described above. In order to simulate a pathological deformation, a tumor mass effect simulation from [Pfarrkirchner et al., 2018] is applied around randomly placed spherical structures. This deformation approach only considers the simplified model assumption that the mass effect is an outward radial force that weakens with growing distance to the tumor center. The following steps are featured by the deformation method: 1) a distance transform of the sphere is applied; 2) a dense vector field is created by extracting the gradients from the distance transformation; 3) the normalized distance transform and the vector field are combined using a weighted element-wise multiplication. Using this method, a synthetic dataset of deformed shapes with their non-deformed correspondences can be generated. Although the presented approach is a drastic simplification and does not deliver reliable realistic results, the generalization ability of neural networks could enable a deduction of healthy shapes from deformed ones rather than simply learning the concrete deformations.

Ensuring Plausible Displacements: The next significant hurdle is the fact that during the inference phase, the displacement field needs to be inverted before it warps the image (see Fig. 5.2). Yet, the inverting of dense displacement fields is not a trivial task. Thus, a network that learns a static velocity field, rather than the deformation field directly, is proposed [Rohé et al., 2017]. This method ensures diffeomorphism and easy displacement field inverting. The deformation field can be estimated from the velocities using the following dependencies: assume v is a static velocity field, then the displacement field φ and its corresponding inverse φ^{-1} can be calculated easily as

$$\begin{aligned} \varphi &= \exp(v) \\ &\Longleftrightarrow \\ \varphi^{-1} &= \exp(-v). \end{aligned} \tag{5.1}$$

Here, $\exp(\cdot)$ is a matrix exponential function. In order to implement the image warping approach in a differentiable manner as a part of the network, $\exp(\cdot)$ needs to be approximated using the scaling and squaring algorithm [Dalca et al., 2018]. Roughly formulated, this method is based on the assumption that a displacement field is the change of pixel positions over time $\varphi^{(t)}$, where $t \in [0, 1]$ is a time point such that $\varphi^{(0)} = Id$ is the identity transformation and $\varphi^{(1)}$ is the final displacement. From this follows $\varphi^{(1)} = \exp(v)$. Derived from the properties of the matrix exponential function for any scalars t_1 and t_2 the following equation can be determined

$$\exp\Big((t_1 + t_2) \cdot v\Big) = \exp(t_1 v) \cdot \exp(t_2 v). \tag{5.2}$$

So,

$$\begin{aligned} \exp\Big((0.5+0.5)\cdot v\Big) &= \varphi^{(1)} \\ \Longleftrightarrow \\ \varphi^{(1/2)} + \varphi^{(1/2)} &= \varphi^{(1)}. \end{aligned} \tag{5.3}$$

To approximate this, a recurrence is used starting with $\varphi^{(1/2^T)} = \mathbf{p} + v(\mathbf{p})$, where $\mathbf{p}$ is a map of spacial locations, e.g. a grid. The scaling factor T is chosen s.t. $||\varphi^{(1/2^T)}||_\infty \approx 1$. Then, iteratively $\varphi^{(1/2^{(t-1)})} = \varphi^{(1/2^t)} + \varphi^{(1/2^t)}$ is calculated.

To further facilitate the plausibility of the displacement field predicted by the probabilistic U-Net, some biophysical properties are captured into a regularization function in addition to the loss. Globally, constant tissue diffusivity is assumed, hence, a diffusion regularizer is integrated into the training process. Furthermore, the locality of the mass effect is captured by a weighted sparsity regularization with weight increasing with distance from the tumor center. So the overall objective of the probabilistic U-Net can be formally expressed as:

$$\mathcal{L}_{pUnet} = \mathcal{L}_{KL}(\mathbf{z}_{prior}, \mathbf{z}_{post}) + \mathcal{L}_{rec}(\mathbf{s}_p \circ \varphi, \mathbf{s}_h) + \alpha \underbrace{\sum_j^m \left\| \nabla v^{(j)} \right\|_2^2}_{\text{diffusion reg.}} + \beta \underbrace{\|\mathbf{w} \odot v\|}_{\text{local sparsity}} \tag{5.4}$$

where $\mathcal{L}_{KL}$ is the Kullback-Leibler loss of the prior vector $\mathbf{z}_{prior}$ and the posterior latent vector $\mathbf{z}_{post}$. The loss $\mathcal{L}_{rec}$ is a pixel-/voxel-wise loss, here L1 loss, between the pathological shape $\mathbf{s}_p$ warped with the predicted deformation field φ and the healthy shape $\mathbf{s}_h$. The upper-script in the diffusion regularization denotes the order of spatial dimension. The local sparsity term $\mathbf{w}$ corresponds to a weight mask with values increasing with distance from the tumor center and $\odot$ denotes an element-wise multiplication. The coefficients α and β are used for weighting the regularization terms [Uzunova et al., 2020a].

Combining the described pathology-induced simulation with the MEGAN image generation approach from the previous chapter yields the following pipeline schematically presented in Fig. 5.3 1) Train MEGAN for the generation of pathological image appearances from an intensity-independent sketch image and an overlayed tumor mask. 2) Train a probabilistic U-Net on synthetically deformed brain shapes to transform the inputs into healthy brain shapes. 3) Automatically extract a sketch from a healthy input image. Eventually, a real or a synthetic pathology mask can be overlayed. The trained MEGAN translates the image into the pathological domain, but the topology and, thus, the given annotations (here: ventricles) of the healthy image are preserved. 4) Automatically extract the shape information of a real pathological image and input it to the trained probabilistic U-Net. It generates an inverse pathology-induced displacement. 5) Invert the displacement field and warp the synthetic pathological image with it.

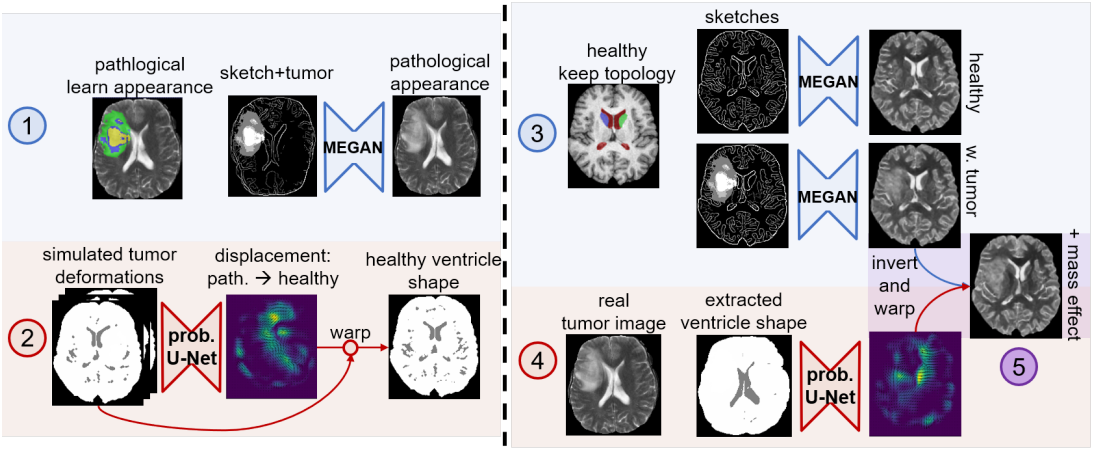


Fig. 5.3: An overview of the proposed pipeline for the generation of pathological images. Left: training; right: inference. 1) & 3) Topology-preserving domain translation with explicit pathology injection (■). 2) & 4) Generation of the pathology-induced tissue deformation field (■). 5) Combination of the pathological image with the deformation (■). Based on [Uzunova et al., 2020a].

5.2.2 Concept Analysis and Parameter Tuning

A significant problem connected to using the synthesized training shapes as inputs to the probabilistic U-Net is the fact that in early experiments, the generated deformations were nearly equivalent to the identity transformation. After an extensive analysis, the thresholded segmentation masks were found to be inappropriate inputs of the network, especially in combination with the used generalized Dice loss [Sudre et al., 2017], which is an intuitive choice when considering binary images. The reason for this is, that when measuring the label loss using a nearest neighborhood interpolation for warping, a relatively small displacement of a single pixel can have a large influence on the loss. This results into displacing the pixel back and forth in an alternating manner during the iterations and impairing the convergence properties of the network. To overcome this issue, as proposed in previous works [Audebert et al., 2019], the binary masks are represented by their signed distance maps enabling the use of (bi-/tri-)linear interpolation for warping, while a simple L1 loss is applied for training. This ensures several advantageous properties during the training. First, smooth input images facilitate the gradient propagation. Second, using linear interpolation enables smaller changes of a pixel positions per iteration and the network's convergence is supported. Additionally, using signed distance maps has shown to ensure topology-preserving properties of the given labels in previous works [Audebert et al., 2019].

This adaptation yields plausible deformation fields. However, the simple warping of synthetically generated images with the inverse generated displacement φ^{-1} delivers infeasible results (Fig. 5.4). To overcome this problem, the generated intensity images

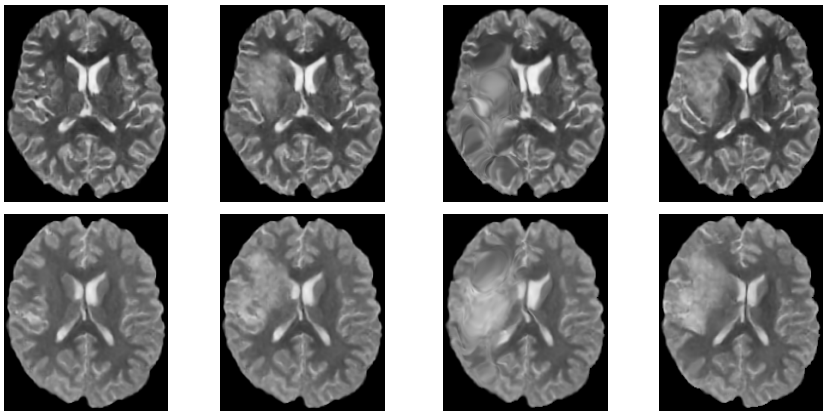


Fig. 5.4: Generated synthetic images without tumor, with an overlayed tumor and with an applied tumor-induced deformation. The naive warping approach delivers infeasible results, where as the indirect sketch warping yield satisfactory appearance of the generated images.

are not warped directly. Instead, a step back is taken and the warping is established on the edge-based sketches used as inputs to MEGAN. Formally, the images are generated as follows: with $\mathbf{e}_h$ being the edge-based sketch of the healthy image $\mathbf{x}_h$ and φ being the deformation field generated from a trained probabilistic U-Net, then the synthetic pathological image can be generated as $g\Big((\mathbf{e}_h \circ \varphi^{-1}) \oplus \mathbf{t}\Big)$, where $g(\cdot)$ is a trained conditional GAN for the sketch-based domain translation (e.g. MEGAN Sec. 4.3) and $\oplus$ denotes the overlaying operation of the tumor mask $\mathbf{t}$. Using the warping of the sketch, rather than the intensity image, delivers visually appealing results and less infeasible distortions can be observed in the area of the pathology (Fig. 5.4).

The proposed improvements generally lead to visually realistic results, where the generated pathology-induced tissue deformations conform to the expectations: the tissue around the tumor is deformed with a radial-like force and the deformation magnitude weakens with growing distance from the tumor (Fig. 5.4). A last property that is explored in the preliminary experiments is whether the generated deformation field is undeterministic as designed by the proposed method. For the use-case presented here, a probabilistic approach is specifically chosen, such that the intuition of multiple possible healthy shapes per given pathological shape can be captured. So, a desirable property is to generate multiple realistic healthy images given a single image containing pathological structures. To this end, different pathological tissue deformations can be deduced from a pathological image. This is important in the case that a large dataset needs to be created, e.g. for network training, where the variability of real pathological images needs to be captured in the entire dataset rather than just considering individual images. To verify this assumption, the same pathological structure is forwarded

Fig. 5.5: An image generated twice with the same tumor does not deliver the same displacement fields. The corresponding predicted inverse displacements are shown using the optical flow field color-coding [Dosovitskiy et al., 2015]. Note that, as required, the generated deformation is not deterministic due to the used probabilistic approach.

through the trained probabilistic U-Net multiple times. As expected, each run delivers different deformations that can be visually assessed as realistic. Fig. 5.5 shows two resulting displacements for the same pathological image, where it can be observed that the displacement fields are substantially different to each other, yet realistic.

Having ensured the most desirable properties of the proposed approach, in the next section, thorough experiments underlining the realism and need for synthetic annotated images containing pathological structures and corresponding pathology-induced deformations are conducted.

5.3 Training and Evaluation of Neural Networks on Synthetic Brain-MRIs with Simulated Tissue Deformations

Pathology simulation in medical images has been explored for many approaches, predominantly targeting the particular pathological structure itself. E.g. in works like [Frid-Adar et al., 2018, Shin et al., 2018, Wu et al., 2018], synthetically generated pathological data can boost neural networks for the segmentation and classification of pathologies. Yet, the presence of pathological structures, typically, also affects the surrounding anatomical structures, especially pathology-induced tissue deformations like the tumor mass effect. Thus, algorithms targeting normal structures could also be impaired by the presence of pathologies. For example, a neural network trained for the segmentation of brain ventricles on healthy brains might not be able to reliably segment the distorted or occluded ventricles in brains of patients with large brain tumors. Pathological images with both ground truth annotations – of normal anatomical regions, as well as the pathologies – are required for the training of algorithms engaging with anatomical structures in pathological images. However, publicly available datasets containing pathological structures usually contain the ground truth

labels of the particular pathological tissue types since they are typically designed for the segmentation or localization of those [Menze et al., 2015]. On the other hand, images containing anatomical annotations most commonly represent exclusively healthy subjects and are used e.g. for atlas generation [Shattuck et al., 2008].

In the following experiments, two main problems induced by this data situation are addressed: 1) algorithms trained on healthy subjects data, but applied on images containing pathologies cannot be reliably evaluated due to the lack of annotations of the normal tissue; 2) no ground truth pathological data are available for the training of neural networks targeting the surrounding anatomical structures, e.g. image registration or segmentation of anatomical objects. Using the method presented in Sec. 5.2, realistic brain tumor MRIs with tumor-induced tissue deformations are synthesized directly with available ground truth pathological and anatomical annotations and used for the evaluation and training of neural networks [Uzunova et al., 2020a].

5.3.1 Architecture and Implementation Details

The generation of medical images containing pathologies is established using the approach visualized in Fig. 5.3. For the simulation of the pathology-induced tissue deformations, the inverse probabilistic U-Net presented in Sec. 5.2 is implemented here. The four main parts of the probabilistic U-Net are the prior net, the posterior net, a U-Net and a small net for the combination of the posterior distribution and the U-Net prediction. The networks are all implemented in a fully-convolutional manner using ReLUs as activations. The prior and posterior nets have four convolutional layers, where average pooling and unpooling are used for down- and upscaling. The U-Net contains four down-convolutional blocks and four up-convolutional blocks, where each block contains three convolutional layers with ReLU activations in between. Downscaling is achieved by average pooling, while upscaling is achieved through bilinear interpolation. For the combination of the last U-Net layer and the latent space sample, four additional convolutional layers using kernels with isotropic side lengths of one are applied. For the 2D scenario, the latent space size is set to 6 and the channel numbers used in the prior, posterior and the U-Net are $[32, 64, 128, 192]$. For the 3D network, the latent space is enlarged to a size of 15 and due to computational constrains, a more shallow architecture is applied using the channel numbers $[32, 64, 128]$. In both scenarios, the objective from Eq. 5.4 is used during training, where α and β are both set to two. The training is implemented using early stopping on a validation dataset of size 10. Further details and the preferred parameters can be seen in the publicly available code[5].

[5] `https://github.com/hristina-uzunova/TumorMassEffect`

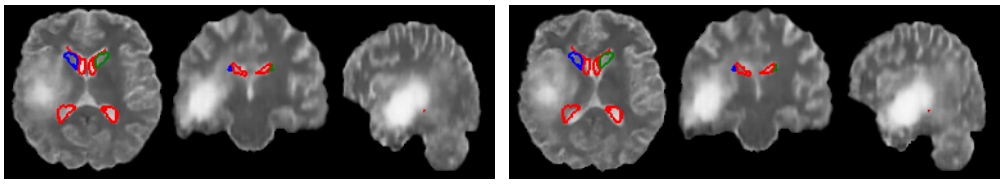
5.3.2 Data and Experimental Setup

Data: *Pathological:* The T2 sequences from the 3D brain tumor MRIs [Menze et al., 2015] commonly used throughout this work are applied in the following experiments. For evaluation purposes, two well-visible structures (ventricles and caudate nuclei) of 20 randomly selected images are segmented manually.

Healthy: In addition to the previously mentioned 40 images from the LPBA dataset [Shattuck et al., 2008], 30 3D brain MR T1 scans with labeled anatomical regions are considered here [Hammers et al., 2003][6].

Atlas: The T1 and T2 sequences of the ICBM152 brain atlas are also employed in the presented experiments [Fonov et al., 2011].

Experimental Setup: Given a set of images containing pathologies and a set of healthy subjects images with given ground truth anatomical annotations, the aim is to generate labeled images that are similar to the real pathological domain. This is indeed beneficial to, firstly, estimate the performance of pre-trained methods on pathological data and, secondly, train algorithms on the synthetic pathological dataset. For all experiments, synthetic data are generated according to the pipeline from Fig. 5.3. First, a 3D GAN is trained on the pathological T2 MRIs. Next, a probabilistic U-Net is trained on ca. 350 synthetically deformed shapes from the healthy images to learn the deformation from pathological to healthy shapes. In the inference phase, three synthetic domain-translated image types can be generated: healthy, with an overlayed tumor as in the previous experiments, and with an injected tumor combined with a predicted tumor-induced deformation. Ca. 440 images of each type are generated, examples can be seen in Fig. 5.6 and Fig. 5.7.



tumor w/o deformation tumor with deformation

Fig. 5.6: Examples of generated 3D brain MRIs: center axial, coronar and sagittal slices. Left: Injected tumor without tumor-induced deformation. Right: applied tumor-induced deformation around the tumor. To enhance the visibility, the edges of the non-deformed segmentations are overlayed.

[6] `https://brain-development.org`

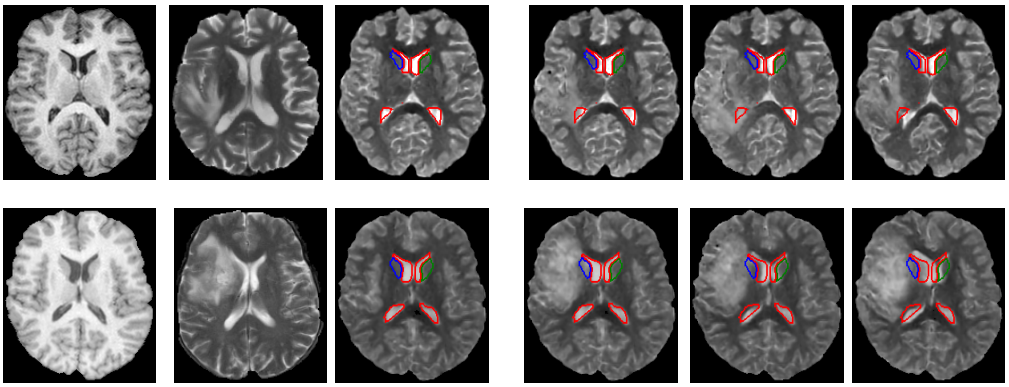


Fig. 5.7: Examples of generated 2D brain MRIs. From left to right: real T1 healthy subjects MRI; real T2 MRI; generated T2 images: without tumor; with an injected tumor and no deformation; deformation created with the naive approach from [Pfarrkirchner et al., 2018] used for shape deformations; deformation predicted by the proposed approach applied on the simulated tumor images. To enhance the visibility, the edges of the non-deformed segmentations are overlayed.

5.3.3 Synthetic Images for Evaluation of Algorithm Accuracy

The quantitative evaluation of pre-trained image processing networks for pathological data without ground truth annotations is strongly impaired. In this experiment, a registration neural network pre-trained on healthy T1 MRIs [Yang et al., 2017] is applied in an atlas-based segmentation. The main assumption is that applying the network on the synthetic pathological images yields results similar to the results on the real pathological data, however they are assumably significantly different to the outcome on healthy subject images. This makes it possible to estimate the algorithm error for pathological cases using synthetic images.

The results from this experiment can be seen in Tab. 5.1. When used for atlas registration on real pathological T2 data, the registration yields mean Dice values of **0.43**/**0.40** (ventricles/caudate nuclei) and for the generated deformed tumor images **0.47**/**0.47**. Those results are not significantly different according to a two-sided t-test, implicating that the generated images are plausible for the estimation of algorithm accuracy in the case of real data. On the contrary, simply overlaying the tumors without using deformations, leads to a significant Dice overestimation **0.62**/**0.63** underlining the importance of the simulation of tumor-induced deformations.

Table 5.1: Results for the error estimation of a pre-trained registration network. Compared are different moving images (synthetic or real) registered to the corresponding real atlas. Superscript (*) indicates no statistical significance ($p > 0.1$) in a hereoscedastic unpaired t-test compared to the last setup (real T2 tumor $\rightarrow$ T2 – marked *italic*).

	moving image	**atlas**	**ventricles**	**caudate n.**
			Dice mean ↑ (± std)	
initial	real T1 healthy	T1	0.47(±0.13)	0.56(±0.15)
	real T2 tumor	T2	0.27(±0.12)	0.36(±0.20)
after registration	real T1 healthy	T1	0.62(±0.14)	0.60(±0.14)
	gen. T2 healthy	T2	0.65(±0.08)	0.66(±0.11)
	gen. T2 tumor no def.	T2	0.62(±0.11)	0.63(±0.10)
	gen. T2 tumor def.	T2	0.47(±0.19)*	0.47(±0.21)*
	real T2 tumor	*T2*	*0.43(±0.14)*	*0.40(±0.20)*

5.3.4 Training of Neural Networks with Synthetic Pathological Images

In the following experiments, the impact of tumors and their mass effect in the training and testing data are investigated in two scenarios: 1) supervised deformable image registration; 2) segmentation of anatomical structures (here: ventricles and caudate nuclei). Suitable neural networks are trained on the different generated and real image datasets and tested on real brain MRIs with or without the presence of tumors.

Image Registration: To underline the necessity of ground truth data for image registration, a supervised approach is required. In this example, the FlowNet [Dosovitskiy et al., 2015] architecture is chosen, where dense displacement fields are estimated directly from a pair of images. Here, the ground truth displacements of the LPBA40 dataset are generated using the pairwise registration approach from [Ehrhardt et al., 2015] and directly transferred to the domain translated images, enabled due to topology preserving. Predicted tumor-induced deformations are integrated directly into these ground truth displacements. Since the registration of two pathological images is infeasible due to the large amount of missing correspondences, a pathological image is registered to a healthy image in the pathological training case. In the inference phase, image-to-atlas registration is established by registering the real T1 healthy images to a T1 atlas and the real pathological T2 images to a T2 atlas. Because of the high computational demand of the FlowNet architecture, this experiment is performed only for 2D axial slices.

In order to ensure the stability of the results, each training setup is executed ten times with different random seeds. The results are evaluated in terms of average Dice

Table 5.2: Registration and segmentation results. The different training and testing scenarios are evaluated in terms of mean Dice coefficients (± standard deviation) over 10 random seed training runs for two structures: ventricles and caudate nuclei. Superscript (^{A}) corresponds to random elastic deformations [Dosovitskiy et al., 2015], "naive" is the naive deformation approach from [Pfarrkirchner et al., 2018], "def" is the proposed deformation approach. *Italic* numbers correspond to the baseline; **bold** indicates the best and statistically significant ($p < 0.005$ in a two-tailed paired t-test) results for each experiment.

	train on	**test on**	**ventricles**	**caudate n.**
			Dice mean ↑ (± std)	
reg. 2D	none (init)	real T1 healthy	0.52(±0.19)	0.52(±0.19)
	real T1 healthy	real T1 healthy	*0.62(±0.17)*	*0.62(±0.17)*
	none (init)	real T2 tumor	0.38(±0.16)	0.39(±0.16)
	real T1 healthy		0.48(±0.15)	0.49(±0.15)
	gen. T2 healthy		0.52(±0.16)	0.52(±0.16)
	gen. T2 tumor		0.50(±0.16)	0.49(±0.16)
	gen. T2 tumor^A		0.44(±0.18)	0.43(±0.18)
	gen. T2 tumor, naive		0.44(±0.19)	0.45(±0.19)
	gen. T2 tumor, def.		**0.55(±0.14)**	**0.55(±0.14)**
	gen. T2 tumor, def.^{A}		0.41(±0.19)	0.41(±0.19)
seg. 2D	real T1 healthy	real T1 healthy	*0.87(±0.13)*	*0.83(±0.23)*
	real T1 healthy	real T2 tumor	0.00(±0.00)	0.01(±0.03)
	gen. T2 healthy		0.64(±0.16)	0.53(±0.23)
	gen. T2 tumor		0.59(±0.21)	0.60(±0.24)
	gen. T2 tumor^A		0.67(±0.14)	0.61(±0.24)
	gen. T2 tumor, naive		0.52(±0.21)	0.49(±0.25)
	gen. T2 tumor, def.		0.67(±0.15)	0.59(±0.24)
	gen. T2 tumor, def.^{A}		**0.71(±0.12)**	**0.64(±0.24)**
seg. 3D	real T1 healthy	real T1 healthy	*0.80(±0.10)*	*0.61(±0.37)*
	real T1 healthy	real T2 tumor	0.01(±0.01)	0.00(±0.01)
	gen. T2 healthy		0.59(±0.14)	0.47(±0.17)
	gen. T2 tumor		0.53(±0.19)	0.51(±0.20)
	gen. T2 tumor^A		0.64(±0.14)	0.56(±0.16)
	gen. T2 tumor,naive		0.62(±0.15)	0.21(±0.14)
	gen. T2 tumor, def.		0.63(±0.16)	0.56(±0.16)
	gen. T2 tumor, def.^{A}		**0.70(±0.11)**	**0.57(±0.17)**

overlaps of the observed labels (ventricles and caudate nuclei) between each test image and the atlas (Tab. 5.2). It can be observed that the registration results for real pathological images can be significantly improved by using synthetic images with tumors and tumor-induced deformations. In contrary, training without simulated tumor-induced deformations or naive tumor-induced deformations [Pfarrkirchner et al., 2018] does not improve the registration abilities of the network since unnatural deformations usually impair registration. Furthermore, the proposed deformation approach leads to better results than simply applying elastic deformation augmentation (as in [Dosovitskiy et al., 2015]). In fact, random deformations significantly deteriorate the network's performance which, consistently with [Uzunova et al., 2017], leads to the conclusion that only carefully modeled realistic deformations improve registration results. Overall, this experiment shows the importance of realistic pathology-induced deformations for the training of registration neural networks.

Semantic Segmentation: In this experiment, a state-of-the-art U-Net [Ronneberger et al., 2015] is used for the semantic segmentation of the brain ventricles and caudate nuclei. Here, both a 2D and a 3D architecture are considered. The training is established in a strictly supervised manner by using the labels of the original healthy IXI data directly as ground truth for the domain translated images. However, when using deformations, the labels are deformed accordingly. Also, anatomical labels covered by tumor tissue are not included in the calculations. The results are evaluated analogously to the image registration.

Tab. 5.2 shows that the 2D segmentation results are similar to the 3D results with the best setup yielding mean Dice coefficients of $\mathbf{0.71}(\pm\mathbf{0.12})/\mathbf{0.64}(\pm\mathbf{0.24})$. Again, naive deformations [Pfarrkirchner et al., 2018] do not deliver significantly better results compared to using no deformations. The segmentation can be significantly improved by using the generated domain-translated images with injected tumors and simulated tumor-induced deformations. It can also be observed, that adding tumors to the generated images improves the segmentation of the smaller structures, yet, the segmentation of the larger ventricles is impaired. This can be explained with the presence of subtle tumors in the test images that do not impact the segmentation of the ventricles (Fig. 5.8). This indicates that mixing the training datasets with and without tumors might be useful. However, when tumor-induced deformations are used for training, the segmentation of all structures is enhanced. For this scenario, the best result is achieved when combining random elastic deformations and the proposed tumor-induced deformations during training. Overall, those experiments show the importance of pathology-induced tissue deformation since it has a significant impact on classical tasks like segmentation and registration.

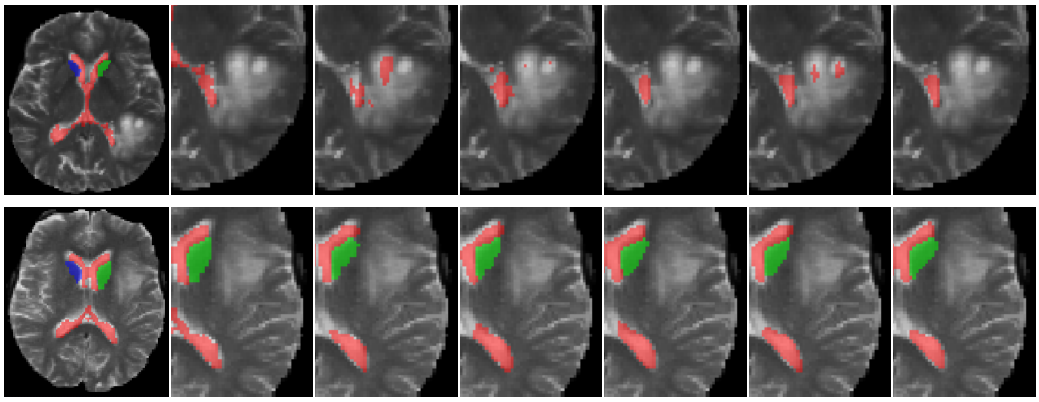


Fig. 5.8: Example segmentations of 3D pathological images with a tumor impacting the considered structures (top) and a tumor not impacting the target structures (bottom). Col. 1–2: ground truth image and zoomed to the tumor region; col. 3–7: segmentations of U-Net trained with different fake images.

5.4 Discussion and Conclusion

Previously, a method for the generation of pathological images by injecting pathologies into the healthy tissue has been proposed. This approach, however, does not consider the healthy tissue deformations around the pathological structures. Yet, the modeling of such deformations is of crucial importance for automatic image processing algorithms, e.g. the deep learning-based segmentation of healthy structures might fail if they are deformed by a pathology and such deformations were not learned during the training process.

In this chapter, the approach is extended by the simulation of pathology-induced deformations to boost the realism of the synthetic pathological images generated by the previously presented approach. The challenge for such approaches most commonly lies in the complex nature of these deformations, tightly connected to the pathology-type and other patient-related factors. Furthermore, longitudinal data are hardly ever available, hence, the real ground truth of the healthy tissue prior to the pathology occurrence is typically unknown. For this reason, a novel approach that features an inverse probabilistic technique for the estimation of pathological deformations directly from a single pathological image is presented here. The intuition that the healthy shape of the anatomical structures is easier to learn than the variety of possible pathology-induced deformations enables an approximation of a possible healthy shape from a given pathological shape. Using a probabilistic approach leads to multiple possible realistic deformations per pathology, outlining the data situation where an absolute ground truth is not available. The estimation of healthy shapes from given pathol-

ogy displaced shapes results into the inverse pathology-induced deformation. Due to a sophisticated velocity-based diffeomorphic displacement modeling, the inverse displacement can be easily calculated and resembles the pathology-induced deformation. Combined with the previous image generation method, a synthetic pathological image can be warped with the resulting displacement field to achieve realistic images featuring a pathology and the corresponding pathology-induced deformation.

In the performed experiments, the focus lies on brain tumor MRIs with the presence of the so-called tumor mass effect, that causes distortions by pushing the surrounding tissue away from the tumor center. Two scenarios for the employment of the synthetically generated images are considered here: 1) evaluate a standard algorithm on synthetic pathological data and 2) train neural networks on synthetically generated pathological images. The presented scenarios are enabled due to the fact that the generated images contain ground truth labels of the tumors as well as the anatomical regions surrounding them. An ablation study investigates the usage of real pathological, real healthy and synthetic pathological images with simply overlaying tumor tissue compared to the impact of the mass effect by using the proposed inverse probabilistic tumor-induced tissue deformation approach.

Overall the experiments show a significant improvement of the training and assessment results when synthetic pathological images are used. Furthermore, the importance of the simulation of the tumor mass effect for common tasks like registration and segmentation is emphasized. For the evaluation, a trained registration network for healthy brains is applied on the real and the synthesized images containing tumors. As expected, the network performs worse on data containing pathological structures, yet, using the synthesized images, this error can be estimated reliably since the network's performance on the real pathological images is statistically equivalent. Furthermore, a standard segmentation and a standard registration network are trained on different datasets and tested on real pathological images. The best results are achieved, when synthetic pathological images with tumor-induced deformations are used for training. These results indicate that the generated images and deformations are of realistic appearance and roughly mirror the real distribution of images containing pathologies. Still, their direct application for clinical use is not desired, i.a. because of the challenging evaluation of the feasibility of the pathology-induced deformations. However, the selected application scenarios show that synthetically generated images are crucial for the enhancement of automatic image processing algorithms and boost the performance of neural networks.

Chapter 6

Summary and Conclusion

Medical images containing pathological structures remain a challenge for many image processing methods. In this work, solutions based on generative deep learning models are presented in order to overcome the difficulties connected to the automatic processing of images with the presence of pathological structures. Several main goals were achieved in this work: reduce the number of required annotated samples for the training of neural networks, facilitate image processing methods like segmentation and registration targeting the normal anatomical structures in pathological data and improve the training and evaluation of neural networks. After introducing the foundations of the work in Chapter 2, the main contributions presented in chapters 3–5 can be grouped into indirect approaches for the modeling for pathological structures and direct approaches for the synthesis of pathological data.

In Chapter 2, the most common generative deep learning methods are presented including their mathematical foundations. Furthermore, a systematic comparison of the deep learning approaches to two baseline statistical generative models shows the clear advantages of the deep learning methods towards conventional statistical models and justifies the choice of the approaches as a backbone of this work. Also, a comparison of the approaches among each other reveals their advantages and useful properties for the further developments of the work: GANs deliver particularly realistically looking images, while VAEs yield various interesting properties in terms of low-dimensional image representation.

Chapter 3 mainly copes with the problem of reducing the number of needed annotated training samples for pathology localization and detection. The general assumption of the developed methodology is that pathological structures can be modeled indirectly by representing them as deviations from a learned healthy norm. This is achieved under the use of VAEs with the utilization of their property to capture the variability of a healthy training set in the image space as well as the latent space. Thus, by training a VAE on healthy subjects images, pathological or otherwise anomalous structures can be detected in the test images. Pathologies can be recognized as regions having the largest distances to their corresponding reconstructions and also to the learned latent space distribution. Indeed, the conducted experiments show that these assumptions hold and a rough detection of pathological structures is enabled. Even

though the method achieves only a rough localization of the pathological structures and not an exact pixel-wise segmentation, it is shown that it can be successfully applied to boost registration of pathological data by integrating a weight mask of the pathologies during registration. Furthermore, the main idea behind the developed methodology can be applied for the explanation of black-box pathology classifiers, delivering two advantages: first, a visual explanation of the classified pathology can support the interpretability of black-box approaches, and second, the regions used for an explanation of a given pathological class roughly resemble an unsupervised pathology detection.

While Chapter 3 concentrates on the indirect modeling of pathological structures, the next Chapter 4 aims for the direct modeling of pathological structures using a GAN-based method. This approach relies on the ability of GANs to generate images of high quality in order to synthesize realistic medical images featuring pathologies. To support the main purpose to reduce the amount of required annotated data, a method that is able to generate pathological data with the direct availability of anatomical and pathological ground truth labels is designed. A significant hurdle in this process is, however, the fact that GANs require an enormous amount to computational resources (mainly GPU RAM). Thus, applying GANs for large 3D medical images is intractable and hardly investigated in the literature. For this reason, the novel MEGAN approach is developed in order to enable the application of GANs for the generation of large high-quality 3D medical images with a constant memory demand regardless the image size. The realistic appearance of the images and the ability to generate artifact-free large medical volumes is shown in multiple experiments.

The developed approach can be used to synthesize realistic large-scale annotated images containing pathological structures, however, the pathological structure is typically only overlayed over the healthy tissue. This is of course not fully realistic since pathologies tend to displace the healthy surrounding tissue and, thus, cause pathology-induced deformations. Such deformations are hard to capture due to the common lack of image acquisitions before the appearance of the pathology and due to the complex and versatile nature of pathology-induced deformations. Thus, in Chapter 5, this problem is approached in an inverse probabilistic manner by predicting a healthy appearance from a pathological one and, thus, approximating possible pathology-induced deformations. The images containing pathological structures and the tissue deformations generated by them are successfully used in a multitude of scenarios to improve the assessment and training of neural networks engaging with tasks like image registration and segmentation.

Overall, several conclusions can be drawn from the developed methods and their thorough investigation in terms of various application scenarios:

- Automatic image processing methods such as registration and segmentation encounter difficulties in the presence of pathologies in the image data. Registration methods, for example, cannot reliably deal with missing correspondences caused by

pathological structures, while segmentation neural networks need a large amount of annotated pathological training data. Furthermore, methods are oftentimes particularly designed for the application on a very specific pathology type or for images of healthy appearance and, thus, fail to generalize to the large variability of available pathology types.

- The presented unsupervised VAE method is not exact and delivers a rough detection of the pathological structures, which is partly due to the blurry images resulting from VAEs. However, no pathological training data and no ground truth annotations are required for training and a rather reliable pathology localization can still be established. Such rough pathology detection approaches are still important in the medical image processing field and can be utilized as a pre-processing step e.g. for a registration method in order to integrate some prior knowledge about the presence of pathological structures.
- In terms of image generation abilities, GANs yield much more realistic results than autoencoder-based methods. Thus, a direct synthesis of pathological images is enabled using GAN-based approaches and even a direct generation of ground truth labels of relevant anatomical and pathological structures is demonstrated. Such generated images can significantly decrease the number of needed real samples for the training of neural networks. However, some real samples are still required for the training of the GANs.
- The lack of annotations of the healthy anatomical structures in medical images is a serious problem, especially when it comes to the assessment of neural networks (or other automatic algorithms) targeting anatomical structures. The error of neural networks trained on healthy subjects but applied on images with the presence of pathologies is typically much higher and its reliable estimation is crucial. This can also be enabled by using synthetic images.
- The consideration of pathology-induced deformations is of crucial importance for the generation of synthetic datasets for training and assessment, since otherwise the significant change of the shapes of the surrounding structures might not be learned by a neural network.
- The presented approaches, especially the image generation method, need to be taken with caution. Synthetic images are not generated for the direct clinical application or consideration by a clinician for diagnostic purposes in any part of this work. Moreover, their aim is to enhance the performance of automatic image processing methods and facilitate the handling of pathological image data in the medical image analysis and processing field.

Overall, this work features several novel methodologies in the field of generative models for medical images featuring pathologies. And even though an extensive ex-

perimental analysis is presented, future work might gain improvement by considering more diverse pathological types than the most frequently observed brain tumors in this work. An especially interesting research question, is e.g. the modeling of the extreme distortions caused by pathological fluids between the retinal layers in OCT acquisitions. Furthermore, more convincing evaluation strategies of the realism of generated images and especially the pathology-induced deformations can be developed in order to underline the implicit evaluation results through using the images as training data. Naturally, it is of great importance to pursue further novel developments in the field of deep learning-based generative models like diffeomorphic autoencoders and consider their advantages for the presented purposes [Uzunova et al., 2021a].

In conclusion, this work develops methods based on deep learning generative models that are able to successfully facilitate the handling of images featuring pathologies, reduce the number of needed training samples and overall improve the training of neural networks for the processing of image data containing pathological structures.

References

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning*, pages 214–223. PMLR.

[Armanious et al., 2019] Armanious, K., Jiang, C., Abdulatif, S., Küstner, T., Gatidis, S., and Yang, B. (2019). Unsupervised Medical Image Translation Using Cycle-MedGAN. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5.

[Armanious et al., 2020] Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., and Yang, B. (2020). MedGAN: Medical image translation using GANs. *Computerized Medical Imaging and Graphics*, 79:101684.

[Audebert et al., 2019] Audebert, N., Boulch, A., Le Saux, B., and Lefèvre, S. (2019). Distance transform regression for spatially-aware deep semantic segmentation. *Computer Vision and Image Understanding*, 189:102809.

[Bakas et al., 2017] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., and Davatzikos, C. (2017). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*, 4:170117.

[Banerjee and Ghose, 2020] Banerjee, R. and Ghose, A. (2020). A Semi-Supervised Approach For Identifying Abnormal Heart Sounds Using Variational Autoencoder. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1249–1253.

[Baur et al., 2020] Baur, C., Wiestler, B., Albarqouni, S., and Navab, N. (2020). Scale-Space Autoencoders for Unsupervised Anomaly Segmentation in Brain MRI. In Martel, A. L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M. A., Zhou, S. K., Racoceanu, D., and Joskowicz, L., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Lecture Notes in Computer Science, pages 552–561, Cham. Springer International Publishing.

[Baur et al., 2021] Baur, C., Wiestler, B., Muehlau, M., Zimmer, C., Navab, N., and Albarqouni, S. (2021). Modeling Healthy Anatomy with Artificial Intelligence for Unsupervised Anomaly Detection in Brain MRI. *Radiol. Artif. Intell.*, 3(3):e190169.

[Boedeker et al., 2004] Boedeker, K., McNitt-Gray, M., Rogers, S., Truong, D., Brown, M., Gjertson, D., et al. (2004). Emphysema: Effect of Reconstruction Algorithm on CT Imaging Measures. *Radiology*, 232(1).

[Bogunović et al., 2019] Bogunović, H., Venhuizen, F., Klimscha, S., Apostolopoulos, S., Bab-Hadiashar, A., Bagci, U., Beg, M. F., Bekalo, L., Chen, Q., Ciller, C., Gopinath, K., Gostar, A. K., Jeon, K., Ji, Z., Kang, S. H., Koozekanani, D. D., Lu, D., Morley, D., Parhi, K. K., Park, H. S., Rashno, A., Sarunic, M., Shaikh, S., Sivaswamy, J., Tennakoon, R., Yadav, S., De Zanet, S., Waldstein, S. M., Gerendas, B. S., Klaver, C., Sánchez, C. I., and Schmidt-Erfurth, U. (2019). RETOUCH: The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge. *IEEE Transactions on Medical Imaging*, 38(8):1858–1874.

[Buades et al., 2011] Buades, A., Coll, B., and Morel, J.-M. (2011). Non-Local Means Denoising. *Image Processing On Line*, 1:208–212.

[Castillo et al., 2013] Castillo, R., Castillo1, E., Fuentes, D., Ahmad, M., Wood1, A. M., Ludwig1, M. S., et al. (2013). A Reference Dataset for Deformable Image Registration Spatial Accuracy Evaluation using the COPDgene Study Archive. *Phys Med Biol*, 58(9):2861–2877.

[Chen et al., 2017] Chen, M., Shi, X., Zhang, Y., Wu, D., and Guizani, M. (2017). Deep Features Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network. *IEEE Transactions on Big Data*, pages 1–1.

[Chen et al., 2018] Chen, Y., Shi, F., Christodoulou, A. G., Zhou, Z., Xie, Y., and Li, D. (2018). Efficient and Accurate MRI Super-Resolution Using a Generative Adversarial Network and 3D Multi-level Densely Connected Network. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 91–99.

[Chitphakdithai and Duncan, 2010] Chitphakdithai, N. and Duncan, J. S. (2010). Pairwise registration of images with missing correspondences due to resection. *Proc IEEE Int Symp Biomed Imaging*, 2010:1025–1028.

[Cohen et al., 2018] Cohen, J. P., Luck, M., and Honari, S. (2018). Distribution Matching Losses Can Hallucinate Features in Medical Image Translation. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 529–536.

[Cootes et al., 1998] Cootes, T. F., Edwards, G. J., and Taylor, C. J. (1998). Active appearance models. In Burkhardt, H. and Neumann, B., editors, *Computer Vision — ECCV'98*, Lecture Notes in Computer Science, pages 484–498, Berlin, Heidelberg. Springer.

[Cootes et al., 1995] Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active Shape Models-Their Training and Application. *Computer Vision and Image Understanding*, 61(1):38–59.

[Dabov et al., 2007] Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. (2007). Image denoising by sparse 3-D transform-domain collaborative filtering. *Trans. Img. Proc.*, 16(8):2080–2095.

[Dalca et al., 2018] Dalca, A. V., Balakrishnan, G., Guttag, J., and Sabuncu, M. (2018). Unsupervised Learning for Fast Probabilistic Diffeomorphic Registration. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*.

[Davatzikos et al., 2003] Davatzikos, C., Tao, X., and Shen, D. (2003). Hierarchical active shape models, using the wavelet transform. *IEEE Transactions on Medical Imaging*, 22(3):414–423.

[Davies et al., 2002] Davies, R. H., Twining, C. J., Cootes, T. F., Waterton, J. C., and Taylor, C. J. (2002). A minimum description length approach to statistical shape modeling. *IEEE Transactions on Medical Imaging*, 21(5):525–537.

[Denton et al., 2015] Denton, E. L., Chintala, S., Szlam, A., and Fergus, R. (2015). Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In *Advances in Neural Information Processing Systems*, pages 1486–1494.

[Doersch, 2021] Doersch, C. (2021). Tutorial on Variational Autoencoders. *arXiv:1606.05908 [cs, stat]*.

[Dosovitskiy et al., 2015] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., and Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision - ICCV 2015*, pages 2758–2766.

[Ehrhardt et al., 2016] Ehrhardt, J., Jacob, F., Handels, H., and Frydrychowicz, A. (2016). Comparison of Post-Hoc Normalization Approaches for CT-based Lung Emphysema Index Quantification. In *Bildverarbeitung Für Die Medizin BVM*, pages 44–49.

[Ehrhardt et al., 2015] Ehrhardt, J., Schmidt-Richberg, A., Werner, R., and Handels, H. (2015). Variational Registration - A Flexible Open-Source ITK Toolbox for Nonrigid Image Registration. In *Bildverarbeitung Für Die Medizin 2015*, Informatik Aktuell, pages 209–214, Lübeck.

[Elazab et al., 2020] Elazab, A., Wang, C., Gardezi, S. J. S., Bai, H., Hu, Q., Wang, T., Chang, C., and Lei, B. (2020). GP-GAN: Brain tumor growth prediction using

stacked 3D generative adversarial networks from longitudinal MR Images. *Neural Networks*, 132:321–332.

[Emami et al., 2021] Emami, H., Aliabadi, M. M., Dong, M., and Chinnam, R. B. (2021). SPA-GAN: Spatial Attention GAN for Image-to-Image Translation. *IEEE Transactions on Multimedia*, 23:391–401.

[Ezhov et al., 2019] Ezhov, I., Lipkova, J., Shit, S., Kofler, F., Collomb, N., Lemasson, B., Barbier, E., and Menze, B. (2019). Neural Parameters Estimation for Brain Tumor Growth Modeling. In Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., and Khan, A., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer Science, pages 787–795, Cham. Springer International Publishing.

[Faktor and Irani, 2012] Faktor, A. and Irani, M. (2012). Clustering by composition – unsupervised discovery of image categories. In *European Conference on Computer Vision – ECCV*, pages 474–487.

[Farsiu et al., 2014] Farsiu, S., Chiu, S. J., O'Connell, R. V., Folgar, F. A., Yuan, E., Izatt, J. A., Toth, C. A., and Age-Related Eye Disease Study 2 Ancillary Spectral Domain Optical Coherence Tomography Study Group (2014). Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology*, 121(1):162–172.

[Fischl, 2012] Fischl, B. (2012). FreeSurfer. *Neuroimage*, 62:774–781.

[Fleitmann et al., 2021] Fleitmann, M., Uzunova, H., Stroth, A. M., Gerlach, J., Fürschke, A., Barkhausen, J., Bischof, A., and Handels, H. (2021). Deep-learning-based feature encoding of clinical parameters for patient specific CTA dose optimization. In *Wireless Mobile Communication and Healthcare: 9th EAI International Conference, MobiHealth 2020, Virtual Event, November 19, 2020, Proceedings 9*, pages 315–322. Springer International Publishing.

[Fong and Vedaldi, 2017] Fong, R. C. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *The IEEE International Conference on Computer Vision (ICCV)*.

[Fonov et al., 2011] Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., and Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313–327.

[Frid-Adar et al., 2018] Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*.

[Gierada et al., 2010] Gierada, D., Bierhals, A., Choong, C., Bartel, S., Ritter, J., Das, N., et al. (2010). Effects of CT Section Thickness and Reconstruction Kernel on Emphysema Quantification: Relationship to the Magnitude of the CT Emphysema Index. *Academic Radiology*, 17(2):146–156.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

[Guan and Loew, 2020] Guan, S. and Loew, M. (2020). An Internal Cluster Validity Index Using a Distance-based Separability Measure. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 827–834.

[Hammers et al., 2003] Hammers, A., Allom, R., Koepp, M. J., Free, S. L., Myers, R., Lemieux, L., Mitchell, T. N., Brooks, D. J., and Duncan, J. S. (2003). Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Human Brain Mapping*, 19(4):224–247.

[Han et al., 2020] Han, X., Shen, Z., Xu, Z., Bakas, S., Akbari, H., Bilello, M., Davatzikos, C., and Niethammer, M. (2020). A Deep Network for Joint Registration and Reconstruction of Images with Pathologies. In Liu, M., Yan, P., Lian, C., and Cao, X., editors, *Machine Learning in Medical Imaging*, Lecture Notes in Computer Science, pages 342–352, Cham. Springer International Publishing.

[Hanif, 2019] Hanif, M. S. (2019). Patch match networks: Improved two-channel and Siamese networks for image patch matching. *Pattern Recognition Letters*, 120:54–61.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

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

[Henry et al., 2021] Henry, T., Carre, A., Lerousseau, M., Estienne, T., Robert, C., Paragios, N., and Deutsch, E. (2021). Brain Tumor Segmentation with Self-ensembled, Deeply-Supervised 3D U-Net Neural Networks: A BraTS 2020 Challenge Solution. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 327–339. Springer International Publishing, Cham.

[Heusel et al., 2017] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge

to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6629–6640, Red Hook, NY, USA. Curran Associates Inc.

[Hinton and Salakhutdinov, 2006] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507.

[Hogea et al., 2007] Hogea, C., Davatzikos, C., and Biros, G. (2007). Modeling Glioma Growth and Mass Effect in 3D MR Images of the Brain. In Ayache, N., Ourselin, S., and Maeder, A., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007*, Lecture Notes in Computer Science, pages 642–650, Berlin, Heidelberg. Springer.

[Hu et al., 2015] Hu, Y., Gibson, E., Ahmed, H. U., Moore, C. M., Emberton, M., and Barratt, D. C. (2015). Population-based prediction of subject-specific prostate deformation for MR-to-ultrasound image registration. *Medical Image Analysis*, 26(1):332–344.

[Huang et al., 2017] Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3.

[Isensee et al., 2021] Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*, 18(2):203–211.

[Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976.

[Jaderberg et al., 2015] Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. (2015). Spatial Transformer Networks. *Advances in Neural Information Processing Systems*, 28:2017–2025.

[Jam et al., 2021] Jam, J., Kendrick, C., Drouard, V., Walker, K., Hsu, G.-S., and Yap, M. (2021). Symmetric Skip Connection Wasserstein GAN for High-resolution Facial Image Inpainting. In *16th International Conference on Computer Vision Theory and Applications*, pages 35–44.

[Jimenez-del-Toro et al., 2016] Jimenez-del-Toro, O., Müller, H., Krenn, M., Gruenberg, K., Taha, A. A., Winterstein, M., et al. (2016). Cloud-Based Evaluation of Anatomical Structure Segmentation and Landmark Detection Algorithms: VISCERAL Anatomy Benchmarks. *IEEE Transactions on Medical Imaging*, 35(11):2459–2475.

[Jin et al., 2019] Jin, H., Heo, C., and Kim, J. H. (2019). Deep learning-enabled accurate normalization of reconstruction kernel effects on emphysema quantification in low-dose CT. *Phys. Med. Biol.*, 64(13):135010.

[Jog et al., 2017] Jog, A., Carass, A., Roy, S., Pham, D. L., and Prince, J. L. (2017). Random forest regression for magnetic resonance image synthesis. *Medical Image Analysis*, 35:475–488.

[Kaji and Kida, 2019] Kaji, S. and Kida, S. (2019). Overview of image-to-image translation by use of deep neural networks: Denoising, super-resolution, modality conversion, and reconstruction in medical imaging. *Radiol Phys Technol*, 12(3):235–248.

[Kamnitsas et al., 2016] Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A. V., Criminisi, A., Rueckert, D., and Glocker, B. (2016). DeepMedic for brain tumor segmentation. In *BrainLes 2016, with the Challenges on BRATS, ISLES and mTOP, MICCAI*, pages 138–149.

[Karaca et al., 2020] Karaca, O., Buyukmert, A., Tepe, N., Ozcan, E., and Kus, I. (2020). Volume estimation of brain ventricles using Cavalieri's principle and Atlas-based methods in Alzheimer disease: Consistency between methods. *J Clin Neurosci*, 78:333–338.

[Karimi et al., 2018] Karimi, D., Samei, G., Kesch, C., Nir, G., and Salcudean, S. E. (2018). Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models. *International Journal of Computer Assisted Radiology and Surgery*, 13(8):1211–1219.

[Karras et al., 2018] Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

[Karras et al., 2019] Karras, T., Laine, S., and Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410.

[Karras et al., 2020] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119.

[Kingma and Welling, 2014] Kingma, D. and Welling, M. (2014). Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, Banff, Canada.

[Kirschner et al., 2011] Kirschner, M., Becker, M., and Wesarg, S. (2011). 3D Active Shape Model Segmentation with Nonlinear Shape Priors. In Fichtinger, G., Martel, A., and Peters, T., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011*, Lecture Notes in Computer Science, pages 492–499, Berlin, Heidelberg. Springer.

[Klein et al., 2009] Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G. E., Collins, D. L., Gee, J., Hellier, P., Song, J. H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R. P., Mann, J. J., and Parsey, R. V. (2009). Evaluation of 14 Nonlinear Deformation Algorithms Applied to Human Brain MRI Registration. *Neuroimage*, 46(3):786–802.

[Kohl et al., 2018] Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K., Eslami, S. M. A., Jimenez Rezende, D., and Ronneberger, O. (2018). A Probabilistic U-Net for Segmentation of Ambiguous Images. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 6965–6975. Curran Associates, Inc.

[Krüger et al., 2015] Krüger, J., Ehrhardt, J., and Handels, H. (2015). Probabilistic appearance models for segmentation and classification. In *International Conference on Computer Vision – ICCV*, pages 1698–1706.

[Krüger et al., 2017] Krüger, J., Ehrhardt, J., and Handels, H. (2017). Statistical appearance models based on probabilistic correspondences. *Medical Image Analysis*, 37:146–159.

[Krüger et al., 2020] Krüger, J., Schultz, S., Handels, H., and Ehrhardt, J. (2020). Registration with probabilistic correspondences — Accurate and robust registration for pathological and inhomogeneous medical data. *Computer Vision and Image Understanding*, 190:102839.

[Kwon et al., 2014] Kwon, D., Niethammer, M., Akbari, H., Bilello, M., Davatzikos, C., and Pohl, K. M. (2014). PORTR: Pre-Operative and Post-Recurrence Brain Tumor Registration. *IEEE Trans Med Imaging*, 33(3):651–667.

[Larsen et al., 2016] Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, pages 1558–1566.

[Lecun et al., 1998] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324.

[Lei et al., 2019] Lei, Y., Wang, T., Liu, Y., Higgins, K., Tian, S., Liu, T., Mao, H., Shim, H., Curran, W. J., Shu, H.-K., and Yang, X. (2019). MRI-based synthetic CT generation using deep convolutional neural network. In *SPIE Medical Imaging*, volume 10949.

[Lotan and Irani, 2016] Lotan, O. and Irani, M. (2016). Needle-match: Reliable patch matching under high uncertainty. In *IEEE Conference on Computer Vision and Pattern Recognition – CVPR*.

[Lu et al., 2019] Lu, D., Heisler, M., Lee, S., Ding, G. W., Navajas, E., Sarunic, M. V., and Beg, M. F. (2019). Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network. *Medical Image Analysis*, 54:100–110.

[Maier et al., 2017] Maier, O., Menze, B. H., von der Gablentz, J., Häni, L., Heinrich, M. P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al. (2017). ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Medical Image Analysis*, 35:250–269.

[McKinley et al., 2016] McKinley, R., Häni, L., Wiest, R., and Reyes, M. (2016). Segmenting the Ischemic Penumbra: A Decision Forest Approach with Automatic Threshold Finding. In Crimi, A., Menze, B., Maier, O., Reyes, M., and Handels, H., editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Lecture Notes in Computer Science, pages 275–283, Cham. Springer International Publishing.

[Menze et al., 2015] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E. R., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., Corso, J. J., Criminisi, A., Das, T., Delingette, H., Demiralp, C., Durst, C. R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K. M., Jena, R., John, N. M., Konukoglu, E., Lashkari, D., Mariz, J. A., Meier, R., Pereira, S., Precup, D., Price, S. J., Raviv, T. R., Reza, S. M. S., Ryan, M. T., Sarikaya, D., Schwartz, L. H., Shin, H.-C., Shotton, J., Silva, C. A., Sousa, N., Subbanna, N. K., Székely, G., Taylor, T. J., Thomas, O. M., Tustison, N. J., Ünal, G. B., Vasseur, F., Wintermark, M., Ye, D. H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., and Leemput, K. V. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging*, 34(10):1993–2024.

[Mohamed and Davatzikos, 2005] Mohamed, A. and Davatzikos, C. (2005). Finite Element Modeling of Brain Tumor Mass-Effect from 3D Medical Images. In Duncan,

J. S. and Gerig, G., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005*, Lecture Notes in Computer Science, pages 400–408, Berlin, Heidelberg. Springer.

[Myronenko, 2019] Myronenko, A. (2019). 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. In Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., and van Walsum, T., editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Lecture Notes in Computer Science, pages 311–320, Cham. Springer International Publishing.

[Nie et al., 2018] Nie, D., Trullo, R., Lian, J., Wang, L., Petitjean, C., Ruan, S., and Wang, Q. (2018). Medical Image Synthesis with Deep Convolutional Adversarial Networks. *IEEE Transactions on Biomedical Engineering*, pages 1–1.

[Pawlowski et al., 2018] Pawlowski, N., Lee, M. C. H., Rajchl, M., McDonagh, S., Ferrante, E., Kamnitsas, K., Cooke, S., Stevenson, S. K., Khetani, A. M., Newman, T., Zeiler, F. A., Digby, R. J., Coles, J. P., Rueckert, D., Menon, D. K., Newcombe, V. F. J., and Glocker, B. (2018). Unsupervised lesion detection in brain CT using bayesian convolutional autoencoders. In *Medical Imaging with Deep Learning – MIDL*.

[Pfarrkirchner et al., 2018] Pfarrkirchner, B., Gsaxner, C., Schmalsteig, D., Egger, J., and Lindner, L. (2018). TuMore: Generation of synthetic brain tumor MRI data for deep learning based segmentation approaches. In Zhang, J. and Chen, P.-H., editors, *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, page 63, Houston, United States. SPIE.

[Qian et al., 2017] Qian, X., Lin, Y., Zhao, Y., Yue, X., Lu, B., and Wang, J. (2017). Objective Ventricle Segmentation in Brain CT with Ischemic Stroke Based on Anatomical Knowledge. *BioMed Research International*, 2017:e8690892.

[Rohé et al., 2017] Rohé, M.-M., Datar, M., Heimann, T., Sermesant, M., and Pennec, X. (2017). SVF-Net: Learning Deformable Image Registration Using Shape Matching. In Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D. L., and Duchesne, S., editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, Lecture Notes in Computer Science, pages 266–274, Cham. Springer International Publishing.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241, Cham. Springer International Publishing.

[Royston, 1982] Royston, J. P. (1982). An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(2):115–124.

[Salimans et al., 2016] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training GANs. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2234–2242, Red Hook, NY, USA. Curran Associates Inc.

[Schlegl et al., 2019] Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., and Schmidt-Erfurth, U. (2019). F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44.

[Schlegl et al., 2017] Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in Medical Imaging – IPMI*, pages 146–157.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

[Shattuck et al., 2008] Shattuck, D. W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K. L., Poldrack, R. A., Bilder, R. M., and Toga, A. W. (2008). Construction of a 3D probabilistic atlas of human cortical structures. *NeuroImage*, 39.

[Shin et al., 2018] Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., Andriole, K. P., and Michalski, M. (2018). Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks. In *Simulation and Synthesis in Medical Imaging*, pages 1–11.

[Sohn et al., 2015] Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems 28*, pages 3483–3491.

[Springenberg et al., 2015] Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2015). Striving for simplicity: The all convolutional net. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[Sudre et al., 2017] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M. (2017). Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In Cardoso, M. J., Arbel, T., Carneiro, G.,

Syeda-Mahmood, T., Tavares, J. M. R., Moradi, M., Bradley, A., Greenspan, H., Papa, J. P., Madabhushi, A., Nascimento, J. C., Cardoso, J. S., Belagiannis, V., and Lu, Z., editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Lecture Notes in Computer Science, pages 240–248, Cham. Springer International Publishing.

[Uzunova et al., 2020a] Uzunova, H., Ehrhardt, J., and Handels, H. (2020a). Generation of Annotated Brain Tumor MRIs with Tumor-induced Tissue Deformations for Training and Assessment of Neural Networks. In Martel, A. L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M. A., Zhou, S. K., Racoceanu, D., and Joskowicz, L., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Lecture Notes in Computer Science, pages 501–511, Cham. Springer International Publishing.

[Uzunova et al., 2020b] Uzunova, H., Ehrhardt, J., and Handels, H. (2020b). Memory-efficient GAN-based domain translation of high resolution 3D medical images. *Computerized Medical Imaging and Graphics*, 86:101801.

[Uzunova et al., 2019a] Uzunova, H., Ehrhardt, J., Jacob, F., Frydrychowicz, A., and Handels, H. (2019a). Multi-scale GANs for Memory-efficient Generation of High Resolution Medical Images. In Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., and Khan, A., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer Science, pages 112–120, Cham. Springer International Publishing.

[Uzunova et al., 2019b] Uzunova, H., Ehrhardt, J., Kepp, T., and Handels, H. (2019b). Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders. In *Medical Imaging 2019: Image Processing*, volume 10949, page 1094911. International Society for Optics and Photonics.

[Uzunova et al., 2018] Uzunova, H., Handels, H., and Ehrhardt, J. (2018). Unsupervised pathology detection in medical images using learning-based methods. In *Bildverarbeitung Für Die Medizin 2018*, pages 61–66. Springer.

[Uzunova et al., 2021a] Uzunova, H., Handels, H., and Ehrhardt, J. (2021a). Guided Filter Regularization for Improved Disentanglement of Shape and Appearance in Diffeomorphic Autoencoders. In *Medical Imaging with Deep Learning*.

[Uzunova et al., 2020c] Uzunova, H., Kaftan, P., Wilms, M., Forkert, N. D., Handels, H., and Ehrhardt, J. (2020c). Quantitative Comparison of Generative Shape Models for Medical Images. In Tolxdorff, T., Deserno, T. M., Handels, H., Maier, A., Maier-Hein, K. H., and Palm, C., editors, *Bildverarbeitung für die Medizin 2020*, Informatik aktuell, pages 201–207, Wiesbaden. Springer Fachmedien.

[Uzunova et al., 2021b] Uzunova, H., Kruse, J., Kaftan, P., Wilms, M., Forkert, N. D., Handels, H., and Ehrhardt, J. (2021b). Analysis of Generative Shape Modeling Approaches: Latent Space Properties and Interpretability. In *Bildverarbeitung Für Die Medizin 2021*, pages 344–349. Springer Fachmedien Wiesbaden.

[Uzunova et al., 2019c] Uzunova, H., Schultz, S., Handels, H., and Ehrhardt, J. (2019c). Evaluation of image processing methods for clinical applications. In *Bildverarbeitung Für Die Medizin 2019*, pages 15–20. Springer.

[Uzunova et al., 2019d] Uzunova, H., Schultz, S., Handels, H., and Ehrhardt, J. (2019d). Unsupervised pathology detection in medical images using conditional variational autoencoders. *International Journal of Computer Assisted Radiology and Surgery*, 14(3):451–461.

[Uzunova et al., 2017] Uzunova, H., Wilms, M., Handels, H., and Ehrhardt, J. (2017). Training CNNs for Image Registration from Few Samples with Model-based Data Augmentation. In Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D. L., and Duchesne, S., editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, Lecture Notes in Computer Science, pages 223–231, Cham. Springer International Publishing.

[Vincent et al., 2008] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, pages 1096–1103, Helsinki, Finland. ACM Press.

[Waheed et al., 2020] Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., and Pinheiro, P. R. (2020). CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection. *IEEE Access*, 8:91916–91923.

[Wang et al., 2018] Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018). High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, Salt Lake City, UT, USA. IEEE.

[Werner et al., 2014] Werner, R., Schmidt-Richberg, A., Handels, H., and Ehrhardt, J. (2014). Estimation of lung motion fields in 4D CT data by variational non-linear intensity-based registration: A comparison and evaluation study. *Physics in Medicine & Biology*, 59(15):4247.

[Wilms et al., 2020] Wilms, M., Ehrhardt, J., and Forkert, N. D. (2020). A Kernelized Multi-level Localization Method for Flexible Shape Modeling with Few Training Data. In Martel, A. L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M. A., Zhou, S. K., Racoceanu, D., and Joskowicz, L., editors, *Medical Image Computing*

and Computer Assisted Intervention – MICCAI 2020, Lecture Notes in Computer Science, pages 765–775, Cham. Springer International Publishing.

[Wilms et al., 2017] Wilms, M., Handels, H., and Ehrhardt, J. (2017). Multi-resolution multi-object statistical shape models based on the locality assumption. *Medical image analysis*, 38:17–29.

[Wu et al., 2018] Wu, E., Wu, K., Cox, D., and Lotter, W. (2018). Conditional Infilling GANs for Data Augmentation in Mammogram Classification. In Stoyanov, D., Taylor, Z., Kainz, B., Maicas, G., Beichel, R. R., Martel, A., Maier-Hein, L., Bhatia, K., Vercauteren, T., Oktay, O., Carneiro, G., Bradley, A. P., Nascimento, J., Min, H., Brown, M. S., Jacobs, C., Lassen-Schmidt, B., Mori, K., Petersen, J., San José Estépar, R., Schmidt-Richberg, A., and Veiga, C., editors, *Image Analysis for Moving Organ, Breast, and Thoracic Images*, Lecture Notes in Computer Science, pages 98–106, Cham. Springer International Publishing.

[Wu et al., 2016] Wu, J., Zhang, C., Xue, T., Freeman, W. T., and Tenenbaum, J. B. (2016). Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90.

[Xing et al., 2019] Xing, Y., Ge, Z., Zeng, R., Mahapatra, D., Seah, J., Law, M., and Drummond, T. (2019). Adversarial Pulmonary Pathology Translation for Pairwise Chest X-Ray Data Augmentation. In Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., and Khan, A., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer Science, pages 757–765, Cham. Springer International Publishing.

[Xue et al., 2017] Xue, W., Islam, A., Bhaduri, M., and Li, S. (2017). Direct multitype cardiac indices estimation via joint representation and regression learning. *IEEE Transactions on Medical Imaging*, 36(10):2057–2067.

[Yan et al., 2016] Yan, X., Yang, J., Sohn, K., and Lee, H. (2016). Attribute2Image: Conditional image generation from visual attributes. In *European Conference on Computer Vision – ECCV*, volume 9908, pages 776–791. Springer.

[Yang et al., 2018] Yang, G., Yu, S., Dong, H., Slabaugh, G., Dragotti, P. L., Ye, X., et al. (2018). DAGAN: Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction. *IEEE Transactions on Medical Imaging*, 37:1310–1321.

[Yang et al., 2020] Yang, Q., Li, N., Zhao, Z., Fan, X., Chang, E. I.-C., and Xu, Y. (2020). MRI Cross-Modality Image-to-Image Translation. *Scientific Reports*, 10(1):3753.

[Yang et al., 2021] Yang, S., Kim, E. Y., and Ye, J. C. (2021). Continuous Conversion of CT Kernel using Switchable CycleGAN with AdaIN. *IEEE Transactions on Medical Imaging*, pages 1–1.

[Yang et al., 2016] Yang, X., Han, X., Park, E., Aylward, S., Kwitt, R., and Niethammer, M. (2016). Registration of Pathological Images. *Simul Synth Med Imaging*, 9968:97–107.

[Yang et al., 2017] Yang, X., Kwitt, R., Styner, M., and Niethammer, M. (2017). Quicksilver: Fast predictive image registration – A deep learning approach. *NeuroImage*, 158:378–396.

[Yao et al., 2019] Yao, R., Liu, C., Zhang, L., and Peng, P. (2019). Unsupervised Anomaly Detection Using Variational Auto-Encoder based Feature Extraction. In *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pages 1–7.

[Yu et al., 2018] Yu, B., Zhou, L., Wang, L., Fripp, J., and Bourgeat, P. (2018). 3D cGAN based cross-modality MR image synthesis for brain tumor segmentation. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 626–630.

[Zhang et al., 2020] Zhang, Z., Sun, L., Zheng, Z., and Li, Q. (2020). Disentangling The Spatial Structure And Style In Conditional VAE. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1626–1630.

[Zhou et al., 2020] Zhou, L., Schaefferkoetter, J. D., Tham, I. W. K., Huang, G., and Yan, J. (2020). Supervised learning with cyclegan for low-dose FDG PET image denoising. *Medical Image Analysis*, 65:101770.

[Zhou Wang et al., 2004] Zhou Wang, Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

[Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251.

[Zimmerer et al., 2019] Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., and Maier-Hein, K. (2019). Unsupervised Anomaly Localization Using Variational Auto-Encoders. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 289–297.

Own Publications

[1] H. Uzunova, H. Handels, and J. Ehrhardt, "Guided Filter Regularization for Improved Disentanglement of Shape and Appearance in Diffeomorphic Autoencoders," in Medical Imaging with Deep Learning – MIDL, Proceedings of Machine Learning Research, 2021 (Code: `https://github.com/hristina-uzunova/GF_DAE`)

[2] H. Uzunova, J. Kruse, P. Kaftan, M. Wilms, N.D. Forkert, H. Handels, and J. Ehrhardt, "Analysis of Generative Shape Modeling Approaches: Latent Space Properties and Interpretability," in Bildverarbeitung für die Medizin, pp. 344-349, Springer Fachmedien Wiesbaden, 2021

[3] M. Fleitmann, H. Uzunova, A.M. Stroth, J. Gerlach, A. Fürschke, J. Barkhausen, A. Bischof, and H. Handels, "Deep-Learning-Based Feature Encoding of Clinical Parameters for Patient Specific CTA Dose Optimization," in EAI International Conference, MobiHealth, pp. 315-322, Springer International Publishing, 2020

[4] H. Uzunova, J. Ehrhardt, and H. Handels, "Generation of Annotated Brain Tumor MRIs with Tumor-induced Tissue Deformations for Training and Assessment of Neural Networks," in Medical Image Computing and Computer Assisted Intervention – MICCAI, pp. 501-511, Springer, Cham, 2020 (Code: `https://github.com/hristina-uzunova/TumorMassEffect`)

[5] H. Uzunova, J. Ehrhardt, and H. Handels, "Memory-efficient GAN-based Domain Translation of High Resolution 3D Medical Images," Computerized Medical Imaging and Graphics, vol. 86, p. 101801, 2020 (Code: `https://github.com/hristina-uzunova/MEGAN`)

[6] H. Uzunova, P. Kaftan, M. Wilms, N. D. Forkert, H. Handels, and J. Ehrhardt, "Quantitative Comparison of Generative Shape Models for Medical Images," in Bildverarbeitung für die Medizin, pp. 201-207, Springer Vieweg, Wiesbaden, 2020

[7] H. Uzunova, J. Ehrhardt, F. Jacob, A. Frydrychowicz, and H. Handels, "Multi-scale GANs for Memory-efficient Generation of High Resolution Medical Images," in Medical Image Computing and Computer Assisted Intervention – MICCAI, pp. 112-120, Springer, Cham, 2019

[8] H. Uzunova, J. Ehrhardt, T. Kepp, and H. Handels, "Interpretable Explanations of Black Box Classifiers Applied on Medical Images by Meaningful Perturbations using Variational Autoencoders," in SPIE Medical imaging: Image processing, vol. 10949, p. 1094911, International Society for Optics and Photonics, 2019

[9] J. Ehrhardt, M. Ahlborg, H. Uzunova, T. M. Buzug, and H. Handels, "Temporal Polyrigid Registration for Patch-based MPI Reconstruction of Moving Objects," International Journal on Magnetic Particle Imaging, vol. 5, no. 1-2, 2019

[10] H. Uzunova, S. Schultz, H. Handels, and J. Ehrhardt, "Evaluation of Image Processing Methods for Clinical Applications," in Bildverarbeitung für die Medizin, pp. 15-20, Springer Vieweg, Wiesbaden, 2019

[11] H. Uzunova, S. Schultz, H. Handels, and J. Ehrhardt, "Unsupervised Pathology Detection in Medical Images using Conditional Variational Autoencoders," International Journal of Computer Assisted Radiology and Surgery, vol. 14, no. 3, pp. 451-461, 2019

[12] H. Uzunova, H. Handels, and J. Ehrhardt, "Unsupervised Pathology Detection in Medical Images using Learning-based Methods," in Bildverarbeitung für die Medizin, pp. 61-66, Springer Vieweg, Berlin, 2018

[13] H. Uzunova, M. Wilms, H. Handels, and J. Ehrhardt, "Training CNNs for Image Registration from Few Samples with Model-based Data Augmentation," in Medical Image Computing and Computer Assisted Intervention – MICCAI, pp. 223-231, Springer, Cham, 2017

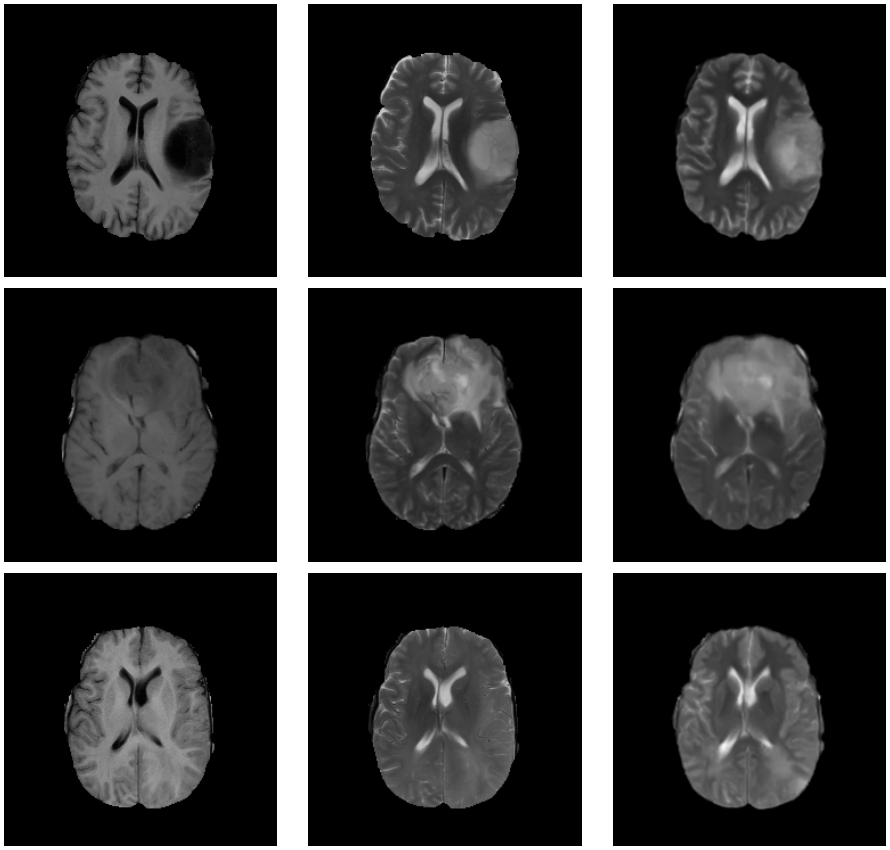
[14] H. Uzunova, H. Handels, and J. Ehrhardt, "Robust Groupwise Affine Registration of Medical Images with Stochastic Optimization," in Bildverarbeitung für die Medizin, pp. 62-67, Springer Vieweg, Berlin, 2017

Appendix A

Example High-Resolution Images Generated with MEGAN

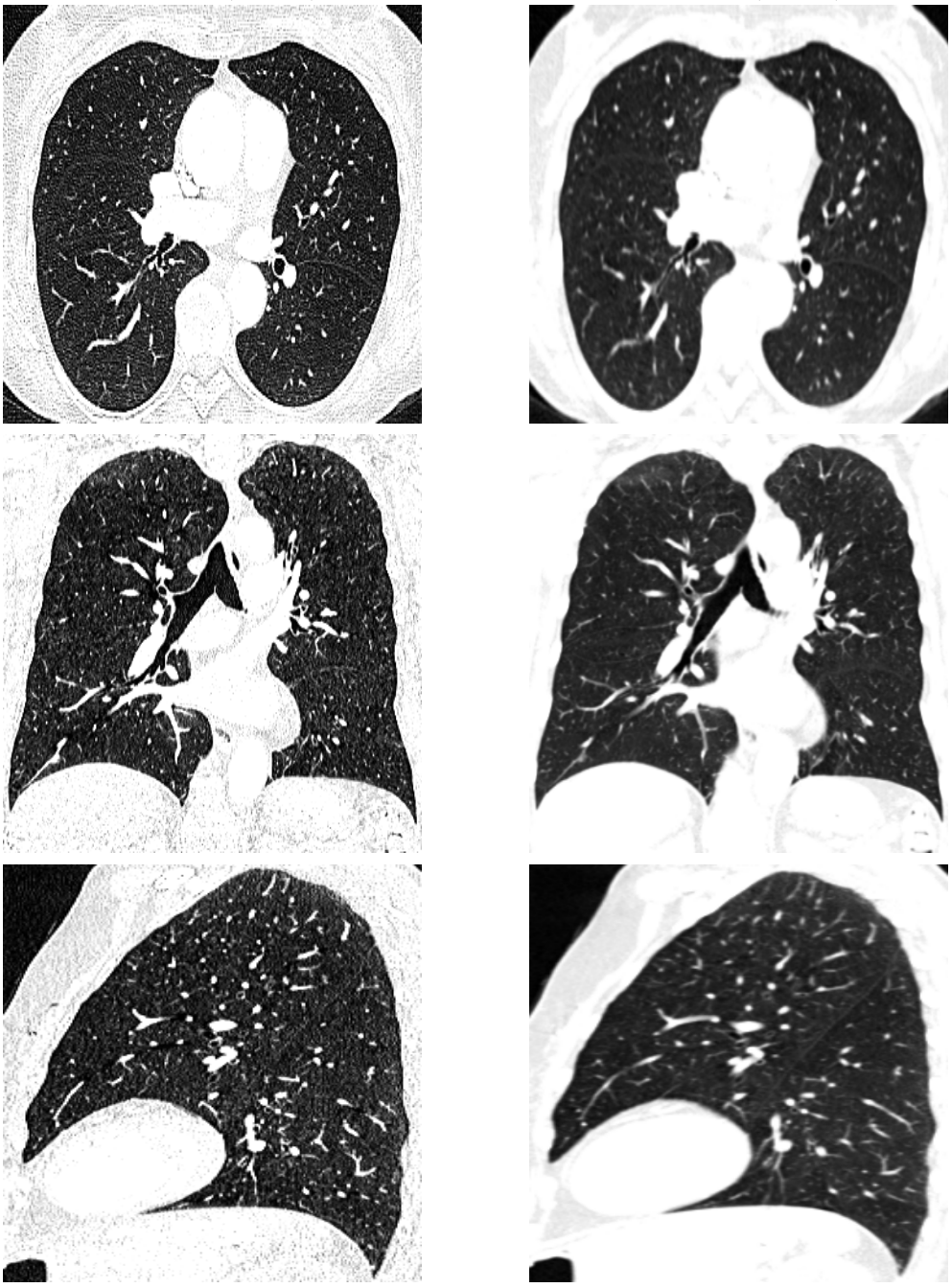
Example 1: Brain MRI T1 to T2 scenario, size $240 \times 240 \times 155$

real T1 (source) | real T2 (target) | fake T2, **MEGAN**



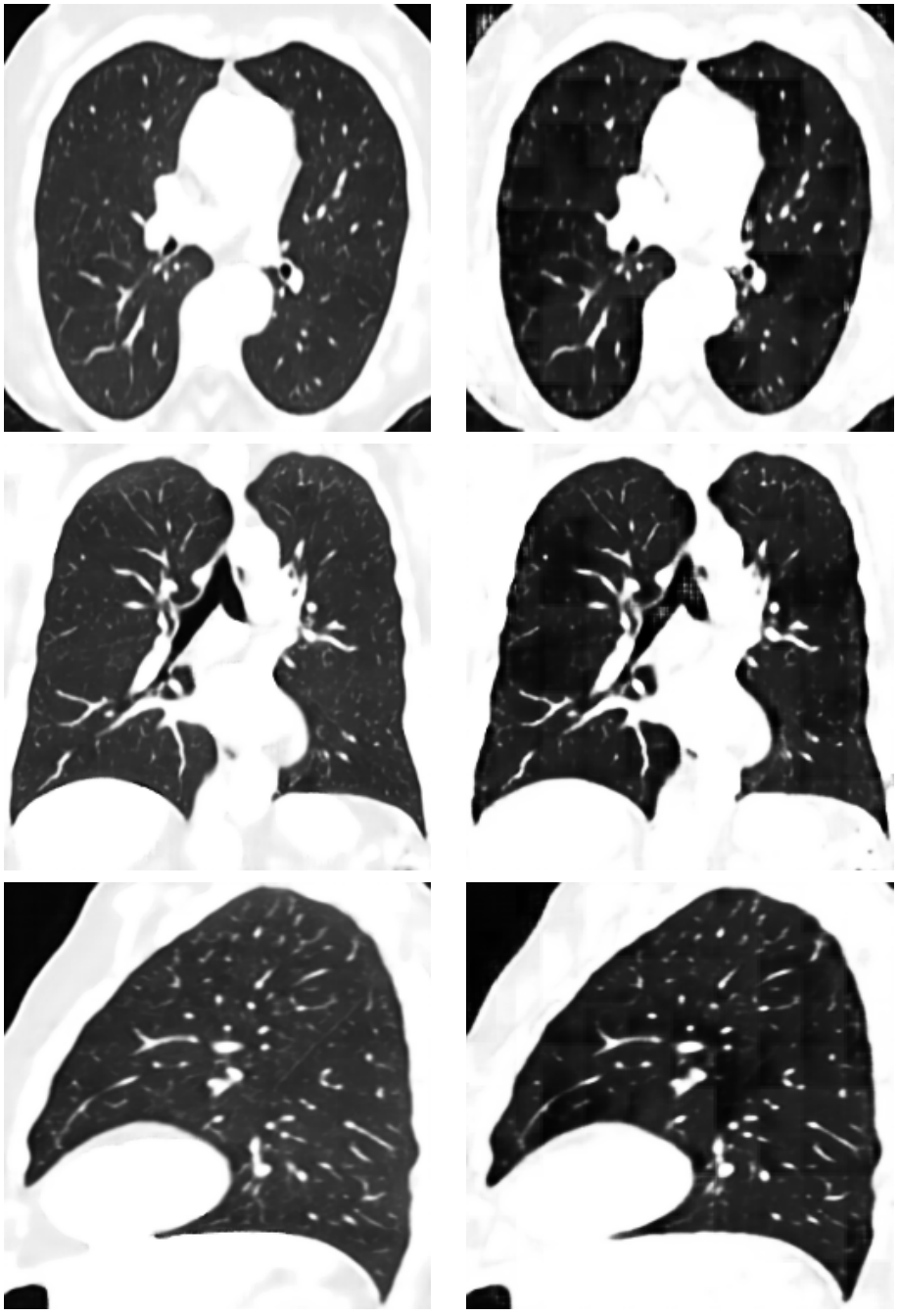
Example 2: Thorax CT sharp (B80f) to soft (B20f) kernel scenario, image size 512^3

real B80f (source) real B20f (target)



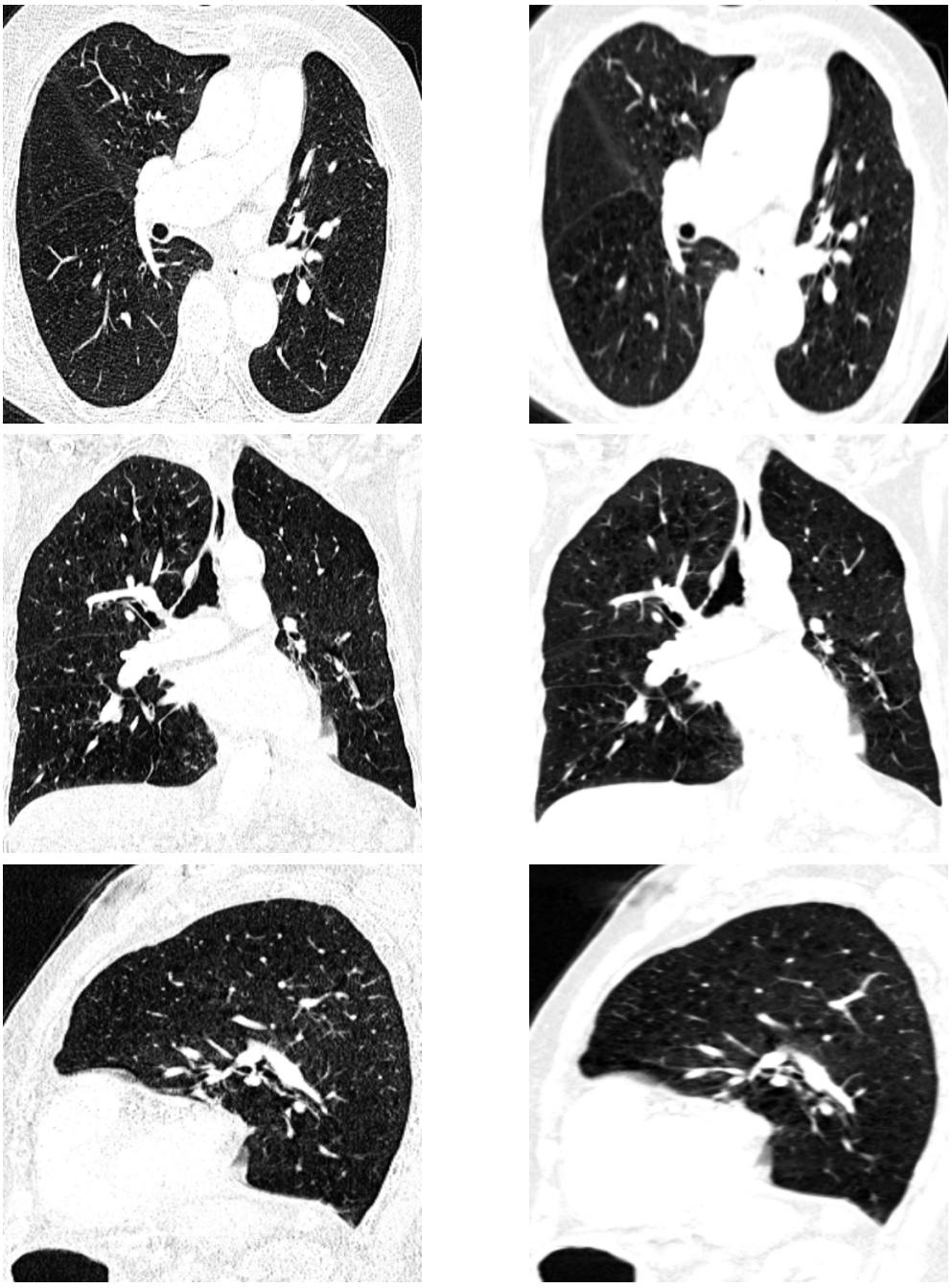
fake B80f→B20f, **MEGAN**

fake B80f→B20f, patchwise (naive)



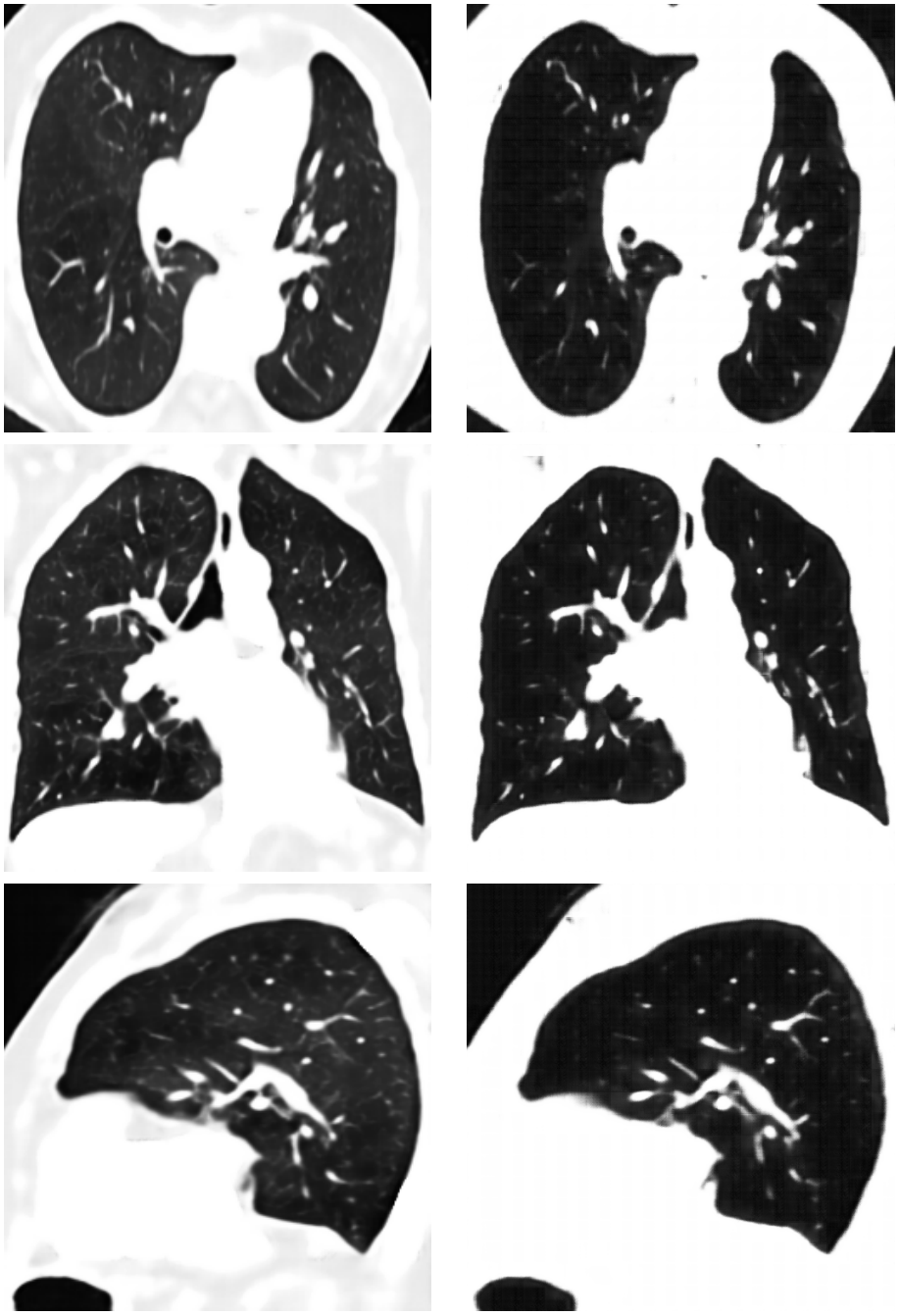
Example 3: Thorax CT sharp (B80f) to soft (B20f) kernel scenario, image size 512^3

real B80f (source) | real B20f (target)



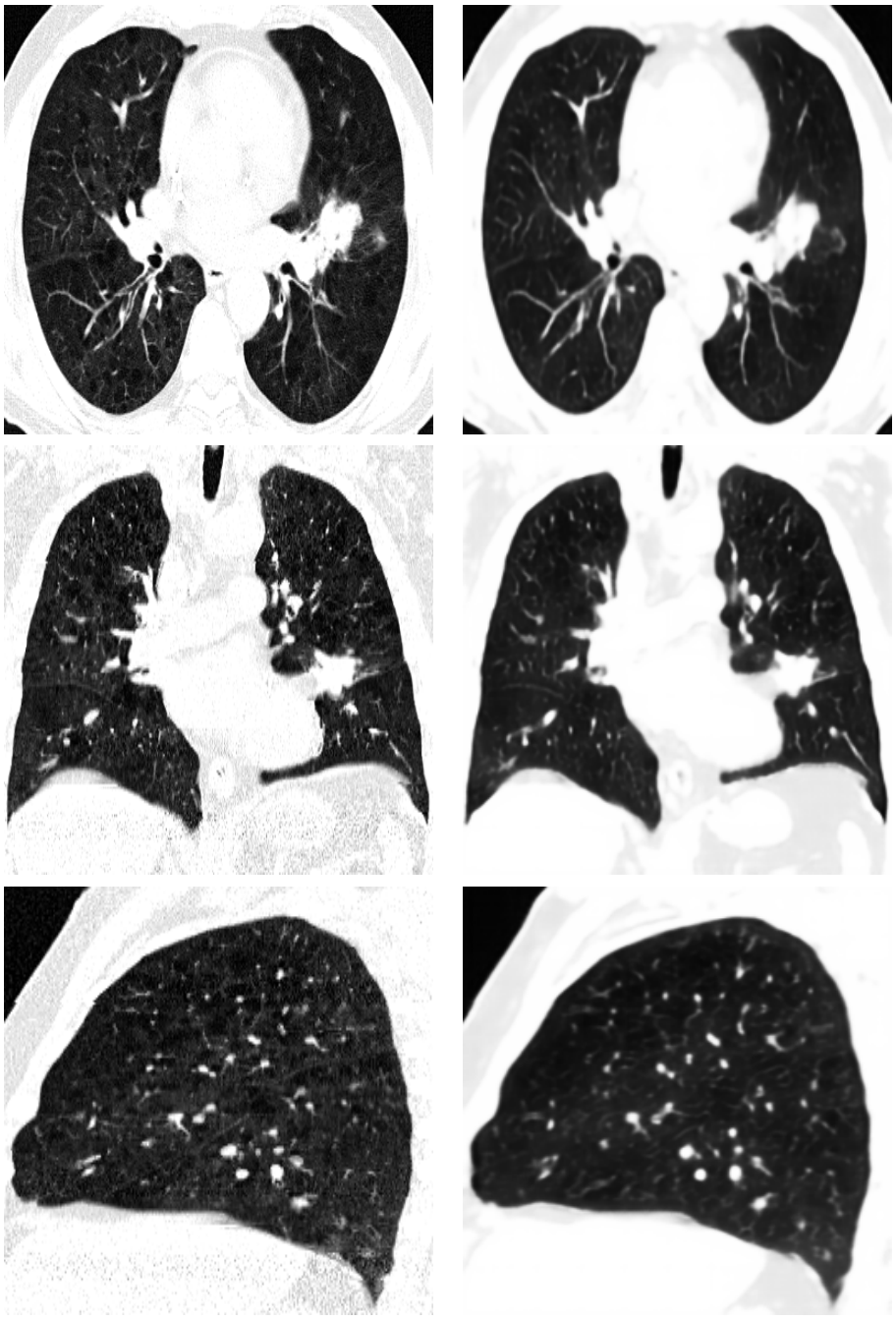
fake B80f→B20f, **MEGAN**

fake B80f→B20f, patchwise (naive)



Example 4: Low-dose to high-dose scenario, image size 256^3

real low-dose fake high-dose, **MEGAN**



Example 5: Low-dose to high-dose scenario, image size 256^3

real low-dose fake high-dose, **MEGAN**