UNIVERSITÄT
BERN

Faculty of Medicine
Artificial Intelligence in Medicine

Master of Science Thesis

Diffusion-Based Filling and Synthesis of Multiple Sclerosis Lesions

by

Vinzenz Uhr

of Menzingen ZG

Supervisor
PD Dr. Richard McKinley

Institution
Support Center for Advanced Neuroimaging (SCAN), University Institute of Diagnostic and Interventional Neuroradiology, University of Bern, Inselspital, Bern University Hospital, Bern, Switzerland

Bern, August 2024

Abstract

White matter (WM) lesions, often associated with neurological conditions like multiple sclerosis (MS), can significantly distort cortical thickness measurements obtained from magnetic resonance imaging (MRI). Traditional methods often rely on lesion filling techniques to address this issue. This thesis explores the potential of deep learning to enhance the accuracy and efficiency of cortical thickness measurement in the presence of WM lesions. One major hurdle in deep learning projects for medical images is the scarcity of large datasets. To overcome this limitation, this thesis investigates in the generation of synthetic data.

Among various approaches, two noise diffusion models with a pseudo-3D U-Net architecture conditioned on binary masks proved to be the most effective in filling and synthesizing WM lesions in MR-images. To assess the quality of the synthesized lesions, two experienced neuroradiologists were asked to identify 20 synthetic lesions among a set of 20 patients. Only three synthetic lesions were correctly identified, highlighting the high realism of the generated lesions.

Furthermore, this thesis compared the robustness of different computational methods for cortical thickness measurement in the presence of WM lesions. Newer deep learning-based methods demonstrated greater robustness, suggesting that lesion filling might eventually become obsolete.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. PD Richard McKinley, for his invaluable guidance and inspiring conversations throughout this project. His expertise, insights, and constructive feedback were instrumental in my research. I am also grateful to the entire SCAN Team, particularly Dr. Ivan Diaz, for their support and fruitful discussions. I would like to thank Lukas Bannwart and Jenny Lauber for their critical reading of my work. I would like to thank my family and friends for their unwavering support throughout my studies. Their love and encouragement have been a constant source of motivation. Finally, I want to thank myself for believing in myself, for persevering through the challenges, and for never giving up on my goals.

Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Als Hilfsmittel habe ich Kunstliche Intelligenz verwendet. Sämtliche Elemente, die ich von einer Kunstlichen Intelligenz uebernommen habe, werden als solche deklariert und es finden sich die genaue Bezeichnung der verwendeten Technologie sowie die Angabe der «Prompts», die ich dafur eingesetzt habe. Mir ist bekannt, dass andernfalls die Arbeit mit der Note 1 bewertet wird bzw. der Senat gemäss Artikel 36 Absatz 1 Buchstabe r des Gesetzes vom 5. September 1996 uber die Universität zum Entzug des auf Grund dieser Arbeit verliehenen Titels berechtigt ist. Fur die Zwecke der Begutachtung und der Ueberprufung der Einhaltung der Selbständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Ueberprufung von Arbeiten Dritter zu verwenden oder hierzu zur Verfugung zu stellen.

Bern, August 31st 2024

Vinzenz Uhr

Contents

List of Figures

List of Tables

List of Abbreviations

BraTS Challenge Brain Tumor Segmentation Challenge

CNS Central nervous system

CSF Cerebrospinal fluid

DDPM Denoising diffusion probabilistic models

DiReCT Diffeomorphic registration-based cortical thickness

GM Gray matter

LPIPS Learned Perceptual Image Patch Similarity

MR(I) Magnetic resonance (imaging)

MS Multiple sclerosis

MSE Mean squared error

PSNR Peak signal noise ratio

ROI Region of interest

RR-MS Relapsing-remitting MS

SSIM Structural similarity index measure

T1w T1-weighted

WM White matter

Chapter 1

Introduction

Multiple sclerosis (MS) is a neurological condition that affects the brain and spinal cord, which control all bodily functions. MS causes damage to the coating that protects the nerves (myelin). This damage interrupts the communication between the brain and the rest of the body. People with MS experience various symptoms, like numbness, weakness, and vision problems. The disease can progress differently for each person, with some experiencing periods of recovery and others facing a steady decline. MS is a common cause of disability in young adults, impacting their lives and those around them. In 2020 the number of MS patients around the globe was estimated at 2.8 million. MS not only affects adults - there are at least 30,000 people living with MS who are under 18 [47] [24] [49].

Magnetic resonance imaging (MRI) is the primary imaging method for diagnosing and monitoring the progression of MS. The development of computational methods that, using conventional MRI scans, are able to provide sensitive and reproducible measures of brain volumes, has allowed an indirect quantification of the brain. These methods have been extensively used in the study of MS to estimate total and regional i.e., white matter (WM) and gray matter (GM) cerebral tissue loss, providing measures able to accurately assess and monitor the pathologic evolution of the disease. In the presence of WM lesions MR-based measurements like cortical thickness can be affected, caused by misclassification of different tissue types. The misclassification varies considerably with lesion size and intensity, especially when the lesion intensity is similar to that of the GM/WM interface. In the past different lesion filling algorithms were used to fill the WM lesions with intensities matching the correct tissue type, leading to more robust measurements [4] [52] [2] [14] [1].

In recent years, deep learning has emerged as a powerful tool in medical image analysis, revolutionizing the field with its ability to automatically learn and extract meaningful features from large datasets. Deep learning techniques, especially convolutional neural networks (CNNs), have shown remarkable success in various medical imaging applications, including lesion filling [1]. In image generation denoising diffusion probabilistic models (DDPM) [19] have shown an impressive performance and experienced increasing popularity in medical image analysis [22]. One current limitation of deep learning projects for medical images is the lack of availability of large datasets. Medical data is costly and laborious to collect, and privacy concerns create challenges to data sharing. This limitation creates a bottleneck on models' generalizability and hampers the rate at which cutting-edge methods are deployed in the clinical routine. Generating synthetic data provides a promising alternative [37].

1.1 Goal

This thesis aims to achieve three primary objectives:

- Develop a method to fill WM lesions in MR-images using DDPMs.
- Assess the impact of lesion filling on cortical thickness measurements using current tools to evaluate their robustness.
- Create a DDPM capable of generating new, synthetic WM lesions within MR-images.

Chapter 2

Background

This chapter provides an overview of MS and discusses the specific areas of lesion filling and lesion synthesis.

2.1 Multiple Sclerosis

MS is a neurological condition that affects the brain and spinal cord, which control all bodily functions.

Epidemiology

Globally, an estimated 2.8 million people live with MS in 2020, translating to roughly 1 in 3,000 individuals. MS occurs worldwide, with a significantly higher prevalence in Europe and the Americas. However, regional variations are substantial. For instance, within Europe, San Marino (337 per 100,000), Germany (303 per 100,000), and Denmark (282 per 100,000) have the most cases, with San Marino and Germany having the highest prevalence worldwide, followed by the USA (288 per 100,000). Conversely, several European countries report prevalence below 40 per 100,000. While MS can develop at any age, the global average diagnosis age is 32. There is currently no cure for MS, leading to individuals managing the disease for decades. Notably, at least 30,000 MS patients are under 18 years old [47].

Symptoms

MS is a complex disease affecting the central nervous system (CNS). It causes a wide range of symptoms, including physical and cognitive issues. These symptoms vary greatly between individuals and can change over time. Common symptoms are fatigue, muscle stiffness, problems with bladder and bowel function, pain, and difficulties thinking and concentrating. These symptoms can interact and worsen each other, making daily life challenging. There is no typical MS case, as each person experiences the disease differently [10].

Diagnostics

MS diagnosis is a complex process lacking a single definitive test. Over time, various diagnostic criteria sets have been established. These criteria emphasize two key principles [33]:

1. Dissemination in Space (DIS): This refers to evidence of lesions, or damaged areas, scattered across different regions of the CNS.

2. Dissemination in Time (DIT): This indicates that the CNS damage occurred at separate points in time, suggesting recurring episodes.

The widely used McDonald criteria, named after the neurologist W. Ian McDonald, serve as a foundation for the diagnosis of MS in both research and clinical settings. In order to make a diagnosis of MS, typical relapse symptoms must be present on the one hand, and on the other hand the temporal and spatial dissemination must be fulfilled on the basis of clinical or imaging findings. Brain and spinal cord MRI remain the most useful paraclinical tests to aid the diagnosis of MS and can substitute clinical findings in determination of DIS and/or DIT in patients with a typical clinically isolated syndrome [48]. MS plaques can develop throughout the CNS, affecting various regions of the WM. In some cases, involvement extends to the GM, where nerve cell bodies reside [33].

Therapy

MS is currently incurable, but there are various treatment options available to manage the disease. A distinction is made between course, relapse and symptom therapy. These follow-up therapies, also known as baseline therapy, aim to minimize MS progression as effectively as possible. Medications used for MS target the immune system. These medications can be broadly categorized into two groups: those that alter the immune system's activity and those that suppress the overall immune function. By regulating the immune system's response, these treatments can help reduce the frequency of relapses, lessen disease severity and activity, and slow down the accumulation of disabilities [41]. The areas of the CNS most affected by MS vary from person to person. This variability in lesion location contributes to the diverse range of symptoms experienced by individuals with MS and necessitates targeted treatment approaches. Treatment strategies typically combine expertise from various disciplines, including medicine, physiotherapy, occupational therapy, speech therapy, rehabilitation specialists, neuropsychologists, and psychotherapists. Additionally, complementary therapies and adjustments to lifestyle habits, such as diet, can offer some relief from symptoms [42].

Course

MS presents itself with a highly variable clinical course. However, four main disease types have been identified [33]:

Relapsing-remitting MS (RR-MS): This is the most common form, affecting roughly 80-85% of initial diagnoses. RR-MS is characterized by distinct episodes of new or returning neurological symptoms. These episodes, often called relapses, are followed by periods of full or partial recovery with no disease progression in between.

Primary-progressive MS (PP-MS): Around 10-15% of MS patients have PP-MS. This form is characterized by steady disease progression from the very beginning, with occasional plateaus or minor temporary improvements possible.

Secondary-progressive MS (SP-MS): Approximately half of RR-MS patients transition to SP-MS after ten years, and this number rises to 90% after 25 years. SP-MS starts with an RR-MS disease course, followed by a gradual progression with or without occasional relapses, minor recoveries, and plateaus.

Progressive-relapsing MS (PR-MS): This form is less common and may be considered a subtype of PP-MS due to its similar progression. PR-MS involves steady disease progression

from the outset, punctuated by distinct acute relapses that may or may not fully resolve. The periods between relapses are still marked by ongoing disease progression.

2.2 Lesion Filling

The presence of MS WM lesions can significantly impact MR-based measurements like cortical thickness due to misclassification of different tissue types [4] [2] [49] [5] [29]. This misclassification is particularly problematic for WM lesions with size and intensity similar to the GM/WM interface and leads to overestimation of GM atrophy [17]. Lesion filling algorithms have been developed to address this issue and improve measurements such as cortical volume, thickness and surface area estimation [2].

Early lesion filling approaches employed various strategies. For instance, [4] utilized lesion filling to enhance brain volume measurements, including normalized brain volume (NBV), normalized white matter volume (NWMV), normalized gray matter volume (NGMV), and percentage brain volume change (PBVC). Their method involved calculating intensity distributions of cerebrospinal fluid (CSF), CSF/GM, GM, and GM/WM from existing brain MR-images and filling WM lesions with pixel intensities randomly sampled from these distributions.

Another approach, proposed by [52], involved refilling WM lesions by replacing lesion voxel intensities with random values drawn from a normal distribution based on the WM signal intensity of each two-dimensional slice. Segmentation of the slices was achieved using the fuzzy c-means algorithm.

Graph theoretical network analysis, a technique used to assess brain connectivity patterns, can also benefit from lesion filling. To reduce the variability in network analysis caused by WM lesions, [53] applied lesion filling by substituting lesion voxel intensities with intensities from nearby voxels. Their study suggests that lesion filling might improve the detection of network alterations in MS patients, but also highlights the potential for introducing artifacts. Therefore, caution is advised, especially for individuals with high lesion loads or lesions located at the WM/CSF or WM/GM interface.

More recent advancements leverage machine learning for lesion inpainting. [14] employed a total variation model to improve registration performance with brain atlases. Inspired by Gated Convolution, [1] introduced a user-guided deep adversarial inpainting model capable of filling irregularly shaped holes in high-resolution T1w MR brain images. Training data generation involved synthesizing lesion masks by sampling and deforming random circles. Additional data augmentation techniques included rotation, cropping, flipping, noise addition, and varying brightness levels.

The emergence of DDPMs [19] offers a novel approach for high-quality image generation. DDPMs exhibit superior distribution coverage and training stability compared to adversarial loss-trained models, achieving state-of-the-art performance in various image synthesis tasks.

The International Brain Tumor Segmentation (BraTS) challenge in 2023 incorporated an inpainting challenge focused on synthesizing healthy brain tissue in glioma-affected regions [23]. Due to the high computational cost of 3D processing, [12] opted for a 2D diffusion model conditioned on glioma masks. While achieving comparable results to other participants, their approach resulted in stripe artifacts due to stacking of the 2D slices. Gaussian filtering was subsequently employed to mitigate these effects at the slice borders.

[1] addressed the high computational demands of 3D diffusion models by proposing several resource-reduction strategies. Notably, they introduced PatchDDM, a memory-efficient patch-based diffusion model that allows for inference on the entire volume while training solely on patches. Additional approaches included reducing self-attention layers,

incorporating additive skip connections, and training on downsampled data.

In pursuit of improved inpainting quality for 3D MR-images, [13] evaluated and modified various diffusion models, including 2D, pseudo-3D, and 3D models operating in image space, 3D wavelet or 3D latent space. Their findings suggest that the pseudo-3D model proposed by [60] achieved the best performance in terms of structural similarity index measure (SSIM), peak signal noise ratio (PSNR), and mean squared error (MSE).

2.3 Lesion Synthesis

Supervised deep learning models rely heavily on vast amounts of training data to effectively capture the inherent variability within a population or disease state. However, acquiring medical data often presents challenges due to cost, time constraints, and privacy concerns. This is particularly true for diseases like MS, which has a relatively low prevalence. Existing MS datasets with manually labeled lesions are typically small. These datasets may also originate from single institutions, leading to a lack of variability in factors like scanners, scanning protocols, and patient demographics. This limited diversity can hinder the generalizability of deep learning models trained on such data. While techniques like regularization, data augmentation, and cross-validation can mitigate the effects of limited training data, they cannot fully address the issue. Pooling data from multiple sources offers a potential solution but introduces new challenges due to inconsistencies in data acquisition methods that necessitate standardization and post-processing steps [55].

In this context, synthetic data generation emerges as a promising alternative. Generative models learn the underlying statistical distribution of the data they are trained on. This allows them to create realistic representations of data points that differ from those present in the training set by sampling from the learned distribution. Synthetic data offers several advantages in the medical field. It can serve as a cost-effective alternative or supplement when manually labeled data is scarce or unavailable. Additionally, it can help preserve patient confidentiality and privacy. Synthetic data generation allows for the creation of a wider range of variations, including edge cases not represented in the original dataset. This facilitates testing the robustness of models under diverse scenarios. However, the trustworthiness of synthetic datasets remains a critical question. This includes aspects like data representativeness, privacy preservation, and potential biases inherited from the training data, pre-processing steps, or the algorithms themselves. For synthetic data to be truly valuable, it must faithfully reflect the statistical properties of the original data while maintaining the inherent variability and structure. The risk lies in creating data that oversimplifies or misrepresents the complexities found in real-world scenarios [27].

Several promising applications of synthetic data generation exist for medical imaging tasks. For example, a recent study created a massive dataset of 100,000 T1w MR brain images for public use, trained on over 31,000 images from the UK Biobank [37]. In another application, researchers proposed a generative adversarial network for synthesizing liver lesions, which were then used to train a lesion segmentation network [26].

Beyond the realm of medical imaging, large-scale text-to-image models are also being explored as training data generators. For instance, one study proposed a method for editing images based on human instructions. This method was trained on instructions sampled from GPT-3 and images generated through stable diffusion [7]. Another study demonstrated that incorporating high-quality synthetic data generated from a text-to-image model could improve performance on established benchmark tasks like ImageNet classification [3].

For generating synthetic skin lesions from a limited dataset, [15] employed a fine-tuned stable diffusion model using LoRA [20] and Dreambooth [40] techniques. Furthermore, [32]

proposed an image editing method based on a diffusion model generative prior, which synthesizes realistic images by iteratively denoising through a stochastic differential equation. Given an input image with a user-defined coarse manipulation, the method first adds noise and then progressively denoises the image using the stochastic differential equation (SDE) prior to incorporate the guidance and enhance authenticity.

These advancements suggest that synthetic data generation has the potential to revolutionize the field of deep learning for medical applications, particularly in areas like WM lesion analysis, where data scarcity remains a significant challenge.

Chapter 3

Methods

This chapter details the technologies, processes, and datasets employed to accomplish the objectives outlined in Chapter 4.

3.1 Denoising Diffusion Probabilistic Models

Diffusion models are a generative deep learning technique that leverage an approach for data synthesis. The core idea lies in progressively transforming a data sample x_0 from its original distribution into a sample x_T from a simple normal distribution. The model then learns to reverse this transformation process [43].

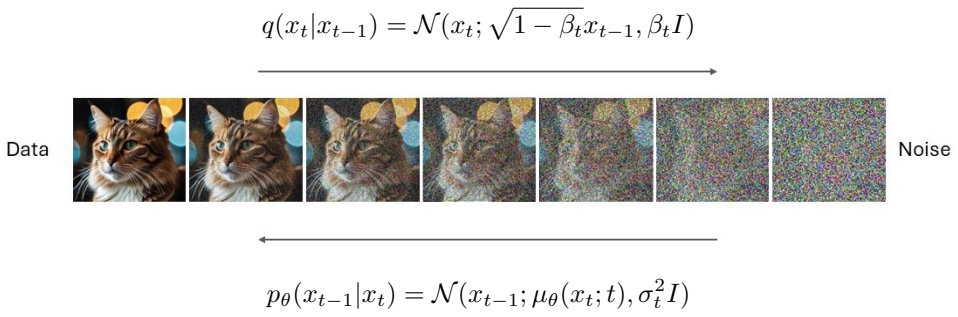

Figure 3.1: Diffusion process from Data to Noise and reverse process from Noise to Data.

The forward process is the transition from a clean image x_0 to a completely noisy one x_T through a series of steps, where each step adds a small amount of random noise. A Markov chain is used to model this process, where the current noisy version x_t only depends on the previous one x_{t-1} in the sequence. At each step, zero-mean Gaussian noise is added, gradually increasing its strength until a maximum level is reached at a predefined endpoint $t = T$. The original DDPM paper suggests using 1000 steps for this process [19]. The mathematical formulation behind this forward noising process q is denoted by

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{3.1}$$

Where I represents the identity matrix and β_t the variance schedule, which controls the amount of noise added at each step based on the current step t. A noisy image x_t can be produced by

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \text{ where } \epsilon \sim \mathcal{N}(0, I) \text{ and } \alpha_t := 1 - \beta_t \text{ and } \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s \tag{3.2}$$

The noise schedule β_t is designed such that $\alpha_T \to 0$ and $q(x_T|x_0) = \mathcal{N}(0, I)$.

The next step involves learning to reverse this entire noising process. This is referred to as the denoising process p_θ, where the goal is to predict a less noisy sample x_{t-1} from a noisy sample x_t. Typically, the equation $q(x_{t-1}|x_t) \propto q(x_t|x_{t-1})q(x_{t-1})$ is intractable, but it can be approximated with a Gaussian for small transitions (small β_t). The equation for the reverse can be written as

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \tag{3.3}$$

The variance σ_t^2 can be fixed, eliminating the need to learn it explicitly. To learn μ_θ it's easier to train a model $\epsilon_\theta(x_t, t)$ which predicts the noise that needs to be removed at each step [19]. This allows for more efficient training. A sample x_{t-1} can be generated from x_t by

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_t z, \text{ with } z \sim \mathcal{N}(0, I) \tag{3.4}$$

To generate entirely new data, for example, a new image, we can start with pure noise $x_T = \mathcal{N}(0, I)$ and iteratively apply equation 3.4 for all timesteps $t \in \{T, ..., 1\}$ to obtain the final prediction x_0. The diffusion model μ_θ usually uses a U-Net architecture [39]. The training objective focuses on minimizing the MSE loss

$$\mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0,I)}[||\epsilon - \epsilon_\theta(x_t, t)||^2] \tag{3.5}$$

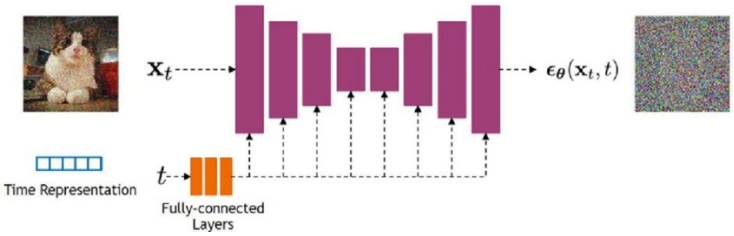

Figure 3.2: Diffusion models often use U-Net architectures with ResNet blocks and self-attention layers to represent $\epsilon_\theta(x_t, t)$. Figure copied from Yufeng Wei 2024 J. Phys.: Conf. Ser. 2711 012005

3.2 Model Selection

High-resolution 3D MR-images require significant storage space and computational resources. To address this challenge, we can slice the 3D MR-image into batches of 2D transversal slices and employ 2D diffusion models for processing, reconstructing a consistent 3D MRI at the end.
This thesis explores two approaches in MR-images using diffusion models: conditional and unconditional. Both approaches utilize the ground truth MR-image x, a binary mask m defining the lesion region, and the masked ground truth image $\hat{x}$.

Conditional Lesion Filling

The conditional approach trains a diffusion model conditioned on the masked ground truth image and the binary mask. The conditioning information is incorporated through channel-wise concatenation. At each timestep t during reverse diffusion, the model receives the concatenated input of the noisy image x_t, the masked ground truth image $\hat{x}$ and the binary mask m. The objective is to predict the noise term for calculating a less noisy image. This leads to the loss function,

$$\mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0,I)}[||\epsilon - \epsilon_\theta(((\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon) \oplus \hat{x} \oplus m), t)||^2] \tag{3.6}$$

For sampling, we employ Denoising Diffusion Implicit Models (DDIM) [44], a computationally efficient class of iterative probabilistic models that share the training procedure of DDPM. DDIM utilizes a non-Markovian sampling process, which is deterministic. The sequence of a training step for the conditional model is described later in Figure 3.9.

Unconditional Lesion Filling

The unconditional approach does not use conditioning information during training. We train an unconditional DDPM as a generative prior, as described in Chapter 3.1. This essentially creates a model that can produce random 2D brain MRI samples. To condition the generation process, we modify the reverse diffusion iterations by sampling masked regions using the provided image information, as proposed in the RePaint paper [28]. This technique does not modify the original DDPM network and is applicable to any inpainting mask distribution.

RePaint

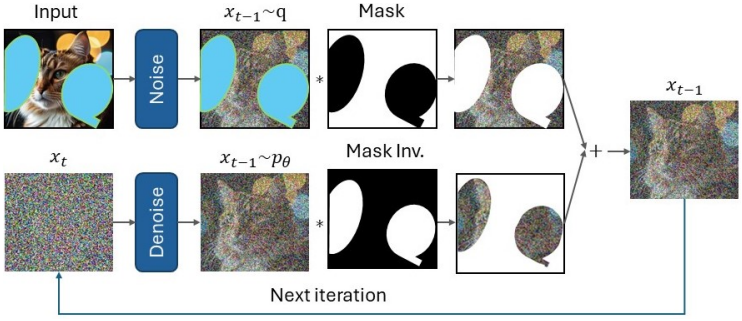

Figure 3.3: RePaint [28] modifies the standard denoising process in order to condition on the given image content. In each step, they sample the known region (top) from the input and the inpainted part from the DDPM output (bottom).

Inpainting aims to predict missing pixels within a masked region based on surrounding image information. Each reverse step from image x_t (noisy) to x_{t-1} (less noisy) depends solely on the noisy image x_t, which consists of unknown pixels within the mask $m \odot x_t$ and known pixels outside the mask $(1-m) \odot x_t$. The known pixels can be calculated for each timestep based on the forward process (Equation 3.2). The RePaint paper [28] proposes separate sampling processes for unknown and known pixels during the reverse step, resulting

in the following expression:

$$x_{t-1}^{known} \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \tag{3.7}$$

$$x_{t-1}^{unknown} \sim \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3.8}$$

$$x_{t-1} = (1-m) \odot x_{t-1}^{known} + m \odot x_{t-1}^{unknown} \tag{3.9}$$

Here, x_{t-1}^{known} is sampled using the known pixels in the given image $(1-m) \odot x_t$, while $x_{t-1}^{unknown}$ is sampled from the model given the previous iteration x_t. These are then combined to form the new sample x_{t-1} using the mask m.

Resampling

Direct application of the RePaint method can lead to generated content that only partially matches the known regions. To address this, the authors propose an additional resampling approach that allows the model more time to harmonize the conditional information x_{t-1}^{known} with the generated information $x_{t-1}^{unknown}$. They achieve this by diffusing the output x_{t-1} back to x_t through a forward diffusion step (Equation 3.1). While this operation introduces noise and scales back the output, it also preserves some information incorporated from the generated region $x_{t-1}^{unknown}$. This leads to a new $x_t^{unknown}$ that is both more harmonized with x_{t-1}^{known} and incorporates conditional information from it. Since this operation can only harmonize information for one step, it might not be able to fully integrate semantic information across the entire denoising process. To overcome this limitation, the authors introduce the concept of jump length j, which defines the time horizon for this operation ($j = 1$ for the previous case). By repeating the resampling step multiple times (denoted by parameter r), the model can progressively improve the harmonization (recommended settings by the authors: $T = 250$ timesteps, $r = 10$ resampling steps with jump length $j = 10$).

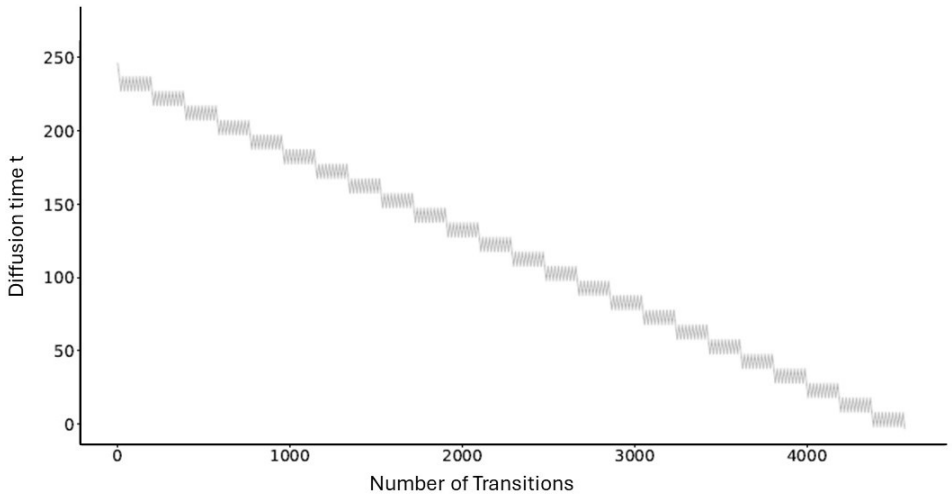

Figure 3.4: [28] visualizes the diffusion time t through which a sample transits during the inference process with jump length ($j = 10$) and resampling ($r = 10$).

Conditional Lesion Synthesis

Building upon the conditional lesion filling approach outlined in Section 3.2, conditional lesion synthesis adopts a similar workflow. The key distinction lies in the target data. While lesion filling methods aim to inpaint masked regions with healthy WM, conditional lesion synthesis focuses on replicating WM lesions within the masked areas. This requires the model to learn a different mapping, where the input combines the original image and a lesion mask from an MS patient, and the output replicates the WM lesion within the masked region.

Unconditional Lesion Synthesis

For the unconditional lesion synthesis task, we're following an approach similar to the SDEDIT paper by [32]. As in the RePaint approach we're training an unconditional DDPM and alter the reverse diffusion process to condition the generation process.

The core idea is to "hijack" the generative process of diffusion-based generative models. We begin by introducing a coarse lesion into the input image by manipulating pixels. To smooth out artifacts and distortions while preserving the overall structure of the lesion, we add a controlled amount of noise using Equation 3.2. We then initialize the sampling process with this noisy input at an intermediate timestep t_0 and progressively remove noise to obtain a denoised result that is both realistic and faithful to the user-provided coarse lesion. To prevent unintended changes to pixels outside the lesion area, we employ the RePaint inference process described in Section 3.2.

A key challenge lies in balancing faithfulness to user input (e.g., hand-drawn coarse lesions) and the authenticity of the synthesized images. This balance can be controlled by adjusting the initial timestep t_0. Higher values of t_0 lead to more realistic but less faithful lesion representations.

Coarse Lesion

Binary lesion masks are used to define the region where the WM lesion will be added. To determine the intensity of the lesion pixels within this mask, the intensity distribution of existing WM lesions in a patient is leveraged. During evaluation, the impact of different intensity values is explored. These values include the 1st quartile, mean, median, 3rd quartile, and the 99th percentile of the intensity distribution.

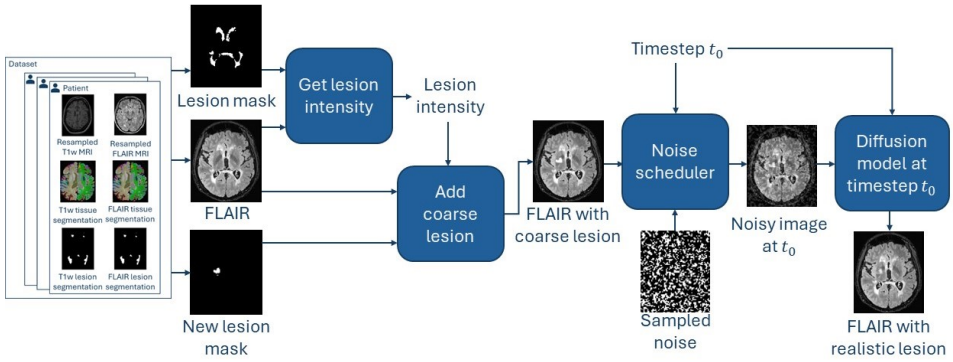

Figure 3.5: Synthesizing a new lesion with an unconditional model

U-Net Architectures

As model we used a 2D U-Net and a pseudo-3D U-Net [60] which achieved high scores in another study comparing different diffusion models architectures for 3D healthy brain inpainting [13]. The 2D U-Net has an architecture similar to [35]. It uses six feature map resolutions with two convolutional residual blocks per resolution level and one self-attention block at the 16x16 resolution after each convolutional block. The architecture of a convolutional residual block is described in Figure 3.7. From highest to lowest resolution the U-Net stages use (128, 128, 256, 256, 512, 512) channels.

Figure 3.6: Pseudo-3D convolution, where b, c, h and w are the batch size, channels, height and width.

The pseudo-3D U-Net has, in addition to the 2D U-Net, a volumetric layer inside the residual block after each 2D convolution. Pseudo-3D convolutions result from 1D convolutions in the z-axis, requiring the batch to be rearranged before and after. Following [13] we apply the model in the image space and directly use the pseudo-3D convolutions without the proposed fine-tuning strategy used by the original paper [60]. To setup the U-Net models and the training environment, we used the python library diffusers from huggingface [54].

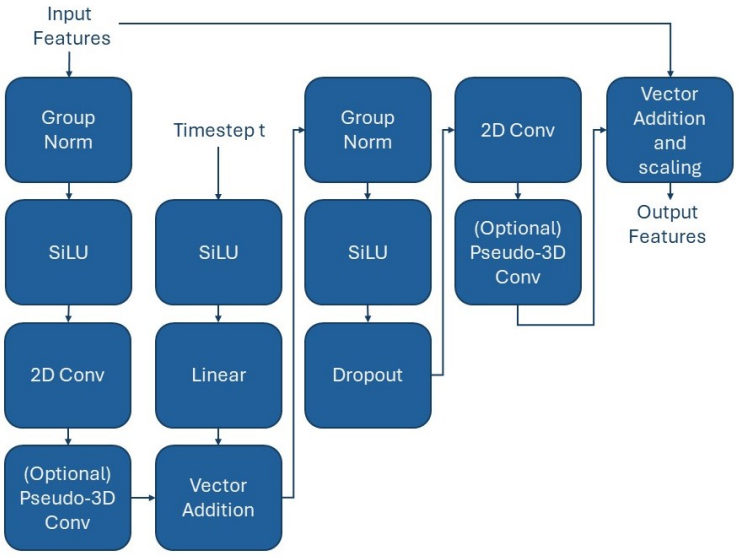

Figure 3.7: Architecture of a convolutional residual block. The inputs are the features from the image and the timestep of the diffusion process.

3.3 Datasets

Our research employed different datasets for training and evaluating our models for lesion synthesis and filling.

The lesion synthesis models were trained using a dataset obtained from the MICCAI challenge [9]. This dataset comprised of MRI scans of 15 patients diagnosed with MS. Each scan included both T1w and FLAIR images, with the FLAIR images containing manually segmented lesion masks. The dataset was divided into training and validation sets, with 13 samples dedicated to training and 2 for validation.

A separate dataset was utilized for training the lesion filling models. This dataset originated from the OASIS project [30] and consisted of T1w MRI scans from 20 healthy subjects. The OASIS dataset was also split into training and validation sets, with 16 samples used for training and 4 for validation.

Section 3.3 explores the use of additional synthetic masks in the form of random circle masks for lesion filling. To evaluate their effectiveness on larger datasets, the BraTS Inpainting Challenge 2023 dataset [23] was employed. The training set of this dataset comprises 1251 brains. Since the challenge has concluded and online analysis of the validation set is no longer possible, the training set was divided into 90% training data and 10% test data.

Evaluation of lesion synthesis models and the impact of lesion filling on cortical thickness measurements was conducted using a test set composed of 65 patients diagnosed with RR-MS. This data originated from an internal longitudinal study conducted at the Insel hospital. All patients had been undergoing Natalizumab treatment for over two years and had at least four MRI scans performed over a period of approximately six months each, with corresponding clinical evaluations. MRI scans included a combination of 1.5T and 3T datasets with a slice thickness of 1mm or less in the T1w sequences. For each patient, the T1w and FLAIR images from their final visit, typically containing the highest lesion burden, were used for testing.

To further explore the impact of lesion synthesis, a scenario where a new MS lesion is inpainted in a previously healthy region should be considered. The MSSEG 2 dataset [8] was utilized for this purpose. This dataset comprises 100 patients with two time points each and a binary mask indicating new MS lesions that emerged between the two time points, but were absent in the initial time point.

Preprocessing

All T1w images undergo resampling to a standardized size of 256x256x256 voxels with a 1.0x1.0x1.0 mm voxel size and are reoriented to RAS orientation. FLAIR images are resampled to 160x256x256 voxels. The resampling process is carried out using nibabel.processing.conform [6]. Values below 0.01 are discarded as noise and the remaining data is scaled to the range [-1, 1]. A deep learning-based tissue segmentation is performed on the T1w images for each patient using the DL+DiReCT model [38]. To accelerate this process, the parallelization program GNU Parallel [46] is employed. The resulting segmentation masks are registered from T1w to FLAIR images using NiftyReg [34]. For datasets with existing lesion masks, these are registered from FLAIR to T1w images. In the absence of lesion masks in the test set, a separate segmentation model DeepSCAN [31] is utilized to identify MS lesions. Only 2D slices containing WM, based on the DL+DiReCT segmentation, were incorporated for training.

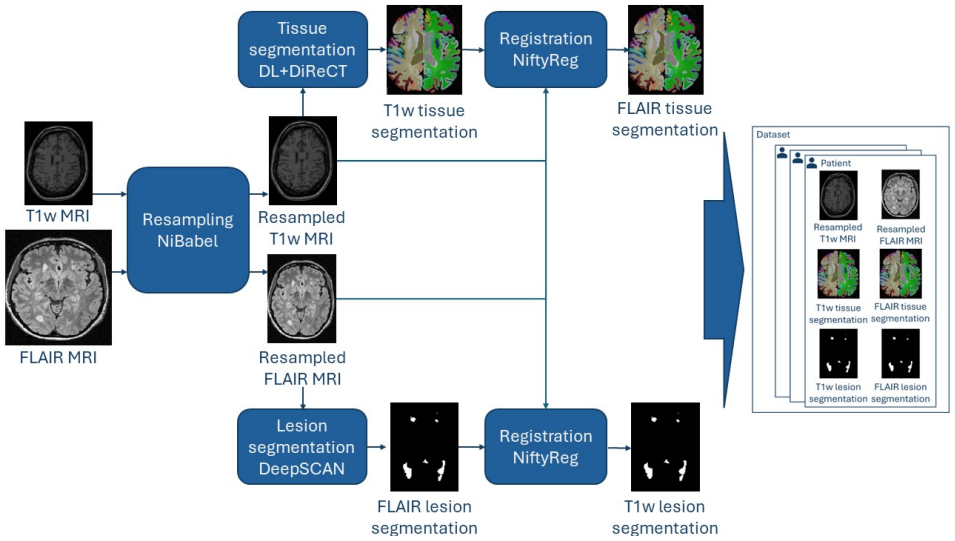

Figure 3.8: Creation of the training dataset

Mask Generation

The distribution of masks employed for training a conditional model can have an impact on the performance of the model [45] [57] [58] [25]. Conditional lesion filling models were trained on the healthy subject images using lesion masks obtained from the MS patients. To achieve this, each lesion mask from the MS patients was registered to every T1w image from the healthy subjects. This resulted in 15 registered lesion masks for each of the 20 healthy patients.
Three registration approaches were evaluated:

1. Non-linear registration of FLAIR lesion masks directly to healthy T1w images.

2. Non-linear registration of FLAIR lesion masks to the T1w image of the corresponding MS patient, followed by registration to the healthy T1w images.

3. Affine registration of the FLAIR lesion mask to the healthy T1w image, followed by skull stripping and then non-linear registration using the affine registration as initialization.

Approaches 1 and 3 yielded visually similar results. For simplicity, approach 1 was adopted for lesion registration. See Appendix B for visual examples. Each lesion mask was restricted to WM tissue by multiplying it with a binary WM mask derived from the DL+DiReCT segmentation. To augment the diversity of the masks, the set of connected lesions was computed for each mask. During training, a different set of connected lesions was sampled and used as the lesion mask.

Given the limited dataset of masks, a secondary approach was explored that utilized a second mask distribution consisting of random circle masks with varying locations and sizes. This resulted in three models being trained: One model trained on the distribution of real lesion masks (conditional lesions model), one model trained on the distribution of random circle masks (conditional circles model) and one model trained with a combination of 50% real lesion masks and 50% random circle masks (conditional mixture model).

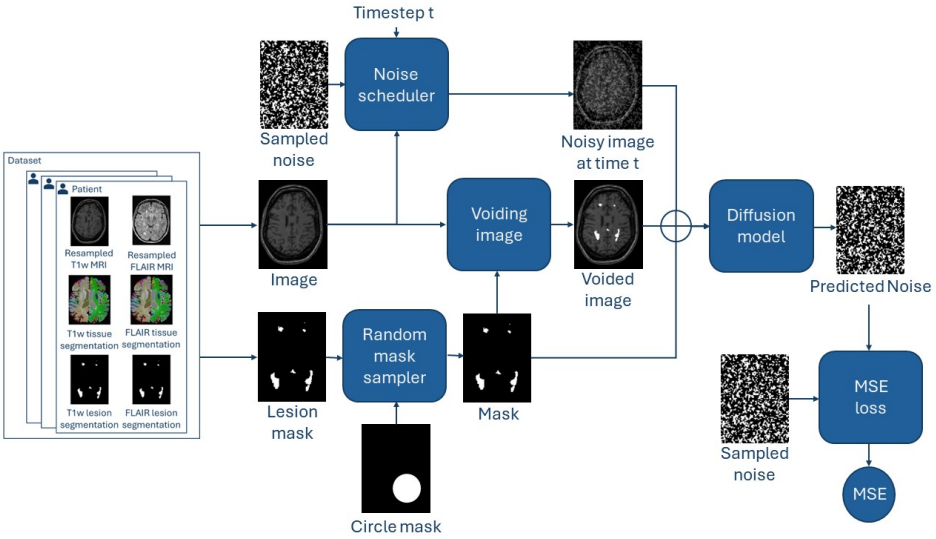

Figure 3.9: Training step involving the conditional mixture model. An MR-image and its corresponding lesion mask are sampled from the dataset. Alternatively, with a 50% probability, a random circle mask is sampled instead. This mask is used to void the portion of the image, which requires inpainting. Additionally, a random timestep t and random noise matching the image shape are sampled. These are used to generate a noisy image as described in Section 3.1. The mask, the noisy image, and the voided image are concatenated and fed into the diffusion model, which aims to predict the sampled noise. The predicted and sampled noise are used to calculate the MSE.

3.4 Training Environment

Number of training diffusion steps	1000
Number of inference steps	50
Batch size	16
Learning rate	1e-4
Optimizer	AdamW
Learning rate scheduler	Cosine schedule with 500 steps warmup
RePaint Jump length	8
RePaint Resample	10

Table 3.1: Hyperparameters for training and evaluation

GPU	3x Nvidia RTX A6000 40GB
CPU	64x Intel Xeon Gold 6226R @ 2.9Ghz
RAM	196 GB

Table 3.2: Hardware

Min-SNR Loss weighting

During the training of the unconditional model for lesion filling, the validation loss indicated faster overfitting with smaller timesteps (e.g., 200 steps) compared to larger ones (e.g., 1000 steps). This could be due to the different levels of difficulty inherent in the time steps of the diffusion models. Predicting added noise becomes progressively easier as the image approaches pure noise. Consequently, bigger timesteps naturally result in lower MSE and correspondingly weaker gradients compared to smaller timesteps. This imbalance leads the training process to prioritize optimization for smaller timesteps.

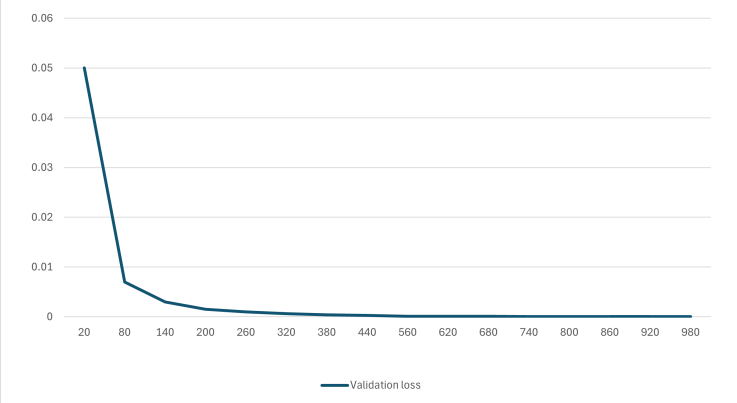

Figure 3.10: Mean validation loss per timestep from 1 to 1000 at the end of training with 1500 epochs

To address this and achieve a more balanced loss function, we explored the min-SNR weighting strategy proposed in [18]. This approach advocates for adapting loss weights assigned to individual timesteps based on clamped signal noise ratios.

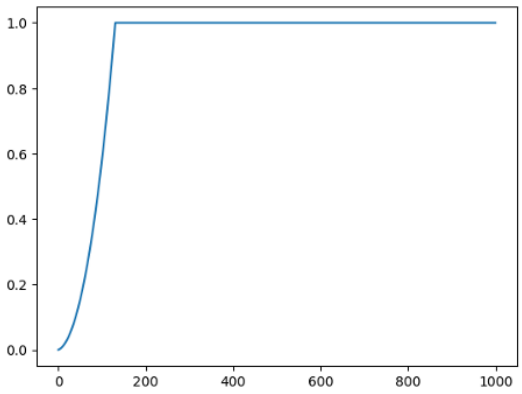

Figure 3.11: Loss weights per timestep

3.5 Evaluation

Metrics

During the training, the model was evaluated at regular intervals using the validation dataset and various metrics were measured (Table 3.3). MSE, PSNR, and SSIM are calculated inside the masks, outside the masks, and across the entire image. Other metrics are solely evaluated over the whole image.

Regarding the performance of perceptual metrics trained on non-medical datasets, [36] indicates that pre-trained models are still useful for medical image evaluation. Additionally, visual inspection is included in the evaluation process, achieved through the creation of inpainted samples. This allows for qualitative assessment of model progress. For unconditional models with RePaint functionality, the evaluation includes sampling unconditional images alongside inpainted samples. Furthermore, to ensure reproducibility, both the initial noise sampling and the validation dataset are seeded.

All metrics are measured on 2D images. The model version with the highest SSIM score is periodically saved to disc.

Mask Dilation

The evaluation of lesion filling revealed the presence of artifacts at the boundaries of inpainted lesions. This occurs because RePaint replaces all areas outside the designated mask with the original image. If the annotated masks don't fully encompass the entire lesion, these small residual areas of the original lesion can lead to border artifacts. Conditional models exhibited similar, though less pronounced, artifacts. To address this issue, we implemented a minor, one-pixel dilation restricted to WM regions during lesion filling. This dilation strongly minimizes the artifacts.

Robustness evaluation cortical thickness

Further evaluations centers on the influence of lesion filling on cortical thickness measurements derived from various processing tools. Cortical thickness assessments are performed on sixty-five patients from the test set, both before and after lesion filling using four distinct methods: ANTs [51], ANTsPyNet [50], Freesurfer [16] and DL+DiReCT [38]. Freesurfer calculates cortical thickness by modeling the cortical band as a surface mesh. ANTs, on the other hand, provides a method based on diffeomorphic registration-based cortical thickness (DiReCT) applied to an atlas-based segmentation. ANTsPyNet extends this approach by incorporating deep learning for segmentation. Similarly, DL+DiReCT is another deep learning-based method that utilizes DiReCT. Freesurfer and DL+DiReCT directly calculate cortical thickness for various regions of interest (ROI). In contrast, ANTs and ANTsPyNet generate a segmentation and a thickness map. To obtain mean thickness per ROI for ANTs and ANTsPyNet, the pipeline and parcellation file from DL+DiReCT were employed. To assess robustness, scans before and after lesion filling are employed, with the assumption that robust models should yield similar results, reflecting reproducibility [21]. The average absolute changes relative to the mean (%) are calculated using the following formula:

$$\epsilon_\mu = \frac{100}{N} \sum_{i=1}^{N} \frac{1}{2} \sum_{t=1}^{2} \frac{|m_{i,t} - \mu_i|}{\mu_i}$$

Where N is the number of patients, m_1 the measurement before lesion filling, m_2 the measurement after lesion filling and $\mu_i = \frac{1}{2} \sum_{t=1}^{2} m_{i,t}$ the within-patient mean.

Training loss	MSE loss of the noise prediction (Equation 3.5) over different diffusion timesteps
Training loss per timestep	MSE loss of the noise prediction (Equation 3.5) for specific diffusion timesteps
Gradient norm	Total norm of the parameter gradients
Validation loss	MSE loss of the noise prediction (Equation 3.5) over different diffusion timesteps on the validation set
Validation loss per timestep	MSE loss of the noise prediction (Equation 3.5) for specific diffusion timesteps on the validation set
MSE (mean squared error)	MSE between the inpainted image and the ground truth. A lower MSE indicates better image quality. $$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$
PSNR (peak signal noise ratio)	PSNR is mostly defined via the scaled inverse of the MSE between the inpainted image and the ground truth. A higher PSNR value indicates better image quality. $$PSNR = 20log_{10}(\frac{MAX_f}{\sqrt{MSE}})$$
SSIM (structural similarity index measure)	SSIM [56] measures the perceptual difference between two images x, y based on the comparison of luminance, contrast, and structure. A SSIM value near 1 indicates better image quality. $$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$ • μ_x the pixel sample mean of x • μ_y the pixel sample mean of y • σ_x^2 the variance of x • σ_y^2 the variance of y • σ_{xy} the covariance of x and y • $c_1 = (k_1 L)^2, c_2 = (k_2 L)^2$ two variables to stabilize the division with weak denominator
LPIPS (Learned Perceptual Image Patch Similarity)	LPIPS [59] computes the similarity between deep network activations of two images. A pretrained AlexNet network was used. A lower LPIPS value indicates better image quality.

Table 3.3: Evaluation metrics

This calculation is performed both for the global mean thickness (averaging the mean thickness of the left and right hemispheres) and for the average of specific ROI defined by the Desikan-Killiany (DK) atlas [11].

We further acknowledged the potential influence of MS lesions located near the cortical surface (juxtacortical lesions). Therefore, we conducted additional analyses excluding patients with such lesions when calculating mean cortical thickness. To identify patients with juxtacortical lesions, the binary lesion masks are dilated by one pixel and multiplied with the tissue segmentations. Patients with lesions outside WM are excluded in the second analysis.

Evaluation of Synthesized Lesions

To evaluate the synthetic lesions, we conducted a qualitative evaluation involving a panel of two trained neuroradiologists who assessed 20 patient scans. Each radiologist was tasked with identifying the synthesized lesion among other real lesions present in the MRI scans of MS patients.

To ensure a realistic scenario, we first calculated the volume of all connected lesions within the test set. The ten largest lesions were then selected, and one of this group was chosen to be replaced by a synthetic lesion. This chosen lesion was filled with zero pixel intensity before running the patient scan through the lesion synthesis model, effectively embedding the synthetic lesion within the existing MRI data.

During the evaluation, each doctor received one minute per patient to locate the synthetic lesion. They were provided with access to the entire 3D MRI scan and the ability to inspect all three axes for a comprehensive analysis. The ITK-SNAP software was used to facilitate this inspection process.

This qualitative evaluation aimed to assess the ability of trained radiologists to distinguish synthetic lesions from real lesions within a clinical setting, providing valuable insights into the authenticity and credibility of the synthesized lesions.

3.6 Report

To enhance the readability and style of the text, AI tools were employed. DeepL was utilized for translations between German and English, while Gemini was tasked with improving the overall flow and style. The following prompt was used to guide Gemini:

“Our task is to write a master’s thesis about deep learning. I have written a draft referencing important papers. The references are inside the square brackets. Your task is to write a consistent text. Don’t invent anything new, use the facts from the text I give you, don’t change the references and don’t use too complicated words. Don’t use any enumerations and don’t repeat the same reference within one section. Do you have any questions? Please repeat your task.”

Chapter 4

Results

This chapter presents the outcomes of the methods detailed in Chapter 3. The results are organized separately for lesion filling and lesion synthesis. Quantitative metrics are evaluated first, followed by qualitative examples. Additional observations are then discussed.

4.1 Lesion Filling

Evaluation

The 3D conditional model trained with a balanced mixture of lesion masks and random circle masks emerges as the top-performing model, attaining a SSIM of 0.96 and LPIPS of 2e-4 on the evaluation set. A history of the metrics measured during training can be viewed in Appendix A. A comparative analysis between 2D and 3D models reveals that the latter consistently outperforms the former across all metrics. Furthermore, within the realm of conditional models, the architecture trained with random circle masks demonstrates superior performance compared to its lesion mask-trained counterpart.

	SSIM	PSNR	MSE	LPIPS
2D unconditional RePaint	0.83	28	8.2e-3	2.0e-3
2D conditional circles	0.9	32	4e-3	2e-3
2D conditional lesions	0.85	28	0.01	5e-3
2D conditional mixture	0.9	33	4e-3	1e-3
3D unconditional RePaint	0.90	32	3e-3	9e-4
3D conditional circles	0.95	38	1e-3	3e-4
3D conditional lesions	0.93	34	3e-3	4e-4
3D conditional mixture	**0.96**	**39**	**8e-4**	**2e-4**

Table 4.1: Metrics measured with validation dataset. SSIM, PSNR and MSE are measured inside the mask and LPIPS over the full image.

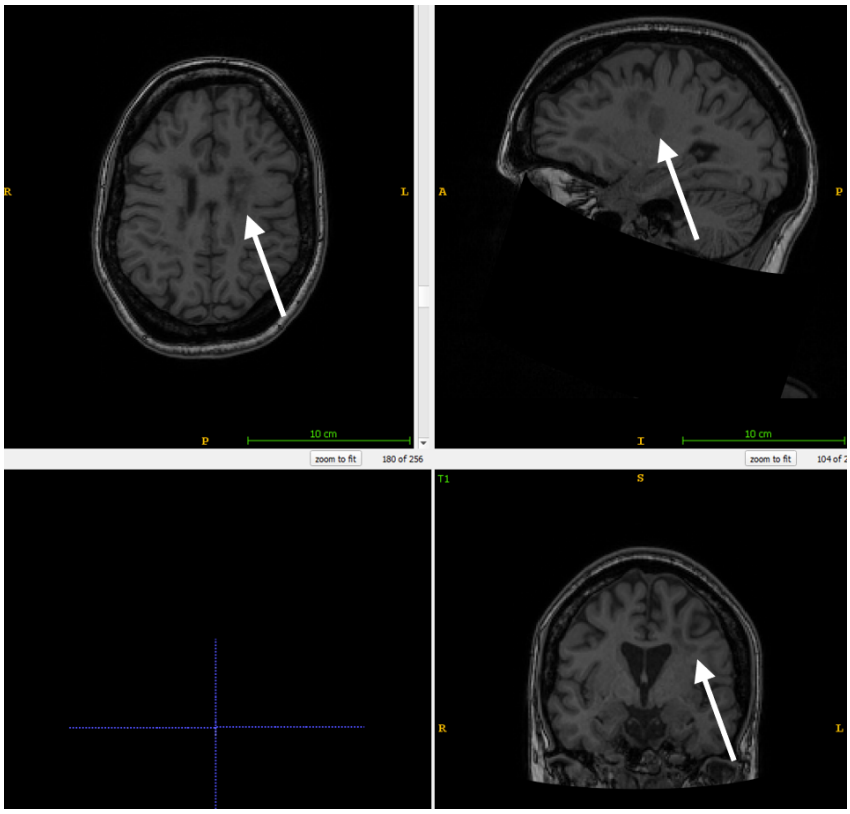

Figure 4.1: T1w before lesion filling with conditional mixture model

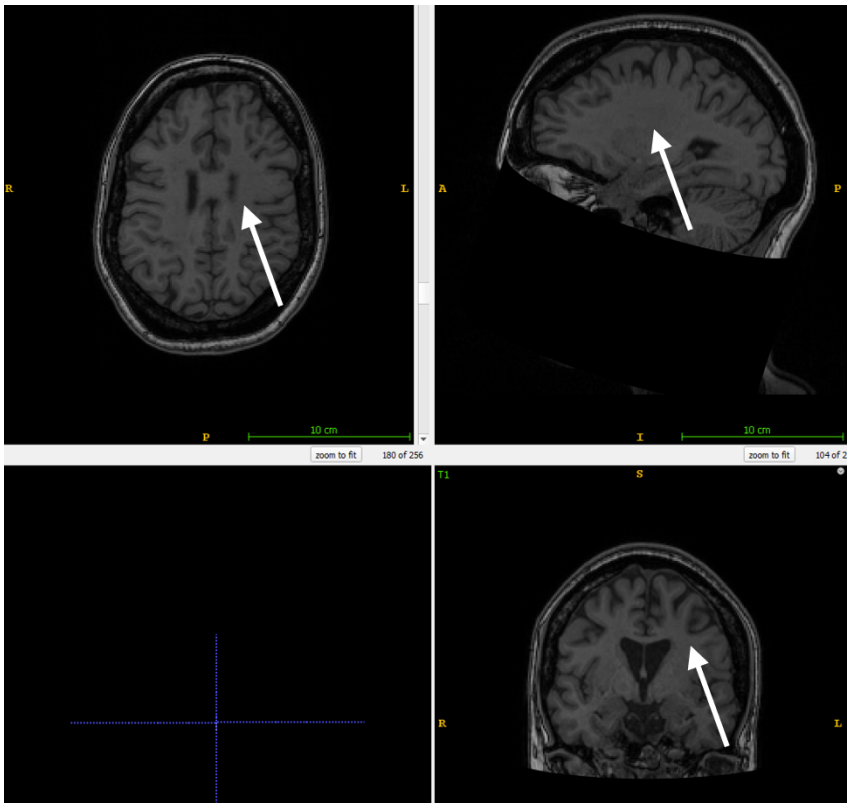

Figure 4.2: T1w after lesion filling with conditional mixture model

A significant difference exists in terms of inference time. Due to the resampling approach, the inference time for the unconditional RePaint model is substantially longer compared to the conditional models. For a batch of 16 samples on a single Nvidia RTX A6000 40GB GPU, the inference time is 45 seconds for the conditional mixture model and 350 seconds for the unconditional RePaint model.

To investigate whether the augmentation with additional random circle masks benefits larger datasets, we trained the 3D conditional mixture model on the BraTS Inpainting Challenge 2023 dataset without further fine-tuning. The model achieved an SSIM of 0.86, a PSNR of 34.4, and an MSE of 0.0063. In terms of SSIM, these results are comparable to those reported in [13], which also employed the pseudo-3D U-Net architecture but was trained solely with the lesion masks provided in the challenge dataset.

Initially, we were uncertain about the effectiveness of standard image quality metrics like SSIM in accurately representing lesion quality compared to more advanced metrics such as LPIPS or FID. However, direct comparisons with training samples shows that they are suitable representatives as presented in Appendix A. Furthermore, when analyzing metric development over time, similar trends for SSIM, PSNR, MSE, and LPIPS can be observed.

Artifacts

Incomplete lesion masking leads to recognizable artifacts in the form of residual borders at the lesion edges. This phenomenon is particularly pronounced in the unconditional RePaint approach, which replaces regions outside the mask with original image content. Although

less obvious, conditional models also exhibit similar artifacts. A small one-pixel dilation limited to the WM effectively mitigates this problem.

The intrinsic two-dimensional nature of 2D models leads to another artifact: inconsistencies along the z-axis, manifesting as visible stripes. The pseudo-3D models successfully mitigate these z-axis irregularities.

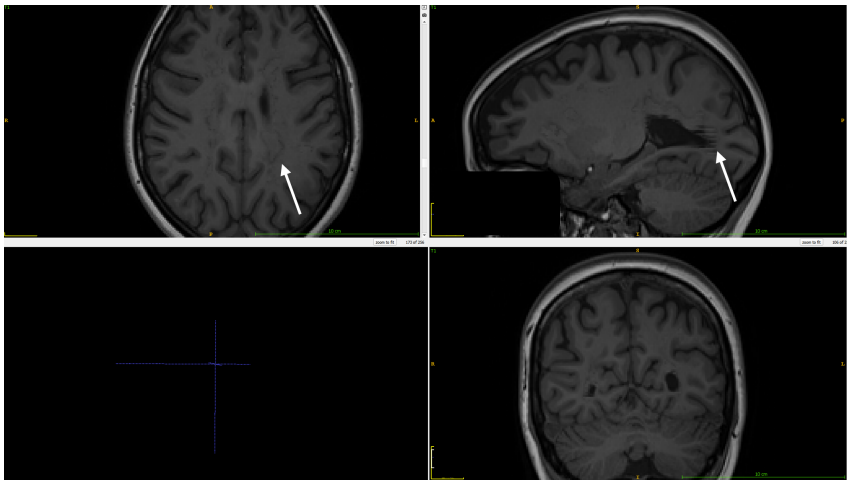

Figure 4.3: Border (left arrow) and stripe artifacts (right arrow)

Training Duration

The unconditional RePaint model exhibits significantly faster convergence compared to conditional models, achieving a peak SSIM of 0.9 after only 6000 training steps, while conditional models require approximately 90,000 steps to reach comparable performance.

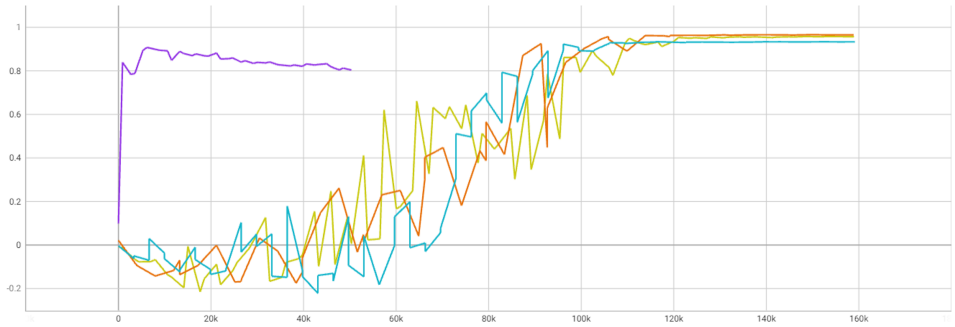

Figure 4.4: SSIM score during training of the 4 3D models unconditional RePaint (violet), conditional mixture (red), conditional circles (yellow) and conditional lesions (blue).

Interestingly, RePaint achieves optimal performance when the underlying unconditional model remains unconverged. This phenomenon is evident when sampling random 2D images using a DDIM sampler instead of the RePaint sampler, resulting in highly noisy outputs. While RePaint's strong guidance produces high-quality inpainting results, DDIM sampling reveals the underlying unconditional model's immaturity. This raises the question

of whether preventing overfitting and refining the unconditional RePaint model can match or surpass the performance of conditional models while requiring substantially less training time.

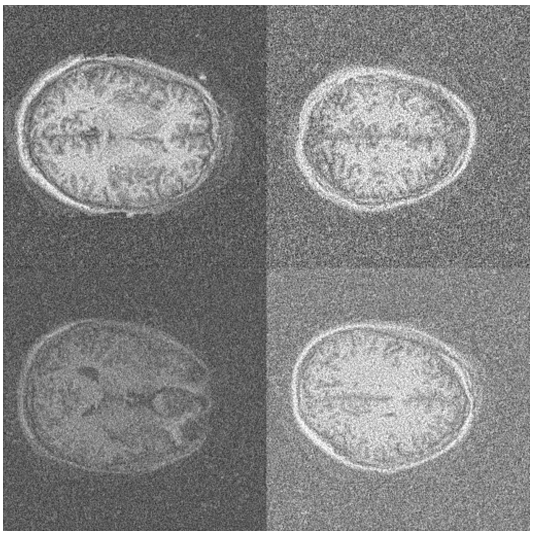

Figure 4.5: Samples of unconditional model with DDIM sampler at a timestep in training, where the model achieves its best scores with the RePaint sampler.

Examining the validation loss per timestep reveals that smaller timesteps begin to overfit while larger timesteps continue learning. To counteract this imbalance, we adopted the min-SNR weighting strategy outlined in Section 3.4.

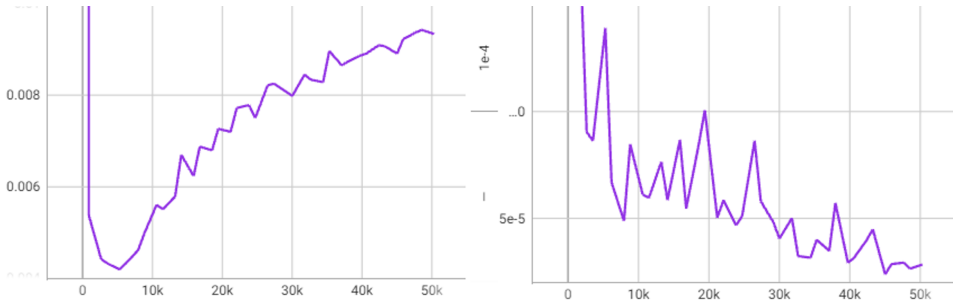

Figure 4.6: Comparison of the validation loss of timestep 200 (left) and 800 (right).

	SSIM	PSNR	MSE	LPIPS
2D Unconditional RePaint	0.83	28	8.2e-3	**2.0e-3**
2D Unconditional RePaint with min-SNR	0.83	**29**	**7.3e-3**	2.8e-3

Table 4.2: Min-SNR metrics measured with validation dataset. SSIM, PSNR and MSE are measured inside the mask and LPIPS over the full image.

Min-SNR loss mitigated overfitting, reducing the overall validation loss across all timesteps to 0.008 compared to 0.013 with unweighted MSE loss. However, metrics such as SSIM, PSNR, MSE, and LPIPS showed no significant improvement. To simplify the training process, min-SNR loss was excluded from the model training.

Robustness Evaluation Cortical Thickness

Table 4.3 presents mean reproducibility errors for both global mean thickness and the average across all 68 ROIs, calculated using data from 65 patients. To account for potential influences of juxtacortical lesions, we excluded patients with such lesions and recalculated the same measurements for the remaining 17 patients, with results displayed in Table 4.4. Lesion filling was performed using the 3D conditional mixture model. ANTsPyNet incorporating deep learning demonstrates significantly improved robustness compared to its predecessor and Freesurfer. Furthermore, the DL+DiReCT approach yields a substantial additional reduction in error.

Comparing robustness across different regions reveals consistent superiority of the newer deep learning-based methods (see Figure 4.7). ANTsPyNet's least robust region is the left frontal pole with a 1.4% error. In contrast, Freesurfer's least robust regions are the left temporal pole and entorhinal cortex, with errors of 2.4% and 2.2% respectively. Finally, DL+DiReCT exhibits the lowest robustness in the right and left pericalcarine regions, with errors of 0.6% and 0.5%.

	Global mean thickness (%)	ROI-average (%)
ANTs	1.31	1.68
ANTsPyNet	0.52	0.84
Freesurfer	0.51	0.92
DL+DiReCT	**0.05**	**0.14**

Table 4.3: Mean reproducibility errors

	Global mean thickness (%)	ROI-average (%)
ANTs	1.34	1.68
ANTsPyNet	0.38	0.81
Freesurfer	0.61	0.90
DL+DiReCT	**0.04**	**0.12**

Table 4.4: Mean reproducibility errors without patients with juxtacortical lesions

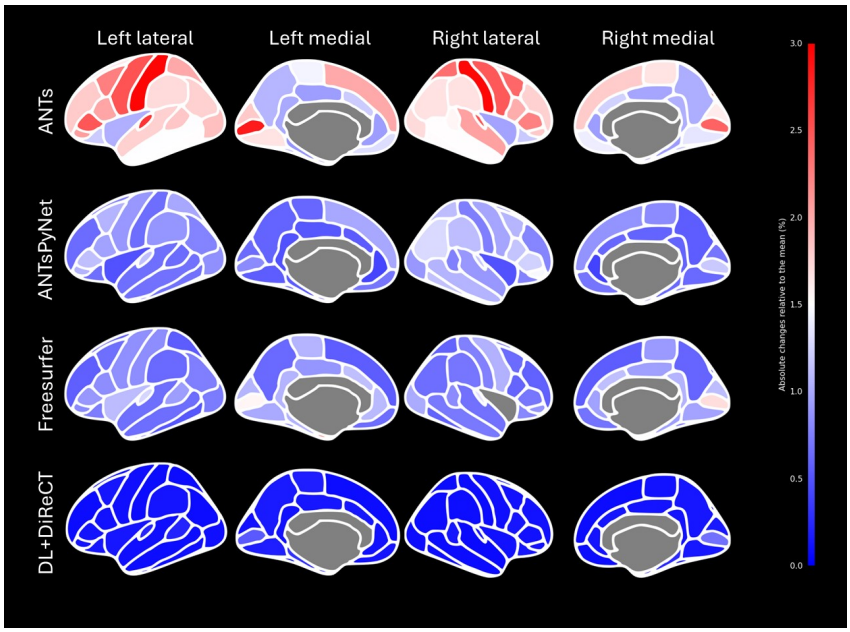

Figure 4.7: Color-coded reproducibility errors of the ROI-wise average cortical thicknesses evaluated on all samples.

4.2 Lesion Synthesis

Evaluation

The 3D conditional model emerged as the top-performing model, achieving an SSIM of 0.79 and an LPIPS value of 1.3e-4.

	SSIM	PSNR	MSE	LPIPS
2D Unconditional (with $t_0 = 1$ and median lesion intensity)	0.69	23.16	0.023	2.3e-3
3D Unconditional (with $t_0 = 3$ and median lesion intensity)	0.69	23.92	0.019	1.6e-4
3D Conditional	**0.79**	**27.13**	**0.009**	**1.3e-4**

Table 4.5: Metrics measured with validation dataset. SSIM, PSNR and MSE are measured inside the mask and LPIPS over the full image.

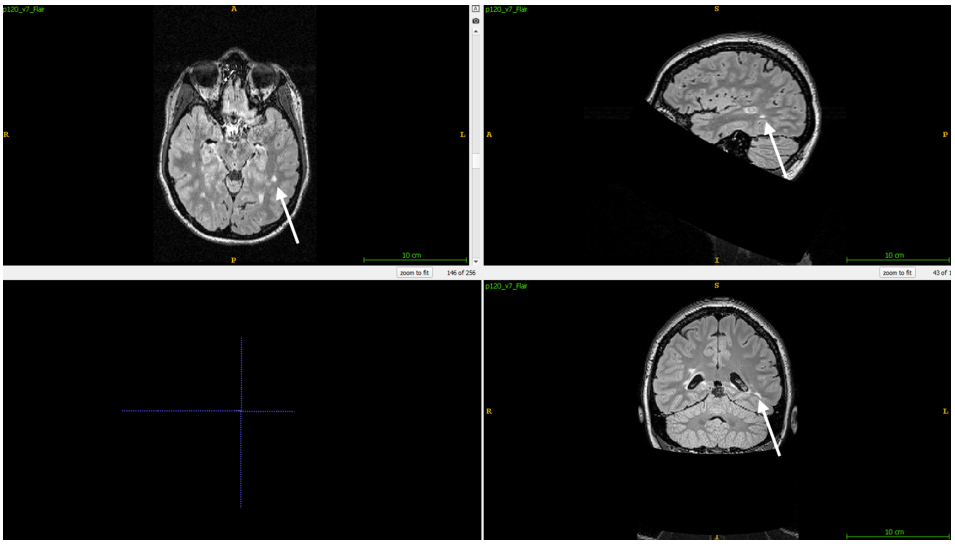

Figure 4.8: First example of synthetic lesion in a MS patient from the RR-MS dataset

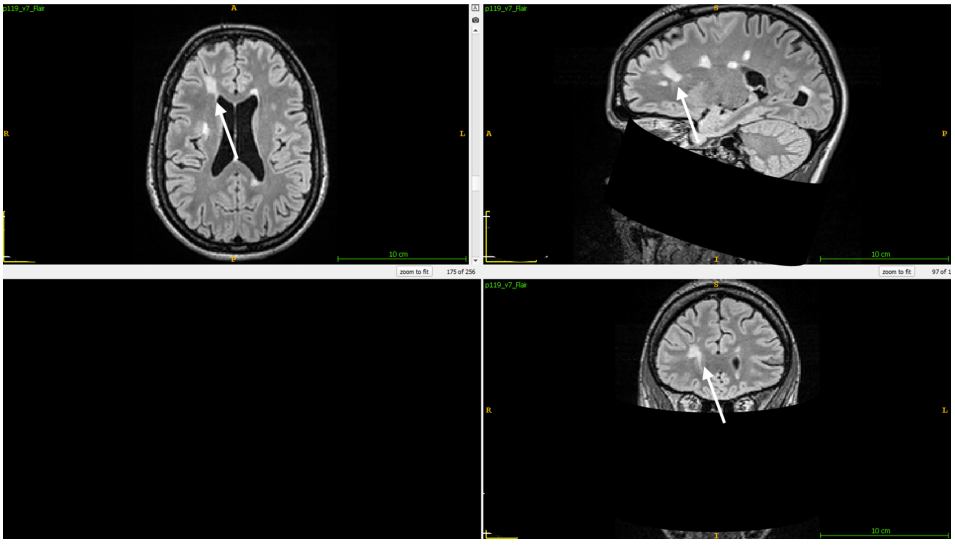

Figure 4.9: Second example of synthetic lesion in a MS patient from the RR-MS dataset

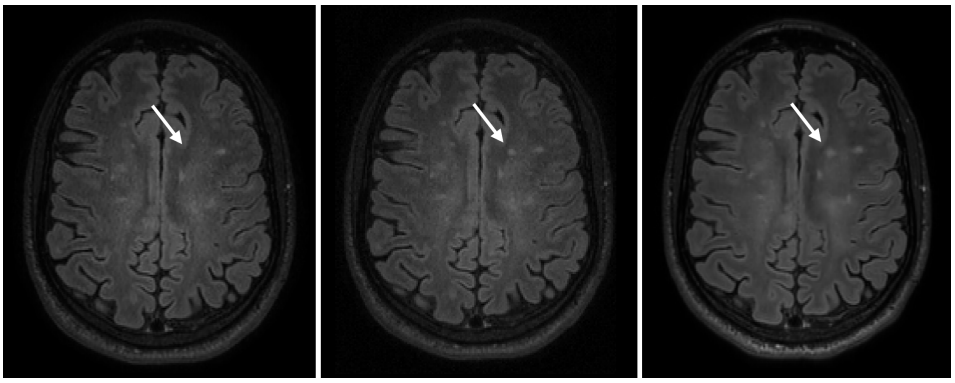

Figure 4.10: Comparison of a patient's brain scans from the MSSEG 2 dataset: (1) before adding a synthetic lesion to healthy tissue, (2) after adding the synthetic lesion, and (3) at a later timepoint with a natural new lesion at the same location.

Qualitative Evaluation

In a set of 20 examples, neuroradiologist number one identified three synthetically added lesions, while neuroradiologist number two detected only one.

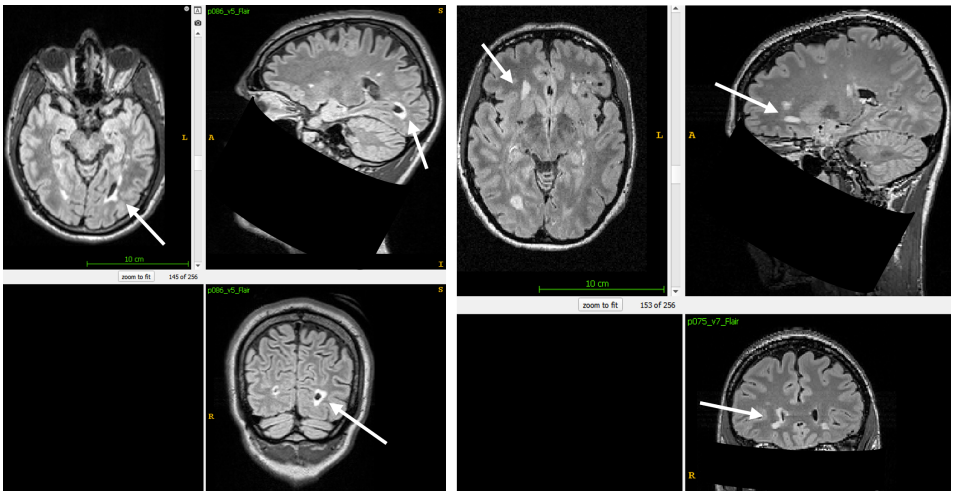

Figure 4.11: Example of two synthetically added lesions, each identified by one neuroradiologist

Artifacts in Unconditional Model

The unconditional model exhibited high sensitivity to the initial timestep and coarse lesion technique parameters. Without precise control, the model attempted to eliminate the coarse lesion. Within just five diffusion steps, the lesion was often completely removed, which is significantly fewer than the total 50 inference steps. Using brighter coarse lesion techniques, such as the 75th percentile, reduced lesion removal.

Although very small timesteps prevented lesion removal, the resulting lesions appeared less realistic.

Further experiments with T1w images revealed that lesions were frequently inpainted as GM, especially when located near the cortex or having larger sizes.

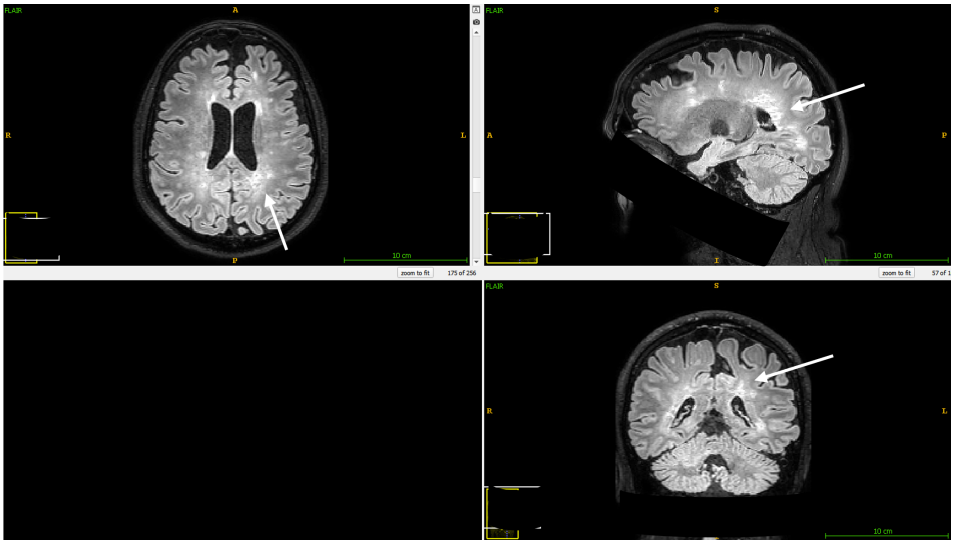

Figure 4.12: Synthetic lesions added to FLAIR images using a 3D unconditional model with parameters set to $t_0 = 4$ and median coarse lesion intensity. Partial lesion removal resulted in diffuse patterns.

Figure 4.13: Synthetic lesions added to FLAIR images using a 3D unconditional model with parameters set to $t_0 = 1$ and 0.75 percentile coarse lesion intensity. The resulting structures are very coarse and lack authenticity.

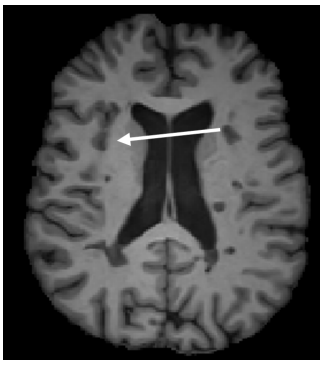

Figure 4.14: Synthetic lesion on a T1w image inpainted as GM

Inaccuracy in mask registration

Achieving realistic lesion inpainting requires the use of realistic masks. However, registration of lesion masks across patients has often yielded poor results, resulting in unrealistic masks and hindering the automated production of synthetic masks.

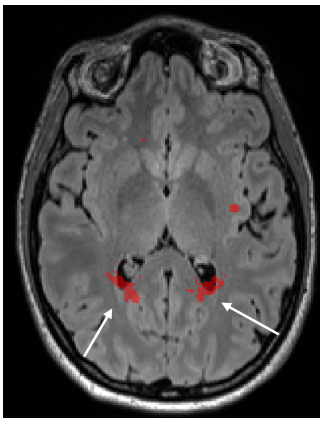

Figure 4.15: Example of a poorly registered lesion mask

4.3 Discarded Methods

Dice Score Evaluation

To objectively evaluate the quality of synthesized lesions, we employed the DeepSCAN segmentation model to automatically segment lesions within the synthesized 3D MRIs. The Dice score, calculated to measure the overlap between the segmented lesion mask and the conditional lesion mask used for synthesis, was used as an assessment metric.

Initial experiments revealed that the segmentation model also labeled MS lesions of very poor quality. While this capability is valuable for assessing model faithfulness to the masks, it is not a reliable indicator of lesion quality. Therefore, this evaluation method was discarded for subsequent experiments.

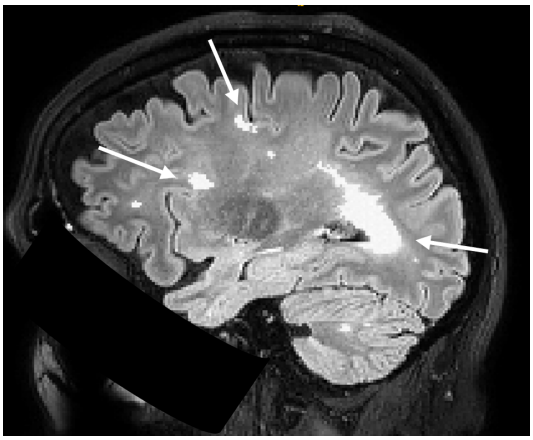

Figure 4.16: Low-quality synthetic lesions labeled by the segmentation model

Axis Augmentation

To augment the training data, we incorporated slices from the vertical and sagittal views in addition to the transversal slices. This approach was tested with a 2D unconditional RePaint model. However, neither the validation loss nor the SSIM, PSNR, or MSE scores improved. As a result, axis augmentation was discarded.

Chapter 5

Discussion

This chapter places the results from Chapter 4 within a broader context, with separate discussions for lesion filling and lesion synthesis.

5.1 Lesion Filling

Conditional models perform better than unconditional models, but at the cost of significantly longer training times (days versus hours). The better performance is intuitive, as conditional models are explicitly trained for inpainting, while unconditional models are reused through the RePaint sample approach. Another downside lays in the longer sampling time. Despite this, the RePaint method yields surprisingly high-quality inpainted images. This is particularly impressive considering the underlying unconditional model's immaturity at peak inpainting performance, characterized by substantial noise in sampled T1w images. Remarkably, the RePaint process effectively guides the generation of high-quality inpainted outputs. Given the potential for further improvement in the underlying unconditional model, we explored reducing overfitting through min-SNR loss. However, these efforts met with limited success and were therefore excluded from the final models in order to reduce the complexity. Despite this, the observed reduction in validation loss suggests that min-SNR loss holds potential as a valuable approach for future exploration. Optimizing hyperparameters could enhance its effectiveness. Additionally, other unexplored strategies for mitigating overfitting might prove beneficial. Further research is necessary to identify and evaluate these alternatives.

Regarding architectures, employing a pseudo-3D architecture, as suggested by [13], improved the performance of all models.

Comparing different conditional models, we found that training with additional circle masks enhanced performance. Surprisingly, training exclusively with random circle masks yielded better results than using only lesion masks. This suggests that a broader range of unrealistic masks distributed across MR-images is beneficial within the given dataset, rather than a smaller set of masks sampled from the true mask distribution. However, the impact of random circle masks is notably reduced with larger datasets, as demonstrated by the results obtained with the BraTS Inpainting Challenge dataset.

Evaluations of lesion filling demonstrate the importance of precise masks. Inaccurate masks result in residual lesion borders, with the unconditional RePaint model exhibiting a more pronounced effect compared to conditional models. Larger masks generally work better than smaller ones.

The impact of lesion filling on cortical thickness measurements varies significantly de-

pending on the morphological tool employed. While the original ANTs model is strongly affected, the newer ANTsPyNet model, incorporating deep learning, shows a much smaller impact and also outperforms Freesurfer. DL+DiReCT, however, exhibits the smallest differences among all models. These findings suggest that deep learning models are more robust to WM lesions.

5.2 Lesion Synthesis

We experimented with different models and image modalities for lesion synthesis. Initially using T1w images, we achieved promising results when lesions were surrounded by WM. However, issues arose when lesions were near the cortex or large, as the model incorrectly inpainted them as GM. To address this, we transitioned to FLAIR images, where WM lesions and GM matter exhibit distinct intensities.

While the FLAIR images prevented GM inpainting with the unconditional RePaint approach, the model showed a tendency to remove the coarse lesions introduced during sampling. Increasing time steps and decreasing intensity enhanced this removal. Extensive parameter tuning did not consistently produce satisfactory results. Whether this behaviour could be effectively utilized as a reliable mask-free lesion filling technique was not further evaluated.

Transitioning to a conditional model yielded significantly improved, more consistent, and reproducible results. Comparing the synthetic lesion to the original lesion they seem to be a bit duller and less diverse. During the evaluation with the neuroradiologists they didn't notice that, respectively it wasn't an indication of a synthetic lesion.

Creating new lesion masks was difficult due to inconsistent registration results when registering masks from different patients. The disadvantage of the conditional approach is that the intensity of the lesion is less controllable because it is only a binary mask. For more control, extended annotated data would be needed.
As with lesion filling the pseudo-3D architecture improved the performance of all models.

Qualitative evaluation by two neuroradiologists confirmed the authenticity of synthetic lesions generated by the conditional model, as professionals struggled to identify them. This suggests the model effectively captured lesion distribution. The next step involves training a new segmentation model using synthetic data augmentation to assess its impact. However, this requires a satisfactory method for producing synthetic masks.

Chapter 6

Conclusions

In this thesis we successfully developed two deep learning models for filling and synthesizing MS lesions in MR-images. We demonstrated the suitability of diffusion models for both tasks.

Comparing unconditional and conditional models for lesion filling, we found that while unconditional models initially produced high-quality results, they overfitted faster. Conditional models ultimately performed better. Additionally, using random circles as masks alongside lesion masks improved performance on smaller datasets compared to using only lesion masks.

Our analysis of the impact of lesion filling on cortical thickness measurement revealed that modern deep learning-based tools are less affected by MS lesions than traditional methods. This raises the possibility that lesion filling might become obsolete with the advancement of these tools.

For lesion synthesis, conditional models again outperformed unconditional ones. The generated lesions were realistic, misleading even trained neuroradiologists.
Employing pseudo-3D architectures significantly enhanced the performance of all models, confirming findings from previous research [13].

In conclusion, this thesis contributes to the development of deep learning-based lesion filling and synthesis methods. We have shown the potential obsolescence of lesion filling and the high quality of synthesized lesions, which can be used to train other deep learning models.

6.1 Limitations

This thesis has limitations that should be considered when interpreting the results. The dataset used in this study is relatively small, which might limit the representativeness of the ethnic groups included. Regarding lesion filling, the current study focused on filling multiple lesions simultaneously. The performance of the model for inpainting single lesions while preserving others remains unexplored and could potentially differ. For lesion filling and synthesis, we utilized lesion masks created by doctors based on their interpretation of MR-images. It's important to note that lesions can also influence the surrounding brain tissue, which may not be readily identifiable by humans on current MR-images. The extent of this influence and its relevance for lesion filling is a separate research question and may vary depending on the specific use case.

6.2 Outlook

While this research successfully achieved its goals, several avenues for further exploration remain.

The lesion filling models, developed for MS lesions, could be applied to other inpainting tasks. However, performance might vary, especially considering the training objectives. It would be interesting to determine if the performance advantage of conditional models over unconditional ones persists in these new applications. Improving the unconditional model to match the performance of the conditional model is another potential area of research. This is desirable due to the unconditional model's shorter training time and independence from mask distribution.

Although we demonstrated that deep learning-based tools are more robust to MS lesions than older methods when measuring cortical thickness, a larger population study is necessary to definitively establish the obsolescence of lesion filling.

With a model capable of generating realistic lesion inpaintings, the next step is to create a synthetic lesion dataset for training a new segmentation model. This dataset can then be used to determine the optimal proportion of synthetic data needed to maximize performance gains. To generate a large dataset, a reliable method for producing synthetic masks is essential.

Bibliography

[1] M. Almansour, N. M. Ghanem, and S. Bassiouny. High-resolution MRI brain inpainting. In BHI 2021 - 2021 IEEE EMBS International Conference on Biomedical and Health Informatics, Proceedings, 2021.

[2] H. Amiri, A. de Sitter, K. Bendfeldt, M. Battaglini, C. A. Gandini Wheeler-Kingshott, M. Calabrese, J. J. Geurts, M. A. Rocca, J. Sastre-Garriga, C. Enzinger, N. de Stefano, M. Filippi, Rovira, F. Barkhof, and H. Vrenken. Urgent challenges in quantification and interpretation of brain grey matter atrophy in individual MS patients using MRI, 2018.

[3] S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, and D. J. Fleet. Synthetic Data from Diffusion Models Improves ImageNet Classification. 4 2023.

[4] M. Battaglini, M. Jenkinson, and N. De Stefano. Evaluating and reducing the impact of white matter lesions on brain volume measurements. Human Brain Mapping, 33(9), 2012.

[5] F. Bieder, J. Wolleb, A. Durrer, R. Sandkühler, and P. C. Cattin. Memory-Efficient 3D Denoising Diffusion Models for Medical Image Processing. In Proceedings of Machine Learning Research, volume 227, 2023.

[6] M. Brett, C. J. Markiewicz, M. Hanke, M.-A. Côté, B. Cipollini, P. McCarthy, D. Jarecka, C. P. Cheng, E. Larson, Y. O. Halchenko, M. Cottaar, S. Ghosh, D. Wassermann, S. Gerhard, G. R. Lee, Z. Baratz, H.-T. Wang, D. Papadopoulos Orfanos, E. Kastman, J. Kaczmarzyk, R. Guidotti, J. Daniel, O. Duek, A. Rokem, M. Scheltienne, C. Madison, A. Sólon, B. Moloney, F. C. Morency, M. Goncalves, R. Markello, C. Riddell, C. Burns, J. Millman, A. Gramfort, J. Leppäkangas, J. J. F. den Bosch, R. D. Vincent, H. Braun, K. Subramaniam, A. Van, K. J. Gorgolewski, P. R. Raamana, J. Klug, R. de Wael, B. N. Nichols, E. M. Baker, S. Hayashi, B. Pinsard, C. Haselgrove, M. Hymers, O. Esteban, S. Koudoro, F. Pérez-García, J. Dockès, N. N. Oosterhof, B. Amirbekian, H. Christian, I. Nimmo-Smith, L. Nguyen, P. Suter, S. Reddigari, S. St-Jean, E. Panfilov, E. Garyfallidis, G. Varoquaux, J. H. Legarreta, K. S. Hahn, L. Waller, O. P. Hinds, B. Fauber, B. Dewey, F. Perez, J. Roberts, J.-B. Poline, J. Stutters, K. Jordan, M. Cieslak, M. E. Moreno, T. Hrnčiar, V. Haenel, Y. Schwartz, B. C. Darwin, B. Thirion, C. Gauthier, I. Solovey, I. Gonzalez, J. Palasubramaniam, J. Lecher, K. Leinweber, K. Raktivan, M. Calábková, P. Fischer, P. Gervais, S. Gadde, T. Ballinger, T. Roos, V. R. Reddam, and freec84. nipy/nibabel: 5.2.0, 12 2023.

[7] T. Brooks, A. Holynski, and A. A. Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2023-June, 2023.

[8] O. Commowick, F. Cervenansky, F. Cotton, and M. Dojat. MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure. MICCAI 2021 - 24th International Conference on Medical Image Computing and Computer Assisted Intervention, 2021.

[9] O. Commowick, M. Kain, R. Casey, R. Ameli, J. C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, S. Camarasu-Pop, T. Glatard, S. Vukusic, G. Edan, C. Barillot, M. Dojat, and F. Cotton. Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset. NeuroImage, 244, 2021.

[10] H. Crayton, R. A. Heyman, and H. S. Rossman. A multimodal approach to managing the symptoms of multiple sclerosis, 2004.

[11] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, M. S. Albert, and R. J. Killiany. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage, 31(3), 2006.

[12] A. Durrer, P. C. Cattin, and J. Wolleb. Denoising Diffusion Models for Inpainting of Healthy Brain Tissue. 2 2024.

[13] A. Durrer, J. Wolleb, F. Bieder, P. Friedrich, L. Melie-Garcia, M. Ocampo-Pineda, C. I. Bercea, I. E. Hamamci, B. Wiestler, M. Piraud, Yaldizli, C. Granziera, B. H. Menze, P. C. Cattin, and F. Kofler. Denoising Diffusion Models for 3D Healthy Brain Tissue Inpainting. 3 2024.

[14] M. R. Farazi, F. Faisal, Z. Zaman, and S. Farhan. Inpainting multiple sclerosis lesions for improving registration performance with brain atlas. In 1st International Conference on Medical Engineering, Health Informatics and Technology, MediTec 2016, 2017.

[15] M. A. Farooq, W. Yao, M. Schukat, M. A. Little, and P. Corcoran. Derm-T2IM: Harnessing Synthetic Skin Lesion Data via Stable Diffusion Models for Enhanced Skin Disease Classification using ViT and CNN. 1 2024.

[16] B. Fischl. FreeSurfer. NeuroImage, 62(2):774–781, 8 2012.

[17] R. Gelineau-Morel, V. Tomassini, M. Jenkinson, H. Johansen-Berg, P. M. Matthews, and J. Palace. The effect of hypointense white matter lesions on automated gray matter segmentation in multiple sclerosis. Human brain mapping, 33(12):2802–14, 12 2012.

[18] T. Hang, S. Gu, C. Li, J. Bao, D. Chen, H. Hu, X. Geng, and B. Guo. Efficient Diffusion Training via Min-SNR Weighting Strategy. In Proceedings of the IEEE International Conference on Computer Vision, 2023.

[19] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems, volume 2020-December, 2020.

[20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. 6 2021.

[21] J. Jovicich, M. Marizzoni, R. Sala-Llonch, B. Bosch, D. Bartrés-Faz, J. Arnold, J. Benninghoff, J. Wiltfang, L. Roccatagliata, F. Nobili, T. Hensch, A. Tränkner, P. Schönknecht, M. Leroy, R. Lopes, R. Bordet, V. Chanoine, J. P. Ranjeva, M. Didic, H. Gros-Dagnac, P. Payoux, G. Zoccatelli, F. Alessandrini, A. Beltramello, N. Bargalló,

O. Blin, and G. B. Frisoni. Brain morphometry reproducibility in multi-center 3T MRI studies: A comparison of cross-sectional and longitudinal segmentations. NeuroImage, 83, 2013.

[22] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof. Diffusion Models for Medical Image Analysis: A Comprehensive Survey. arXiv, 2022.

[23] F. Kofler, F. Meissen, F. Steinbauer, R. Graf, E. Oswald, E. de da Rosa, H. B. Li, U. Baid, F. Hoelzl, O. Turgut, I. Horvath, D. Waldmannstetter, C. Bukas, M. Adewole, S. M. Anwar, A. Janas, A. F. Kazerooni, D. LaBella, A. W. Moawad, K. Farahani, J. Eddy, T. Bergquist, V. Chung, R. T. Shinohara, F. Dako, W. Wiggins, Z. Reitman, C. Wang, X. Liu, Z. Jiang, A. Familiar, G.-M. Conte, E. Johanson, Z. Meier, C. Davatzikos, J. Freymann, J. Kirby, M. Bilello, H. M. Fathallah-Shaykh, R. Wiest, J. Kirschke, R. R. Colen, A. Kotrotsou, P. Lamontagne, D. Marcus, M. Milchenko, A. Nazeri, M.-A. Weber, A. Mahajan, S. Mohan, J. Mongan, C. Hess, S. Cha, J. Villanueva-Meyer, E. Colak, P. Crivellaro, A. Jakab, J. Albrecht, U. Anazodo, M. Aboian, J. E. Iglesias, K. Van Leemput, S. Bakas, D. Rueckert, B. Wiestler, I. Ezhov, M. Piraud, and B. Menze. The Brain Tumor Segmentation (BraTS) Challenge 2023: Local Synthesis of Healthy Brain Tissue via Inpainting. 5 2023.

[24] H. Lassmann. Multiple sclerosis pathology, 2018.

[25] G. Liu, F. A. Reda, K. J. Shih, T. C. Wang, A. Tao, and B. Catanzaro. Image Inpainting for Irregular Holes Using Partial Convolutions. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 11215 LNCS, 2018.

[26] Y. Liu, F. Yang, and Y. Yang. A partial convolution generative adversarial network for lesion synthesis and enhanced liver tumor segmentation. Journal of Applied Clinical Medical Physics, 24(4), 2023.

[27] Y. Lu, M. Shen, H. Wang, X. Wang, C. van Rechem, T. Fu, and W. Wei. Machine Learning for Synthetic Data Generation: A Review. 2 2023.

[28] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2022-June, 2022.

[29] S. Magon, L. Gaetano, M. M. Chakravarty, J. P. Lerch, Y. Naegelin, C. Stippich, L. Kappos, E. W. Radue, and T. Sprenger. White matter lesion filling improves the accuracy of cortical thickness measurements in multiple sclerosis patients: A longitudinal study. BMC Neuroscience, 15(1), 2014.

[30] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner. Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. Journal of Cognitive Neuroscience, 19(9), 2007.

[31] R. McKinley, R. Wepfer, F. Aschwanden, L. Grunder, R. Muri, C. Rummel, R. Verma, C. Weisstanner, M. Reyes, A. Salmen, A. Chan, F. Wagner, and R. Wiest. Simultaneous lesion and brain segmentation in multiple sclerosis using deep neural networks. Scientific Reports, 11(1), 2021.

[32] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J. Y. Zhu, and S. Ermon. SDEDIT: GUIDED IMAGE SYNTHESIS AND EDITING WITH STOCHASTIC DIFFERENTIAL EQUATIONS. In ICLR 2022 - 10th International Conference on Learning Representations, 2022.

[33] R. Milo and A. Miller. Revised diagnostic criteria of multiple sclerosis, 2014.

[34] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin. Fast free-form deformation using graphics processing units. Computer Methods and Programs in Biomedicine, 98(3):278–284, 6 2010.

[35] A. Nichol and P. Dhariwal. Improved Denoising Diffusion Probabilistic Models. 2 2021.

[36] J. A. Oareilly and F. Asadi. Pre-trained vs. Random Weights for Calculating Fréchet Inception Distance in Medical Imaging. In BMEiCON 2021 - 13th Biomedical Engineering International Conference, 2021.

[37] W. H. Pinaya, P. D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso. Brain Imaging Generation with Latent Diffusion Models. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 13609 LNCS, 2022.

[38] M. Rebsamen, C. Rummel, M. Reyes, R. Wiest, and R. McKinley. Direct cortical thickness estimation using deep learning-based anatomy segmentation and cortex parcellation. Human Brain Mapping, 41(17), 2020.

[39] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. 5 2015.

[40] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. 8 2022.

[41] Schweizerische Multiple Sklerose Gesellschaft. Behandlung der Multiplen Sklerose.

[42] Schweizerische Multiple Sklerose Gesellschaft. MS-Symptome und ihre Behandlung.

[43] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In 32nd International Conference on Machine Learning, ICML 2015, volume 3, 2015.

[44] J. Song, C. Meng, and S. Ermon. DENOISING DIFFUSION IMPLICIT MODELS. In ICLR 2021 - 9th International Conference on Learning Representations, 2021.

[45] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky. Resolution-robust Large Mask Inpainting with Fourier Convolutions. In Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, 2022.

[46] O. Tange. GNU Parallel 2018. Ole Tange, 3 2018.

[47] The Multiple Sclerosis International Federation. Atlas of MS, 3rd edition (September 2020). The Multiple Sclerosis International Federation (MSIF), September 2020, 2020.

[48] A. J. Thompson, B. L. Banwell, F. Barkhof, W. M. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M. S. Freedman, K. Fujihara, S. L. Galetta, H. P. Hartung, L. Kappos, F. D. Lublin, R. A. Marrie, A. E. Miller, D. H. Miller, X. Montalban, E. M. Mowry, P. S. Sorensen, M. Tintoré, A. L. Traboulsee, M. Trojano, B. M. Uitdehaag, S. Vukusic, E. Waubant, B. G. Weinshenker, S. C. Reingold, and J. A. Cohen. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria, 2018.

[49] V. E. Tiu, I. Enache, C. A. Panea, C. Tiu, and B. O. Popescu. Predictive MRI Biomarkers in MS—A Critical Review, 2022.

[50] N. J. Tustison, P. A. Cook, A. J. Holbrook, H. J. Johnson, J. Muschelli, G. A. Devenyi, J. T. Duda, S. R. Das, N. C. Cullen, D. L. Gillen, M. A. Yassa, J. R. Stone, J. C. Gee, and B. B. Avants. The ANTsX ecosystem for quantitative biological and medical imaging. Scientific reports, 11(1), 2021.

[51] N. J. Tustison, P. A. Cook, A. Klein, G. Song, S. R. Das, J. T. Duda, B. M. Kandel, N. van Strien, J. R. Stone, J. C. Gee, and B. B. Avants. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. NeuroImage, 99, 2014.

[52] S. Valverde, A. Oliver, and X. Lladó. A white matter lesion-filling approach to improve brain tissue volume measurements. NeuroImage: Clinical, 6, 2014.

[53] C. W. van der Weijden, M. S. Pitombeira, Y. R. Haveman, C. A. Sanchez-Catasus, K. R. Campanholo, G. D. Kolinger, C. M. Rimkus, C. A. Buchpiguel, R. A. Dierckx, R. J. Renken, J. F. Meilof, E. F. de Vries, and D. de Paula Faria. The effect of lesion filling on brain network analysis in multiple sclerosis using structural magnetic resonance imaging. Insights into Imaging, 13(1), 2022.

[54] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, D. Nair, S. Paul, S. Liu, W. Berman, X. Xu, and T. Wolf. Diffusers: State-of-the-art diffusion models.

[55] H. Vrenken, M. Jenkinson, D. L. Pham, C. R. Guttmann, D. Pareto, M. Paardekooper, A. de Sitter, M. A. Rocca, V. Wottschel, M. Jorge Cardoso, and F. Barkhof. Opportunities for Understanding MS Mechanisms and Progression With MRI Using Large-Scale Data Sharing and Artificial Intelligence, 2021.

[56] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 2004.

[57] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang. Free-form image inpainting with gated convolution. In Proceedings of the IEEE International Conference on Computer Vision, volume 2019-October, 2019.

[58] Y. Zeng, Z. Lin, J. Yang, J. Zhang, E. Shechtman, and H. Lu. High-Resolution Image Inpainting with Iterative Confidence Feedback and Guided Upsampling. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 12364 LNCS, 2020.

[59] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018.

[60] L. Zhu, Z. Xue, Z. Jin, X. Liu, J. He, Z. Liu, and L. Yu. Make-A-Volume: Leveraging Latent Diffusion Models for Cross-Modality 3D Brain MRI Synthesis. 7 2023.

Appendices

Appendix A

Quantitative and Qualitative Training Progression

This chapter presents both quantitative and qualitative visualizations of the different model's training progression.

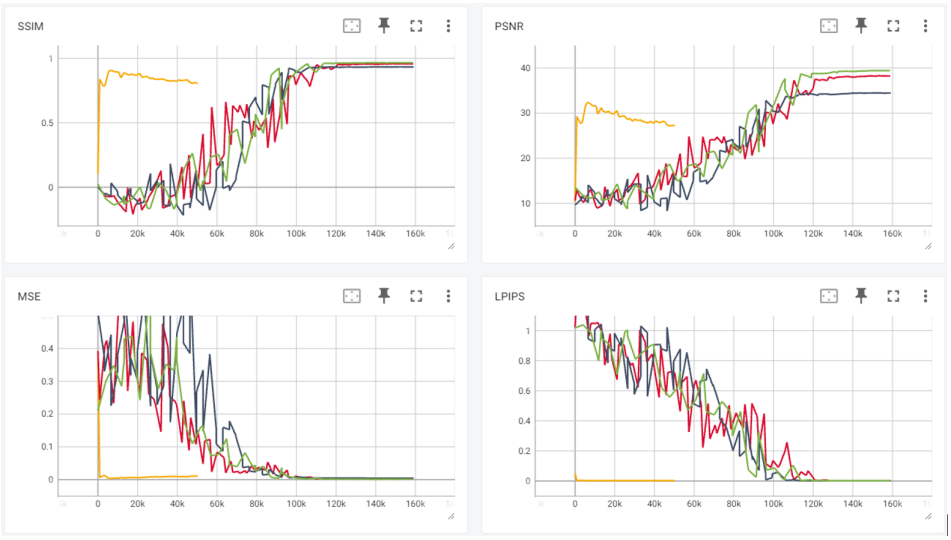

Figure A.1: Training metrics of the lesion filling 3D model's conditional mixture (green), conditional circles (red), conditional lesions (black) and unconditional RePaint (orange).

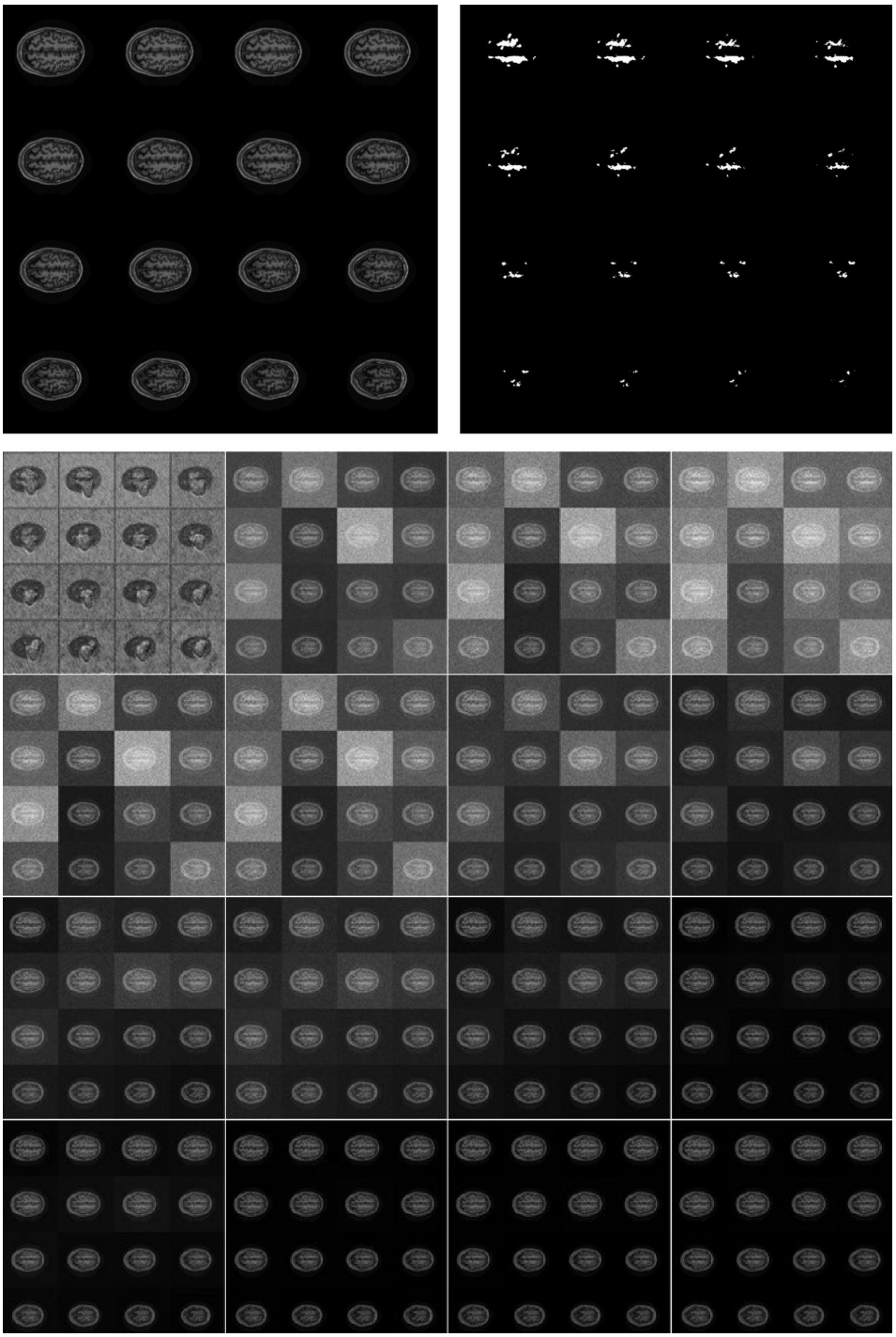

Figure A.2: 16 image-mask pairs for evaluation at the top right and left corners (Top). Below, the results of the lesion filling 3D conditional mixture model training in a grid ordered from left to right and top to bottom at specific timesteps: 000001, 011'914, 017'210, 026'476, 034'419, 039'715, 051'628, 056'924, 066'190, 074'133, 079'429, 091'342, 096'638, 105'904, 113'847, 119'143

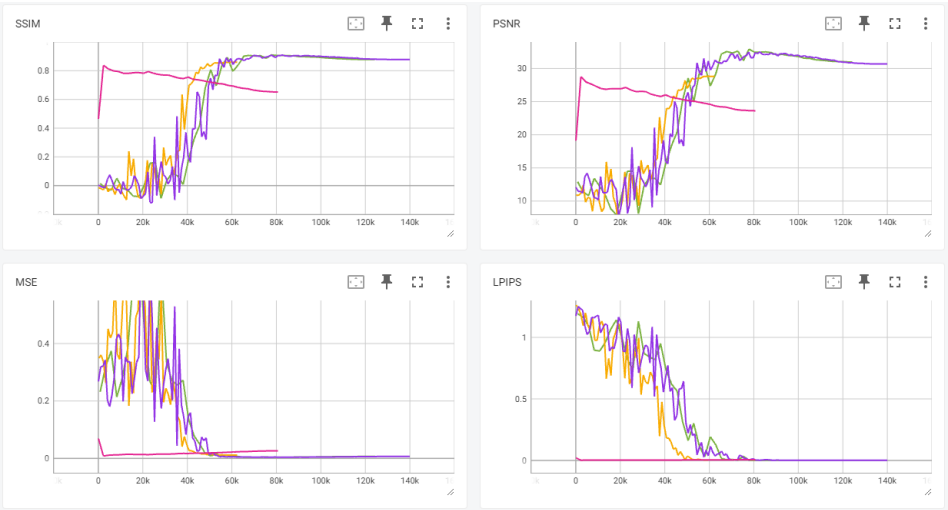

Figure A.3: Training metrics of the lesion filling 2D model's conditional mixture (green), conditional circles (purple), conditional lesions (orange) and unconditional RePaint (red).

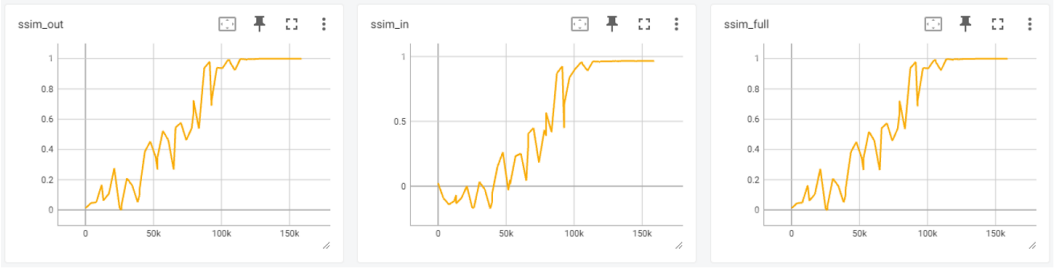

Figure A.4: SSIM metric of the 3D conditional mixture model outside and inside the mask and over the entire image.

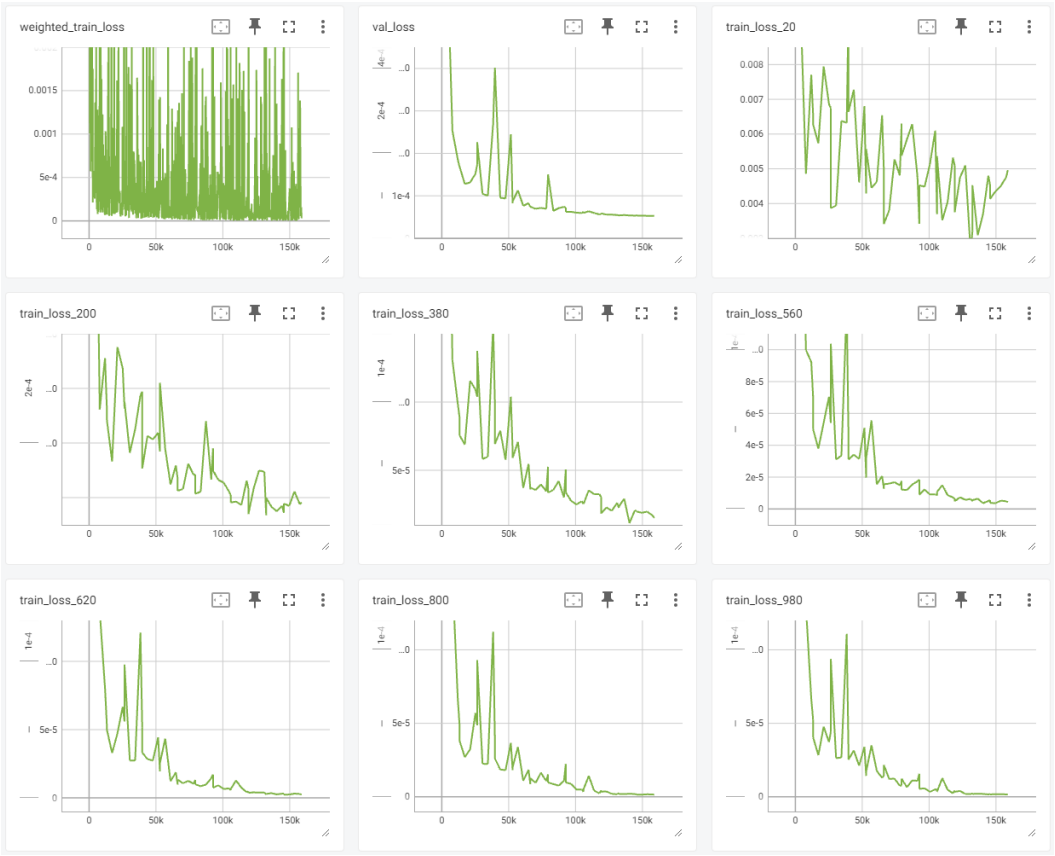

Figure A.5: Training metrics of 3D conditional mixture model: Training loss, validation loss and training loss (MSE) for timesteps 20, 200, 380, 560, 620, 800 and 980 (from left to right and top to bottom). The different MSEs between the timesteps are noteworthy, which leads to an oscillating training loss function.

Appendix B

Mask Registration Examples

This section provides visual examples of how lesion masks were aligned with images from healthy patients. The process for registration is detailed in Section 3.3.

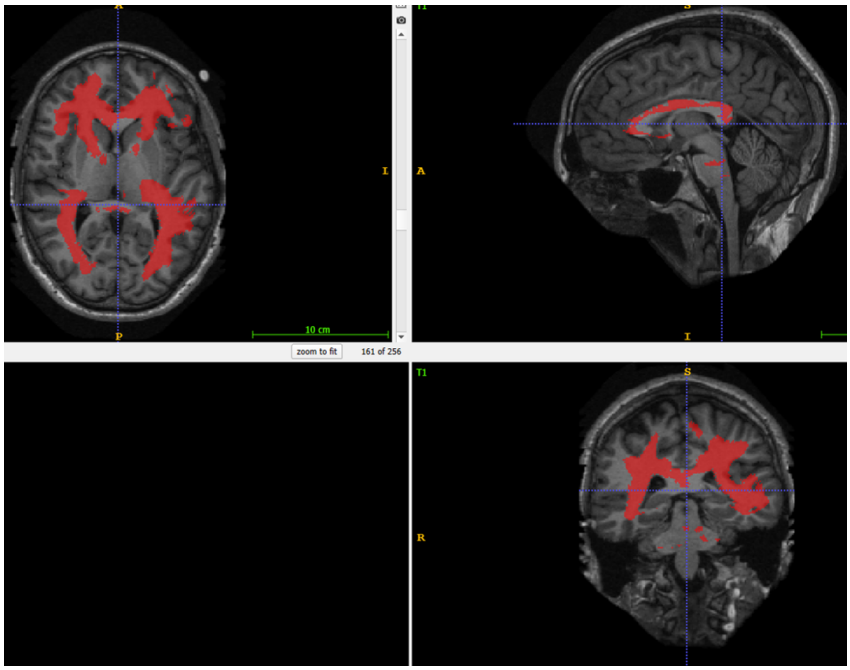

Figure B.1: Example of non-linear registration of FLAIR lesion masks directly to healthy T1w images.

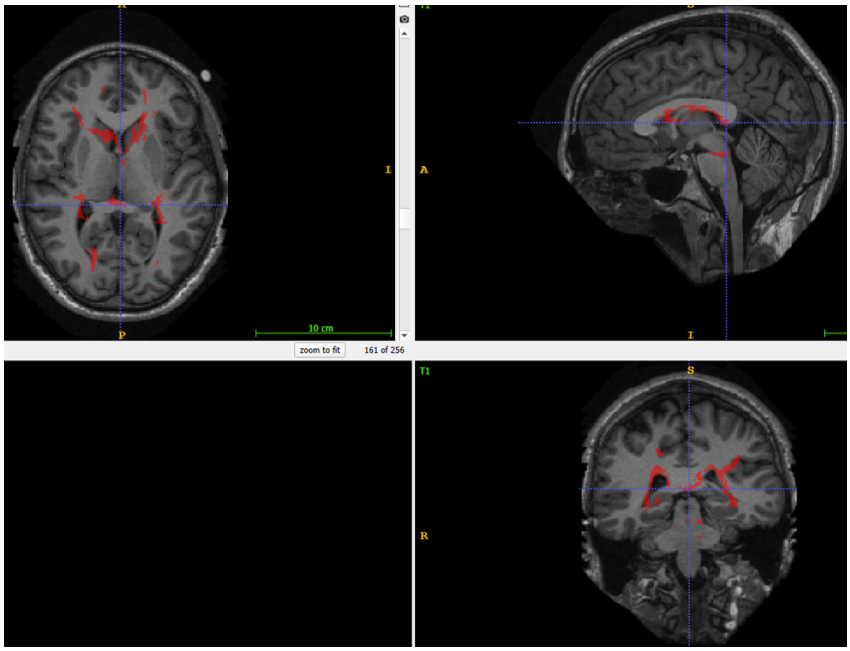

Figure B.2: Example of non-linear registration of FLAIR lesion masks to the T1w image of the corresponding MS patient, followed by registration to the healthy T1w images.

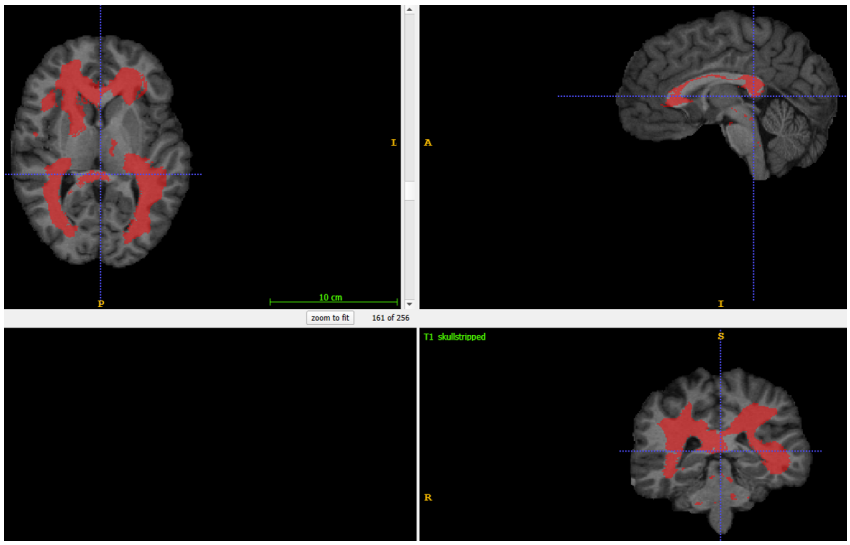

Figure B.3: Example of affine registration of the FLAIR lesion mask to the healthy T1w image, followed by skull stripping and then non-linear registration using the affine registration as initialization.