Imperial College of Science, Technology and Medicine
Department of Computing

IMPERIAL

Machine Learning for Brain Lesion Segmentation

Berke Doğa BAŞARAN

Main Supervisor:
Dr. Wenjia BAI

Second Supervisor:
Prof. Paul M. MATTHEWS

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of Imperial College, September 2024

Declaration of Originality

I, Berke Doga Basaran, hereby declare that the work described in this thesis is my own, except where specifically acknowledged.

Copyright Statement

Abstract

Deep learning techniques hold significant promise in advancing medical image processing. This is especially true for brain lesion segmentation, which is crucial for diagnosing and monitoring neurological conditions like multiple sclerosis (MS). However, challenges such as lack of clinically relevant machine learning models, data scarcity, and model generalisation remain major obstacles. This thesis addresses these challenges by proposing innovative machine learning methodologies for improving brain lesion segmentation.

The first contribution of this work focuses on a clinically overlooked task: segmentation of new lesions in MS using longitudinal data. We develop an end-to-end pipeline leveraging a convolutional neural network with specialised preprocessing and data augmentation techniques to segment new lesions appearing in follow-up scans. We validate our method against a public challenge dataset, and demonstrate superior performance against state-of-the-art biomedical segmentation models.

The second focus is on overcoming the scarcity of labelled data, which hampers the training of effective segmentation models. Two approaches are proposed. The first utilises an adversarial technique to generate synthetic brain images with lesions from healthy subjects, enhancing training dataset diversity. The second approach augments datasets at the lesion-level by altering the lesion load of images with respect to dataset load distributions. Both methods enhance brain lesion segmentation performance, especially in data-limited scenarios.

Lastly, this thesis tackles the challenge of model generalisation by introducing a segmentation model capable of handling heterogeneous input data and segmenting multiple lesion types. Our model incorporates domain-specific anatomical constraints and benefits from the augmented datasets produced by LesionMix. Results show that allowing for diverse input data and incorporating domain knowledge, we can achieve enhanced segmentation accuracy and generalisability across different datasets.

Collectively, these contributions provide a robust framework for advancing brain lesion segmentation, with significant implications for clinical applications and the broader field of medical image analysis.

Acknowledgements

Completing this PhD has been a transformative journey, filled with challenges and growth. This work is the product of years of dedication and the invaluable support of many individuals. As I present this thesis, I am deeply grateful for the guidance and encouragement that has brought me to this point.

First and foremost, I am profoundly grateful to my supervisors, Wenjia Bai and Paul M. Matthews, for their invaluable guidance and mentorship throughout this process. Your insights, patience, and commitment to my development have been instrumental in shaping this work. Thank you for giving me the freedom to pursue my research interests. I have learned immensely from your expertise and have been truly fortunate to have you as mentors. Wenjia, you have always been available whenever I have required your counsel, and could always provide me with the answers that I needed. I couldn't have wished for a better, understanding, and supporting supervisor. 谢谢!

I would also like to thank my dear friends, Will Bolton, Simon Hanassab, and Josh Southern, for their unwavering friendship and support. Our 'sitcom'-ing in the CDT space, long coffee breaks, and 2 hours lunches were a driving force for me to come into the office. Whether it was offering a listening ear, providing a much-needed distraction, or simply being there through the highs and lows, your presence has made this journey not only bearable but also truly enjoyable. I am lucky to have such a solid group of friends by my side.

I want to express my deepest appreciation to my girlfriend, Marielena Marcu, whose love and understanding have been my anchor during this demanding journey. Your endless patience as I went through my ups and downs, words of encouragement, and belief in me have been a constant source of motivation. Thank you for standing by my side. I am incredibly grateful for your support, which has meant more to me than words can express.

Most of all, I would like to express my deepest gratitude to my parents, Eser and Kaan, and my sister, Yağmur, whose continuous support and inspiration have been the foundation of my journey. Your belief in my potential and your sacrifices over the years have been a constant source of strength. Without your love and guidance, I would not have had the courage or perseverance to pursue this PhD. And now, I am fortunate enough to complete a PhD at the same university as my father's. Teşekkür ederim, sizi çok seviyorum.

If I have seen further it is by standing on the shoulders of giants.

Isaac Newton

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The field of computer vision has seen significant advancements in recent years, largely due to the development of sophisticated machine learning techniques. Among these, deep learning has gained considerable attention for its ability to automatically learn hierarchical representations of data [79]. Unlike traditional methods that require handcrafted features, deep learning models build multi-layered structures where complex features emerge from simpler ones, such as edges, blobs, and basic shapes identified in earlier feature-selective layers [111]. This capability not only reduces the need for manual feature engineering but also significantly enhances the ability of the model to capture intricate patterns, making deep learning particularly effective for tasks in medical imaging, such as brain segmentation, where brain substructures or abnormalities take on distinctive three-dimensional shapes [3].

Biomedical image applications of deep learning are slowly going into clinical practice. Deep learning aims to provide better and faster medical diagnoses despite its lengthy time for development and dataset curation [175], an example process visualised in Figure 1.1. However, the power of deep learning in medical imaging extends beyond its accuracy and adaptability to different datasets, organs, or imaging modalities; it also offers significant time savings compared to manual analysis [4]. Currently, medical images are often analysed by human experts, but given the sheer volume of scans, manual annotation and reading have become the bottleneck of the image analysis pipeline. For example, in the case of 3D brain scans, radiologists must analyse them slice by slice, which is both time consuming and prone to human error [185]. Additionally, the process is inherently subjective, with different experts potentially providing varying interpretations of the same scan [28, 182,

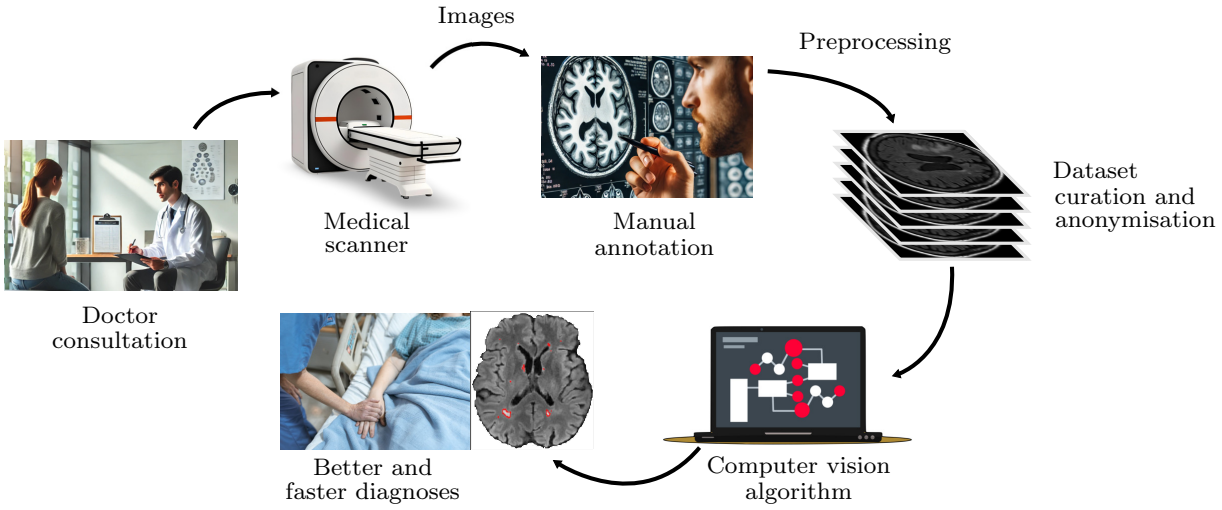


FIGURE 1.1: **An example overview of biomedical application of computer vision.** Process begins with a doctor consultation, which requests a medical scan. Acquired medical images are manually annotated by radiologists or expert professionals. Labelled data is anonymised and curated into a dataset. A computer vision algorithm utilises the dataset to train a model for segmentation or other relevant task. Robust model provides better and faster diagnoses of diseases. Adapted from [175].

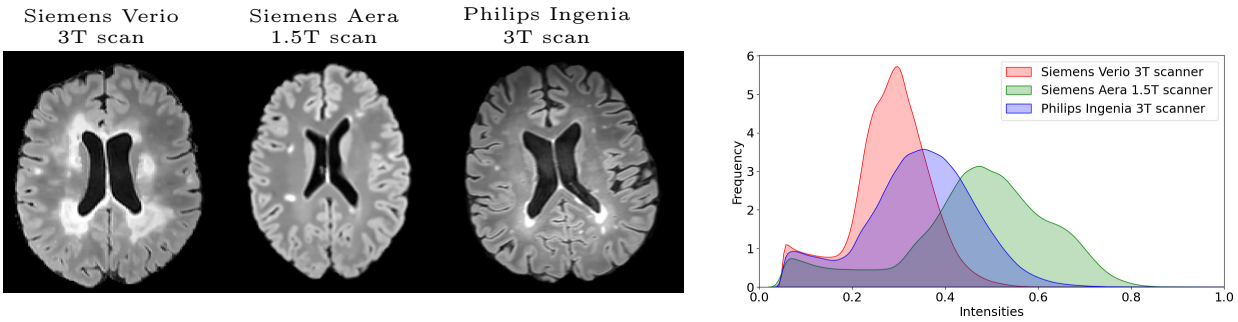
102]. In this context, a deep learning pipeline capable of segmenting brain lesions quickly, accurately, and objectively would be an invaluable support tool for radiologists and doctors, enhancing decision-making and improving patient outcomes by providing consistent and reliable analyses.

Within biomedical applications of deep learning, medical image segmentation, the process of partitioning medical images into distinct regions or objects, plays a crucial role in various clinical applications. Accurate and efficient segmentation enables precise quantification of anatomical structures, localisation and characterisation of lesions, and assessment of disease progression [156]. It facilitates computer-aided diagnosis, surgical planning, radiotherapy treatment planning, and quantitative analysis of medical images for research purposes [85]. By automating the segmentation process, deep learning models have the potential to significantly improve the efficiency and accuracy of these critical tasks, leading to better patient care and improved clinical outcomes.

Despite the promising advancements deep learning has brought to medical image segmentation, its widespread adoption remains limited due to several distinct, yet overlapping, obstacles. One of the foremost challenges is the limited availability of labelled medical datasets [156]. High-quality, annotated medical data is essential for training deep learning

models, yet obtaining such data is difficult due to the labor-intensive nature of manual annotation, the specialised expertise required, as well as privacy concerns. Furthermore, with medical imaging modalities such as magnetic resonance imaging (MRI), the situation is exacerbated by the cost and length of the imaging procedure [60]. Data scarcity coupled with potentially high variability in unseen data can lead to poorly performing models [181]. For this reason, it is essential to explore strategies to increase the variability and diversity in the labelled datasets which we possess, such that the models trained on them generalise better to unseen data.

Data scarcity not only hampers the ability to develop robust models but also aggravates a domain shift problem, where models trained on data from one source may fail to generalise well to data from different sources [107]. These "domains" can be data which are acquired in different countries, different medical scanners, parameters and pulse sequences, or even data belonging to different age ranges or sexes. The simplest and most effective method of approaching the domain shift problem is by incorporating more and diverse data into the model, thus capturing the different distributions the data may possess. For example, analysing the brain regions in Figure 1.2 (a) can be straightforward for medical experts, however, the variation in intensity and shape of brain structures can vary widely, and can prove troublesome for computer vision algorithms which don't capture a diverse range of brain shape and intensities. In practice, models trained on specific MRI machines drastically suffer in performance when implemented on other machines, as the images can take very different intensity distributions [109], see Figure 1.2 (b) for an example. Therefore, it becomes paramount that models are trained using as much of the available data that there is, in order to reduce domain shift. However, it is common practice to omit relevant datasets from model training in medical image segmentation due to differences in image and label formats. This, inevitably, leads to missed opportunities in improving model robustness. For example, in brain lesion segmentation, some datasets include only a single timepoint scan [44, 108], while others provide scans at multiple timepoints [28]. Furthermore, there can also be variations in the types of labels they offer, such as all or new lesion segmentation. Therefore, it is of paramount interest to develop methods which can adapt to the heterogeneous nature of medical imaging datasets. This will maximise utilisation of information and build well generalising frameworks.



(a) (b)

FIGURE 1.2: **Visualisation of the domain shift problem.** (a) Brain scans from three different MRI scanners. Scanner choice affects resolution of image and can alter intensities of respective brain structures. For example, the gray matter region is significantly brighter on the Siemens Aera scan compared to the rest. (b) Averaged histogram of normalised brain MRI scans from three different MRI scanner models. Intensity distribution of a scan is greatly affected by the choice of the scanner. A model which incorporates images from all domains will improve model generalisability.

Machine learning for medical imaging analysis requires the cross-disciplinary collaboration of two very different disciplines: computer vision and medicine. This unique combination has occasionally led to divergences in the focus of developed methods due to the differing motivations of each discipline. While computer vision researchers often prioritise technical benchmarks, such as improving segmentation accuracy or optimising model efficiency, medical experts are more concerned with the clinical relevance and utility of these methods. For instance, a model that achieves high accuracy on a standard dataset may still fall short if it does not address specific clinical needs, such as the precise delineation of lesion boundaries critical for surgical planning, or the ability to handle the variability found in real-world patient data. Additionally, computer vision models sometimes focus on achieving results in idealised conditions, which may not translate well to the complex environments of clinical practice. Unfortunately, many recent academic works on medical image segmentation focus solely on improving a segmentation metric by a small, non-statistically significant, fraction of a percent. This disparity has caused a lack of focus on clinically relevant machine learning tasks, with much of the research concentrated on technical benchmarks rather than practical, real-world applications. As an example, the majority of brain lesion segmentation datasets, which are of neuro-degenerative nature, contain only single timepoint scans [44, 108, 114, 137, 160], despite the temporal progression of the disease and

its significance to clinical decision-making. Bridging the efforts of machine learning scientists and medical professionals is crucial for overcoming these obstacles and the successful integration of deep learning into routine clinical workflows.

Finally, adoption of machine learning models in clinical settings face resistance partly because many of these models operate as "black boxes", making it difficult for doctors to understand or explain their decision-making processes. Clinicians are understandably hesitant to rely on tools they cannot fully interpret, especially when patient outcomes are at stake [123]. This lack of transparency hinders trust and integration of deep learning models into routine practice, emphasising the need for models that are not only accurate but also interpretable. One method of making deep learning methods approachable to medical professionals is by incorporating anatomical constraints, which align the model predictions with known medical domain knowledge and structures. Furthermore, the lack of transparency can exacerbate issues of artificial intelligence (AI) fairness [147], as biases present in the training data may be implicitly learned and propagated by the model without any clear explanation. This can lead to biased predictions and potentially discriminatory outcomes for certain patient subgroups, further eroding trust and hindering the equitable adoption of AI in healthcare.

1.2 Objectives, Outline of Thesis, and Contributions

This thesis aims to provide solutions in three overarching topics in machine learning for brain lesion segmentation: clinical relevance, data scarcity, and generalisability. We first tackle the problem of the lack of a clinically relevant and overlooked longitudinal segmentation task for multiple sclerosis. As such, we develop a method which utilises longitudinal data for the segmentation of new lesions. Our solution is covered in Chapter 3. The second obstacle is the scarcity of labelled data for training robust brain lesion segmentation models. For that reason, we provide two very different approaches for augmenting data to increase data variability and improve brain lesion segmentation. Our solutions are presented in Chapter 4. Finally, we improve model generalisability by maximising the information that we can use. We approach this by developing a comprehensive brain lesion segmentation model with heterogeneous input capabilities, which can also segment multiple types of brain lesions. Our solution increases model interpretability by incorporating

domain knowledge in multiple sclerosis. It is presented in Chapter 5.

Chapter 2 provides an introduction into the domain of deep learning. A comprehensive description of the most recent deep neural network architectures and implementations is given. We also explain the shortcomings of deep neural networks, as well as common methods of overcoming them. Finally, we provide a brief review on some of the state-of-the-art deep learning methods in medical imaging segmentation and techniques on optimizing these networks for use in the medical imaging domain.

Chapter 3 investigates the lack of use of longitudinal data in multiple sclerosis brain lesion segmentation. We develop a segmentation pipeline in accordance to a 2021 MICCAI challenge for new lesion segmentation. We implement a comprehensive preprocessing stage along with a fully convolutional neural network architecture and imaging and lesion-aware data augmentation to accurately segment new lesions which appear the second timepoint of a longitudinal public dataset. Unlike other challenge participants, our method does not see a trade-off between evaluation metrics, and achieves very promising performance. We provide an in-depth analysis of the evaluation and comparison to other challenge participants as well as state-of-the-art segmentation methods.

Chapter 4 focuses on alleiviating the problem of data scarcity in medical image segmentation. We provide two different approaches for tackling the problem of limited data. First, we propose an adversarial method for generating pseudo-healthy images from pathological images, and unhealthy brain images from healthy images. Our method incorporates an attention mechanism, such that it only alters the intensity of the lesion regions. Second, we provide a fast and effective non-deep learning method, LesionMix, for enlarging training dataset size. Unlike previous disease-specific augmentation techniques, LesionMix augments lesions at the lesion-level, providing diverse shape and intensities of lesions even when there are very few images. We also investigate optimal lesion load distribution for training a segmentation model.

Chapter 5 introduces a novel method for overcoming the difficulty of training unified segmentation models. We propose a 3D segmentation model, which can be trained using a variety of different image and label availabilities, for segmenting new, all, and vanishing multiple sclerosis lesions. We improve our method by implementing anatomical constraints into the model, and by incorporating LesionMix to significantly increase image and label

diversity. Our method demonstrates that notable improvements in segmentation performance can be achieved by allowing heterogeneous data input and anatomically plausible constraints into the training procedure.

Chapter 6 concludes and summarises the contributions of our work. We review our main findings and discuss open challenges and future outlook of the adoption of deep neural networks in medical image analysis.

1.3 Publications

A list of publications and collaborations achieved throughout the PhD, in chronological order:

1. **Basaran, B.D.**, Matthews, P.M. and Bai, W., 2021. MSSEG-2 Challenge, Team New Brain: Cascaded networks for new MS lesion detection. MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure, p.77.

2. **Basaran, B.D.**, Qiao, M., Matthews, P.M. and Bai, W., 2022, September. Subject-Specific Lesion Generation and Pseudo-Healthy Synthesis for Multiple Sclerosis Brain Images. In International Workshop on Simulation and Synthesis in Medical Imaging (pp. 1-11). Cham: Springer International Publishing.

3. Qiao, M., **Basaran, B.D.**, Qiu, H., et al., 2022, September. Generative modelling of the ageing heart with cross-sectional imaging and clinical data. In International Workshop on Statistical Atlases and Computational Models of the Heart (pp. 3-12). Cham: Springer Nature Switzerland.

4. **Basaran, B.D.**, Matthews, P.M. and Bai, W., 2022. New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation. Frontiers in Neuroscience, 16, p.1007453.

5. Eisenmann, M., Reinke, A., Weru, V., et al., 2022. Biomedical image analysis competitions: The state of current participation practice. arXiv preprint arXiv:2212.08568.

6. Zhang, W., **Basaran, B.**, Meng, Q., et al., 2023, October. MoCoSR: Respiratory Motion Correction and Super-Resolution for 3D Abdominal MRI. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 121-131). Cham: Springer Nature Switzerland.

7. **Basaran, B.D.**, Zhang, W., Qiao, M., et al., 2023, October. LesionMix: A Lesion-Level Data Augmentation Method for Medical Image Segmentation. In Data Augmentation, Labelling, and Imperfections. MICCAI 2023. Lecture Notes in Computer Science, vol 14379. Springer, Cham.

8. **Basaran, B.D.**, Zhang, X., Matthews, P.M., Bai, W., 2025, October, SegHeD: Segmentation of Heterogeneous Data for Multiple Sclerosis Lesions with Anatomical Constraints. In International Workshop on Longitudinal Disease Tracking and Modelling with Medical Images and Data (LDTM). Cham: Springer International Publishing.

9. Zhang, X., Ou, N., **Basaran, B.D.**, et al., 2024. A Foundation Model for Brain Lesion Segmentation with Mixture of Modality Experts. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. MICCAI 2024. Lecture Notes in Computer Science, vol 15012. Springer, Cham.

10. **Basaran, B. D.**, Matthews, P. M., & Bai, W. (2024). SegHeD+: Segmentation of Heterogeneous Data for Multiple Sclerosis Lesions with Anatomical Constraints and Lesion-aware Augmentation. arXiv preprint arXiv:2412.10946.

Chapter 2

Background Theory

2.1 Fundamentals of Deep Learning

Deep learning is a subset of machine learning which utilises artificial neural networks with multiple layers. Unlike traditional machine learning methods, which often require manual feature extraction, deep learning models can learn hierarchical representations of data, making them particularly powerful for tasks like image segmentation, speech recognition, natural language processing, and more. In this chapter, we will delve into the fundamentals of deep learning, with a focus on image segmentation. We begin with an introduction to the building blocks of deep neural networks, and cover the different layers and functions involved in training deep neural networks. We will also explore various architectures that have emerged, such as convolutional neural networks and generative adversarial networks, highlighting their applications and their strengths and weaknesses. Finally, we will examine recent advancements in deep learning applications in medical image segmentation.

2.1.1 Deep Neural Networks

Deep Neural Networks (DNNs) are a subset of machine learning models inspired by the structure and function of the human brain. They consist of multiple layers of interconnected artificial neurons, where each layer transforms the input data through learned weights and biases. The use of multiple layers along with non-linear functions allow DNNs to model complex, hierarchical patterns in data. In this subsection, we start by defining the building blocks of DNNs. We then introduce state-of-the-art and recent deep neural networks, discuss limitations of deep networks, and techniques for overcoming these challenges.

2.1.1.1 Building Blocks of Deep Neural Networks for Image Segmentation

To better understand how deep neural networks work for image segmentation tasks, we start from defining and explaining the core elements which build deep networks. Here, we introduce fully connected layers, convolutional layers, activation layers, pooling layers, and the notion of receptive fields.

2.1.1.1.1 Fully Connected Layers

Fully connected layers, also known as dense layers, connect each neuron to every neuron from the previous layer. These layers compute a dot product between a weight matrix and the output from a previous layer plus a bias term. This dense connectivity allows the network to integrate features extracted by convolutional and pooling layers, facilitating tasks such as classification or providing latent representations.

Latent representations are high-dimensional feature vectors that capture the underlying structure or essential features of the input data in a compact and abstract form. In the context of fully connected layers, these representations are learned as the network maps raw input data (e.g., an image) to a more meaningful and task-specific feature space. For instance, in an image classification task, the latent representation might encapsulate abstract concepts like shapes, textures, or object types that are crucial for distinguishing between classes.

However, fully connected layers lack inherent invariances, such as translation invariance, which are present in convolutional layers (see next section). This means that a fully connected layer treats all input features independently, without considering their spatial or temporal relationships. For example, if an input image is shifted slightly, the outputs of a fully connected layer will change significantly unless the model has explicitly learned to account for such shifts during training. In contrast, convolutional layers leverage inductive biases, such as weight sharing and locality, which make them naturally suited to handle structured data like images. Despite this limitation, fully connected layers remain crucial for tasks that require the integration of high-level features learned from earlier layers or when working with data without clear spatial or temporal structure.

2.1.1.1.2 Convolutional Layers

Convolutional layers are one of the core elementary units in deep neural networks. A convolutional layer consists of a set of learnable filters (also known as kernels), which are used to extract features from the input data. Each filter convolves (slides) across the image, performing element-wise multiplication and summing the results to produce a feature map, also known as activation maps. This process is repeated for multiple filters, capturing various features such as edges, textures, and patterns.

The convolution operation works as follows. Given a 2D input image, $\mathbf{I}$, of size $H_{\text{in}} \times W_{\text{in}}$, and a 2D convolutional with a filter, $\mathbf{K}$, of size $N \times N$, the feature response, $\mathbf{O}$, at a pixel position i, j can be expressed as

$$\mathbf{O}[i,j] = (\mathbf{I} * \mathbf{K})[i,j] = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \mathbf{I}[i+m, j+n] \cdot \mathbf{K}[m,n], \tag{2.1}$$

where $*$ is the convolution operation, and m, n are the positions of the weights in the filter. Here, $\mathbf{I}[i+m, j+n]$ refers to the pixel value in the input image that the filter is currently covering. The size of the output feature map, $H_{\text{out}} \times W_{\text{out}}$, is given by

$$H_{\text{out}} = \left\lfloor \frac{H_{\text{in}} - N + 2P}{S} \right\rfloor + 1 \qquad W_{\text{out}} = \left\lfloor \frac{W_{\text{in}} - N + 2P}{S} \right\rfloor + 1, \tag{2.2}$$

where H_{in} is the height of the input image, W_{in} is the width of the input image, P is the amount padding applied to the input image before convolution, and S is the stride. Padding is required to control the spatial dimensions of the output feature map and can be performed in varying techniques, most common being zero-padding. The stride determines by how much the filter moves over the image per operation. A stride of 1, where the filter moves one pixel at a time, is the most common value. Thus, the convolution operation over the whole image can be expressed as

$$\mathbf{O} = \sum_{i=0}^{H_{\text{out}}-1} \sum_{j=0}^{W_{\text{out}}-1} \left(\sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \mathbf{I}[i \cdot S + m, j \cdot S + n] \cdot \mathbf{K}[m,n] \right). \tag{2.3}$$

This operation is repeated for however many kernels and layers the neural network may have. A visual example of the convolution operation is shown in Figure 2.1.

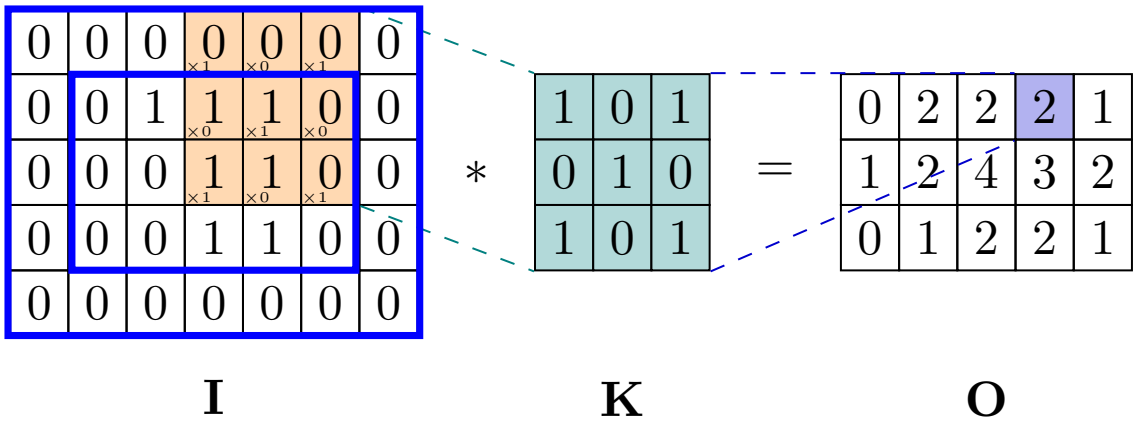


FIGURE 2.1: **Illustration of a convolution operation**. A 3x3 kernel, **K**, is applied to a 3x5 input image, **I**, to produce a feature map, **O**. The cells between the blue borders in **I** are the introduced zero-padding, which results in an output with the same dimensions as the input. In this example, $P = 1$ and $S = 1$.

2.1.1.1.3 Activation Layers

Real-world data and the relationships between their features are often non-linear. Activation layers contain functions which introduce non-linearity to neural network models. Without non-linearity, neural networks would be equivalent to single-layer linear models, and unable to solve complex tasks. Among the many activation functions, the most popular is the Rectified Linear Unit (ReLU) which was introduced by Kunihiko Fukushima in 1969 for visual feature extraction in hierarchical neural networks [68]. ReLU applies a non-saturating activation function,

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

where x is the output from a convolutional layer. ReLU removes negative values from an activation map by setting them to zero, while retaining positive values. Other common activation functions include the sigmoid, hyperbolic tangent, and leaky ReLU [5]. Activation layers are prevalent in many deep neural network models, such as transformer networks, fully connected networks, autoencoders, and GANs.

2.1.1.1.4 Pooling Layers

Pooling layers are responsible for reducing the spatial dimensions of the feature maps generated by convolutional and activation layers. This is achieved by dividing the input data into non-overlapping tiles (regions), and combining the values obtained from each region

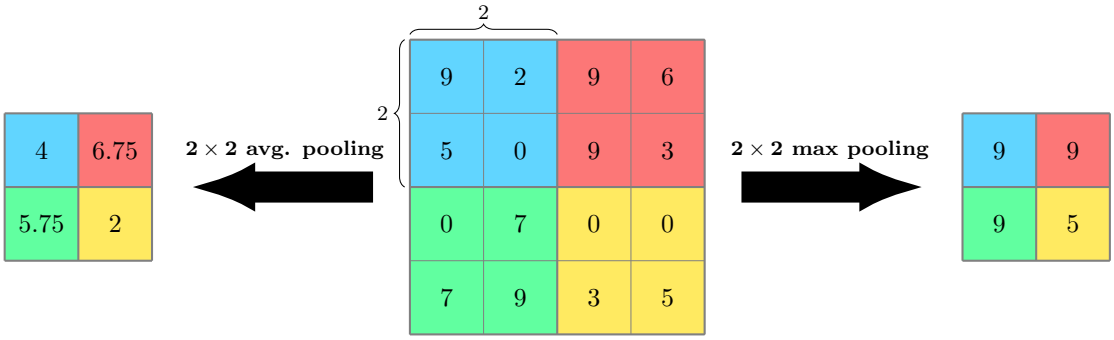


FIGURE 2.2: **Visualization of common pooling operations**. 2x2 average pooling and max pooling operations, where average pooling calculates the mean value, and max pooling selects the highest value within each region, both reducing the spatial dimensions while retaining key features.

to create a downsampled representation of the input data. Aside from dimensionality reduction, pooling also helps in extracting dominant features, and provides local translation invariance by ensuring small spatial shifts in the input do not significantly affect the output.

There are two prominent types of pooling operations: max pooling and average pooling, typically using tiling sizes of 2×2. An example is provided in Figure 2.2. Max pooling selects the maximum value within a defined window. In contrast, average pooling computes the mean value within the window, resulting in a smoother representation of the feature map. Several efforts have gone into developing more advanced pooling techniques, such as strided pooling, spatial pyramid pooling, and others [186].

2.1.1.1.5 Receptive Field

Receptive field refers to the region of the input space (such as a region of an image) that a particular neuron in a network is responsive to, i.e. how far and large of a region a neuron is looking to decide what its label will be. In a convolutional network, each neuron in a convolutional layer is connected to a localised region of the previous layer, termed the neuron's receptive field, which is typically a small, square area of 3×3 neurons. A large receptive field suggests that the neuron also takes into account information from more global features. Therefore, it might be intuitive to have a large convolutional kernel of 7×7 or 9×9, as it would have a larger receptive field than a 3×3 kernel. However, large kernels drastically increase the amount of computations made per output. Furthermore, we can achieve the same receptive field by stacking multiple convolutional layers. For example, in 2D convolution,

a large 7x7 kernel can be effectively replaced by three successive 3x3 convolutional layers, all while effectively reducing the number parameters nearly by half ($(3 \times 3^2)/7^2 \approx 0.55$). Stacking convolutional layers in 3D when dealing with 3D images results in reducing the parameters by approximately 25%!

As convolutional layers stack and the network becomes deeper, the receptive field of neurons increases. This allows the network to integrate information from larger areas of the input image and capture more complex features, such as shapes or objects, rather than just edges or textures. Understanding and calculating the receptive field is crucial for designing networks, particularly for tasks requiring spatial context and hierarchical feature extraction.

2.1.1.2 Convolutional Neural Networks

The development of convolutional neural networks (CNNs) was significantly influenced by Kunihiko Fukushima's work in 1980 on the Neocognitron, a deep neural network model inspired by the structure and function of the visual cortex in the human brain [69]. The Neocognitron was designed to recognise patterns through a hierarchical structure of layers, similar to how the visual system processes information in our brains. Fukushima's Neocognitron laid the groundwork for CNNs by demonstrating the effectiveness of hierarchical feature extraction and local receptive fields. Soon after, LeCun et al. developed models using neural networks for zip code and number classification, thus extending CNNs to image classification [112].

In a CNN, the input layer receives raw data, typically an image, which is then passed through a series of convolutional layers that apply various filters to extract spatial features such as edges, textures, and shapes. These convolutional layers are often followed by activation layers, such as ReLU, to introduce non-linearity, and pooling layers, like max pooling, to reduce the spatial dimensions and provide slight shift invariance while also retaining important information. As the data progresses through these layers, the feature maps become increasingly complex, capturing higher-level patterns. The output from the final layers is typically flattened and fed into fully connected layers, which act as a classifier or regressor. Finally, a softmax activation function normalises the outputs, and provides the predictions as class probabilities for classification tasks. A visual example of a CNN architecture for a classification task is shown in Figure 2.3.

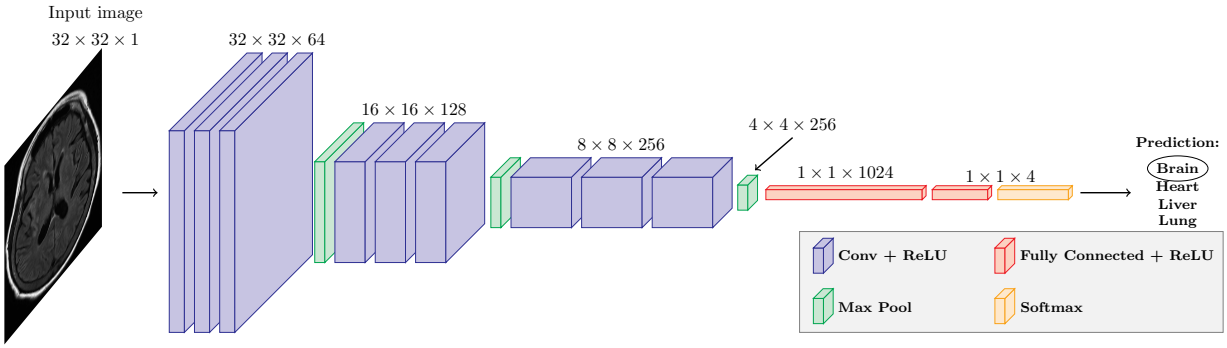


FIGURE 2.3: **A generic convolutional neural network framework for medical image classification**. The input image is processed through a series of convolutional, activation, and pooling layers to extract features, followed by fully connected layers that classify the image into its category. The dimensions above the layers denote the height, width, and number of feature maps of the image. (Conv=Convolution)

CNNs can be applied to image segmentation tasks with modifications to the input and the architecture mentioned above. A naive way for CNN-based image segmentation is to implement a patch-based approach, where we divide the input image into patches and individually pass the patches through the network, we can obtain class-labels for the centre pixel of each patch. Repeating this process for all patches yields a pixel-wise classification, or segmentation. A clear disadvantage of this approach is that it requires every patch to be passed through the network, making it inefficient and computationally expensive. It would be much more preferable to have network which could output pixel-wise labels without needing a patch-based approach, like fully convolutional networks, explained in the following section.

2.1.1.3 Fully Convolutional Networks

Fully Convolutional Networks (FCNs) were initially proposed by Long et al. [124] for the task of semantic image segmentation. Unlike CNNs, which utilise fully connected layers for classification purposes, FCNs use convolutions to output a prediction heatmap for the input image. This structure allows FCNs to generate outputs of the same spatial dimensions to the input image, making them suitable for tasks requiring pixel-wise predictions. An FCN typically adopts an encoder-decoder architecture. The encoder compresses the input image into a high-level feature representation through a series of convolutional, activation, and pooling layers (similar to CNNs), while the decoder reconstructs the spatial details through convolutional and activation layers, and upsampling operations.

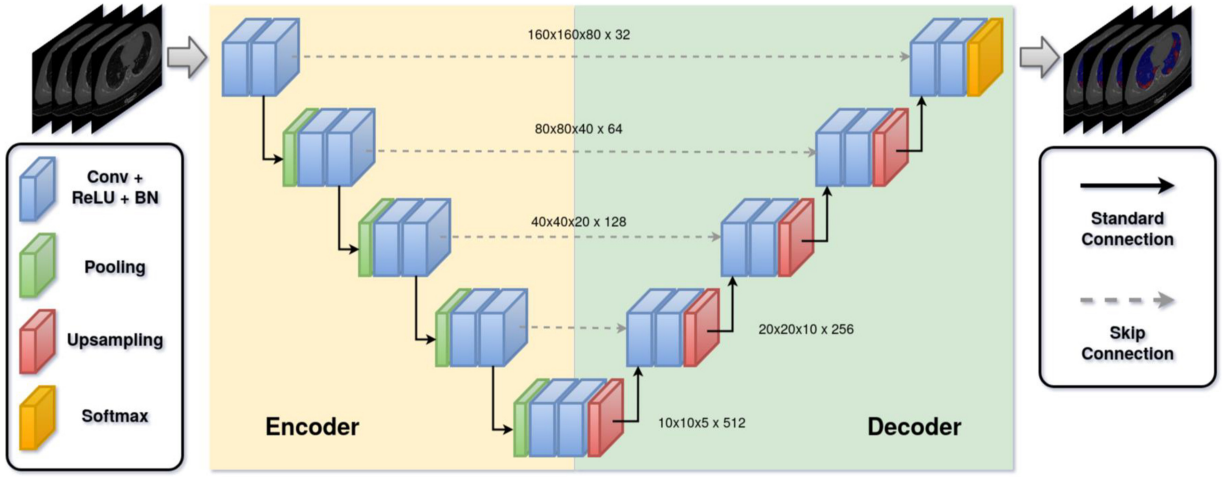


FIGURE 2.4: **Example 3D U-Net architecture for liver and tumor segmentation**. The framework features the characteristic encoder-decoder structure with skip connections. The skip connections link corresponding layers in the encoder and decoder paths, enabling the model to combine detailed spatial information from the encoder path with the upsampling layers, improving segmentation accuracy for both liver and tumor regions. Image source [138].

In contrast to CNNs, which require a patch-based approach for segmentation, FCNs can operate on entire images, simplifying the image analysis computational process and improving efficiency. However, a basic encoder-decoder FCN may face challenges in preserving fine details and contextual information due to the reduction in spatial resolution caused by pooling layers. Various enhancements have been proposed to address this shortcoming. The U-Net architecture [148], introduces progressive upsampling and skip connections to the FCN, which link the encoder and decoder layers at different resolution levels and help in retaining and recovering spatial context that might otherwise be lost, thus leading to more accurate segmentation results. Extensions of the U-Net architecture, such as the 3D U-net [41] and the 3D V-net [131], have become widely adopted in medical image segmentation, demonstrating substantial improvements in segmentation accuracy by capturing both detailed features and broader contextual information. An example 3D U-Net with skip connections for liver and liver tumor segmentation is provided in Figure 2.4.

2.1.1.4 Autoencoders

Autoencoders (AEs) originate from work in the 1980s for dimensionality reduction, and data reconstruction and compression [151]. Since, they have been adapted to image denoising, image reconstruction, image segmentation, anomaly detection, and more [188]. Autoencoders consist of two main parts: the encoder, which compresses the input image into a

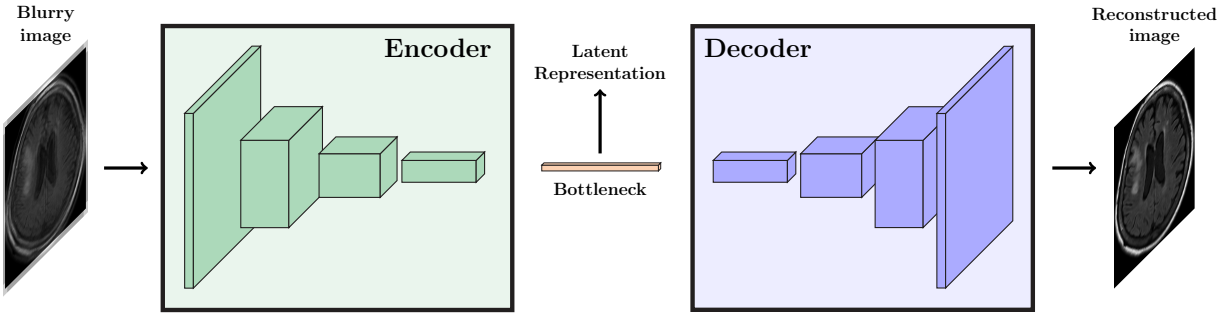


FIGURE 2.5: **A generic autoencoder structure**. The autoencoder processes a blurry and poor quality image, where the encoder compresses the input into a low-dimensional latent representation, and the decoder reconstructs a clean, deblurred and denoised version of the image.

latent representation, also known as *bottleneck*, and the decoder, which reconstructs the image from this representation. A visualisation is provided in Figure 2.5.

Modern AEs for image analysis tasks are commonly composed of convolutional layers, pooling layers, activation layers, and fully connected layers. The convolutional layers are able to better capture spatial hierarchies and local patterns, and the fully connected layers reduce to input to a latent representation at the network bottleneck. The encoder reduces the image to a lower-dimensional feature space, capturing essential structures, while the decoder upsamples these features to produce a pixel-wise segmentation map. Upsampling can be performed using transposed convolutions [59], such as 3×3 transposed convolutional kernels with a stride of 2, which effectively double the size of the feature maps. Alternative methods for upsampling include max-unpooling and bilinear or trilinear interpolation.

However, AEs can be time-consuming and computationally expensive to train, as well as be prone to overfitting, especially in low-data settings [116]. Furthermore, choosing an appropriate size for the bottleneck layer is crucial for two reasons: (i) an excessively large bottleneck can result in redundant dimensions and increased computational cost, while (ii) an overly small bottleneck can cause significant information loss [80]. Despite these limitations, autoencoders remain a valuable tool in reducing dimensionality through latent representations, especially when dealing with large unlabelled 3D imaging data [10].

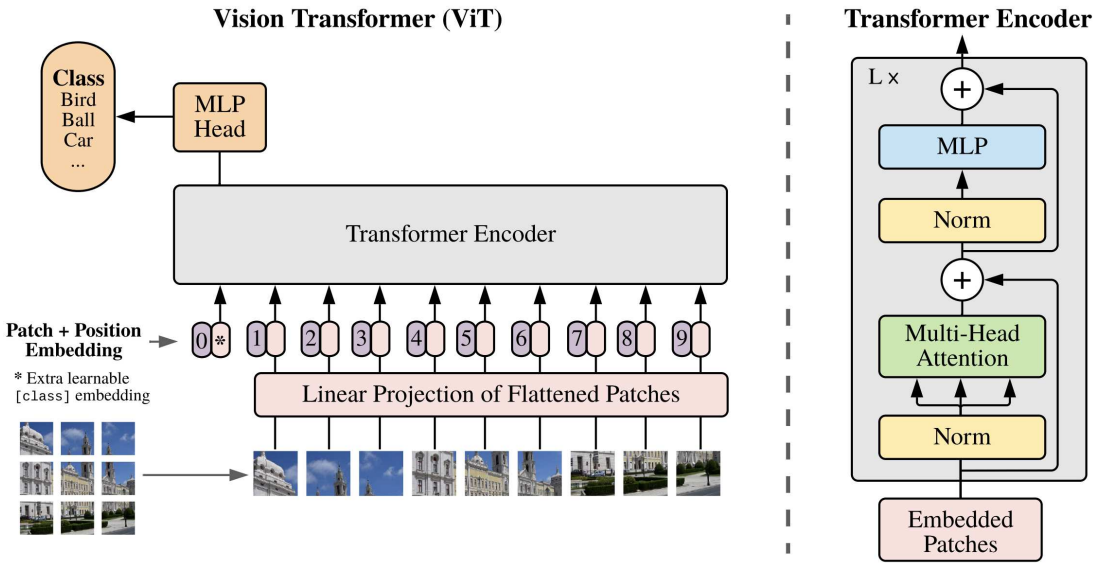


FIGURE 2.6: **The Vision Transformer model.** An input image is divided into fixed-size patches, which are then linearly embedded. Positional embeddings are added to to store global location of patches.The resulting sequence of vectors is fed into a standard Transformer encoder. The encoder features normalisation, multi-head attention, and multi-layer perception layers. Image source [57].

2.1.1.5 Transformers

Transformer networks were originally introduced in Vaswani et al.'s *Attention is All You Need* paper in 2017 for natural language processing (NLP) purposes [171]. In NLP tasks, transformers leverage self-attention mechanisms to capture long-range dependencies within text, enabling better understanding between word relationships. In text data, words or subwords are tokenized into quantitative representations, termed tokens. Each token is then mapped to a learned vector in a fixed-dimensional space using an embedding matrix. The embedding matrix captures information about the token's meaning in the context of the training data. Soon after application of transformers on text data, they were extended to images using the Vision Transformer (ViT) by dividing images into smaller patches and treating each patch as a token, similar to how words are treated in NLP tasks [57]. Transformers are utilised in architectures for image classification, segmentation, and generative models [104]. The ViT model is presented in Figure 2.6.

The workflow of a transformer network for image classification or segmentation starts with the input image being divided into a sequence of patches. Each patch is embedded into a fixed-dimensional vector, and coupled with positional embeddings which provide information about their positions in the image. This embedding is then passed through the

transformer layers which have multiple stacks of transformer encoders. The self-attention mechanism within these layers enables the network to model relationships between different parts of the image, effectively capturing both local and global features. Finally, a multi-layer perceptron (MLP) classification head can be used for image classification, or a segmentation map can be output using upsampling and deconvolutional layers.

The Vision Transformer has since been improved. For example, the Swin Transformer decreases computational complexity and introduced hierarchical structure by limiting self-attention computation to non-overlapping local windows during patch formation [122]. On the other hand, MaskFormer provides a unified framework for image segmentation which combines mask prediction and classification in a single model through a pixel and transformer decoder [38]. Transformers architectures have shown several advantages over FCNs. The self-attention mechanisms enables transformers to excel at capturing long-range dependencies and contextual information, which can lead to improved accuracy in tasks where understanding the global context is crucial [83].

2.1.1.6 Generative Adversarial Networks

Generative Adversarial Networks (GANs) were introduced by Goodfellow et al. in 2014 for the purpose of generating images. Since then it has seen substantial adoption in natural and medical imaging tasks due to its wide range of applications [77]. As shown in Figure 2.7, a GAN consists of two neural networks: a generator network, which creates synthetic images, and a discriminator network, which distinguishes between real and synthetic images. During training, the generator learns to produce increasingly realistic images by attempting to fool the discriminator, while the discriminator improves its ability to detect the fake images synthesised by the generator.

GANs are used in a multitude of techniques for assisting in image segmentation. One common method is for augmenting datasets where acquiring large labelled data is challenging. By leveraging the adversarial training process, GANs can learn the underlying distribution of a given dataset and produce new, high-quality images that closely resemble real samples. These generated samples can then be incorporated into training datasets, increasing image diversity, thus improving model generalisability and performance [14, 23, 98].

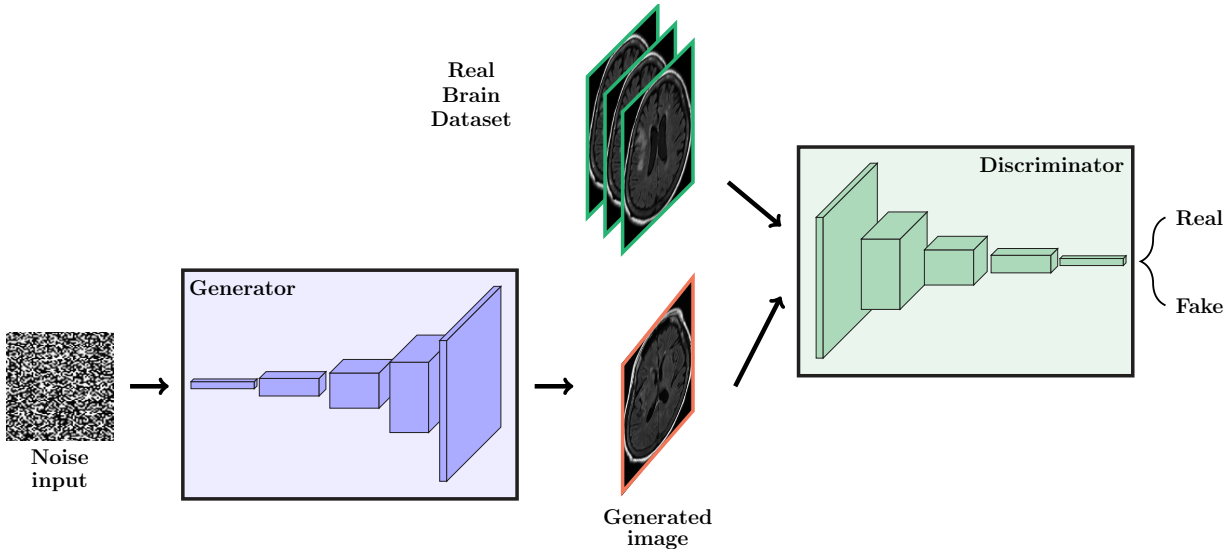


FIGURE 2.7: **A generic generative adversarial network for brain image synthesis**. A generator produces images from noisy input, attempting to fool a discriminator into classifying them as real. Meanwhile, the discriminator is iteratively improving its ability to distinguish between real and fake images. This relationship between the generator and discriminator leads to network to iteratively generate more realistic images.

Another utilised method is pairing a discriminator with a segmentation network to distinguish generated segmentation maps from ground truth maps. The segmentation network generates pixel-wise maps, while the discriminator evaluates these maps against real annotations, providing feedback on how to improve the segmentation networks performance. This adversarial setup encourages the segmentation network to produce more realistic segmentations, as it learns to minimise the discrepancies identified by the discriminator.

However, GANs can be challenging to train due to instability problems [74] and require careful tuning of the adversarial balance. Methods such as Wasserstein GANs [6] have improved on the issue of mode collapse by enforcing smoother gradients and better convergence, however, the problem is still not fully solved for complex data. Additionally, GANs demand substantial computational resources for 3D image data and large datasets to achieve optimal performance. Despite these challenges, GANs offer a powerful approach for generating realistic images and segmentation maps, contributing significantly to recent advancements in image segmentation techniques.

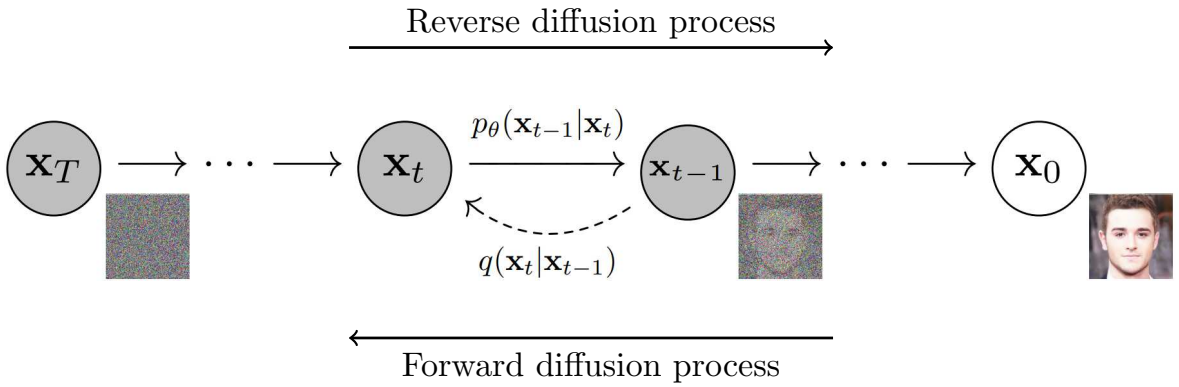


FIGURE 2.8: **Illustration of the denoising diffusion probabilistic model.** $x_1, ..., x_T$ are latents of the same dimensionality as the data $x_0 \sim q(x_0)$. The forward process in a Markov chain, where the output of each step is used as the input for the next step, and involves gradually corrupting the image by adding small amounts of Gaussian noise. In the reverse process p_θ, a Markov chain in the opposite direction, the model learns to reverse the noise added during the forward process. p_θ is approximated by minimizing a variational lower bound on the negative log-likelihood of the data. Image source [87].

2.1.1.7 Diffusion Models

Diffusion models have made a significant contribution in the sphere of generative methods. Inspired by non-equilibrium thermodynamics, the 2020 paper *Denoising Diffusion Probabilistic Models* (DDPMs) introduced a process to produce high-quality synthetic images [87]. The fundamental idea behind DDPMs, and their variants [50], is to learn a generative process that reverses a fixed forward diffusion process which gradually corrupts the data. The *forward diffusion process* is a fixed Markovian process where the image is gradually corrupted by adding Gaussian noise over a series of timesteps. Once the image is corrupted, the *reverse diffusion process*, also known as the generative process, learns how to gradually denoise a sample starting from noise to the original image. The diffusion process is shown in Figure 2.8.

Diffusion models have shown great contribution in various imaging application. Some areas which they excel in include image denoising, super-resolution, synthetic image generation for data augmentation purposes, image-to-image translation, and image segmentation [50].

While diffusion models and GANs are both considered generative models, they each

have their pros and cons. Unlike GANs, diffusion models do not suffer from the mode collapse problem [74], and are much more stable during training. On the other hand, the iterative nature of the diffusion process leads to diffusion models being highly computationally intensive, and thus slower to train and infer from, especially in 3D data.

2.1.2 Training Neural Networks

Here, we cover the essential ingredients for training a segmentation model in a supervised training setting, where images are provided with their respective pixel/voxel-wise labels. A well-trained neural network will have minimal error when outputting a prediction for a task. This process involves selecting a loss function to calculate the error between the model prediction and the ground truth, and an optimizer to adjust network parameters according to the loss calculated. The problem can be posed as

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(\mathbf{x},\mathbf{y})} \mathcal{L}[\mathbf{y}, f(\mathbf{x}, \theta)], \tag{2.5}$$

where θ^* are the optimal parameters of the network, $\mathcal{L}$ is the selected loss function, $\mathbf{y}$ is the ground truth label, function $f()$ is the neural network, $\mathbf{x}$ is the input image which we want to predict from, and θ are the parameters of the network.

Here, we first introduce loss functions and optimizers, followed by the backpropagation algorithm.

2.1.2.1 Loss functions

The objective of training a machine learning model is to find the values of the network parameters which minimise the loss function. Loss functions quantify the difference between the predicted outputs of the neural network and their true values. Commonly used loss functions for image analysis include mean squared error (MSE) loss, cross-entropy loss, Dice loss, and weighted cross-entropy loss. The MSE loss, typically used for regression tasks, is defined as

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{y}_n - \hat{\mathbf{y}}_n)^2, \tag{2.6}$$

where N is the total number of samples in the dataset, $\mathbf{y}_n$ is the vector of target values, and $\hat{\mathbf{y}}_n$ is the vector of predicted values. One of the most common losses for classification and

segmentation tasks is the cross-entropy loss, which is calculated using each pixels' probability of it belonging in given class. The cross-entropy loss for multi-class classification is given by

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\mathbf{y}_{n,c}\cdot\log(\mathbf{p}_{n,c}), \tag{2.7}$$

where C is the total number of classes, $\mathbf{y}_{n,c}$ is the ground truth one-hot segmentation map, and $\mathbf{p}_{n,c}$ is the predicted *probabilistic* output from the network for image n for each class c. A one-hot segmentation map is a pixel-wise representation where each pixel is encoded as a binary value with a single "1" indicating its class and "0"s elsewhere.

A frequent issue in biomedical image segmentation tasks is the class imbalance problem. Image patches which are used for training neural networks do not have the same ratio of healthy to unhealthy tissue. Often times, the unhealthy tissue represents a small percentage of pixels/voxels in the patch. This imbalance biases the network towards more seen classes, leading them to perform poorly on the lesser seen unhealthy tissue class. A go-to loss in medical imaging and segmentation tasks which alleviates this problem to some extent is the Dice loss [27], which is derived from the Dice Similarity Coefficient (see Equation 2.16). The Dice loss measures the overlap between predicted and ground truth segmentations. There are two variants of Dice loss [92], one which employs squared terms in the denominator [131], expressed as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{1}{N}\sum_{n=1}^{N}\frac{2\sum_{c=1}^{C}\mathbf{y}_{n,c}\cdot\hat{\mathbf{y}}_{n,c}}{\sum_{c=1}^{C}\mathbf{y}_{n,c}+\sum_{c=1}^{C}\hat{\mathbf{y}}_{n,c}}. \tag{2.8}$$

and a variant which doesn't [58]. Similarly to $\mathcal{L}_{CE}$, soft-Dice loss operates on predicted probabilities rather than binary predictions, and can show higher sensitivity for segmenting smaller structures [75]. It is characterised as:

$$\mathcal{L}_{\text{soft-Dice}} = 1 - \frac{1}{N}\sum_{n=1}^{N}\frac{2\sum_{c=1}^{C}\mathbf{y}_{n,c}\cdot\mathbf{p}_{n,c}}{\sum_{c=1}^{C}\mathbf{y}_{n,c}+\sum_{c=1}^{C}\mathbf{p}_{n,c}}. \tag{2.9}$$

Several other losses have been proposed as a solution to the class-imbalance problem, such as the weighted cross-entropy loss, weighted soft-Dice loss, and focal loss [166]. These losses assign different weights to different classes to help the model focus more on minority classes.

The weighted soft-Dice, and weighted cross-entropy losses are formulated as

$$\mathcal{L}_{\text{weighted soft-Dice}} = 1 - \frac{1}{N}\sum_{n=1}^{N}\frac{2\sum_{c=1}^{C} w_c \cdot \mathbf{y}_{n,c} \cdot \mathbf{p}_{n,c}}{\sum_{c=1}^{C} w_c \cdot (\mathbf{y}_{n,c} + \mathbf{p}_{n,c})}, \tag{2.10}$$

$$\mathcal{L}_{\text{weighted-CE}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c \cdot \mathbf{y}_{n,c} \cdot \log(\mathbf{p}_{n,c}), \tag{2.11}$$

$$\mathcal{L}_{\text{Focal}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c \cdot (1 - \mathbf{p}_{n,c})^{\gamma} \cdot \mathbf{y}_{n,c} \cdot \log(\mathbf{p}_{n,c}), \tag{2.12}$$

where w_c is the weight for class c, and γ is a focusing parameter which controls the strength of focusing on hard-to-classify examples (a higher γ value increases the focus on less prevalent classes). Similarly, the weight, w_c, is set higher for a less prevalent class than more seen classes, allowing them to have a greater influence on the loss calculation.

2.1.2.2 Optimization of Parameters

A good neural network model should have little to no error when inferring for a task. How you go about reducing the error in a model is a crucial consideration, and is where choosing the right optimizer comes into play. Optimizers adjust the models weights and biases to minimise the loss function, thereby improving the networks performance. As opposed to classical optimization problems, neural networks cannot optimize the full loss computed over the full dataset due to memory constraints, which is why stochastic approaches that compute gradients using only a subset of data points are often used [150]. Various optimization algorithms exist, each with unique mechanisms and advantages. For instance, stochastic gradient descent (SGD) updates parameters using the gradient of the loss function with respect to each parameter,

$$\theta_{t+1} = \theta_t - \eta\frac{\partial \mathcal{L}}{\partial \theta_t}, \tag{2.13}$$

where θ_t is the parameter at step t, η is the learning rate, a value that defines the size of the update, and $\frac{\partial \mathcal{L}}{\partial \theta_t}$ is the gradient of the loss function $\mathcal{L}$ with respect to the parameter θ at step t.

However, SGD can be slow to converge and sensitive to the choice of learning rate. Many

adaptive and faster converging optimizers have been proposed [162]. One widely used optimizer is Adaptive Moment Estimation (Adam), which adjusts learning rates dynamically based on bias-corrected first and second moments of the gradients, leading to faster convergence and better performance on complex tasks [105].

2.1.2.3 Backpropagation

Backpropagation is a gradient estimation algorithm fundamental to the training process of neural networks. The process of iteratively updating network parameters using loss functions and optimizers is termed backpropagation. The process involves calculating the gradient of the loss function with respect to each parameter in the network, and updating the parameters to reduce the error. The key steps of the backpropagation algorithm can be summarised as follows:

1. **Forward Pass and Evaluation**: Input data, $\mathbf{x}$, is passed through the network, layer by layer, until the output, $f(\mathbf{x}, \theta)$, is generated. The output is compared to the ground truth to compute the loss, $\mathcal{L}[y, f(\mathbf{x}, \theta)]$, using a predefined loss function, as in 2.1.2.1.

2. **Backward Pass (Backpropagation)**: Propagate the error back through the network, layer by layer. Compute the gradient of the loss, $\frac{\delta \mathcal{L}}{\delta \theta_t}$, one layer at a time and iterate backward from the last layer. This process avoids redundant calculations of intermediate terms in the chain rule, significantly improving computational efficiency.

3. **Parameter Update**: Update each parameter using an optimizer, e.g., SGD: $\theta_{t+1} \leftarrow \theta_t - \eta \frac{\partial \mathcal{L}}{\partial \theta_t}$

The backpropagation algorithm is computed and updated for each sample in the dataset. After each sample has been passed, the process is repeated until a certain number of iterations, epochs, or until the loss is not decreasing. However, performing backpropagation can be computationally expensive and time consuming when working with large datasets and deep networks with millions of parameters. To overcome this, it is common to select optimizers which sample data from the dataset to update the parameters, such as batch gradient descent and mini-batch gradient descent [150].

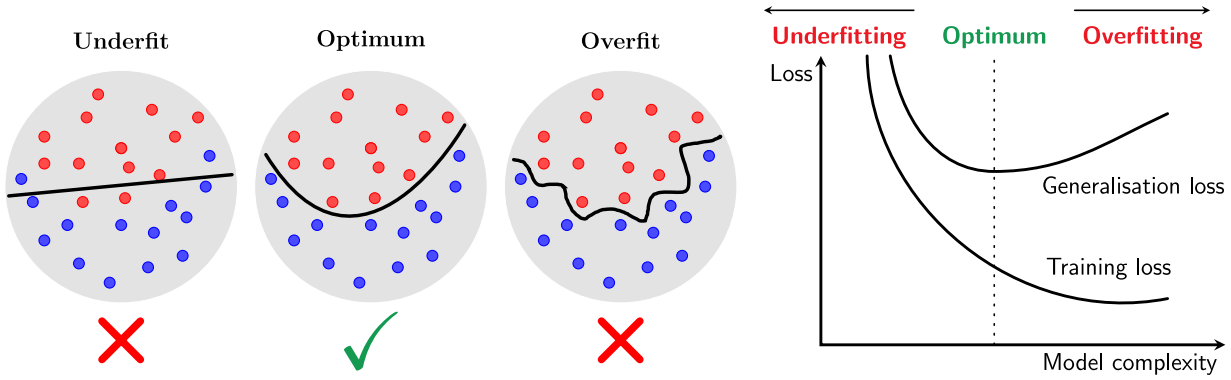


FIGURE 2.9: **Illustration of underfitting, optimal fitting, overfitting, and their relationship to training and generalisation loss**. (a) Decision boundary drawn in black. Underfitting may incorrectly classify samples. Overfitting results in a decision boundry that matches the training samples too closely. Optimal fitting shows the best trade-off, fitting the data well with an optimal complexity. (b) Graphical relation between the three fittings and training and generalisation loss. Low model complexities underfit the data and result in high training and generalisation losses. High model complexities overfit on data with low training loss but high generalisation loss. Optimal model complexity is found when generalisation loss is minimal.

2.1.3 Limitations of Neural Networks in Image Segmentation

Deep learning methods have immensely improved natural and medical image segmentation through more accurate and automated processes, leading to faster diagnostics. Many improvements have been made since the first deployment of these models, however, many challenges still remain. Here, we cover some of the limitations of neural networks.

2.1.3.1 Requirement of Large-scale Labelled Datasets

Neural networks need data to train models. The more data a neural network can expose itself to, the more diversity and variety it can capture and learn from. However, just as importantly, selecting the right degree of model complexity is crucial for representing the underlying distribution of the data correctly. An imbalanced relationship between the data and the model complexity can result underfitting or overfitting of the data, illustrated in Figure 2.9.

Underfitting occurs when a model fails to capture the underlying structure of a dataset due to using a limited model. In such cases, more complex models are required. Overfitting can occur when a model learns from the data too well. This can happen due to too complex

of a model being used, where the model can essentially memorise the training data samples and not adapt to unseen data, resulting in high generalisation loss. Alternatively, overfitting can also happen due to too small of a training dataset being used, with the model, again, fitting too well to the training dataset which has limited diversity. The optimal point for a machine learning model is one that generalises well for unseen data, which lies somewhere between underfitting and overfitting.

In medical imaging applications, the scarcity of data often leads to poor generalisation of the model, inadequate performance of unseen data points, and can overfit on the small, available training data. The top performing natural image segmentation models train using millions, sometimes billions, of labelled data [34, 106, 141]. In contrast, labelled medical imaging datasets can be formed of only thousands of images, or even hundreds, depending on the organ and disease [20, 44, 108]. This is due to a multitude of reasons. The process of taking a picture of a car takes seconds and costs nothing, whereas taking a brain MRI scan, for example, can take up to an hour, accompanied with hundreds in medical bills. As importantly, medical images can be in three or four dimensions (temporal scans), which require slice-by-slice annotations or interactive/semi-automated segmentation methods [184], which substantially increase time for accurate delineations. Not to mention that those annotations need to be made by time-limited expert professionals. This reality often leads to medical image segmentation models not generalising well to unseen data. While overfitting is a major problem for instances where a large dataset is not available, modern deep learning is almost immune to overfitting owing to the contributions covered in 2.1.4.

2.1.3.2 Sensitivity to Small Changes

State-of-the-art deep neural networks can exhibit surprising brittleness when faced with nearly imperceptible changes to the human eye. These networks are not only sensitive to the noise in the input images, but also to geometric alterations such as translations (except for convolution-based networks) and rotations of the images. This means that even minute changes in the input can affect final decisions. This sensitivity raises significant concerns about the safety of AI, particularly in safety-critical applications like medical diagnosis.

There are many factors which may influence these small changes in medical imaging. For

example, there are a plethora of parameters to take into consideration when taking a medical scan using magnetic resonance imaging. The strength of an MRI machine, whether it is 1.5, 3, 5, or 7 Tesla, profoundly alters the resolution and detail of the produced image. Furthermore, the lack of consensus on tuning machine parameters for specific diseases [62, 153] results in subtle variations in noise and intensity between identical machines across different hospitals. Hence, deep learning models which are trained using images from specific hospitals or machines exhibit poor generalisability and performance when deployed on images acquired from different sites.

2.1.3.3 Privacy Considerations

Large and diverse datasets are needed for training robust and well-performing machine learning models. In the medical domain, this is often accomplished through voluntary data sharing by data owners and the accumulation of datasets across multiple institutions and countries. The time consuming processes of data anonymisation and removal of personally identifiable information have proven insufficient protection against re-identification attacks [61]. Coupled with the varying medical data protection laws across countries present a difficult hurdle for curating large medical datasets. As a response, privacy-preserving methods, such as federated learning [115], are becoming more popular field of research.

2.1.3.4 Computational and Resource Constraints

Training deep learning models on medical images, especially 3D or 4D scans, requires substantial computational power. State-of-the-art models are trained using multiple GPUs, and have long training times which can range from days to weeks [177]. Rising GPU prices and large dataset sizes make training robust machine learning models a difficult task for independent researches, and often require financial support of institutions.

2.1.4 Overcoming Limitations and Improving Performance

There have been substantial improvements in the robustness and performance of deep neural networks since their inception in imaging tasks. Here, we cover some of the most widely used methods for training and deploying well-generalisable deep neural networks for image segmentation.

2.1.4.1 Regularisation Techniques

The goal of regularisation is to reduce the complexity of machine learning models, encouraging them to learn simpler functions that generalise better to unseen data. Regularisation prevents overfitting on training data, and can be achieved by dropout, or weight regularisation.

2.1.4.1.1 Dropout

Dropout [159] involves randomly omitting, or "dropping out", certain neurons in the neural network during training. A visual example is seen in Figure 2.10. Dropout discourages co-adaptations of feature detectors, where a detector can be useful only when combined with other specific feature detectors. This technique incentivizes the network to learn sparse representations and prevents overfitting. The most common dropout rate used is 50%, such that half of the neurons are not being utilised at any given training iteration.

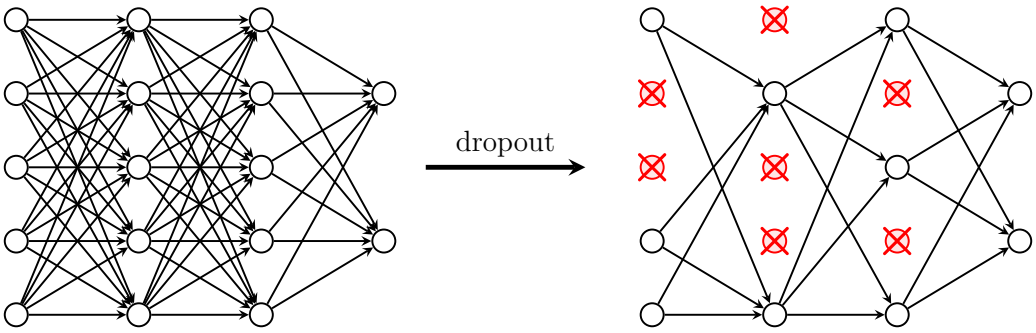


FIGURE 2.10: **Implementation of dropout for regulatisation in a multilayer perceptron network.** After dropout, the red neurons are not utilised in a given training iteration. In this example, 50% of the neurons are omitted.

2.1.4.1.2 Weight Regularisation

Large network weights can lead to unstable networks where small changes in the input can cause large changes in output. Weight regularisation helps keep weights small or zero for less relevant inputs. This helps prevent overfitting of the data and improves the generalisability of the model. Two of the most common weight regularisation methods are L1 and L2 regularisation.

- **L1 regularisation**, also known as Lasso regression (least absolute shrinkage and selection operator), penalises large weights by adding the absolute value of magnitude of the coefficient as a penalty term to the loss function,

$$\mathcal{L} = Error(y, \hat{y}) + \lambda \sum_i |w_i|, \tag{2.14}$$

  where λ is the regularisation parameter. L1 regularisation enforces a sparse model by setting unimportant features' weights to zero, effectively performing feature selection. However, it can be unstable in the presence of highly correlated features.

- **L2 regularisation**, also known as ridge regression, penalises the sum of the squared weights,

$$\mathcal{L} = Error(y, \hat{y}) + \lambda \sum_i w_i^2. \tag{2.15}$$

  In contrast to L1 regularisation, L2 does not force unimportant feature weights to zero, and distributes importance among features.

2.1.4.2 Data Splitting and K-Fold Cross-Validation

Correct data splitting is a simple yet crucial element in making sure your model generalises well. There are three non-overlapping subsets that a dataset is split into: the training set, the validation set, and the test set. The training set is the set that the model learns from throughout training. During training, the model is evaluated on the validation set. The error computed from the validation set, the validation error, serves as an estimate of the error on unseen test data. The model with the lowest validation error is often selected as the optimal model. Finally, the test set is the group of unseen images which the model is ultimately evaluated on. Given a dataset of images, it is common to split it such that 75-80% is used for training set, and the rest is used for the validation set. An illustation of k-fold cross-validation is presented in Figure 2.11.

K-fold cross-validation allows for all of the images to be used for training, rather than the 75-80% stated above. The dataset is split into k smaller sets (folds), *k-1* of them are used for training, and the remaining fold is used for validation. This is repeated k times such that each fold is left out once for validation, resulting in k trained models. The performance

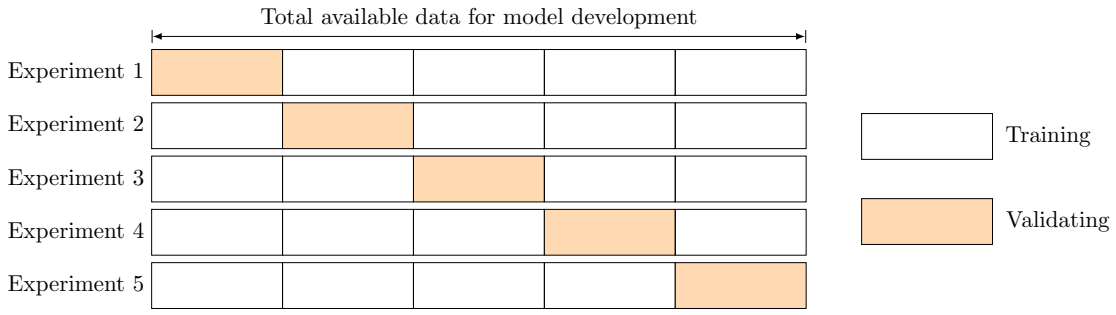


FIGURE 2.11: **Visualisation of k-fold cross-validation implementation.** k=5 in this example.

metric obtained from k-fold cross-validation is then the average of the values across all k models. It is common to use a k of 4 or 5.

2.1.4.3 Ensemble Learning

Ensemble learning [133] is a method of combining multiple trained models in order to improve robustness and better predictive performance compared to using a single model. Ensemble learning can be achieved by combining trained models which use different dataset splits, such as in k-fold cross-validation, using the same model with different hyperparameter settings, utilising multi-head architectures, or even combining predictions from different neural networks architectures all together.

2.1.4.4 Transfer Learning

Transfer learning [174] is a technique which allows knowledge obtained in one task to be reused in another. This often achieved by utilising a pretrained model, and using its learned weights as initialisation to finetune it to the desired task. Transferring information from previously learned tasks to new tasks greatly increase learning efficiency.

2.1.4.5 Data Augmentation

Data augmentation is a strategy for artificially increasing the variety in the training data. This is achieved by applying transformations on the data such that they are different from the original samples. Traditional data augmentation techniques can be separated into three: spatial, colour, and noise. Spatial augmentations include flipping, rotating, scaling, cropping, and elastic deformations of the image. Colour augmentations alter the the brightness,

contrast, and gamma levels of an image. Finally, noise augmentations can be applied by adding or multiplying the image with Gaussian, Rician, or salt and pepper noise.

These augmentation techniques originate from deep learning methods which were designed for natural image segmentation. Lately, there is a growing interest in developing augmentation methods tailored for medical image segmentation. For example, image resampling augmentation increases image diversity by changing the resolution of the input image, mimicking images with lower spatial resolution [93]. Similarly, for MR image augmentation, bias field augmentation simulates the artefacts produced during the imaging process [32]. On the other hand, disease-specific methods have been developed which increase the variability of disease classes [195]. These efforts have been combined with recent developments in generative methods to provide synthetic images and corresponding labels, considerably increasing training data diversity [37, 67].

2.1.4.6 Domain Adaptation

Domain adaptation addresses the critical challenge of applying machine learning models trained on one dataset (source domain) to another dataset with different characteristics (target domain). In the context of brain lesion segmentation, this is highly relevant due to the inherent variability observed across different imaging modalities (e.g., MRI, CT), scanners, acquisition protocols, and patient populations. Traditional deep learning models often exhibit suboptimal performance when confronted with such variations, as they tend to overfit to the specific characteristics of the training data. Domain adaptation techniques aim to mitigate this issue by learning domain-invariant representations or by adapting the model parameters to better suit the target domain. This can be achieved through various approaches, including adversarial domain adaptation, domain confusion, and self-supervised learning [76, 172].

Domain adaptation techniques offer promising solutions to mitigate the impact of domain shifts in brain lesion segmentation. Adversarial domain adaptation, as demonstrated by Tzeng et al., has shown significant potential by training a discriminator to distinguish between source and target domain data while simultaneously training the segmentation model to fool the discriminator [168]. This encourages the model to learn domain-invariant features that are relevant for both source and target domains. On the other hand, domain

confusion aims to minimize the discrepancy between the feature distributions of the source and target domains [169]. Self-supervised learning, as explored in Liu et al., leverages unlabeled data from the target domain to learn domain-invariant representations [121].

2.2 Applications of Deep Learning to Lesion Segmentation in Brain MRI

Here, we provide a literature review of important developments and applications of deep learning in brain lesion segmentation. We focus our review on brain MRI, one of the most commonly used imaging modalities for detailed visualisation of brain substructures and abnormalities. MRI machines use powerful magnets to create a strong magnetic field, aligning protons in the body with this field. A radiofrequency pulse then stimulates the protons, causing them to spin out of equilibrium. When the pulse stops, MRI sensors detect the energy released as the protons realign with the magnetic field. The MRI machine can then differentiate various tissues depending on how quickly they release energy after the pulse is turned off. In comparison to brain computed tomography (CT), MRI provides much superior resolution of anatomical structures, allowing for improved delineation of organ substructures and abnormalities. In the following sections, we specifically cover the application of deep learning for segmenting abnormalities in brain MRI.

2.2.1 Fully Convolutional Networks

Fully convolutional networks (FCNs) are the most widely used deep learning method for medical image segmentation. It gained traction in medical image segmentation in 2015, after the introduction of a 2D fully convolutional network, U-Net [148], aptly named after the models "U" shape, represented in Figure 2.4. Since 2015, the U-Net architecture has been improved and adapted to various tasks, including classification, reconstruction, image synthesis and more [158].

The initial U-Net framework was refined in many ways. One year after its inception, the 3D U-Net was developed, which was more capable of capturing the spatial context of volumes in medical images, and introduced on-the-fly elastic deformations for data augmentation [41]. Milletari et al. introduced the V-Net for 3D medical image segmentation,

proposing residual connections to help alleviate the vanishing gradient problem [131]. Oktay et al. investigated the use of attention gates which learn to focus on particular structures in medical images [140].

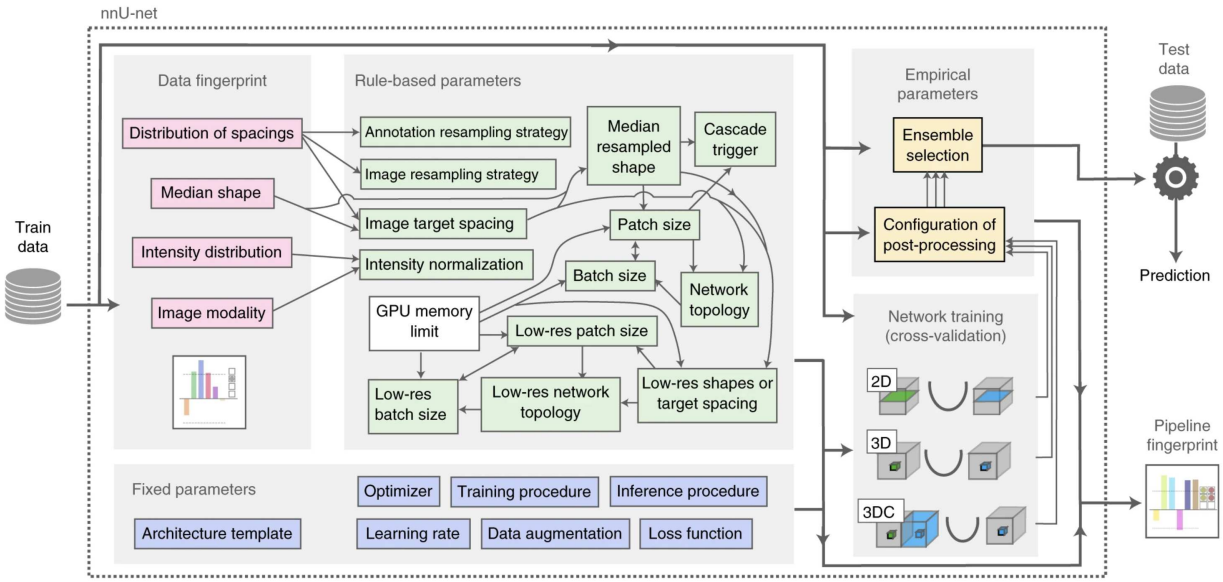


FIGURE 2.12: **The automated method configuration of nnU-Net for biomedical image segmentation**. Given a segmentation task, the framework extracts dataset properties (pink), termed 'dataset fingerprint', and operates rule-based parameters (green) with respect to the dataset fingerprint. This is complemented with fixed, predefined parameters (blue). nnU-Net can train in 3 network configurations with five-fold cross-validation. Finally, the framework can empirally select the optimal ensemble of the trained models for inference and determines whether post-processing is required (yellow). Image source [92].

Currently, one of the most popular and adopted networks for medical image segmentation is the nnU-Net (no-new-net) framework [92], visualised in Figure 2.12. As the authors express, nnU-Net takes "away the superfluous bells and whistles" of recently proposed network designs, and focuses on optimizing the steps of preprocessing, training, inference, and post-processing in accordance to image geometry and GPU resources. Its ease of use, 2D, 3D, and 3D Cascade network configurations, and impressive dataset-agnostic performance has led it to become a basis framework and solid benchmarking method.

2.2.2 Usage of Transformers

It took less than a year for the transformer encoders in Dosovitskiy et al.'s *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* paper [57] to be widely featured in

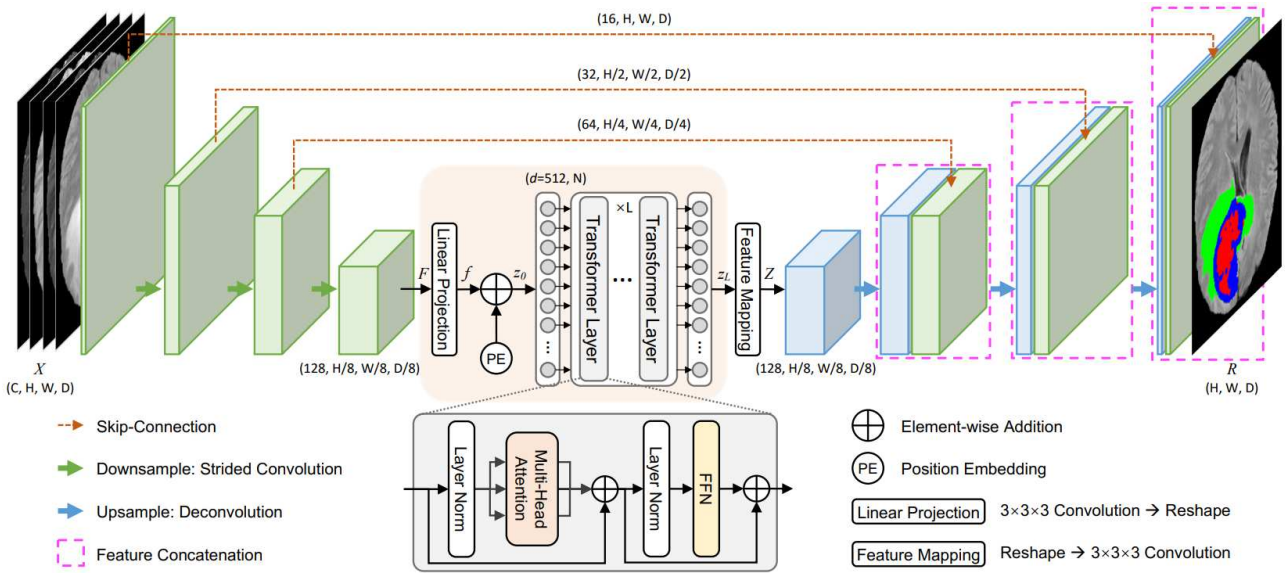


FIGURE 2.13: **The TransBTS framework**. TranBTS features transformer layers and convolutional layers in the encoder, followed by deconvolutional layers to upsample the image and produce a 3D segmentation. The transformer layers are composed of normalisation functions, multi-headed attention, and feed forward networks (FFN). Image source [173].

medical image segmentation applications. The self-attention mechanisms in transformer encoders became a common module in segmentation frameworks, allowing models to weigh the importance between different regions of the image. This ability to focus on the most relevant regions enhances the segmentation of small, critical features found in medical images, such as tumors, lesions, or other pathological structures. Furthermore, the ability to capture long-range dependencies and global context in transformer networks provide a possible solution to the limited receptive field problem of CNNs and FCNs.

Transformers encoders became particularly effective in medical image segmentation networks when they were incorporated in U-Net architectures featuring convolutions. The hybrid approach took advantage of both the local feature extraction strength of CNNs and the global context modelling mechanisms of transformers. Hatamizadeh et al. proposed UNETR (U-NEt TRansformers), a U-Net framework consisting of a stack of transformers as the encoder, and a FCN-based decoder with skip connections [82]. On the other hand, the 2D TransUNet and 3D TransBTS models proposed combining convolutional layers with self-attention mechanisms in the encoder, thus embedding global self-attention [33, 173]. The 3D TransBTS model is visualised in Figure 2.13. Isensee et al., however, have shown that at controlled validation settings transformer networks do *not* outperform CNN networks [94].

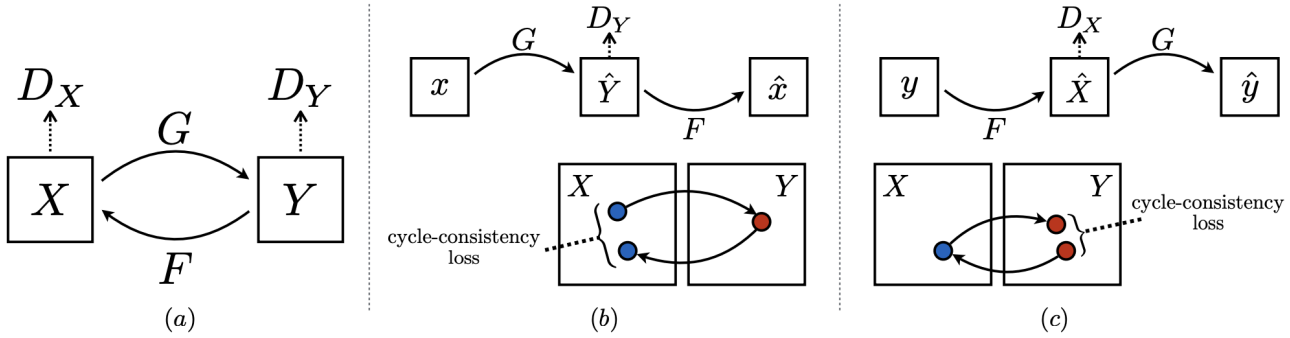


FIGURE 2.14: **Illustration of the CycleGAN model**. (a) The model proposes two mapping functions, $G : X \rightarrow Y$ and $F : Y \leftarrow X$, and two discriminators D_X and D_Y for domains X and Y (for example MRI and CT), respectively. D_Y drives the mapping function G to generate indistinguishable samples from X, while D_X pushes F to generate outputs resembling X from domain Y. The CycleGAN framework is able to iteratively improve translating between two image domains. Two cycle-consistency losses are proposed based on the notion that mapping from one domain to another and back should return the original input. (b) forward cycle-consistency loss: $x \rightarrow G(x) \rightarrow F(G(x) \approx x$ (c) backward cycle-consistency loss: $y \rightarrow F(y) \rightarrow G(F(y) \approx y$. Image source [197].

2.2.3 Usage of Generative Methods

GANs, and more recently diffusion models, have influenced the field of medical image segmentation in a multitude of domains. One frequent application has provided a possible solution to the data scarcity and high annotation cost problem in medical image segmentation. By synthesising medical images from noisy inputs, GANs and diffusion models can drastically increase training dataset size, and improve model generalisability and performance [14, 23, 98].

An alternative approach to improving data diversity and model robustness has been through implementing domain adaptation techniques. Frameworks such as CycleGAN [197], see Figure 2.14, enable image translation from one modality to another (e.g., from MRI to CT), which provides opportunity for expanding dataset sizes.

Another application of generative methods is improving image quality. This can be acheived by denoising (noise reduction) and/or super-resolution (improving resolution). Medical images, such as those acquired from MRI, can be significantly impacted by noise. Denoising and, thus, improving the clarity of obtained images aid in precise diagnoses. On the other hand, Super-resolution GANs (SRGANs) [113] or denoising diffusion probabilistic models (DDPMs) [87] can improve the resolution and quality of images, making it easier for segmentation models to delineate boundaries accurately. Recently, diffusion models have

been introduced for anomaly detection, proving useful in detecting abnormalities such as tumors or lesion in brain scans, see Figure 2.15. However, the majority of generative approaches operate on 2D images rather than 3D due to high computational requirements, making it an obstacle for adoption in 3D medical imaging.

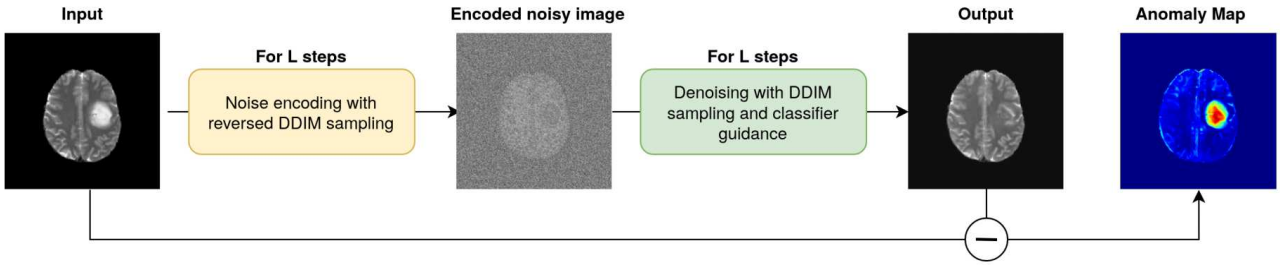


FIGURE 2.15: **Example use of a diffusion model for anomaly detection in brain scans.** A denoising diffusion implicit model with classifier guidance [55] is used to translate a diseased image into a healthy one while preserving overall structure. An anomaly map is produced by subtracting the generated image from the original. Image source [176].

2.2.4 Incorporating Spatial and Temporal Context

The deep learning methods mentioned above have been adapted to medical image segmentation from tasks for natural images, such as animal or car segmentation, where 2D models are implemented. The drawback of using such 2D models for medical image segmentation, where a 3D MRI or CT scan would be segmented slice by slice, is the inability to capture anatomical inter-slice dependencies. It is crucial to accurately segment 3D volumes in medical imaging, as it not only provides volumetric information, but also insight on disease development and progression. Adopting a 3D or 2.5D network for improved spatial consistency in medical image segmentation is now a routine implementation. While 3D networks process entire volumetric data to leverage spatial context across all dimensions, a 2.5D network processes multiple adjacent 2D slices to balance spatial context and computational efficiency.

Diseases are not static. The longitudinal nature of healthcare necessitates periodic scanning and monitoring of patients. In multiple sclerosis, for example, increase of brain lesion volumes and prevalence of newly forming lesions are important biomarkers for disease progression, and are crucial to accurately detect with respect to previous scans [89]. However, methods for detecting growing, shrinking, new and vanishing lesions are not unified and require separately trained models. There is a relative dearth of work focusing on temporal

segmentation of medical images. Many analyses constitute of basic arithmetical operations, such as thresholding and subtracting the intensity images from consecutive time points for segmenting new or disappearing white matter lesions [17, 64, 71].

2.2.5 Integrating Anatomical Constraints

Many segmentation methods are trained using pixel-wise loss functions, such as cross-entropy or Dice loss. Pixel-wise loss functions optimize performance by operating on individual pixels, by comparing predicted values of pixels with their respective ground truths. While this sounds sufficient for a well working framework, in practice, implementing these functions stand-alone may not capture underlying anatomical structures, and lead to incorrect segmentations. For this reason, it can be sensible to incorporate constraints into the model for anatomical plausibility in order reduce false positive and false negative segmentations. These constraints can be represented in the model as regularisers in the loss functions [139], priors which provide probabilistic information [35], or as topological functions [42].

For example, Oktay et al. proposes an Anatomically Constrainted Neural Network (ACNN) for cardiac image segmentation [139]. The ACNN features both the cross-entropy loss and a shape regularisation loss as the training objective function, where the latter encourages the model to follow global anatomical properties of the heart. Another method of decreasing false positive segmentations is to ensure that those segmentations are in realistic locations. For example, Dalca et al. proposes a variational auto-encoder with a location-specific probabilistic prior layer enabling accurate brain substructure segmentation [51].

2.2.6 Deep Learning for Multiple Sclerosis Brain Lesions

There have been numerous contributions to machine learning methods specifically for MS lesion segmentation. Many methods were developed following the 2015 ISBI Longitudinal Multiple Sclerosis Lesion Segmentation Challenge [28]. Valverde et al. employed a cascade of two 3D patch-wise CNNs, where the first CNN proposed candidate lesion voxels and the second one reduced falsely classified voxels [170]. Birenbaum et al. developed a multi-view longitudinal CNN and utilised priors about lesion intensities and spatial distribution to extract candidate lesions [22]. Similar to Valverde et al., Birenbaum et al. also used

3D patches for model training. Contrary to patch-based training, Aslani et al. proposed a multi-branch CNN which takes whole slices of the brain as input. Three 2D ResNets were separately trained for the axial, sagittal, and coronal planes, the outputs of which were fused to generate a final 3D segmentation [7]. Zhang et al. developed a fully convolutional densely connected network (Tiramisu) using a 2.5-dimensional input where slices were stacked from three anatomical planes, providing both global and local context in segmentation [192].

MS lesions can significantly disrupt brain anatomy, hindering accurate brain atlas registration and subsequent substructure analysis. To address this, numerous lesion inpainting techniques have been proposed, aiming to reconstruct the underlying brain tissue which lesions occupy. By "filling in" the missing anatomical information, inpainting facilitates more accurate brain atlas registration, improving the alignment of individual MS brains with a standardised template, and enabling more precise delineation of brain substructures for understanding functional impairments in MS. Many inpainting techniques for this application use classical methods, such as non-local means inpainting [78] and fast marching [154]. With the improvements in deep learning, inpainting techniques which employ encoder-decoder structures [189] and inpainting-specific loss functions [164] have also been developed.

2.2.7 Data Augmentation

Data augmentation can be classified into four categories: affine transformations, elastic transformations, intensity alterations, and incorporation of synthetic data. Examples of the first three categories are described in 2.1.4.5. Synthetic data augmentation methods utilise either non-generative models, often image mixing-techniques, or generative models which often employ GANs or diffusion models.

2.2.7.1 Non-Generative Data Augmentation

The traditional data augmentation (TDA) techniques which are widely used for training medical image segmentation models [92] do not dramatically change the lesion properties, such as the shape and location of lesions with respect to the surrounding tissue. MixUp [190] and related methods such as MixMatch [19] and CutMix [183] designed specific operations on two or more images to generate new samples. Zhang et al. proposed CarveMix, derived from CutMix [183], which uses a lesion-aware Mix-based technique to carve lesion regions

from one image and insert them into another image [195]. Zhu et al. developed another Mix-based data augmentation method, SelfMix, which performs augmentation by mixing tumours with non-tumour regions [143]. Zhang et al. presented ObjectAug, an object-level augmentation method for semantic image segmentation [193]. These methods directly employ foreground masks from other subjects to compose an augmented image with already present lesion shapes and characteristics, and provide valuable insights for lesion-level data augmentation. However, they directly mix original lesion masks for augmentation without augmenting individual lesion volumes, with no attention to location of the augmentation, or the lesion load of augmented images.

2.2.7.2 Generative Data Augmentation

Generative methods have received a lot of attention recently for providing an alternative way for data augmentation by performing abnormality synthesis [24, 161]. Salem et al. uses an encoder-decoder U-Net structure to fuse lesion masks acquired from MS patients to healthy subjects [152]. Bissoto et al., Jin et al., and Li et al. utilise generative adversarial networks (GANs) to synthesise new image samples [23, 98, 117]. Yet, [23] and [117] use semantic maps from pathological subjects to generate images, which lead to augmented images which have identical foreground shapes and labels to the samples that they are drawn from. Jin et al. uses a user-specified tumour mask to generate synthetic tumours on healthy scans. These methods do not generate diverse data samples with high foreground variance (lesion/tumour masks) which differ from already present samples in the training dataset. Reinhold et al. employs a casual model using variational autoencoders to generate lesions with predefined lesion load [146]. As for subject-specific pathological to healthy image synthesis, known as pseudo-healthy synthesis, Xia et al. uses a framework that consists of a generator, a segmentor and a reconstructor, along with a mask discriminator which distinguishes segmented masks from pathology masks [179]. Lin et al. proposes InsMix, a data augmentation method for nuclei segmentation, by employing a Copy-Paste-Smooth principle with a smooth-GAN for achieving contextual smoothness [119]. While generative augmentation methods have potential for diverse abnormality generation, they are often disease-specific and are not easily extended to different applications and datasets.

2.3 Evaluation Metrics

Here, we cover two main evaluation metrics for assessing model performance for brain lesion segmentation. One common method is measuring the overlap between the obtained segmentation from the model with the ground truth, which can be quantified using the Dice similarity coefficient, formulated as

$$\text{Dice} = 2\frac{\mid A \cap G \mid}{\mid A \mid + \mid G \mid}, \tag{2.16}$$

where A represents the automated segmentation and G represents the ground truth segmentation. The ideal Dice score is 1, and the worst possible score is 0. Alternatively, lesion segmentation performance can be evaluated by quantifying lesion detection rate. The performance of lesion detection is assessed using the lesion-wise F_1 score, which quantifies the accuracy of lesion detection irrespective of the precision of their boundaries. The F_1 score is calculated as

$$F_1 = 2\frac{S_L \cdot P_L}{S_L + P_L}, \tag{2.17}$$

where S_L represents the sensitivity of lesion detection (true positive rate or recall) and P_L represents the precision of positive predictive value. Similarly, the ideal F_1 score is 1, with the worst possible score being 0.

2.4 Conclusion

In this chapter, we have introduced the foundational elements of deep neural networks. We provided insight into the different methods, along with their building blocks and implementation details. We discussed the limitations of these methods and common means of overcoming them. We then provided a brief review on applied deep learning in medical image segmentation, focusing on brain MRI and lesion segmentation. The following chapters will focus on our recent works which tackle challenges in utilisation of longitudinal data, overcoming limited data scenarios, and incorporation of heterogeneous data for brain lesion segmentation.

Chapter 3

Learning from Longitudinal Data

This chapter contains material from:

1. Basaran, B.D., Matthews, P.M. and Bai, W., 2021. MSSEG-2 Challenge, Team New Brain: Cascaded networks for new MS lesion detection. MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure, p.77.

2. Basaran, B.D., Matthews, P.M. and Bai, W., 2022. New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation. Frontiers in neuroscience, 16, p.1007453.

3.1 Introduction

Multiple sclerosis (MS) is an inflammatory and demyelinating neurological disease of the central nervous system. Image-based biomarkers, such as lesions defined on magnetic resonance imaging (MRI), play an important role in MS diagnosis and patient monitoring. The detection of newly formed lesions provides crucial information for assessing disease progression and treatment outcome. Here, we propose a deep learning-based pipeline for new MS lesion detection and segmentation, which is built upon the nnU-Net framework. In addition to conventional data augmentation, we employ imaging and lesion-aware data augmentation methods, axial subsampling and CarveMix, to generate diverse samples and improve segmentation performance. The proposed pipeline is evaluated on the MICCAI 2021 MS new lesion segmentation challenge (MSSEG-2) dataset. It achieves an average Dice score of 0.510 and F_1 score of 0.552 on cases with new lesions, and an average false positive

lesion number n_{FP} of 0.036 and false positive lesion volume V_{FP} of 0.192 mm3 on cases with no new lesions. Our method outperforms other participating methods in the challenge and several state-of-the-art network architectures.

3.1.1 Contributions

Formation of new lesions is an important biomarker for the analysis of multiple sclerosis disease progression [89]. Unfortunately, majority of efforts are focused on segmentation of cross-sectional images, providing annotations of all lesions. We have three contributions in this work: (1) We present an end-to-end pipeline for segmenting the overlooked and clinically relevant formation of new lesions in follow-up brain MRI scans. (2) We improve performance by incorporating both imaging and lesion specific data augmentation techniques on top of conventional ones. (3) We provide an in-depth analysis of sources of failure and present an alternative metric for evaluating newly formed lesions.

3.2 Methods

Here, we present our proposed method for new lesion segmentation. We first provide an overview of our pipeline and data preprocessing steps, the details of the network architecture and hyperparameters, and explain our applications of imaging and lesion-aware data augmentation.

3.2.1 Proposed Pipeline

The proposed pipeline consists of a brain image preprocessing step, followed by nnU-Net [92] for lesion segmentation, which is trained with imaging and lesion-aware data augmentation. An overview of the pipeline is presented in Figure 3.1.

3.2.1.1 Preprocessing

Skull is stripped using an atlas-based brain extraction tool [56] followed by N4 bias field correction [167]. N4 bias correction reduces the low frequency intensity non-uniformity present in MRI image data. In practice, N4 bias leads to tissues of the same class to have different intensities. This is implemented using the MSSEG-2 longitudinal preprocessing script on

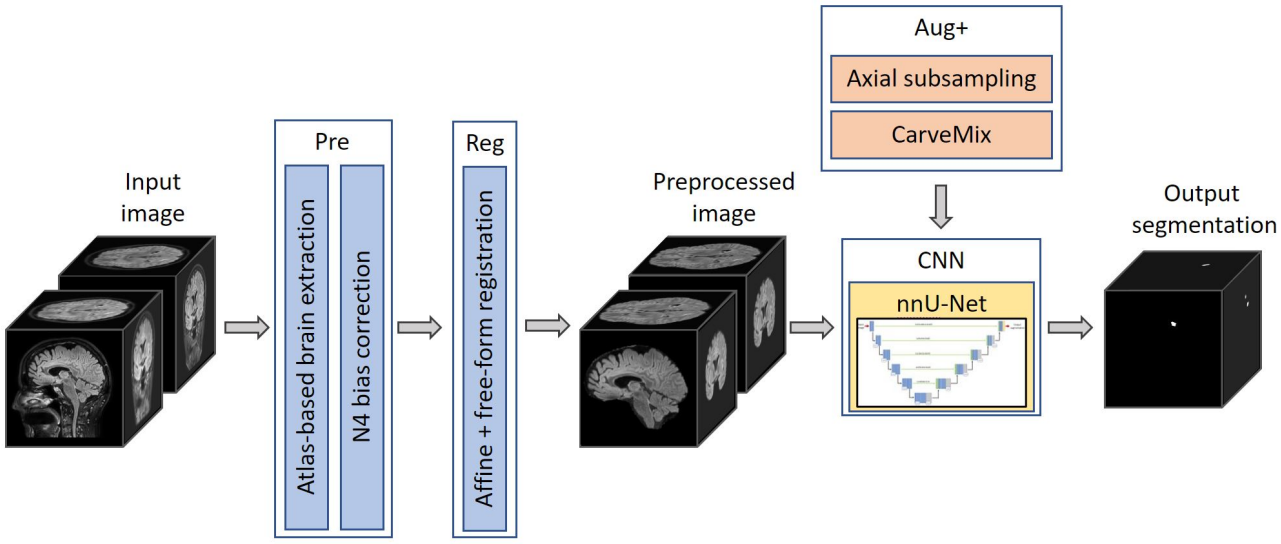


FIGURE 3.1: **The proposed 3D framework for multiple sclerosis new lesion detection and segmentation**. The framework comprises of brain extraction and N4 bias correction (Pre), intra-subject registration (Reg), imaging and lesion-aware data augmentations (Aug+), and nnU-Net. The pipeline takes input the baseline and follow-up FLAIR MRI scans, and outputs the proposed binary segmentation for new lesions.

Anima[1] provided by the challenge organisers. In addition, as the segmentation problem concerns imaging scans taken at different time points, we also perform intra-subject image registration so that scans of the same subject can be aligned and new lesions can be better differentiated. Since new lesions are defined on the follow-up scan, we register the baseline scan to the space of the follow-up scan. Affine image registration is performed, followed by free-form deformation, implemented using the MIRTK toolbox [132] using normalised mutual information as the loss function. Free-form deformation assists lesion segmentation in two ways: (1) brain structures, such as gyri, ventricles etc., are better registered; (2) lesions which slightly grow between scans are elastically registered so that the subsequent segmentation network can focus more on newly formed lesions.

3.2.1.2 Segmentation Network

We adopt nnU-Net [92] as the segmentation network, with a two-channel input: preprocessed baseline scan and preprocessed follow-up scan. The output of the network is a binary 3D prediction of new lesions which have formed in the follow-up scan. The network

[1] Anima scripts: RRID:SCR_017072, https://anima.irisa.fr

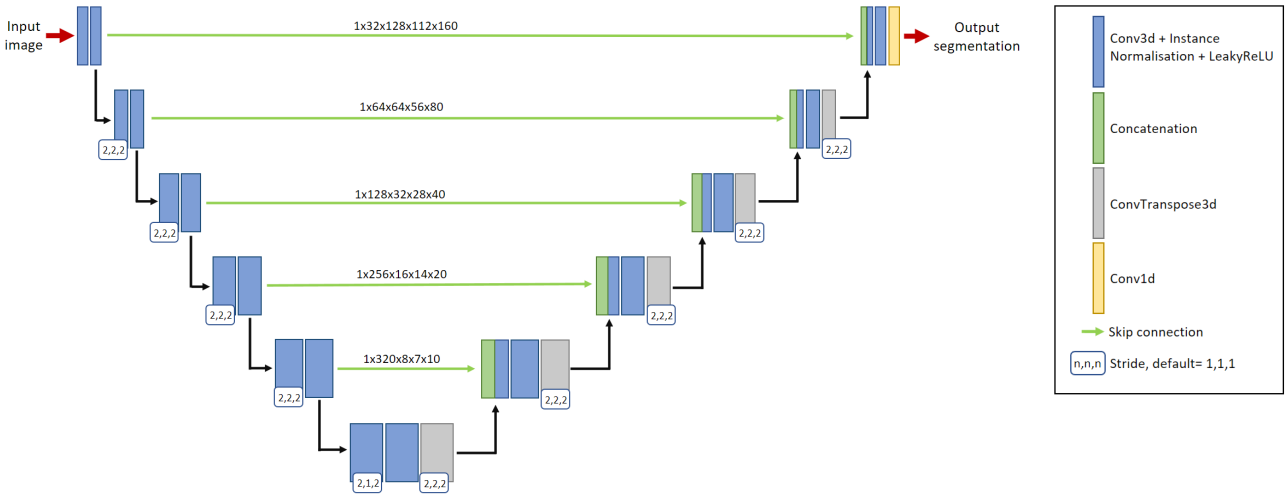


FIGURE 3.2: **The proposed 3D U-Net architecture for segmenting new lesions**. The network contains six resolution levels formed from contracting and expanding paths, and skip connections to recover fine-grain detail from the contracting path. The input to the network is a pair of baseline image and follow-up image. The output is the prediction of the new lesions.

consists of six resolution levels, formed from contracting and expanding paths. On the contracting path, each resolution level consists of two convolutional layers, each with a $3 \times 3 \times 3$ convolution kernel, followed by instance normalisation and LeakyReLU operation. At the start of each resolution level, the first convolution has a stride of (2,2,2), which effectively downsamples the feature map. At the lowest resolution level, the first convolution has a stride of (2,1,2).

On the expanding path, each resolution level consists of two convolutional layers with a $3 \times 3 \times 3$ convolution kernel, followed by instance normalisation and LeakyReLU operations, followed by an additional transposed $2 \times 2 \times 2$ convolution operation. The transposed convolution has a stride of (2,1,2) at the lowest resolution level, and a stride of (2,2,2) at all other resolution levels. By utilising skip connections, features extracted from the contracting path are concatenated with features at the the expanding path at their respective resolution level. The network uses 32-dimensional features maps at the highest resolution layer, which is increased to 320 feature maps at the lowest resolution layer. Please refer to Figure 3.2 for a graphical representation of the architecture.

3.2.1.3 Hyperparameters and Implementation Details

We implement the 3D full-resolution U-Net model of nnU-Net, using the `3d_fullres` configuration, utilising PyTorch. A single NVidia Tesla T4 GPU with 16GB RAM is used. Due to the GPU memory limit, 3D patches of size 128×112×160 are extracted from the original 3D images for model training. Images are not resized so that resolution is not lost. Patches are drawn randomly from the image with a 67% probability, which may or may not include the foreground class (lesions), and are selected with a 33% probability to guarantee inclusion of the lesion region [92]. The network is trained using a combination of Dice and cross-entropy loss, formulated as

$$\mathcal{L} = -\frac{2}{|C|}\sum_{c\in C}\frac{\sum_{i\in I}\hat{y}_{i,c}y_{i,c}}{\sum_{i\in I}\hat{y}_{i,c}+y_{i,c}} - \sum_{i\in I}\sum_{c\in C} y_{i,c}\log\hat{y}_{i,c} \tag{3.1}$$

where c denotes the class, C denotes the number of classes ($C = 2$ in our method), i denotes a given voxel, I denotes the set of voxels over the image, $\hat{y}$ is the softmax output of the segmentation network, y is the one-hot encoding of the ground truth label for the new lesions, and subscript i, c is the number of voxels in the training patch for class c.

We use the stochastic gradient descent optimizer with Nesterov momentum of 0.99, an initial learning rate of 0.01, a polynomial learning rate decay and a batch size of 2 patches. When developing the model on the training data, five-fold cross-validation is used. Each model is trained for 1,000 epochs. After training, for each fold, we select the model which produces the highest Dice score. For inference, we ensemble outputs from the five models from each fold by averaging their softmax outputs. No post-processing step is applied.

3.2.1.4 Imaging and Lesion-aware Data Augmentation

Incorporation of data augmentation methods increases model generalisability and robustness, and decreases overfitting. We utilise multiple data augmentation techniques, including the default augmentations that nnU-Net provides in the batchgenerators data augmentation framework [93]. These augmentations include mirroring, rotating, scaling, input channel

translation, elastic deformations, linear downsampling, brightness and contrast augmentation, gamma augmentation, Gaussian and Rician noise augmentation, and random cropping. Input channel translation simulates registration errors which may have occured during the curation or preprocessing of the dataset. This is achieved by slightly shifting the first or second input image in one of the three dimensions, making the model robust to such errors.

In addition to these augmentations, inspired by [100, 101], we introduce axial subsampling to simulate artifacts during the image acquisition process on the axial plane. Brain MRI typically acquires a stack of 2D image slices in the axial plane to form a 3D volume, which can be of a high resolution within the axial plane but subject to low resolution across the plane [30]. We simulate this by applying axial subsampling augmentation by applying a median filter of size $[1 \times 1 \times n]$ where $n \in 2, 3, 4$ to the axial image slices. This effectively blurs the image in the sagittal and coronal planes as a function of n. Figure 3.3(a) illustrates an example of axial subsampling.

Finally, to increase the diversity of lesion images, a lesion-aware data augmentation method, CarveMix [195], is used. CarveMix extracts a 3D region of interest (ROI) according to the lesion location and shape from one subject and mixes it with the brain image of another subject, thus creating augmented training samples. To increase diversity in augmentation, the lesion-aware ROI is generated by thresholding the distance transform of the lesion using a random threshold [195]. A synthetic image, $\mathbf{X}$, and its label, $\mathbf{Y}$, is generated by

$$\mathbf{X} = \mathbf{X}_i \odot \mathbf{M}_i + \mathbf{X}_j \odot (1 - \mathbf{M}_i), \tag{3.2}$$

$$\mathbf{Y} = \mathbf{Y}_i \odot \mathbf{M}_i + \mathbf{Y}_j \odot (1 - \mathbf{M}_i), \tag{3.3}$$

where $\{\mathbf{X}_i, \mathbf{Y}_i\}$ denotes one pair of image and label, $\{\mathbf{X}_j, \mathbf{Y}_j\}$ denotes a second pair of image and label, $\mathbf{M}_i$ denotes the binary mask of the ROI, and $\odot$ represents voxel-wise multiplication. We randomly select two subjects for CarveMix augmentation when training. Incorporation of CarveMix data augmentation increases the total volume which the lesion class covers in an image, thus reducing the effect of class imbalance caused from the foreground class making up a small percentage of the overall image. Figure 3.3(b) illustrates an example

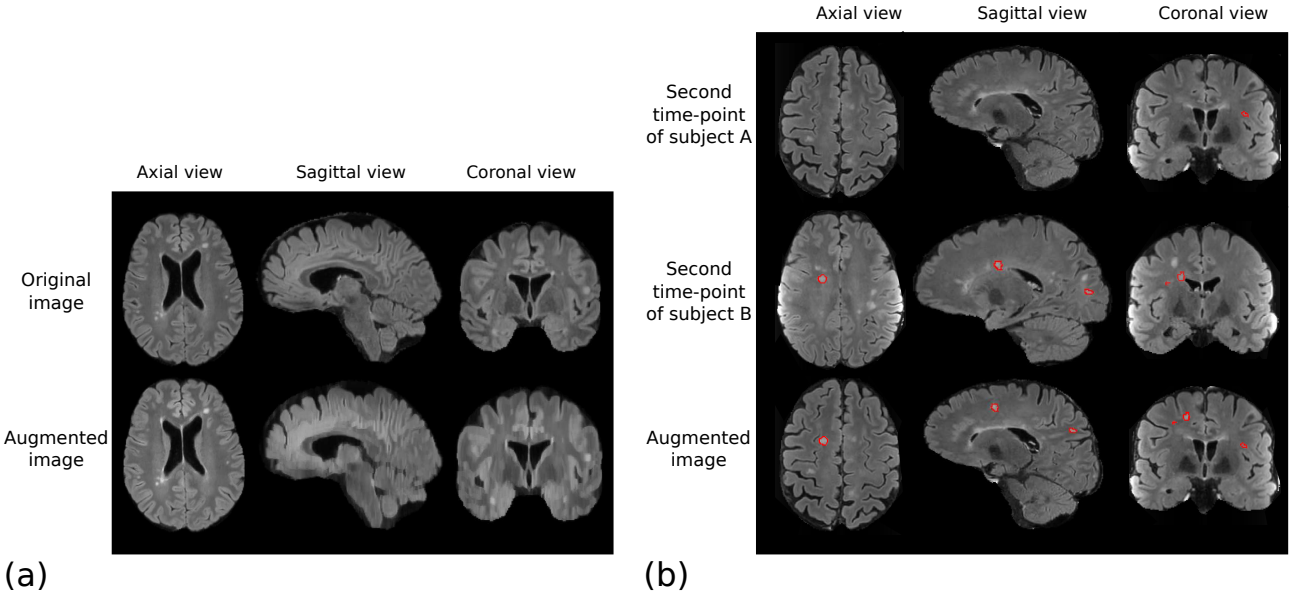


FIGURE 3.3: **Imaging and lesion-aware data augmentations applied on the MSSEG-2 training set**. **(a)** Example of axial subsampling ($n = 4$) to simulate the blurring in image acquisition. **(b)** Example of the CarveMix augmentation. Lesions from subject B are carved out and fused onto scan from subject A. Contours delineate lesion labels.

of CarveMix augmentation. It is important to note that while CarveMix augmentation increases the diversity of the dataset, it does not guarantee the generation of realistic images, as 'mixed' images may contain lesions crossing the white and gray matter, or even the ventricles. However, recent works show that augmented images do not need to look realistic for increase in model performance [73].

3.3 Experiments

3.3.1 Data

We evaluate the pipeline on the *MICCAI 2021 MS new lesion segmentation challenge* dataset (MSSEG-2) [43], which provides 3D FLAIR images of 100 MS patients. The images were acquired from 15 different scanners, six of them 1.5T and nine of them 3T, including three GE scanners, six Philips scanners, and six Siemens scanners. Dataset scanner information can be found at the MICCAI 2021 MSSEG-2 challenge demographics data [46]. The images have varying image size and voxel spacing, which we resample to the median spacing of the dataset, $0.977 \times 0.977 \times 0.530 mm^3$, before model training. Each patient was scanned twice,

with 1 to 3 years between the two time points, constituting for a total of 200 images. Only new lesions at the second time point were annotated. Existing lesions, growing or shrinking lesions were not delineated. Each patient was annotated by four neuroradiologist and one consensus new lesion mask was provided. We use the consensus lesion masks as ground truth for model training and evaluation.

The dataset has been partitioned into 40 training and 60 test subjects by the challenge organisers. Of the 40 training subjects, 11 of them do not exhibit new lesions, which are referred to as “no-new lesion cases". We exclude these 11 no-new lesion cases from the training set, utilising the remaining 29 cases for model training. Of the 60 test subjects, 28 of them do not exhibit new lesions. We use all 60 subjects for testing.

3.3.2 Evaluation Metrics

The method is evaluated using the Anima analyser tool's `animaSegPerfAnalyzer`[2] function, provided by the the MSSEG-2 challenge organisers in order to provide a fair comparison with other participating methods. In line with the MSSEG-2 evaluation, we use the default configuration of `animaSegPerfAnalyzer`, which excludes lesion volumes smaller than 3mm^3. The performance is evaluated separately for patients with new lesions and those with no-new lesions on the test set. For the new lesion cases, we report new lesion detection and segmentation performance, true positive lesion count, false positive lesion count, and false negative lesion count; for the no-new lesion cases, we calculate the number of new lesions detected (false positive lesions) and the volume of these false positive lesions.

3.3.2.1 Performance on New Lesion Cases

New lesion detection and new lesion segmentation performances are evaluated using the F_1 score and Dice similarity coefficients, respectively, described in 2.3. Using the Anima software, a lesion is considered as being detected or true positive if the automatic detection overlaps with at least 10% of the ground truth lesion volume and does not go outside by more than 70% of the volume [44].

In addition to the metrics used in the MSSEG-2 challenge, we also present results for average true positive lesion count, n_{TP}, average false positive lesion count, n_{FP}, and average

[2] Anima scripts: RRID:SCR_017072, https://anima.irisa.fr

false negative lesion count, n_{FN}. Average true positive lesion count evaluates the average correctly detected lesions by the automated method. n_{FP} evaluates the average incorrectly detected lesions by the automated method. Finally, n_{FN} evaluates the lesions not detected by the automated method. These metrics are averaged over the 32 new lesion cases in the test set. The consensus ground truth segmentation contains a total of 224 new lesions, therefore the optimal average true positive lesion count, n_{TP}, is 7 ($\frac{224}{32}$). The optimal score for n_{FP} or n_{FN} is 0.

3.3.2.2 Performance on No-New Lesion Cases

For no-new lesion cases, the number and volume of falsely predicted lesions are evaluated. To count the number of false positive lesions, the Anima tool, `animaConnectedComponents` [3] function is used with default parameters. The volume of false positive lesions is calculated by multiplying the number of lesion voxels by voxel spacing. We denote number of false positive lesions as n_{FP}, and volume of false positive lesions as V_{FP}. The optimal scores for both false positive lesion number and volume are 0.

3.3.3 Results

3.3.3.1 Comparison against Participating Methods in the Challenge

The proposed pipeline is compared against MSSEG-2 participating methods and also four expert raters [45], reported in Table 3.1. The performance of MSSEG-2 participating methods and four expert raters is acquired from [45]. Table 3.1 shows that the proposed pipeline ranks competitively against methods submitted to the MSSEG-2 challenge. For the new lesion cases, it outperforms the other methods in terms of both the average DSC and the average F_1 scores. Also, our method outperforms three of the experts in n_{TP} and n_{FN} metrics. Our method correctly identifies 24 of the 32 new lesion cases as having new lesions. We achieve comparable performance to Experts 1, 2, 3 and 4, which correctly identify 26, 25, 27, and 22 of the 32 new lesion cases as having new lesions, respectively. A non-zero F_1 score is regarded as a method having correctly identified a new lesion case.

For the no-new lesion cases, the proposed pipeline achieves the lowest metrics for false positive lesions, including the average number n_{FP} and the average volume V_{FP}. It correctly

[3] Anima scripts: RRID:SCR_017072, https://anima.irisa.fr

identifies 27 out of 28 no-new lesion cases as subjects with no-new lesions. When comparing to expert raters, on the new lesion cases, the proposed pipeline outperforms Expert 4 in terms of DSC and F_1 scores and approaches the performance of Expert 2. On the no-new lesion cases, the proposed pipeline outperforms or achieves a comparable performance to Experts 1, 2 and 4.

3.3.3.2 Sensitivity and Specificity Analysis

Figure 3.4 and Table 3.1 show that there is a reverse correlation between the results for new lesion cases versus no-new lesions cases, especially in the participating methods for the MSSEG-2 challenge. In Figure 3.4, we plot the average DSC and F_1 scores against the average n_{FP} and V_{FP} metrics. It shows that methods which perform well in new lesion cases do not perform as well in no-new lesion cases. Contrary to other methods, the proposed pipeline does not suffer from severe negative correlation, which performs well in both new lesion and no-new lesion cases. This is visualised by our proposed method, depicted as a black star in Figure 3.4, having both high DSC and F_1 scores as well as low n_{FP} and V_{FP} metrics.

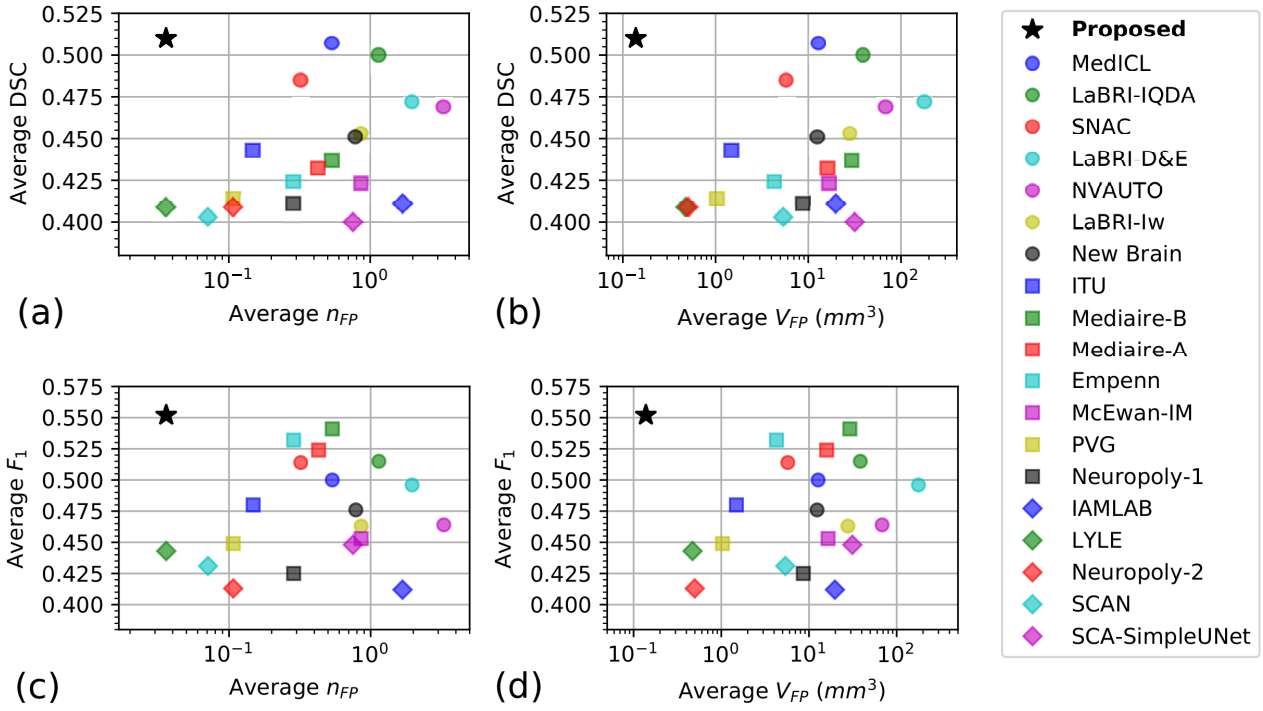


FIGURE 3.4: **Comparison of different methods in new lesion metrics (DSC and F_1) vs no-new lesion metrics (n_{FP} and V_{FP}).** X-axis denotes one of the no-new lesion metrics in logarithmic scale and Y-axis denotes one of the new lesion metrics. Star denotes the proposed pipeline. **(a)** Plot of average DSC vs average false positive lesion count n_{FP}. **(b)** Plot of average DSC vs average false positive lesion volume V_{FP}. **(c)** Plot of average F_1 vs average false positive lesion count n_{FP}. **(d)** Plot of average F_1 vs average false positive lesion volume V_{FP}.

TABLE 3.1: **Comparison of the proposed method to the challenge participating methods in DSC, F_1 scores, the number of false positive lesions n_{FP} and their volume V_{FP} (unit: mm^3), averaged across cases.** The methods are sorted in the descending order of DSC. Best results are in bold. Asterisks indicate statistical significance (*: p$\leq$ 0.05, **: p $\leq$ 0.01, ***: p $\leq$ 0.005) when using a paired Student's *t*-test compared to the proposed method. We implement the Mann-Whitney U test for n_{TP}, n_{FP}, and n_{FN} metrics due to their non-normal distribution. The dagger symbols indicate statistical significance (†: p$\leq$ 0.05, ††: p $\leq$ 0.01, †††: p $\leq$ 0.005) when using a Mann-Whitney U test compared to the proposed method.

		New lesion cases (n = 32)				No-new lesion cases (n = 28)	
Method	DSC	F_1	n_{TP}	n_{FP}	n_{FN}	n_{FP}	V_{FP}
Expert 1	0.629	0.709	6.063	1.281	1.094	0.036	1.453
Expert 3	0.597	0.637	4.500	0.844	2.375	0.000	0.000
Expert 2	0.535	0.601	4.313	1.094	2.500	0.107	3.981
Expert 4	0.459	0.519	4.469	0.594	2.375	0.036	0.623
Proposed	**0.510**	**0.552**	4.969	2.031	2.281	**0.036**	**0.192**
MedICL	0.507	0.500	5.344	5.063††	1.875	0.536†††	12.713
LaBRI-IQDA	0.500	0.515	5.563	6.094††	1.656	1.143††	38.486*
SNAC	0.485	0.514	5.219	3.689†	2.031	0.321	5.726
LaBRI-D&E	0.472	0.496	5.500	9.156†††	1.750	1.964††	177.131
NVAUTO	0.469	0.464*	5.344	12.000†††	1.906	3.286†††	68.211*
LaBRI-Iw	0.453*	0.463*	5.000	6.719††	2.250	0.857†	27.761*
New Brain	0.451***	0.476***	4.032	2.903	3.355	0.786†	12.371
ITU	0.443	0.480	4.688	3.094	2.438	0.148	1.487
Mediaire-B	0.437**	0.541	**5.688**	4.469††	**1.500**	0.536†††	29.235*
Mediaire-A	0.432***	0.524	5.156	3.500	2.031	0.429†	15.908*
Empenn	0.424*	0.532	4.178	2.719	3.031	0.286††	4.258*
McEwan-IM	0.423***	0.453*	5.469	8.531†††	1.781	0.857†	16.504
PVG	0.414***	0.449*	4.032	2.903	3.355	0.107	1.031
Neuropoly-1	0.411***	0.425***	3.625	2.813	3.563	0.286†	8.615
IAMLAB	0.411***	0.412***	5.094	6.844†††	2.156	1.679†††	19.753*
LYLE	0.409***	0.443**	3.406	**1.250**	3.594	**0.036**	0.470
Neuropoly-2	0.409***	0.413***	3.656	1.906	3.469	0.107	0.498
SCAN	0.403***	0.431**	4.156	2.406	3.031	0.071	5.373
SCA-SimpleUNet	0.400***	0.448*	5.406	6.344†††	1.813	0.750†††	31.232*
I3M	0.398***	0.358***	4.250	4.313†	3.000	0.393	14.800
Neuropoly-3	0.379***	0.416***	3.719	2.625	3.500	0.321†	19.240
The NoCoDers	0.365***	0.381***	4.750	7.594†††	2.500	1.370†††	25.848*
Vicorob	0.357***	0.369***	3.906	4.094††	3.156	0.964†	88.402
HufsAIM	0.346***	0.407***	2.938	1.979	4.156	0.444†	17.128*
CMIC	0.330***	0.362***	3.906	6.094†††	3.344	4.714†††	123.442
MIAL	0.309***	0.332***	4.516	6.097†	2.774	1.464††	177.861
SCA-withPriors	0.223***	0.216***	2.750	6.719††	4.219	2.464†††	302.121
LIT	0.214***	0.242***	2.406	11.063	4.469	0.607†	35.404
IBBM	0.155***	0.145***	1.906††	7.625†††	5.188†	3.786†††	123.309***
Optimal score	1.000	1.000	7.000	0.000	0.000	0.000	0.000

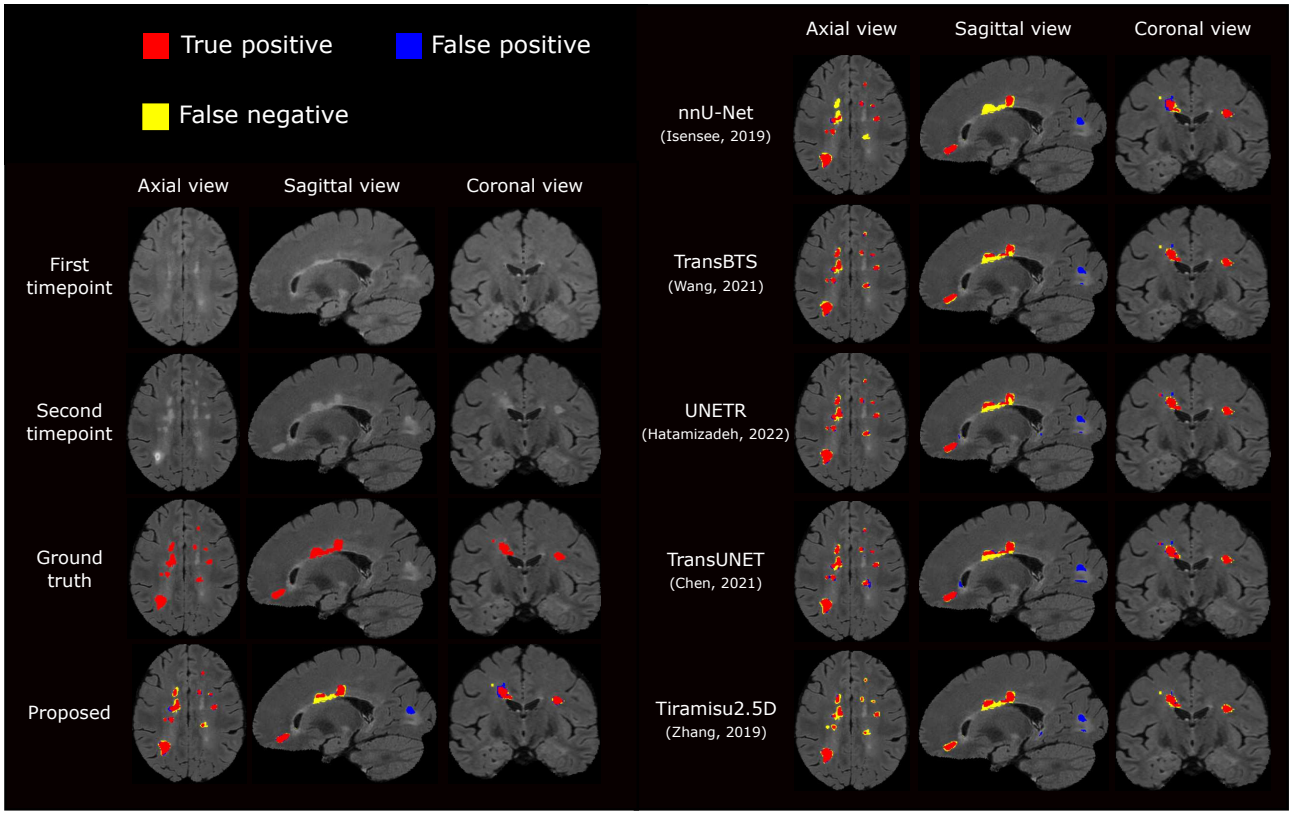


FIGURE 3.5: **Visual comparison of the proposed segmentation pipeline to other methods**. The proposed method produces segmentation maps closest to the ground truth annotation.

3.3.3.3 Comparison against State-of-the-Art Architectures

We also compare the proposed pipeline to a number of state-of-the-art convolutional and transformer-based architectures, which have demonstrated excellent performance in biomedical image segmentation tasks. These architectures include the standard nnU-net [92], TransBTS [173], UNETR [82], TransUNet [33] and Tiramisu 2.5D [192]. In order to evaluate methods fairly, we train these methods using the same preprocessed data, described in Sec. 3.2.1.1, which includes atlas-based brain extraction, N4 bias field correction, and free-form deformation registration, and use the standard data augmentation. The quantitative comparison results are reported in Table 3.2, and an example segmentation for visual comparison is provided in Figure 3.5. Table 3.2 shows that nnU-Net with standard data augmentations performs favourably against these state-of-the-art methods, and the proposed pipeline further improves performance possibly due to the additional data augmentation that we have introduced.

TABLE 3.2: **Comparison of the proposed method to recent state-of-the-art deep learning architectures in DSC and F_1 scores, the number of false positive lesions n_{FP} and their volume V_{FP} (unit: mm^3), averaged across cases.** The methods are sorted in the descending order of DSC. Best results are in bold. Asterisks indicate statistical significance (*: p$\leq$ 0.05, **: p $\leq$ 0.01, ***: p $\leq$ 0.005) when using a paired Student's *t*-test compared to our proposed method. We implement the Mann-Whitney U test for n_{TP}, n_{FP}, and n_{FN} metrics due to their non-normal distribution. The dagger symbols indicate statistical significance (†: p$\leq$ 0.05, ††: p $\leq$ 0.01, †††: p $\leq$ 0.005) when using a Mann-Whitney U test compared to the proposed method.

	New lesion cases (n = 32)					No-new lesion cases (n = 28)	
Method	DSC	F_1	n_{TP}	n_{FP}	n_{FN}	n_{FP}	V_{FP}
Proposed	**0.510**	**0.552**	4.969	2.031	2.281	**0.036**	0.192
nnU-Net [92]	0.490	0.548	4.562	**1.281**	2.688	**0.036**	**0.138**
TransBTS [173]	0.477	0.470*	**5.492**	5.718††	**1.848**	0.939†	12.238
UNETR [82]	0.462	0.468*	5.343	9.031†††	1.906	4.214†††	23.705*
TransUNet [33]	0.428**	0.434**	4.491	4.043††	2.102	1.081††	9.620
Tiramisu 2.5D [192]	0.363***	0.365***	4.313	4.625††	2.938	1.384††	15.120
Optimal score	1.000	1.000	7.000	0.000	0.000	0.000	0.000

3.3.3.4 Ablation Study

We carry out an ablation study to evaluate the impacts of different components of the pipeline, including brain extraction and N4 bias correction (Pre), affine and free-form image registration (Reg) and additional data augmentation methods including axial subsampling and CarveMix (Aug+). By default, standard data augmentation methods are used which come with nnU-Net, described in 3.2.1.4. The ablation study results are presented in Table 3.3.

Interestingly, adding pre-processing alone or registration alone does not drastically change performance metrics. However, when they are combined, for new lesion cases, the DSC score is increased from 0.476 to 0.490 and the F_1 score is increased from 0.533 to 0.548. When imaging-related and lesion-aware data augmentations (Aug+) are introduced,

TABLE 3.3: **Ablation study for preprocessing, registration and augmentation steps.** Results present DSC and F_1 scores, the number of false positive lesions n_{FP} and their volume v_{FP} (unit: mm^3), averaged across cases. Asterisks indicate statistical significance (*: p$\leq$ 0.05, **: p $\leq$ 0.01) when using a paired Student's *t*-test compared to our proposed method. Best results are in bold.

			New lesion cases (n = 32)					No-new lesion cases (n = 28)	
Pre	Reg	Aug+	DSC	F_1	n_{TP}	n_{FP}	n_{FN}	n_{FP}	V_{FP}
			0.476*	0.533	4.250**	**1.281**	3.000**	**0.000**	**0.000**
✓			0.475*	0.524*	4.688	**1.281**	2.563	**0.000**	**0.000**
	✓		0.473*	0.525*	4.188**	1.343	3.062**	0.036	0.083
✓	✓		0.490*	0.548	4.562	**1.281**	2.688*	0.036	0.138
✓	✓	✓	**0.510**	**0.552**	**4.969**	2.031	**2.281**	0.036	0.192

the DSC score is further increased to 0.510 and the F_1 score is increased to 0.552. This demonstrates that all the three components play an important role in the proposed pipeline. We also observe that when DSC and F_1 scores are increased, metrics concerning no-new lesion cases become poorer. The undesired increase in false positive lesion count and lesion volume for the no-new lesion cases is discussed in detail in Sec 3.3.3.6.

3.3.3.5 Exclusion of No-New Lesion Cases During Training

The MSSEG-2 training dataset is composed of 40 subjects. 11 subjects do not exhibit new lesions in their follow-up scans. These subjects were removed from the training dataset, thus we only utilised 29 subjects. We carry out an additional study to investigate the impact of the exclusion of these images, by comparing the performance on the test set when utilising all 40 subjects for segmentation model training against using the 29 subjects with new lesions. Results are presented in Table 3.4. Interestingly, removing the no new lesion subjects result in slightly higher DSC and F_1 score, without compromising performance in the no new lesion cases. This is likely due to higher average representation of the foreground class (new lesions) in the altered training set. In addition, reducing the training set from 40 to 29 subjects decreased the model training time by 27.5%.

3.3.3.6 Sources of Failure

We carry out a qualitative investigation on the test set to better understand where our method fails. In the no-new lesion cases, the proposed pipeline correctly classifies 27 out

TABLE 3.4: **Comparison between using the complete MSSEG-2 training set (40 subjects) against using 29 subjects which excludes the no new lesion cases.** We present the DSC and F_1 scores, the number of false positive lesions n_{FP} and their volume V_{FP} (unit: mm^3), averaged across cases. Asterisks indicate statistical significance (*: p$\leq$ 0.05) when using a paired Student's *t*-test compared to our proposed method. Best results are in bold.

	New lesion cases (n = 32)		No-new lesion cases (n = 28)	
Training set	DSC	F_1	n_{FP}	V_{FP}
29 subjects with new lesions	**0.510**	**0.552**	**0.036**	**0.192**
All 40 subjects	0.502	0.530*	**0.036**	**0.192**

of the 28 subjects. The one misclassified (subject ID: 004) is incorrectly segmented to have 1 new lesion, which amount to 14 false positive voxels (3.875 mm^3), shown in Figure 3.6. The segmented region has a higher intensity compared to surrounding regions and we suspect that it is likely to be a new lesion. Two of four expert raters also delineate this region as a new lesion, although the consensus segmentation does not regards this as a lesion, which leads to the misclassification of our method. This test set subject is the cause of the undesired increase in false positive lesion count and false positive lesion volume in Table 3.3.

In the new lesion cases, when assessing against the DSC and and F_1 score, there is still room for improvement in performance. There are possibly three sources of failure that affect the DSC and F_1 scores. The first is the incorrect segmentation of growing lesions. The pipeline employs affine and non-rigid registration to align the baseline scan to the follow-up and thus suppresses the detection of growing lesions. However, the remaining misalignment for some growing lesions still leads to the boundary voxels, i.e. the grown regions of lesions, being incorrectly segmented as new lesions. Secondly, the proposed pipeline may miss some tiny and less apparent new lesions. In some cases, new lesions which form in the follow-up scan are very small and less hyperintense compared to large new lesions. This makes the detection of these lesions very difficult and leads to misclassifications. Finally, new lesion segmentation is a generally challenging task even for human raters and there are indiscrepancies between annotations from different human experts. The effect of inter-rater variability in biomedical annotations can be seen in Yang et al. and Kang et al. [182, 102]. The noise in the annotations may limit what an automated method can achieve [194]. We present examples of all three sources of failure in Figure 3.7.

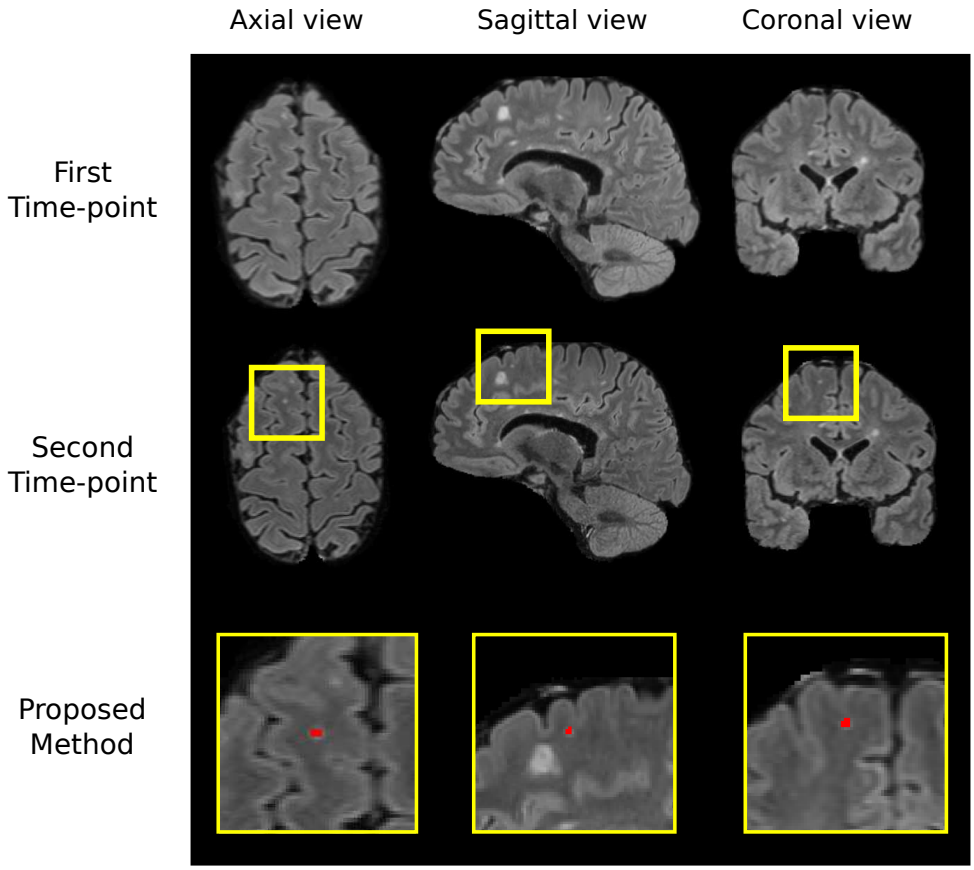


FIGURE 3.6: **A qualitative analysis of a false positive segmentation**. The region which we incorrectly classify as a lesion in the no-new lesion cases in the MSSEG-2 test set (subject ID: 004). We suspect the segmented region to be a new lesion, although the consensus segmentation provided by the organiser classifies this as normal tissue. Two of the four human raters also classify this region as a new lesion.

3.4 Discussion and Conclusion

Here we demonstrate that by incorporating appropriate preprocessing steps, an nnU-net segmentation network, imaging and lesion-aware data augmentation techniques, we can achieve promising performance in new MS lesion segmentation tasks. The proposed pipeline outperforms other challenge participating methods in both new lesion cases and no-new lesion cases, in terms of DSC, F_1, n_{FP} and V_{FP} scores. We also observe that in terms of network architecture, the recently popular transformer architectures may not necessarily outperform convolutional neural network architectures, such as nnU-net (Table 3.2). This finding is aligns with Isensee et al. [94].The design of proper pre-processing steps and problem-specific augmentations may play a more important role in this particular lesion segmentation task (Table 3.3).

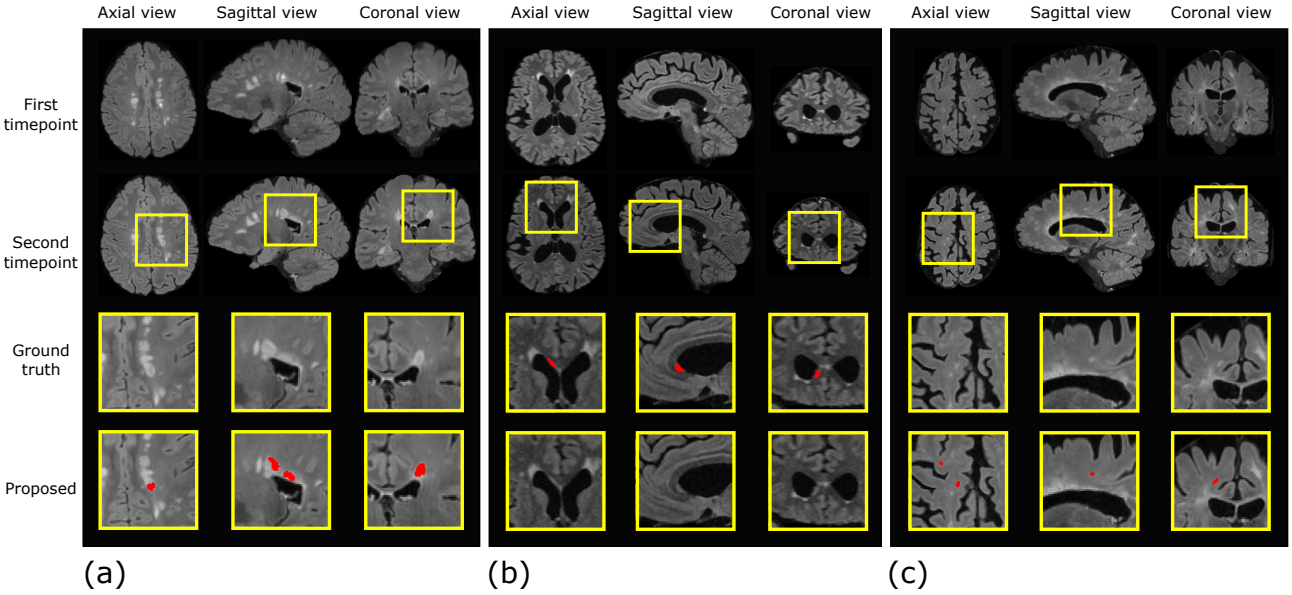


FIGURE 3.7: **Examples of three different sources of error**. Ground truth and predicted lesions are delineated in red. **(a)** (subject ID: 012) False positive segmentation of a growing lesion. **(b)** (subject ID: 078) False negative classification of a new lesion. **(c)** (subject ID: 036) False positive segmentation of a region classified as healthy/not-new in the consensus label, but annotated as a new lesion in two of the four provided expert annotations.

In addition to the DSC and F_1 score used by the MSSEG-2 challenge, we introduce extra evaluation metrics, n_{TP}, n_{FP}, and n_{FN}, for the new lesion cases to understand the method performance. While many methods have a high n_{TP} and a lower n_{FN} score, the results suggest that a lower n_{FP} is what differentiates our method and the Experts to the other methods, thus providing a higher DSC and F_1 score. The n_{TP}, n_{FP}, and n_{FN} results also suggest that they should be analysed with respect to each other, as evaluating a method solely with one of these metrics can be misleading. For example, the top performing method in correctly identified average true positive lesions, n_{TP}, ranks 10th in DSC score, and the top performing method in fewest average false positive lesion count ranks 17th in both DSC and F_1 score. Methods with higher n_{TP} score also high n_{FP} scores, with respect to Experts' performance. The results on the new metrics show that methods differ on their approach to achieve optimal DSC and F_1 scores, and suggest that extra thought should be considered when evaluating a method solely on one metric.

Future efforts to improve the proposed method include further investigation of the sources of failure described in Sec. 3.3.3.6 and bridging the gap between automatic segmentation and expert raters. The current MSSEG-2 challenge dataset only contains annotations

of new lesions. To discriminate new lesions from growing lesions, future works may include curating a dataset of both lesion types and training automated methods for detecting and differentiating these lesions. Also, additional post-processing steps could be developed to inspect local neighbourhoods of detected new lesions and check whether they are connected to existing lesions or not, thus decreasing false positives for new lesion detection. However, too large of a local context may come at the cost of decreasing true positives too. Furthermore, the proposed pipeline only focuses on lesions in the brain region and the preprocessing step removes the spinal cord region. Despite the MSSEG-2 testing dataset not featuring any new lesions in the spinal cord, MS lesions can form in this region. Inclusion of the spinal cord into the preprocessing step and training data will extend the application of the proposed pipeline. Finally, despite the long temporal nature of MS, the current dataset and method only considers two timepoint data. For cases with more than two timepoints, a sliding window approach can be implemented, described in 5.2.1.

In conclusion, we propose an nnU-Net-based pipeline for multiple sclerosis new lesion segmentation. A contribution of the pipeline is that it incorporates task-specific data augmentation methods, including axial subsampling, which simulates MRI acquisition-based image artefacts, and CarveMix, which increases the diversity of lesion images. When evaluating on the MSSEG-2 dataset, the proposed pipeline achieves excellent performance in evaluation metrics for both new lesion and no-new lesion cases.

Chapter 4

Learning from Limited Data

This chapter contains material from:

1. Basaran, B.D., Qiao, M., Matthews, P.M. and Bai, W., 2022, September. Subject-Specific Lesion Generation and Pseudo-Healthy Synthesis for Multiple Sclerosis Brain Images. In International Workshop on Simulation and Synthesis in Medical Imaging (pp. 1-11). Cham: Springer International Publishing.

2. Basaran, B.D., Zhang, W., Qiao, M., Kainz, B., Matthews, P.M., Bai, W., 2024. LesionMix: A Lesion-Level Data Augmentation Method for Medical Image Segmentation. In Data Augmentation, Labelling, and Imperfections. MICCAI 2023. Lecture Notes in Computer Science, vol 14379. Springer, Cham.

4.1 An Adversarial Method for Realistic Brain Lesion Data Augmentation

4.1.1 Introduction

Understanding the intensity characteristics of brain lesions is key for defining image-based biomarkers in neurological studies and for predicting disease burden and outcome. Being able to model the lesion intensity patterns also enables us to generate synthetic images for data augmentation purposes and to perform counter-factual inference to understand disease progression. Although many efforts have been dedicated in the past to the general image synthesis problem [90, 103, 145], little has been investigated for subject-specific MS lesion synthesis. In this chapter, we present a novel foreground-based generative method for

modelling the local lesion characteristics that can both generate synthetic lesions on healthy images and synthesize subject-specific pseudo-healthy images from pathological images. An attention mechanism is designed so as to only modify the intensities in the lesion region while maintaining the structure in other brain regions. Furthermore, the proposed method can be used as a data augmentation module to generate synthetic images for training brain image segmentation networks. Experiments on multiple sclerosis (MS) brain images acquired on magnetic resonance imaging (MRI) demonstrate that the proposed method can generate highly realistic pseudo-healthy and pseudo-pathological brain images. We demonstrate that the proposed method achieves high realism in both lesion generation and pseudo-healthy synthesis. In addition, using the synthetic images for training image segmentation networks, we can improve the lesion segmentation performance even in a low-data setting.

4.1.2 Contributions

We present a novel brain lesion synthesis method for multiple sclerosis images. Our method has four contributions: (1) Differing from other techniques, our method is able to preserve subject identity and generate a new lesion foreground class without directly sampling a lesion mask from another subject. (2) An attention-based foreground discriminator is introduced to generate realistic images focusing at the lesion region. (3) The method is able to generate both synthetic lesions as well as pseudo-healthy images, achieved via a cyclic structure. This enables generating a diverse set of brain images for data augmentation purposes. (4) Our method is stochastic, and is able to generate multiple augmented images with different lesion appearances from a single input image.

4.1.3 Methods

The goal is to generate pathological images with lesions, given healthy images as input, while preserving the subject identity, and vice versa, create pseudo-healthy images without the lesions in the pathological subjects. Inspired by CycleGAN [197] and Attention-GAN [163], we adopt a two-branch architecture to learn both the lesion generation and pseudo-healthy synthesis tasks. We employ the attention and content decoders from [163], which provide maps on *where* and *what* the important features of the foreground and background classes of the image are, respectively. In our application, the foreground refers

to the brain regions with lesions, and background refers to the remaining healthy tissue. However, we encourage the foreground decoder to learn lesion specific features. In particular, we introduce a foreground-based discriminator to focus on the generation of the lesion region. We assume that two unpaired datasets are available for learning: a healthy dataset, $X_H = \{x_H\}$, and a pathological dataset with lesions or white matter hyperintensities (WMHs), $X_P = \{x_P\}$. Lower case x_H and x_P denote single image samples from these datasets. For the pathological dataset, a corresponding foreground dataset, $X_F = \{x_F\}$, is constructed using the lesion masks. Each foreground image is a masked lesion intensity image, generated by multiplying the original pathological image from X_P with its respective lesion mask.

Figure 4.1 illustrates the proposed framework, which consists of two branches, one generator, G_P, for synthetic lesion generation and the other generator, G_H, for pseudo-healthy synthesis. We represent the lesion generation branch as $H \rightarrow P$, and the pseudo-healthy synthesis branch as $P \rightarrow H$, where H denotes healthy images, and P denotes pathological images. The lesion generator, G_P, takes in a healthy brain image, x_H, and generates a pathological image with lesions, $\hat{x}_P$, while also aiming to preserve subject identity. It is composed of an encoder, Enc, which maps the input image into a latent code z, a content decoder, Dec_C, that decodes z into a foreground content map, C^{fore}, and an attention decoder, Dec_A, that decodes z into the two attention maps, namely A^{fore} for foreground and A^{back} for background. Subsequently, the foreground content, C^{fore}, in our case the lesion-related regions, and attention map, A^{fore}, are fused to generate a foreground output, O^{fore}. The input image, x_H, and background attention, A^{back}, is fused to generate a background output, O^{back}. The summation of O^{fore} and O^{back} produces the generated synthetic image, $\hat{x}_P$. The pseudo-healthy synthesis branch, G_H, takes an image with lesions, x_P, as input and produces an pseudo-healthy image, $\hat{x}_H$. This branch is constructed in a similar way as the lesion generation branch.

4.1.3.1 Generators

The two generators, G_P and G_H, are both constructed using the ResNet architecture with nine blocks. Each generator features an encoder with three downsampling paths, and two decoders, one for generating content masks, and one for attention masks. For a given input

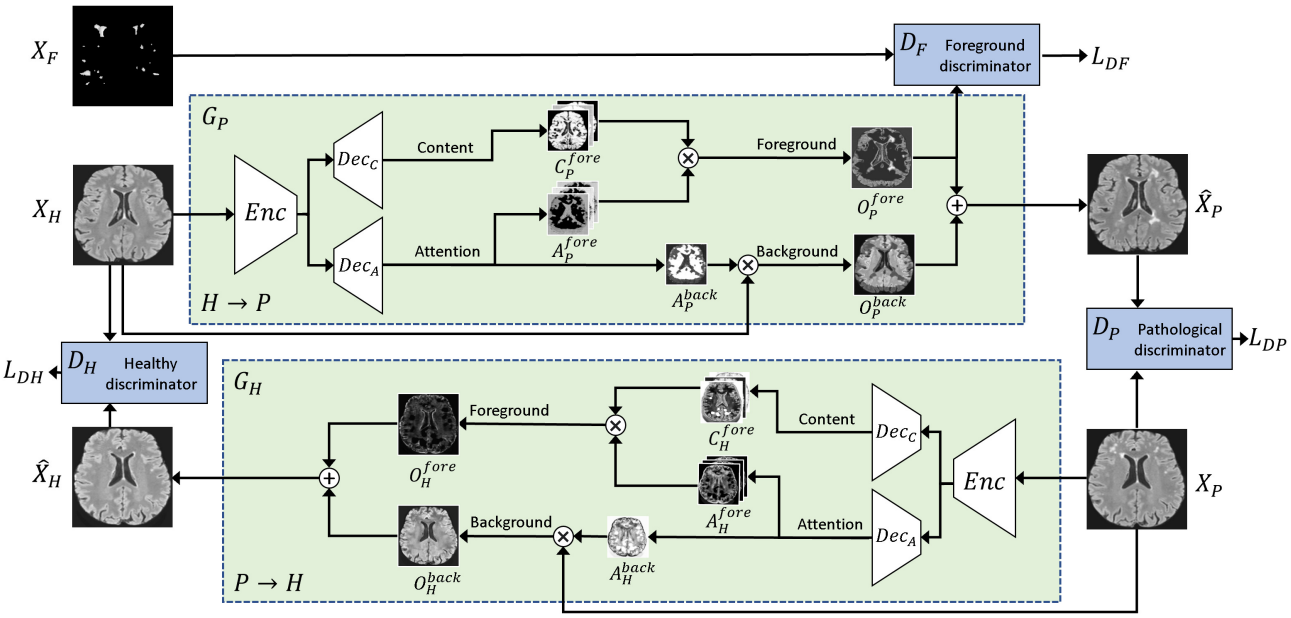


FIGURE 4.1: **Proposed framework for subject-specific lesion generation and pseudo-healthy synthesis of brain images**. Our model consists of two generators, G_P and G_H, and three discriminators, D_H, D_P, and D_F. We refer the reader to the text for more detail.

image, our model generates n attention masks, 1 for background attention, A^{back}, and $n-1$ for foreground attention, A^{fore}, and also generates $n-1$ foreground content masks, C^{fore}. Here foreground broadly refers to regions relevant to lesions and background refers to the rest of the brain image. Producing multiple foreground attention masks allows lesions of varying shape and location characteristics to be decoded, thus producing synthetic images with different lesion distributions. Similar ideas have been explored in [31], where an input image is disentangled into multiple channels representing different anatomical structures. The synthetic image $G(x)$ is generated by

$$G(x) = C^{fore} * A^{fore} + x * A^{back}, \tag{4.1}$$

which combines the foreground content, C^{fore}, multiplied with their respective attention masks, A^{fore}, and the original input image, x, multiplied with the background attention mask, A^{back}. The separation of foreground and background allows us to explicitly focus on the lesions in the generating process by developing a lesion-aware foreground discriminator.

4.1.3.2 Discriminators

To generate realistic brain images, three discriminators are introduced. The pathological discriminator, D_P, distinguishes synthetic images with lesions from real pathological images. The healthy discriminator, D_H, distinguishes pseudo-healthy images from real healthy subjects. The third discriminator, D_F, which is lesion-aware, encourages the foreground (synthetic lesions) to look realistic compared to real lesion masks in the pathological dataset.

4.1.3.3 Losses

The total generator loss, $\mathcal{L}_{G_{total}}$, consists of three loss terms: generation loss, $\mathcal{L}_G$, cycle-consistency loss, $\mathcal{L}_{CC}$, and identity loss, $\mathcal{L}_{idt}$. We employ the least squares GAN loss [129] for the generation loss, also known as $\mathcal{L}_2$ loss. For pseudo-healthy synthesis ($P \rightarrow H$), the generation loss is formulated as

$$\mathcal{L}_{G_H} = \mathbb{E}_{x \sim X_P}\left[(D_H(G_H(x)) - 1)^2\right], \tag{4.2}$$

where D_H denotes the healthy discriminator for the pseudo-healthy image $G_H(x)$. For lesion generation ($H \rightarrow P$), the generation loss is formulated as

$$\mathcal{L}_{G_P} = \mathbb{E}_{x \sim X_H}\left[\left(\tfrac{1}{2}(D_P(G_P(x)) + D_F(O_P^{fore})) - 1\right)^2\right], \tag{4.3}$$

where D_P denotes the pathological discriminator for the lesion image $G_P(x)$ and D_F denotes the foreground discriminator for the foreground output O_P^{fore}, which penalises non-realistic lesion regions.

The cycle-consistency loss, $\mathcal{L}_{CC}$, evaluates a full cycle of an image through the network, $P \rightarrow H \rightarrow P$ or $H \rightarrow P \rightarrow H$, against its original input. We encourage the output of these cycles to be equivalent to the input as we wish to preserve subject identity. This also ensures that the generated synthetic lesions do not appear at unrealistic locations in the brain structure. The cycle-consistency loss, $\mathcal{L}_{CC}$, is given by:

$$\mathcal{L}_{CC} = \mathbb{E}_{x \sim X_H}\left[\|\lambda_H G_H(G_P(x)) - x\|_1\right] + \mathbb{E}_{x \sim X_P}\left[\|\lambda_P G_P(G_H(x)) - x\|_1\right]. \tag{4.4}$$

Finally, we incorporate the identity loss, $\mathcal{L}_{idt}$, with the objective of getting an identical image out of a generator if its healthiness is not changed, such that $G_H(x_H)$ would return an image identical to x_H. We use the $\mathcal{L}_1$ loss for this task:

$$\mathcal{L}_{idt} = \mathbb{E}_{x\sim X_H}\left[\|\lambda_H\lambda_{idt}G_H(x) - x\|_1\right] + \mathbb{E}_{x\sim X_P}\left[\|\lambda_P\lambda_{idt}G_P(x) - x\|_1\right]. \tag{4.5}$$

We utilise hyperparameters λ_H, λ_P, and λ_{idt} for weighting these losses. The total loss for the generator is:

$$\mathcal{L}_{G_{total}} = \mathcal{L}_{G_H} + \mathcal{L}_{G_P} + \mathcal{L}_{CC} + \mathcal{L}_{idt}. \tag{4.6}$$

The discriminators aim to minimise the sum of the squared difference between predicted and expected values for real and synthetic images. We use the least squares loss for the three discriminators, D_H, D_P, D_F, given by:

$$\mathcal{L}_{D_H} = \mathbb{E}\left[(D_H(x_H) - 1)^2 + (D_H(G_H(x_P)))^2\right], \tag{4.7}$$

$$\mathcal{L}_{D_P} = \mathbb{E}\left[(D_P(x_P) - 1)^2 + (D_P(G_P(x_H)))^2\right], \tag{4.8}$$

$$\mathcal{L}_{D_F} = \mathbb{E}\left[(D_F(x_F) - 1)^2 + (D_F(O_P^{fore}))^2\right]. \tag{4.9}$$

4.1.4 Experiments

4.1.4.1 Data

A private dataset is used for training the proposed method, which consists of FLAIR MRI scans of 120 subjects, including 60 healthy scans and 60 scans with MS lesions. For evaluation, we train the segmentation networks using the 2016 Multiple Sclerosis Lesion Segmentation dataset (MS2016) [44], and test on the 2008 MICCAI MS Lesion Segmentation dataset (MS2008) [160] and 2015 ISBI Longitudinal MS Lesion Segmentation dataset (ISBI2015) [28]. We resample all images into $1 \times 1 \times 1$ mm^3 voxel spacing. We extract and select the centre axial slice of each image for use in the 2D model, as the central slice contains the ventricles where periventricular MS lesions often occur. These images are reshaped into 256×256 dimensions to pass through the network. The private, MS2008, ISBI2015, and MS2016 datasets provide 120, 20, 21, and 15 2D FLAIR images, respectively.

4.1.4.2 Evaluation

We evaluate the method performance in data augmentation using the synthetic data for downstream segmentation task as well as evaluate the image quality of synthetic data. To assess performance we conduct two experiments, employing our method as a data augmentation tool and comparing segmentation performance when trained on only synthetic data. To evaluate quality we ask three raters to assess healthiness and realism of generated images.

The proposed method is compared with traditional data augmentation (TDA) techniques which come default with nnU-Net [92], including rotation, scaling, mirroring, elastic deformation, intensity perturbation, and simulation of low resolution. We also compare to a state-of-the-art lesion-aware augmentation method, CarveMix [195]. We utilise offline versions of our method and CarveMix to double the training dataset for those experiments. When performing data augmentation using the proposed method, lesion masks are required for the foreground discriminator. After lesion generation, we perform free-form deformation registration of the healthy subject image, x_H, to the corresponding synthesised pathological image, $\hat{x}_P$, threshold the difference of the two images and generate the lesion masks. We report the mean and variance of the Dice scores and Hausdorff distances of the segmentations.

4.1.4.3 Implementation details

4.1.4.3.1 GAN Network

We implement the generative model with a single channel input and single channel output. We set the generated mask number n to 10, as our experiments show more variety of realistic lesion characteristics and intensities are synthesised with a higher n. Our generative framework method also features data augmentation in the form of sagittal mirroring and elastic deformation.

4.1.4.3.2 Training

The proposed method is developed on PyTorch and trained on one NVIDIA GeForce RTX 3080. We use the Adam optimizer with an initial learning rate of 0.001 and 0.5 dropout. The network is trained for 400 epochs and with a linear decay of the learning rate after

200 epochs with a batch size of 1, due to GPU memory limit. We set the generator loss trade-off hyperparameters λ_H, λ_P, and λ_{idt} to 10, 10, and 0.5, respectively. For evaluating segmentation performance, nnU-Net is trained for 1,000 epochs using stochastic gradient descent.

4.1.4.4 Results

4.1.4.4.1 Data Augmentation Performance

We train twelve separate segmentation models using different dataset sizes and data augmentation methods. We use nnU-Net in 3D full resolution, utilising 100%, 53.3%, 26.6% and 12.3% of the training dataset, corresponding to 15, 8, 4, and 2 subjects.

Table 4.1 indicates that the average Dice score is increased using the proposed method in three of the four dataset sizes when compared to TDA and CarveMix, while also decreasing the standard deviation in most cases. We further report the mean and variance of the Hausdorff distances in Table 4.2. Figure 4.2 visualises the segmentation performance of our method against the benchmark. Figure 4.3 demonstrates different stages of the proposed method and augmentation examples from CarveMix. [195].

TABLE 4.1: **Mean and standard deviations of the Dice scores (%), at different sizes of training data.** Best results are in bold. Asterisks indicate statistical significance (*: p$\leq$ 0.05, **: p $\leq$ 0.001) when using a paired Student's *t*-test compared to baselines.

Size	TDA		TDA+CarveMix[195]		TDA+Proposed	
	MS2008	ISBI2015	MS2008	ISBI2015	MS2008	ISBI2015
100%	$33.06^{**}_{17.37}$	$63.20_{16.70}$	$32.32^{**}_{15.23}$	$57.82_{16.56}$	$\mathbf{44.77_{15.87}}$	$\mathbf{63.90_{16.45}}$
53.3%	$34.03_{17.83}$	$58.84_{24.54}$	$35.39_{17.36}$	$61.25_{19.06}$	$\mathbf{39.70_{18.44}}$	$\mathbf{62.35_{13.21}}$
26.6%	$31.55_{18.62}$	$\mathbf{64.43_{12.64}}$	$\mathbf{39.21_{18.86}}$	$60.28_{21.22}$	$34.58_{15.47}$	$60.83_{10.03}$
13.3%	$28.96_{23.55}$	$37.34^{**}_{23.81}$	$23.57*_{23.18}$	$38.94^{**}_{21.81}$	$\mathbf{32.29_{16.16}}$	$\mathbf{56.64_{13.23}}$

The generator of the method is stochastic, which employs dropout to generate multiple diverse augmented images for each given sample, shown in Figure 4.4. This increases the diversity of training image samples.

4.1.4.4.2 Synthetic data segmentation performance

We compare segmentation performance when training a model using a fully synthetic dataset against a model trained on a real dataset. We generate 15 synthetic images from

TABLE 4.2: **Mean and standard deviations of the Hausdorff distances, at different sizes of training data.** Best results are in bold. Asterisks indicate statistical significance (*: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$) when using a paired Student's *t*-test compared to baselines.

Size	TDA		TDA+CarveMix [195]		TDA+Proposed	
	MS2008	ISBI2015	MS2008	ISBI2015	MS2008	ISBI2015
100%	$3.52^{*}_{1.61}$	$3.35^{*}_{0.94}$	$3.24_{1.09}$	$3.33_{0.89}$	$\mathbf{3.09_{1.54}}$	$\mathbf{3.02_{0.80}}$
53.3%	$3.16^{**}_{1.10}$	$3.12_{0.93}$	$3.30^{***}_{1.26}$	$3.28_{1.00}$	$\mathbf{2.62_{0.98}}$	$\mathbf{3.07_{1.08}}$
26.6%	$3.12_{0.89}$	$\mathbf{3.38_{1.01}}$	$\mathbf{3.08_{1.01}}$	$3.37_{0.94}$	$3.09_{0.72}$	$3.86_{1.61}$
13.3%	$2.92_{1.13}$	$3.86^{**}_{1.62}$	$2.95_{1.03}$	$3.81^{**}_{1.62}$	$\mathbf{2.90_{0.69}}$	$\mathbf{3.33_{1.17}}$

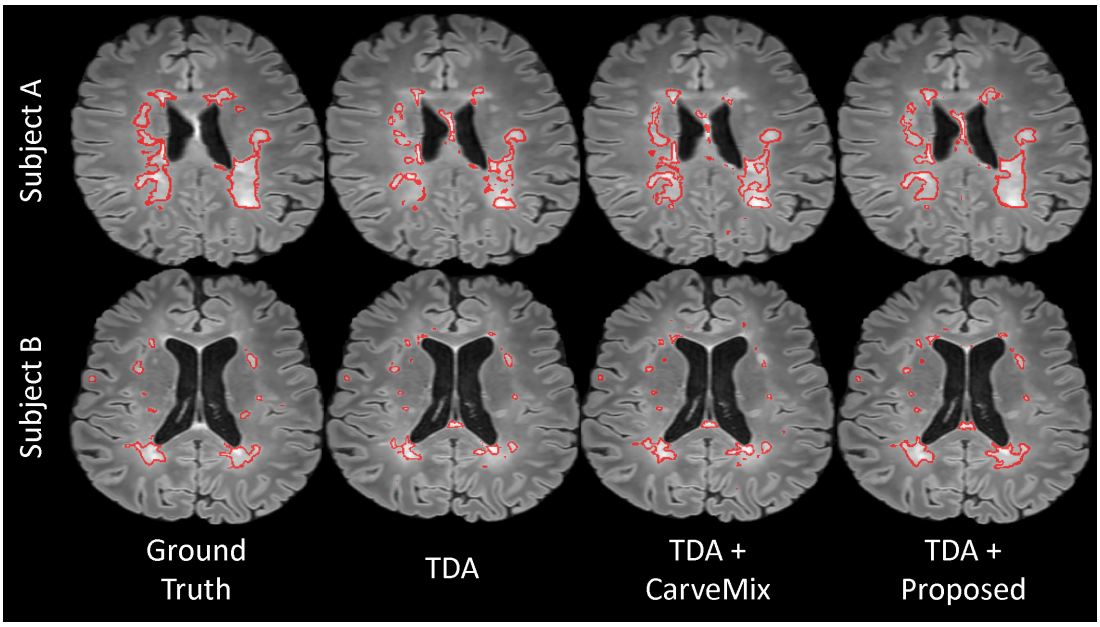


FIGURE 4.2: **Comparison of segmentation performance using different data augmentation techniques at 100% dataset usage.** Lesions are contoured in red. **Left to right:** Test subjects with ground truth lesion masks; segmentation using traditional data augmentation (TDA); segmentation using TDA+CarveMix [195]; segmentation using TDA+proposed.

the MS2016 dataset and test on the MS2008 and ISBI2015 datasets. We compare our results to nnU-Net with TDA at 100% dataset size, as per in Table 4.1. Results in Table 4.3 suggest a model trained on our generated data performs comparable to a model trained on a real dataset.

TABLE 4.3: **Segmentation performance using fully synthetic data for model training, compared to using real data, 15 training images for each.** Mean and standard deviations of the Dice scores (%) are reported with best results in bold.

Test data	Real training data	Synthetic training data
MS2008	$33.06_{17.37}$	$\mathbf{35.74_{19.85}}$
ISBI2015	$\mathbf{63.20_{16.70}}$	$52.72_{18.61}$

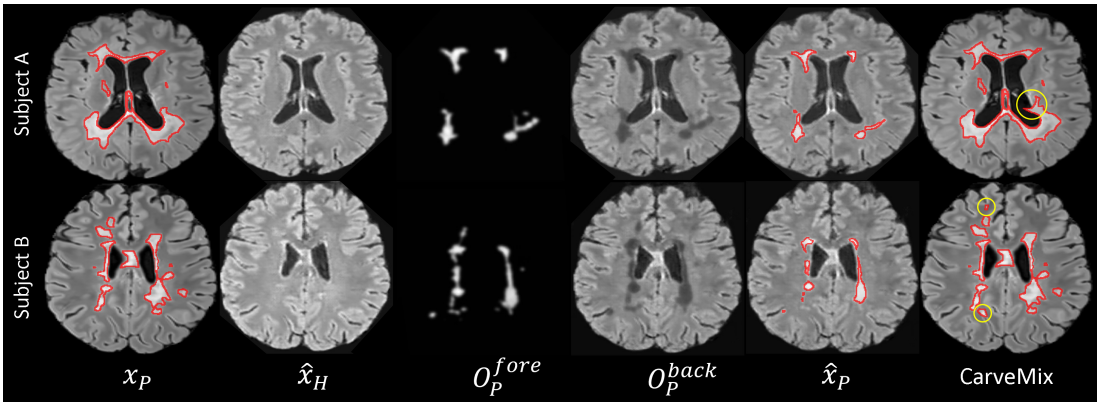


FIGURE 4.3: **Visualisations from different stages of training of the proposed method.** Lesions are contoured in red. **Left to right**: Real pathological image, x_P; pseudo-healthy image $\hat{x}_H$; foreground output for lesion synthesis on pseudo-healthy image O_P^{fore}; background output for lesion synthesis O_P^{back}; the synthetic pathological image generated from the pseudo-healthy image $\hat{x}_P$; augmented image using CarveMix, unrealistic lesions circled in yellow.

4.1.4.4.3 Structural similarity

We measure subject-identity of 60 images acquired from the MS2016 dataset after they have been through a inpainting-lesion generation cycle, $P \rightarrow H \rightarrow P$. We report a raw score, Si_{raw}, and a registered score Si_{reg}, where the synthetic image is first registered to the original image using free-form deformation [132]. For Si_{raw}, we achieve a mean score of 73.83 ± 12.82. For Si_{reg}, we achieve a mean score of 82.30 ± 11.16. Si_{raw} scores are lower than their registered counterparts due to the ventricular volume changing during the generation process. The structural similarity scores reported here are relatively high, which means the subject identity is preserved. However, we are not able to compare the structural similarity results to other methods. To the best of our knowledge, this is the first work which performs subject-specific brain lesion synthesis.

4.1.4.4.4 Expert Evaluation

In order to assess quality of generated images, we ask three human raters to evaluate images in healthiness and realism. We define "healthiness" as a brain MR image with no visible pathological features, and "realism" as an image which possesses the same quality and authentic look as one obtained by a scanner. We collate 50 real healthy, 50 real pathological, 50 pseudo-healthy, and 50 pseudo-pathological images. Raters are asked to classify each image as healthy (1) or pathological (0), and real (1) or fake (0). We produce the results

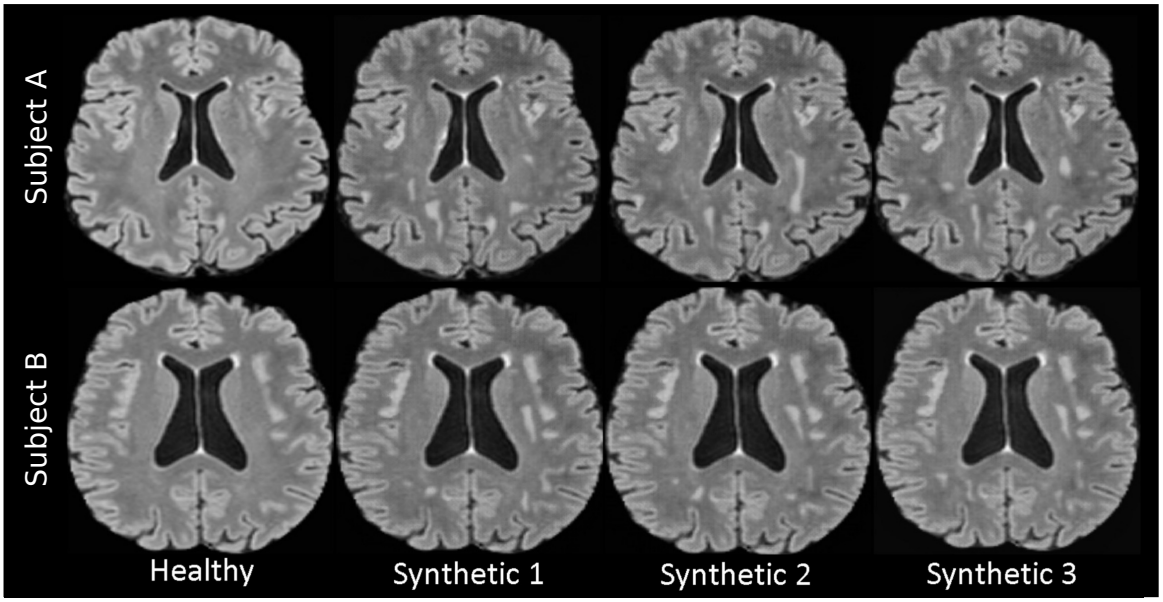


FIGURE 4.4: **Visualisations of stochasticity exemplified by the model**. Setting the generative network's dropout level to 0.5 at training and inference allows for generating a diverse set of synthetic images (second to fourth columns) given a single image as input (first column).

for the real images as a benchmark to compare our generated images. Table 4.4 shows that the generated pseudo-healthy and pseudo-pathological images closely follow real images for the assessed metrics. Regarding realism, 80% of our proposed healthy and pathological images were classified as "realistic" compared to the 85% and 84% of real healthy and real pathological images, respectively. Similarly, our generated healthy images were classified as healthy slightly behind real healthy images, 83% to 85%. Notably, raters classified generated pathological images as pathological (less healthy) more than real pathological images, 1% to 5%, respectively.

TABLE 4.4: **Rater classification of real and generated images on realism and healthiness.** Asterisk (*) indicates that a lower score is better.

Realism		
	Real	Proposed
Healthy images	85%	82%
Pathological images	84%	82%

Healthiness		
	Real	Proposed
Healthy images	85%	83%
Pathological images*	5%	1%

4.1.4.4.5 Correlations with Radiological Findings

The spatial heatmap of the synthetic lesions closely resemble those of the real lesions from clinical data, including the private dataset and the MS2016 dataset. A comparison between the spatial heatmap of synthetic lesions and real lesions is visualised in Figure 4.5.

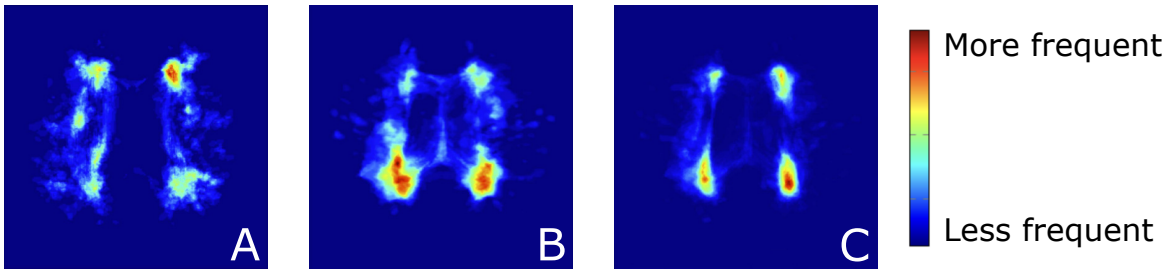


FIGURE 4.5: **Comparison of spatial heatmaps of real lesions and synthetic lesions. A:** Real lesions from private dataset (200 subjects). **B:** Real lesions in MS2016 dataset (60 subjects). **C**: Synthetic lesions generated using the proposed method by training on the MS2016 dataset (200 masks).

Furthermore, generated pseudo-pathological images show an increase in the ventricular volume, while pseudo-healthy images tend to have a decreased ventricular volume. The difference in ventricular volume is shown in Figure 4.6. The change in the ventricular volume is consistent with literature [52, 53, 81, 126].

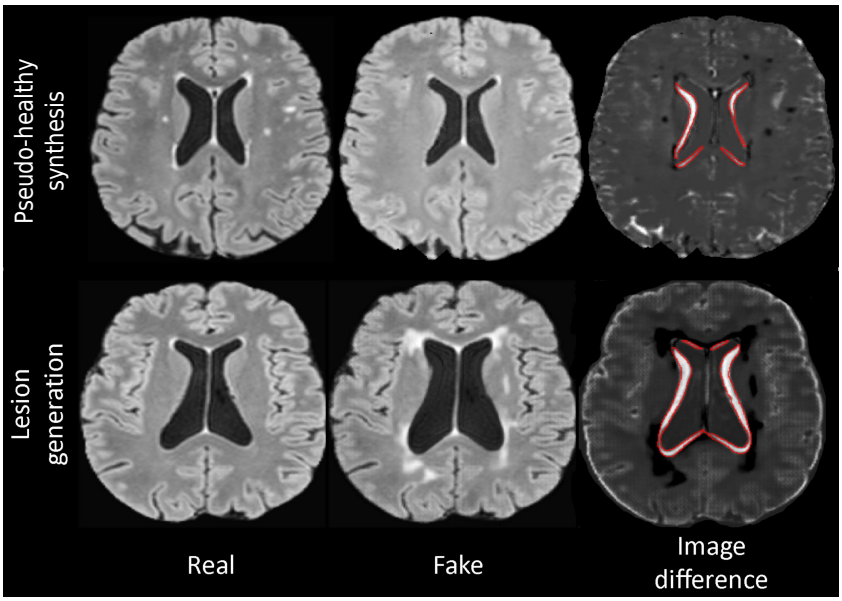


FIGURE 4.6: **Changes the ventricular volume during image generation, consistent with clinical literature. Top:** Pseudo-healthy synthesis leads to decrease in ventricular volume, delineated in red in the difference image. **Bottom:** Lesion generation leads to an increase in ventricular volume, delineated in red in the difference image.

4.1.5 Discussion and Conclusion

We propose a novel method for generating lesions and pseudo-healthy images while preserving subject-specific features, providing a useful tool for data augmentation in brain image analysis to complement real patient datasets. Our method is lesion-aware in the generating process, being able to distinguish lesion regions as the foreground. In quantitative evaluation, our method improves lesion segmentation performance in downstream tasks, showcasing statistically significant improvements in Dice score and Hausdorff distance especially in low data setting. In our human rater assessment, the generated images achieved a high realism score close to the real images, and generated lesions follow a spatial distribution consistent with real clinical datasets. Furthermore, we identified changes in the ventricular volume of generated images, which were inline with clinical literature.

Our method for adversarial data augmentation shows promising performance and high realism. However, it suffers from several limitations. Foremost, the usage of a two generative adversarial networks in the framework exacerbates the difficulties of training GANs, as the slight changes in hyperparameters can lead to instabilities during training. This nature causes obstructions for reproducibility. Furthermore, due to resource constraints, our method applied to two dimensional images, requiring the preprocessing of three dimensional brain MR images.

Despite these shortcomings, our framework can be extended to a multitude of medical applications. In this study, we generate lesions on healthy brain scans as diseased and healthy brain images, despite their scarcity, are easy to find. However, the framework can be extended to longitudinal image generation by replacing X_H with a unhealthy first timepoint scan, and X_P with a follow-up scan. Using the attention mechanism, we could then learn the possible patterns in the progression of lesions and predict future lesion formations. Alternatively, the novel foreground discriminator allows the model to control what features and substructures in the image to employ as the foreground, and specifically introduce or alter those features. As such, we could set it to alter other features, such as for modeling brain atrophy by decreasing white matter or gray matter in the brain.

4.2 A Lesion-Level Data Augmentation Method for Medical Image Segmentation

4.2.1 Introduction

Availability of labelled medical imaging data has been a long-term challenge for developing robust machine learning methods for medical image segmentation. In particular, when dealing with lesions or abnormality detection, datasets often follow a long-tail distribution [70, 149], which means there can be a variety of categories for abnormal cases but with each category only containing very few samples. In medical imaging, most data augmentation methods are developed at the image level, aiming to increase the diversity of the full image [40]. They often lack the capability to model specific abnormalities in the images. Recently, several disease-specific augmentation methods have been proposed for brain tumors, multiple sclerosis, and skin lesions [1, 11, 134]. Unfortunately, the majority of these methods are either disease or organ-specific, or are difficult to train and implement due to their complexity.

In this work, we propose LesionMix, a novel and simple lesion-level data augmentation method for medical image segmentation. LesionMix is able to populate lesions with various properties, including shape, location, intensity and lesion load, as well as inpaint existing lesions by using a dual-branch iterative 3D framework. With LesionMix, we are able to train lesion segmentation models in a low-data setting, even if there are very few samples of lesion images. We perform a comprehensive evaluation of LesionMix using different imaging modalities and datasets, including four brain MR lesion datasets and one liver CT lesion dataset. Experiments show that LesionMix achieves promising lesion segmentation performance on various datasets and outperforms several state-of-the-art (SOTA) data augmentation methods.

4.2.2 Contributions

There are three main contributions of this work: (1) We propose a novel non-deep data augmentation method, which augments images at the lesion level and accounts for lesion shape, location, intensity as well as load distribution. (2) The method is easy to implement

and can be added to a segmentation pipeline to complement traditional data augmentations. (3) It is generic and can be applied to datasets of various modalities (MRI, CT, etc).

4.2.3 Methods

The objective is to develop an efficient and easy-to-implement augmentation method that is aware of lesions in medical images, accounting for the spatial locations of lesions and the overall lesion load distribution within a dataset. Here, *lesion load* refers to the amount of lesion volume present in a particular dataset. This is an important factor to consider, given the fact that total lesion volume can significantly vary from dataset to dataset. Figure 4.7 illustrates the proposed LesionMix method, which consists of two branches, namely for lesion populating and lesion inpainting. LesionMix takes a lesion image, X, and its corresponding lesion mask, Y, as input, and generates an augmented lesion image, X', and lesion mask, Y', as output, which achieves a target lesion load, v_{tar}. If the target lesion load, v_{tar}, is greater than the current load, v_{cur}, the lesion load is increased via the populating branch. Otherwise, the lesion load is decreased via the inpainting branch. To generate diverse lesion samples, lesion-level augmentations are performed during populating. To maintain the fidelity of the samples, lesions are augmented according to learnt spatial and load distributions. LesionMix can be performed both on-the-fly and offline. We evaluate LesionMix offline for fair comparison to benchmark methods.

4.2.3.1 Lesion Populating

4.2.3.1.1 Lesion-level Augmentation.

Given the input image, X, and its lesion mask, Y, a lesion is randomly selected and augmented. We apply 3D spatial augmentations, brightness augmentations (multiplicative), and Gaussian noise augmentations. Lesion-level spatial augmentations include flipping, rotating, resizing, and elastic deformation. Augmentations are applied by extracting the selected 3D lesion volume, applying the augmentation, and inserting the augmented lesion back into the image. By iteratively inserting lesions into the images, augmented lesions can overlap one another, allowing for unique lesion formations. Augmentation parameters are set empirically and provided in Table 4.5.

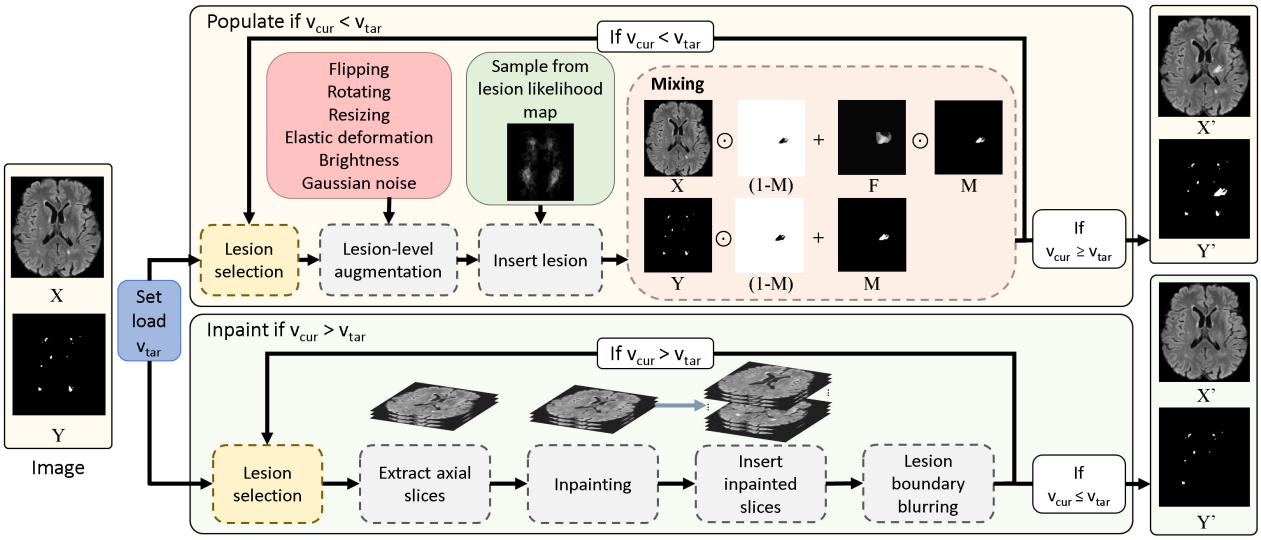


FIGURE 4.7: **Illustration of the augmentation process of LesionMix.** It consists of lesion populating (top) and lesion inpainting (bottom) branches to iteratively augment images to a desired lesion load.

TABLE 4.5: **Data augmentation parameters for LesionMix.** p denotes the probability of augmentation. For augmentations where a range is given, the parameter is determined by selecting a value by uniformly sampling from the range.

Augmentation	Details
Flipping	In X,Y,Z dimensions, $p = 0.5$ for each dimension
Rotating	In X,Y,Z dimensions, $p = 0.5$ for each dimension, range =[1°, 89°]
Resizing	Dimension multiplication, range=[0.5, 1.8]
Elastic deformation[1]	Random deformation grid, σ range= [3, 7]
Brightness	Intensity value multiplication, range = [0.9, 1.1]
Gaussian noise	Addition of $\mathcal{N}(0, 1)$
Inpainting	Fast marching method

The original image, lesion mask, and the augmented lesion region are mixed using the following equations to generate the augmented image and mask:

$$X' = X \odot (1 - M) + F \odot M, \tag{4.10}$$

$$Y' = Y \odot (1 - M) + M. \tag{4.11}$$

F denotes the augmented lesion intensity image, M denotes the mask of the augmented lesion region, and $\odot$ denotes element-wise multiplication. To generate X′ we use a soft mask M, in which the boundary pixels of the lesion mask are weighted by 0.66 and the inner pixels are weighted by 1. This allows the lesion boundary to blend more naturally with the

input image. We opt for this approach rather than using a signed distance function with a sigmoid due to the small size lesions may have. Our approach ensures that at most 1 voxel in the boundary is weighted. Lesion populating can be performed iteratively. At each iteration, lesion-level augmentation is applied to a randomly selected lesion and inserted into the image, until the lesion load, v_{cur}, reaches the target lesion load, v_{tar}.

4.2.3.1.2 Lesion Likelihood Map

The augmented lesion is inserted into the original image at a location sampled from a spatial heatmap, termed the lesion likelihood map, which describes the probability that a lesion appears at a specific spatial location in the anatomy. The map is learnt by summing the labels of the images to produce a lesion heatmap and normalising it into a probability map. The map is computed once for each organ dataset before model training. For brain datasets, augmented lesions can occur on both the white matter and gray matter. Although white matter lesions may be more common, gray matter lesions have been recorded in clinical literature [130].

4.2.3.2 Lesion Inpainting

Given the input image and lesion mask, a 3D lesion volume is randomly selected. 2D axial slices of the volume are inpainted using the fast marching method [165], which fills in the intensities within the lesion mask with neighbouring intensities from the normal region, formulated by

$$I(p) = \frac{\sum_{q \in N(p)} w(p,q)[I(q) + \nabla I(q)(p-q)]}{\sum_{q \in N(p)} w(p,q)}, \tag{4.12}$$

where I denotes the intensity, p denotes a pixel within the lesion mask, $q \in N(p)$ denotes pixels in the neighbourhood of q that belong to the normal region, $\nabla I(q)$ denotes the image gradient at q and $w(p,q)$ denotes a weighting function determined by the distance and direction from q to p [165]. After inpainting, we insert the inpainted slices back into the original image. 2D inpainting is implemented on axial slices of the lesions, due to the simplicity in implementation and fast computation. To ensure 3D continuity of the inpainted lesion, Gaussian blurring is performed on the boundary of inpainted lesions along all three dimensions,

$$\mathrm{X}' = G(f(\mathrm{X},\mathrm{M})) \odot \partial\mathrm{M} + f(\mathrm{X},\mathrm{M}) \odot (1 - \partial\mathrm{M}), \tag{4.13}$$

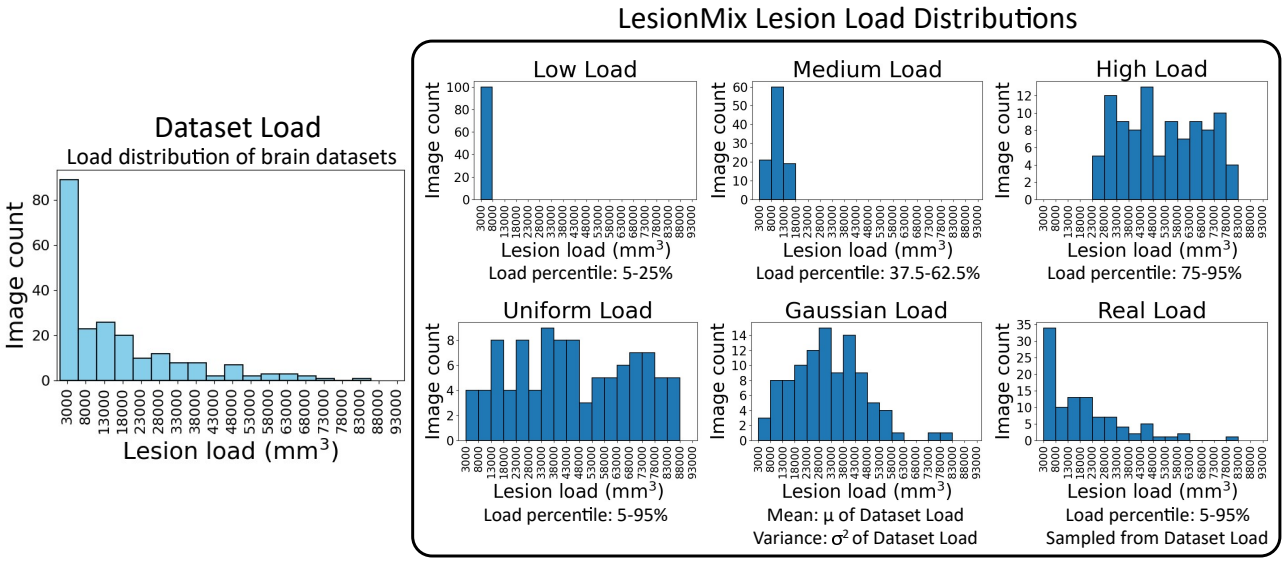


FIGURE 4.8: **Graphs of the different distributions of lesion loads.** The brain lesion datasets' lesion load distribution (light blue), named Dataset Load, and the six load distributions (dark blue) for augmentation used by LesionMix. Low load is a uniform distribution sampled between 5 and 25 percentiles of the Dataset Load. Medium load is a uniform distribution sampled between 37.5 and 62.5 percentiles of the Dataset Load. High load is a uniform distribution sampled between 75 and 95 percentiles of the Dataset Load. Uniform load is a uniform distribution sampled between 5 and 95 percentiles of the Dataset Load. Gaussian load samples from a Gaussian distribution with the same mean and variance as the Dataset load. Real load samples directly from the Dataset load distribution. This process is repeated for the LiTS dataset.

$$Y' = Y - M, \tag{4.14}$$

where $f(X, M)$ denotes the inpainting function using fast marching, G denotes the Gaussian blurring function, and ∂M denotes the boundary of the lesion mask. Lesion inpainting can be performed iteratively for randomly selected lesions, until the lesion load, v_{cur}, reaches the target lesion load, v_{tar}.

4.2.3.3 Lesion Load Distribution

Unlike other Mix-based methods, LesionMix allows for the generation of datasets with varying lesion load distribution. The load distribution, $P(v)$, is a probability distribution function for lesion volume, v. It characterises the degree of severity of the disease. We experiment with six different lesion load distributions: low, medium, high, uniform, Gaussian, and Real. Real denotes the real distribution learnt from the data. The other five are parametric distribution functions, with parameters described in Figure 4.8.

For each image to be augmented, we sample the target lesion load, v_{tar}, from the distribution and apply lesion populating or inpainting iteratively to achieve this target. If v_{tar} is lower than v_{cur}, lesion inpainting is applied; if v_{tar} is greater than v_{cur}, lesion populating is applied. We present examples of inpainting and populating as examples of LesionMix in low load and high load distribution setting examples, respectively, in Figure 4.9. We present the algorithm of LesionMix in Algorithm 1.

Algorithm 1: LesionMix: Lesion-level augmentation

Input Training images and annotations $\{(X_1, Y_1), ..., (X_N, Y_N)\}$; the desired number of augmented images T, the desired load distribution $P(v)$.
Output Augmented training data $\{(X'_1, Y'_1), ..., (X'_T, Y'_T)\}$

```
for t=1,2,...,T do
    Sample target load from the distribution, v_tar ~ P(v)
    if v_cur < v_tar then
        while v_cur < v_tar do
            1) Randomly select lesion from (X_i, Y_i)
            2) Sample lesion location from the lesion likelihood map
            3) Apply lesion-level augmentations and generate F and M
            4) Apply mixing in Eq. 4.10 and 4.11
        end
    else
        while v_cur > v_tar do
            1) Randomly select lesion from (X_i, Y_i) and extract axial slice
            2) Apply inpainting in Eq. 4.13 and 4.14 and reinsert slices
        end
    end
    return (X'_i, Y'_i)
end
```

4.2.3.4 Properties of LesionMix

We compare LesionMix with other Mix-based data augmentation methods, including CutMix [183], CarveMix [195] and SelfMix [143]. LesionMix offers greater control in augmentation, compared to CarveMix and SelfMix. LesionMix is spatially-aware, utilising the lesion likelihood map for drawing locations and thus mixes different backgrounds with the lesion. LesionMix performs shape and intensity augmentations at the lesion level, thus increasing sample diversity. Apart from populating lesions, it is also able to inpaint lesions and control the lesion load distribution. We further summarise in Table 4.6, and present a qualitative comparison against CutMix and CarveMix in Figure 4.10.

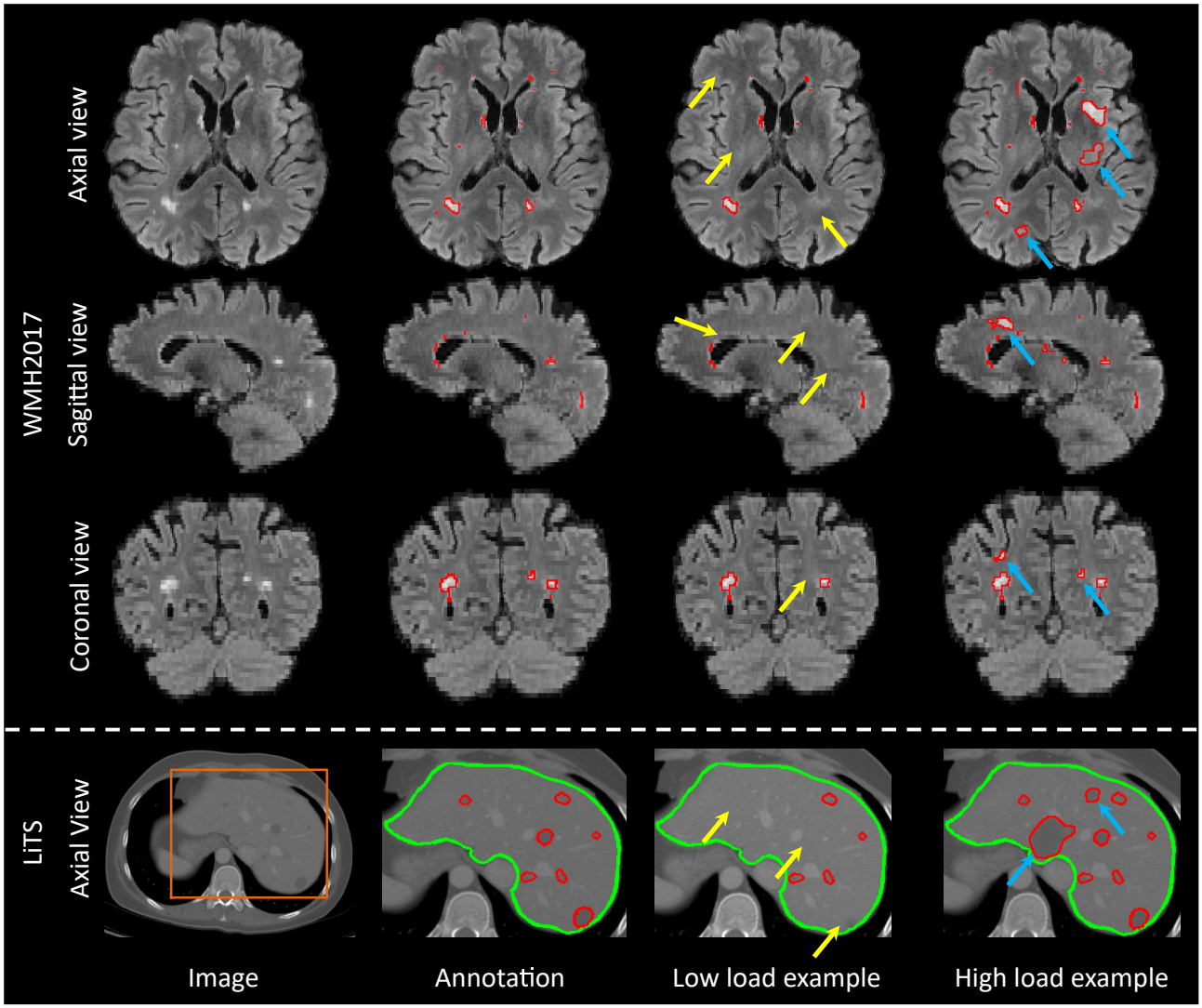


FIGURE 4.9: **Original image, annotation and augmented data by LesionMix with low and high load image examples for both brain (first three rows) and liver (fourth row) datasets.** Red denotes brain or liver lesions. Green denotes the liver. Low load example demonstrates performance of inpainted lesions, indicated by yellow arrows. High load example shows populated lesions, indicated by blue arrows.

TABLE 4.6: **Qualitative comparison of the properties of LesionMix with other Mix-based augmentation methods.** LesionMix exhibits more spatial and lesion-level augmentation properties compared to competitors.

Property	CutMix [25]	CarveMix [26]	SelfMix [27]	LesionMix
Lesion-aware		✓	✓	✓
Spatially-aware				✓
Lesion-background mixing			✓	✓
Lesion-level augmentation				✓
Lesion inpainting				✓
Lesion load control				✓

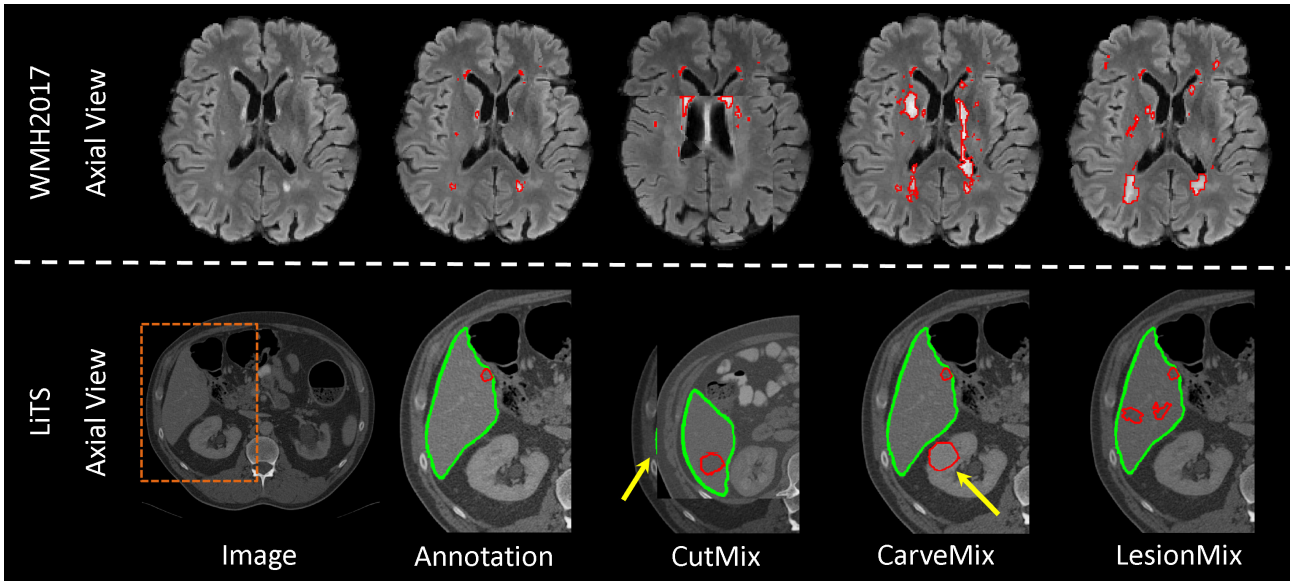


FIGURE 4.10: **Original image, annotation and augmented data by Mix-based methods**. Red delineations denote lesions, green denotes the liver organ. CutMix produces discontinuities in the image. CarveMix can place lesions outside the organ, indicated by arrows.

4.2.4 Experiments

4.2.4.1 Data

As a generic method for lesion data augmentation, LesionMix is evaluated on brain lesion MR images and liver lesion CT images.

4.2.4.1.1 Brain Lesion Data

Four brain lesion datasets are used, the MICCAI 2008 multiple sclerosis (MS) lesion dataset (MS2008, n=20) [160], ISBI 2015 longitudinal MS lesion dataset (MS2015, n=21) [28], MICCAI 2016 MS lesion dataset (MS2016, n=15) [44], and MICCAI 2017 white matter hyperintensity dataset (WMH2017, n_{train}=60, n_{test}=110) [108]. We use the WMH2017 training set for training a lesion segmentation network with the proposed augmentation method and evaluate its performance on the WMH2017 test set and MS2008, MS2015, MS2016 datasets. For all datasets, FLAIR images are used and resampled to $1 \times 1 \times 1$ mm^3 voxel spacing, followed by brain extraction using FSL [96] and rigid registration into the MNI space [66].

4.2.4.1.2 Liver Lesion Data

We use the MICCAI 2017 liver tumor segmentation dataset (LiTS) [20]. The training set contains CT scans for 131 subjects, which are split into batch 1 (n=28) and batch 2 (n=103) by the challenge organisers. We use the LiTS batch 1 dataset for training a liver lesion segmentation network and evaluate its performance on the LiTS batch 2 dataset. The in-plane image resolution ranges from 0.56mm to 1.0mm, and 0.45mm to 6.0mm in slice thickness. The LiTS dataset has high variance of data size and organ shape, therefore we use the normalised label map of the liver as the lesion likelihood map. This ensures the placed lesion is within the liver.

4.2.4.2 Implementation Details

The proposed method is developed on PyTorch. All augmentation methods are evaluated with the same segmentation model, nnU-Net with 3D full resolution configuration, and trained for 1,000 epochs on NVIDIA Tesla T4 GPUs.

4.2.4.3 Results

4.2.4.3.1 Lesion Load Distribution

We perform an ablation study to select the optimal lesion load distribution for LesionMix. We simulate a low-data setting by selecting just one training image from WMH2017, perform data augmentation using LesionMix to generate 100 augmented images, and train nnU-Net, a segmentation network. Table 4.7 reports the lesion segmentation performance when six different lesion load distributions are used, and compared against the *None* method, which is trained with a single image without augmentation.

4.2.4.3.2 Comparison to Other Data Augmentation Methods

Following the ablation study, we choose the uniform lesion load distribution for the remaining experiments. We compare LesionMix to SOTA data augmentation methods, including traditional data augmentations (TDA), which come default with nnU-Net [92], CutMix [183]

TABLE 4.7: **Mean and standard deviations of lesion segmentation Dice scores (%), when different lesion load distributions are used for LesionMix.** Trained using 1 image for None, and 100 generated images for the rest. Best results are in bold.

Test set	None	Low	Medium	High	Uniform	Gaussian	Real
MS2008	$16.46_{15.43}$	$22.90_{18.59}$	$22.23_{17.63}$	$23.57_{18.72}$	$\mathbf{24.07_{19.04}}$	$22.22_{17.59}$	$22.66_{17.84}$
MS2015	$37.58_{16.05}$	$\mathbf{40.51_{9.05}}$	$37.09_{12.20}$	$40.37_{15.64}$	$39.38_{15.68}$	$36.64_{14.34}$	$37.68_{14.21}$
MS2016	$24.35_{20.07}$	$36.75_{21.33}$	$38.10_{20.37}$	$49.08_{18.05}$	$\mathbf{50.72_{18.96}}$	$41.16_{20.04}$	$50.32_{16.43}$
WMH2017	$39.30_{25.16}$	$49.79_{24.88}$	$51.59_{23.80}$	$58.99_{21.26}$	$\mathbf{59.42_{21.25}}$	$53.23_{22.85}$	$55.24_{21.56}$
LiTS	$3.34_{4.94}$	$9.40_{6.25}$	$12.18_{10.34}$	$13.33_{5.94}$	$\mathbf{13.65_{8.02}}$	$12.88_{7.20}$	$11.99_{7.82}$

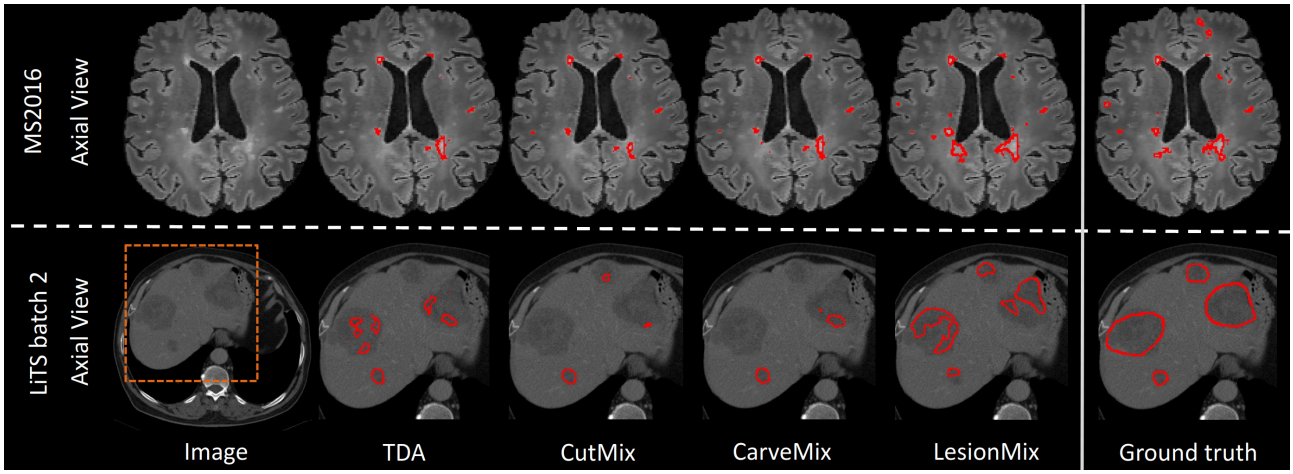


FIGURE 4.11: **Qualitative comparison of segmentation performance when 10% of dataset size is used.** Models with LesionMix detect more lesions and segment them more accurately.

and CarveMix [195]. TDA includes rotation, scaling, mirroring, elastic deformation, intensity perturbation and simulation of low resolution. We add CutMix, CarveMix, or the proposed LesionMix onto TDA. We re-implement CutMix [183] for 3D medical images, and use the public code for CarveMix [195]. We are unable to compare against SelfMix [143] due to unavailability of public code. For fair comparison, all methods use nnU-net as the segmentation network and augment the WMH2017 training set for brain lesions and the LiTS batch 1 dataset for liver lesions by five times. We conduct experiments when different sizes of the training data is used. Table 4.8 reports the lesion segmentation Dice scores for different data augmentation methods. LesionMix improves lesion segmentation against SOTA methods in the majority of experiments. We notice greater statistical significance in experiments with smaller dataset sizes. We present example segmentations against benchmark methods in Figure 4.11.

TABLE 4.8: **Mean and standard deviations of lesion segmentation Dice scores (%), at different sizes of training data.** Best results are in bold. Asterisks indicate statistical significance (*: p$\leq$ 0.05, **: p $\leq$ 0.01, ***: p $\leq$ 0.005) when using a paired Student's *t*-test comparing LesionMix's performance to baseline methods.

Size	Test set	TDA[92]	CutMix[183]	CarveMix[195]	LesionMix
100%	MS2008	$36.72^{*}_{18.07}$	$37.67_{19.09}$	$36.69^{*}_{17.82}$	$\mathbf{38.30_{17.67}}$
	MS2015	$71.59_{11.14}$	$\mathbf{72.81_{7.53}}$	$71.33_{9.96}$	$72.33_{11.41}$
	MS2016	$57.85_{17.50}$	$63.22_{15.04}$	$\mathbf{65.85_{14.26}}$	$65.62_{14.56}$
	WMH2017	$79.13_{10.59}$	$80.14_{9.69}$	$79.74_{9.94}$	$\mathbf{80.95_{9.45}}$
	LiTS	$61.93_{24.30}$	$58.39^{*}_{29.40}$	$60.20_{29.04}$	$\mathbf{63.51_{24.97}}$
50%	MS2008	$35.11_{21.16}$	$35.83_{19.55}$	$32.71^{*}_{19.95}$	$\mathbf{36.04_{18.93}}$
	MS2015	$70.59_{11.14}$	$71.77_{7.16}$	$67.44^{**}_{9.78}$	$\mathbf{71.82_{7.40}}$
	MS2016	$55.75^{***}_{17.82}$	$57.90^{**}_{14.24}$	$62.27_{16.64}$	$\mathbf{62.64_{16.26}}$
	WMH2017	$73.65_{17.30}$	$74.26_{12.69}$	$72.45^{*}_{18.25}$	$\mathbf{75.65_{17.60}}$
	LiTS	$52.40_{29.21}$	$49.60^{*}_{31.32}$	$51.98_{27.99}$	$\mathbf{52.59_{29.18}}$
25%	MS2008	$34.35_{21.49}$	$33.86_{20.21}$	$30.64^{*}_{21.17}$	$\mathbf{34.51_{20.04}}$
	MS2015	$66.09^{**}_{8.82}$	$67.69^{*}_{6.10}$	$67.12^{*}_{8.13}$	$\mathbf{70.73_{8.20}}$
	MS2016	$55.72^{**}_{18.78}$	$58.24_{15.67}$	$60.64_{17.67}$	$\mathbf{60.94_{16.66}}$
	WMH2017	$71.50^{*}_{17.88}$	$72.02_{14.29}$	$72.12_{16.95}$	$\mathbf{73.92_{15.81}}$
	LiTS	$29.88^{***}_{26.52}$	$36.35^{*}_{30.12}$	$36.77_{29.45}$	$\mathbf{39.25_{30.01}}$
10%	MS2008	$28.77^{**}_{17.77}$	$31.32_{18.28}$	$29.57^{*}_{21.66}$	$\mathbf{32.28_{22.11}}$
	MS2015	$63.97^{**}_{9.97}$	$64.97^{*}_{10.15}$	$58.10^{***}_{13.35}$	$\mathbf{67.13_{10.73}}$
	MS2016	$43.07^{***}_{17.03}$	$49.58^{***}_{13.65}$	$50.34^{***}_{13.71}$	$\mathbf{61.02_{15.56}}$
	WMH2017	$71.16_{16.70}$	$70.18^{*}_{14.79}$	$67.80^{**}_{19.97}$	$\mathbf{72.03_{15.48}}$
	LiTS	$24.23^{***}_{25.76}$	$27.99^{*}_{29.26}$	$20.53^{***}_{26.84}$	$\mathbf{30.12_{27.47}}$

4.2.5 Discussion and Conclusion

We present LesionMix, a simple lesion-level data augmentation method which is capable producing diverse lesion distributions even when only one image is provided. LesionMix is aware of the lesion likelihood distribution and augments lesions with respect to that spatial distribution. Furthermore, our framework can produce augmented data with varying lesion load. LesionMix improves segmentation performance against other mix-based augmentation methods across datasets of different modalities and organs. Our method is modality- and organ-agnostic and can serve as a useful tool for medical image segmentation.

LesionMix improves on other mix-based augmentation methods with its increased attention to organ and disease-specific features, such as lesion and organ awareness, and regard for spatial distribution of lesions. However, our framework can be improved to generate even more anatomically aware augmented samples. For example, multiple sclerosis lesions

occur more frequently in the white matter region of the brain. This can be used as a condition for lesion population after using tools such as FSL [96] or SynthSeg [21] to extract the white matter region. However, the required computation for each subject would come at a cost of increasing the duration of the augmentation process.

An important finding of this work is the study on different lesion load distributions. Augmenting data while keeping in mind of the effect on load distribution and therefore class balance is a novel attribute of our work. We draw attention to Table 4.7, and see that a dataset which has uniform distribution of lesion volumes outperforms others across the majority of datasets, including the real distribution of the original data. Our work sets an example on a possible avenue that can be taken for reducing the effect of the class imbalance problem in medical imaging tasks.

Chapter 5

Learning from Heterogeneous Data

This chapter contains material from:

1. Basaran, B.D., Zhang, X., Matthews, P.M., Bai, W., 2024, October, SegHeD: Segmentation of Heterogeneous Data for Multiple Sclerosis Lesions with Anatomical Constraints. In International Workshop on Longitudinal Disease Tracking and Modelling with Medical Images and Data (LDTM). Cham: Springer International Publishing.

2. Basaran, B.D., Matthews, P.M. and Bai, W., 2024. SegHeD+: Segmentation of Heterogeneous Data for Multiple Sclerosis Lesions with Anatomical Constraints and Lesion-aware Augmentation. arXiv preprint arXiv:2412.10946.

5.1 Introduction

The accurate identification and measurement of MS lesions in brain MR images is essential for clinical diagnosis and research. While many lesion segmentation methods have been proposed in recent years, such as those based on machine learning, they are often trained on well-curated datasets with a consistent data format [28, 43, 44]. Clinical studies on MS involve gathering imaging data from different sources, resulting in a variety of dataset styles, being cross-sectional or longitudinal, or segmenting specific types of lesions. This diversity makes it challenging to use off-the-shelf segmentation models effectively [82, 92]. There is a lack of specialised approaches designed to handle lesion segmentation when dealing with such diverse dataset types and annotations. Efforts to bridge this gap are crucial for advancing the performance and applicability of lesion segmentation models in MS research and

clinical practice.

In this study, we introduce SegHeD, a novel brain lesion segmentation model that learns from multiple datasets and for multiple tasks simultaneously, leveraging heterogeneous data for model training. SegHeD enables the incorporation of information from both single-timepoint (cross-sectional) and multiple-timepoint (longitudinal) images, as well as from annotations for various types of lesions, including all, new, and vanishing lesions [142]. We highlight the importance of the overlooked segmentation task of vanishing lesions, as they can be a critical biomarker for differentiating MS from other neurodegenerative diseases [8, 84], and provide insight into the biological processes of MS at different age groups [29]. Our model integrates temporal and volumetric constraints, along with anatomical constraints, to improve the plausibility of the segmentated lesions. In model training, we incorporate a lesion-level data augmentation method, which allows the generation of synthetic lesion images to enlarge the training set. Experimental results demonstrate that SegHeD achieves competitive performance compared to current state-of-the-art (SOTA) approaches for MS lesion segmentation.

5.1.1 Related Work

5.1.1.1 Longitudinal Lesion Segmentation

Several advancements have been made in machine learning techniques for segmenting cross-sectional brain lesion images [15, 99]. However, longitudinal lesion segmentation, which capitalises on temporal information within longitudinal imaging data, remains relatively underexplored. Elliott et al. utilises the difference map between two timepoints to identify new lesions at the second time point [63]. Jain et al. also leverages the difference map and devises an expectation maximisation framework to jointly segment lesions at both timepoints [95]. Denner et al. incorporates a displacement field to capture spatiotemporal changes between timepoints [54]. Additionally, we propose our new lesion segmentation method using an nnU-Net model [92] in Chapter 3.

5.1.1.2 Learning from Heterogeneous Data

Publicly available datasets for brain lesion segmentation come in various styles. Some datasets are cross-sectional, consisting of scans from a single timepoint, while others are

longitudinal, including scans from two or more timepoints. This diversity poses a significant challenge when developing a universal machine learning model that can accommodate different data formats. Wu et al. suggests learning from diverse data for both all-lesion and new-lesion segmentation by incorporating relation regularisation into the prediction of all lesions [178]. Liu et al. develops a multi-organ segmentation model by training it on 13 different organ datasets and integrating CLIP-inspired label encoding techniques [120]. Shi et al. introduces a marginal loss and an exclusion loss to train a unified multi-organ segmentation model using multiple partially annotated datasets [157]. Iglesias et al. proposes a generative label fusion algorithm that unifies atlases with different labelling protocols, enabling accurate automatic segmentation and the creation of new labels by leveraging overlaps in underlying annotations [91]. With the increasing amount of data and available resources, training a universal model from multiple heterogeneous datasets is becoming more and more popular whilst showcasing impressive results [26, 118].

5.1.1.3 Enforcing Anatomical Plausibility

The incorporation of anatomical plausibility within machine learning models enhances the reliability and trustworthiness of the trained models for clinical use. Anatomical constraints have been introduced into machine learning frameworks in a variety of techniques. Dalca et al. introduces a model incorporating a variational auto-encoder to acquire location-specific priors for segmenting brain structures [51]. Strumia et al. restricts lesion segmentation to the white matter by leveraging a geometric brain model [127]. On the other hand, Agn et al. proposes priors on brain tumuors using convolutional restricted Boltzmann machines for accurate brain tumour segmentation [2]. Lastly, Hirsh et al. utilises a multi-prior network alongside tissue probability maps for segmenting MRI head anatomy [86]. Producing anatomically plausible results is critical in ensuring the minimisation of false positives, and promotes the clinical adoption of the developed models.

5.1.2 Contributions

In this work, we propose SegHeD, a multi-task segmentation model that can segment all, new, and vanishing lesions by learning from heterogeneous datasets while leveraging

anatomical constraints and lesion-aware data augmentation. The proposed model demonstrates promising results in all and new lesion segmentation, and sets a benchmark for the overlooked task of vanishing lesion segmentation. The main contributions in this chapter are summarised as follows:

- We offer a comprehensive framework that takes into consideration diverse data types and annotation protocols, accounting for cross-sectional and longitudinal images, and all, new, and vanishing lesion labels.
- We tackle various tasks concurrently, while most state-of-the-art brain lesion segmentation methods take a homogeneous segmentation approach, segmenting a single type of lesions. In addition to all and new lesions, we also account for vanishing (disappearing) lesions, an aspect overlooked in previous studies.
- We incorporate domain knowledge about MS lesion segmentation via a novel combination of losses, accounting for the longitudinal relation between newly forming and vanishing lesions between timepoints, the volumetric change of lesions, and constraining lesion segmentation to anatomically plausible regions.
- We utilise a lesion-level augmentation method to expand the training dataset, which is able to augment both cross-sectional and longitudinal images with new label availabilities.

5.2 Methods

5.2.1 Overall Architecture

SegHeD+ is a versatile model designed to learn from heterogeneous MS imaging data. Regarding data heterogeneity, it considers two types of image dataset styles: cross-sectional and longitudinal. Regarding label heterogeneity, the model accommodates three annotation protocols: all-lesion annotation, new-lesion annotation (which annotates only new lesions in the second scan), and vanishing-lesion annotation (which annotates lesions that disappear in the second scan). The latter two protocols concentrate on the longitudinal changes of lesions and refrain from annotating all lesions to optimize time efficiency. SegHeD+ is

designed to tackle all three annotation tasks simultaneously. The overall framework of the method is depicted in Figure 5.1.

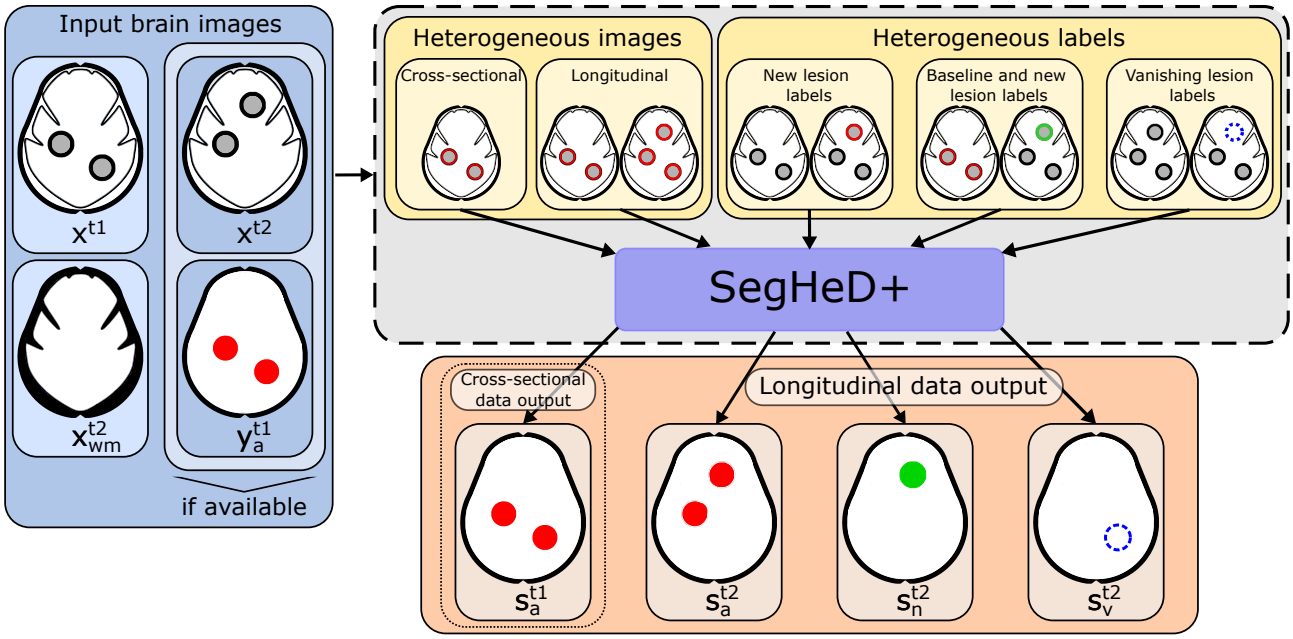


FIGURE 5.1: Visualisation of the proposed framework. SegHeD+ learns from heterogeneous datasets varying in image and label formats. SegHeD+ takes up to four inputs, and can analyse cross-sectional and longitudinal data. Four segmentation heads provide binary-class segmentation to all lesions (red) in two timepoints, and new (green) and vanishing (dashed blue) lesions in a second timepoint.

We make the assumption that the imaging data consists of two timepoints. SegHeD+ requires 4 input channels: a scan at the first timepoint (baseline), x^{t1}, a scan at the second timepoint (follow-up), x^{t2}, a label map of all lesions for the first timepoint scan, y_a^{t1}, and a white matter mask for the second timepoint scan, x_{wm}^{t2}. In cases of heterogeneous datasets, not all of the aforementioned inputs may be present. If data for the second timepoint is missing, the first timepoint data is duplicated as the second, $x^{t2} = x^{t1}$. If the label map for all lesions at the first timepoint is unavailable, y_a^{t1} is substituted with a zero matrix of the same dimensions as x^{t1}. The white matter mask for the second timepoint, x_{wm}^{t2}, can be acquired using a pre-trained brain structure segmentation tool, such as SynthSeg, which is robust in white matter segmentation even in the presence of lesions [21]. SegHeD+ is trained to produce 4 outputs: segmentation of all lesions at the first timepoint, s_a^{t1}, segmentation of all lesions at the second timepoint, s_a^{t2}, segmentation of new lesions at the second timepoint, s_n^{t2}, and segmentation of vanishing lesions at the second timepoint, s_v^{t2}. The segmentation

process is formulated by

$$s_a^{t1}, s_a^{t2}, s_n^{t2}, s_v^{t2} = F(x^{t1}, x^{t2}, y_a^{t1}, x_{wm}^{t2}), \tag{5.1}$$

where F represents the SegHeD+ model. The all-lesion map, y_a^{t1}, is not utilised in cases of cross-sectional data input. It is only input, when available, to provide temporal context for second timepoint lesion prediction, and not used for predicting s_a^{t1}.

Due to computational constraints, the model handles a maximum of two timepoints as inputs. In cases where a subject has longitudinal scans with more than two timepoints [**?**], a sliding window technique is applied across the timepoints during training and inference. For example, for case with three timepoints, [t1, t2, t3], we successively provide two timepoints as input to the model, formulated as [t1, t1], [t1, t2], [t2, t3]. The model first processes images at [t1, t1], i.e. using two channels of identical images as input. After the segmentation output for t1 is computed, the model processes images at [t1, t2], so that the previous timepoint t1 can provide temporal information in processing t2. After the output for t2 is computed, the model further processes two channels of images at [t2, t3]. During inference for longitudinal images, the all-lesion segmentation of a timepoint, for example s_a^{t1}, is used as an input channel, representing y_a^{t1}, for temporal guidance to obtain the second timepoint all-lesion segmentation s_a^{t2}. SegHeD+ utilises a 3D V-Net architecture [131] and a combined training loss function that incorporates the Dice loss, a longitudinal constraint loss, a spatial constraint loss, and a volumetric constraint loss.

5.2.2 Anatomical constraints

SegHeD+ is trained using a mix of loss functions, which we term as *anatomical constraints*. This approach is inspired by the way radiologists examine longitudinal images to detect lesions [88]. Anatomical constraints formulate the relationships in terms of time, space, and volume for segmenting brain lesions at two different time points.

5.2.2.1 Longitudinal constraints

We utilise the existing information embedded in the lesion annotation protocol: (1) Predicted new-lesions at the second timepoint, denoted as s_n^{t2}, should not exist in the all-lesion label at

the initial timepoint, y_a^{t1}, but must be included in the all-lesion label at the second timepoint, y_a^{t2}. (2) Predicted disappearing lesions at the second timepoint, s_v^{t2}, must be present in the all-lesion label at the first timepoint, y_a^{t1}, but should not be present in the all-lesion label at the second timepoint, y_a^{t2}. These longitudinal constraints are formulated using a mean squared error loss,

$$\begin{aligned}\mathcal{L}_{Long} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \|y_{a_i}^{t1} \wedge p_{n_i}^{t2}\|^2 + \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \|y_{a_i}^{t2} \oplus p_{n_i}^{t2}\|^2 + \\ \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \|y_{a_i}^{t1} \oplus p_{v_i}^{t2}\|^2 + \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \|y_{a_i}^{t2} \wedge p_{v_i}^{t2}\|^2,\end{aligned} \tag{5.2}$$

where N represents the total number of training images, i denotes the index of the image, $\wedge$ represents the element-wise logical AND operation, $\oplus$ denotes the element-wise logical XOR operation, and $||\cdot||^2$ represents the squared L2-norm of a flattened image. The superscripts, [t1, t2], for the prediction, p, and label, y, denote the two timepoints used by the sliding window. In contrast to the regularisation proposed in [178], we apply longitudinal constraints to predictions of new and vanishing lesions rather than all-lesion predictions. In cases where datasets are cross-sectional, longitudinal constraints are not enforced.

5.2.2.2 Volumetric constraints

Changes in brain lesion volumes evolve over time. Lesions can grow, shrink, appear, and/or vanish. Both change in lesion volume and lesion count provide important biomarkers for disease progression [65, **?**, **?**]. Unfortunately, a subject with MS can have both hundreds of small lesions or few large lesions, showing high variation in lesion count per subject and over time. This variation can be addressed by measuring total lesion volume change, also reducing the need for three-dimensional connected component analysis during training. To address the correlation between lesion volumes at two different timepoints, a volumetric constraint loss is proposed. This loss penalises significant discrepancies in lesion volume. It permits variations in lesion volume to either rise or fall [155, 72, 65] within specified thresholds, denoted as α_{high} and α_{low}. The penalty is enforced only when the volume alteration

exceeds these predefined thresholds. The volumetric constraint loss is formulated as

$$\mathcal{L}_{Vol} = \begin{cases} \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} (V_{a_i}^{t2} - \alpha_{high} \cdot V_{a_i}^{t1})^2 & \text{if } V_{a_i}^{t2} \geq \alpha_{high} \cdot V_{a_i}^{t1} \\ \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} (V_{a_i}^{t2} - \alpha_{low} \cdot V_{a_i}^{t1})^2 & \text{if } V_{a_i}^{t2} \leq \alpha_{low} \cdot V_{a_i}^{t1} \\ 0 & \text{otherwise,} \end{cases} \tag{5.3}$$

where $V_{a_i}^{t1}$ and $V_{a_i}^{t2}$ represent the total volume of lesions at the first timepoint and second timepoint for the i-th image, respectively. Lesion volumes are calculated by taking a sum of the lesion maps, as 0's represent healthy, and 1's represent lesion tissues. To our knowledge, there is no research on the extent to which lesion volume might *decrease* over time. We determine α_{low} to be 0.8 based on an examination of the largest reduction in lesion volume in a longitudinal MS dataset [**?**]. Likewise, using the same MS dataset, we established α_{high} to be 1.2. These values are in line with existing medical literature [135] for annual rate of white matter lesion change. Therefore, values α_{high} and α_{low} are defined for the annual rate of change, and can be modified when using for subjects that are scanned at a non-annual rate. A sensitivity study on α_{high} and α_{low} supported setting these hyperparameters to 1.2 and 0.8, respectively. When working with a cross-sectional dataset, as input $x^{t2} = x^{t1}$, we expect outputs s_a^{t1} and s_a^{t2} to be equal. As such, in cases of cross-sectional data input, we set α_{high} and α_{low} to 1 in order to guide the network to generate identical outputs.

5.2.2.3 Spatial constraint

MS affects the central nervous system of the body, where lesions may present themselves in the spinal cord, white matter, and rarely in the gray matter of the brain [**?**], alongside leading to gray matter atrophy. Studies demonstrate cognitive impairment in neurodegenerative diseases is significantly associated to white matter lesions in the brain [**?**]. As such, we do not consider gray matter atrophy, or lesions in the spinal cord during training or inference. Additionally, the spinal cord is excluded in most MS datasets. White matter lesions appear as hyperintense regions on FLAIR MR images [110]. To integrate this information into our framework, we adopt two approaches: first, by utilizing the white matter mask, x_{wm}^{t2}, as an input channel for SegHeD+ to benefit from the information of anatomical structure; and secondly, by creating a spatial relation loss function, where we further penalises false positive

segmentations which lie outside of the white matter region. This function is defined as the mean squared error between the segmented lesions, s, which are located outside the white matter, x_{wm}^{t2}, and a zero matrix of the same dimensions as the image:

$$\mathcal{L}_{Spat} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \| p \oplus (\mathbf{1} - x_{wm_i}^{t2}) \|^2. \tag{5.4}$$

The total loss, $\mathcal{L}$, is defined as

$$\mathcal{L} = \begin{cases} \mathcal{L}_{Dice} & \text{if } n < \dfrac{n_{epochs}}{2}, \\ \mathcal{L}_{Dice} + \lambda_L \cdot \mathcal{L}_{Long} + \lambda_V \cdot \mathcal{L}_{Vol} + \lambda_S \cdot \mathcal{L}_{Spat} & \text{if } n \geq \dfrac{n_{epochs}}{2}, \end{cases} \tag{5.5}$$

where $\mathcal{L}_{Dice}$ represents the Dice loss, n stands for the epoch, n_{epochs} represents the total number of epochs for training, and λ_{Long}, λ_{Vol}, and λ_{Spat} are parameters used for weighting. We employ a curriculum learning strategy [18] and introduce the complete loss after half the number of epochs.

5.2.3 Lesion-aware data augmentation

Apart from learning from heterogenous datasets, we also enlarge the training set and lesion diversity by performing data augmentation. On top of traditional data augmentation techniques, we employ LesionMix, our lesion-aware data augmentation method introduced in Section 4.2.

LesionMix was originally proposed for augmenting cross-sectional brain images. Here we employ it to augment both cross-sectional and longitudinal brain images. LesionMix complements SegHeD+ in three manners. First, it can generate both new lesions and vanishing lesions, thus increasing the diversity of the the various labels SegHeD+ can analyse in heterogeneous datasets. This both increases dataset variance, and reduces the effect of label imbalances which may occur from one type (new lesions) appearing more than another one (vanishing lesions). Second, its lesion populating and inpainting strategy along with lesion mask output allows for synthetic generation of longitudinal data. Lesions can be populated at the second timepoint as new lesions, or inpainted as vanishing lesions. Finally, LesionMix operates by populating or inpainting lesions to a predetermined lesion load, thus can

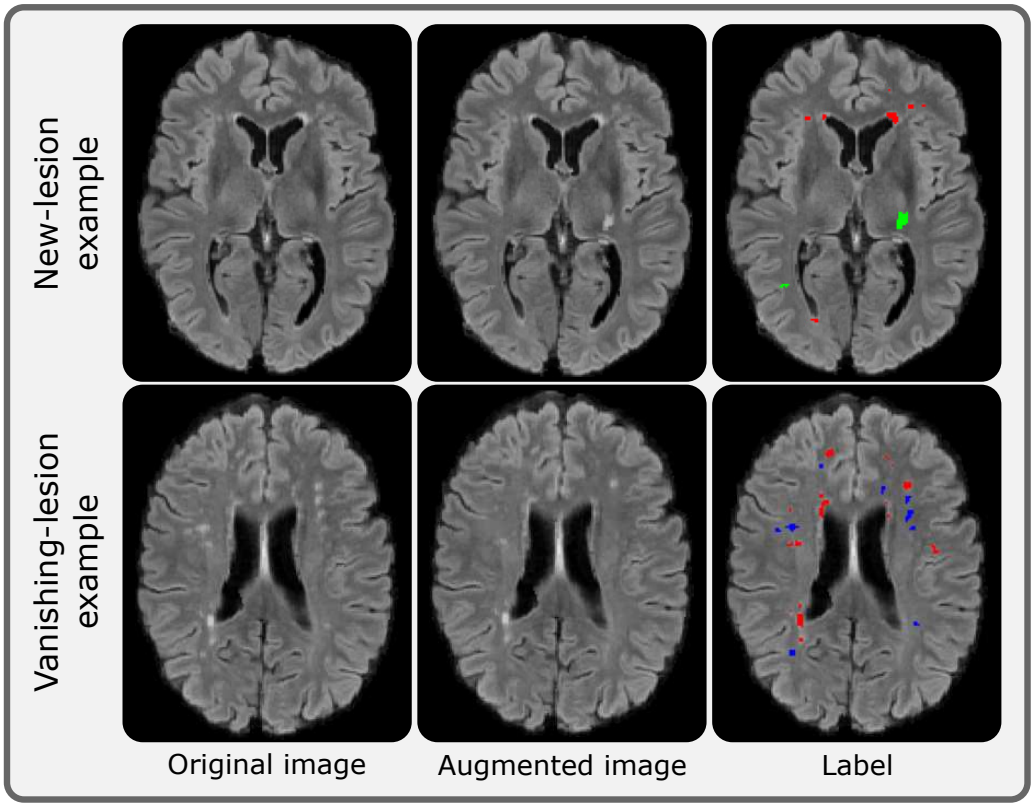


FIGURE 5.2: Example outputs of LesionMix augmentation. Red labels denote lesions which are present in the original image; green labels denote new-lesion generated in the augmented image; blue labels denote vanishing-lesions which have been inpainted from the original image. Images best viewed online.

control lesion volume for augmented images [15]. By integrating LesionMix with the volumetric constraint knowledge and uniformly sampling between the lesion volume increase and decrease thresholds, α_{high} and α_{low}, we set a target load for an image, and augment it to generate an augmented second timepoint. We refer the reader to [15] for more implementation details of LesionMix. Here, we present example outputs of LesionMix augmentation in Figure 5.2.

5.3 EXPERIMENTS

5.3.1 Datasets

We curate five datasets of brain lesions from MS patients that exhibit heterogeneities in data format and annotation. The first three datasets are publicly available. For all datasets, we utilise the FLAIR modality as it is the common modality among the datasets.

ISBI 2015 MS lesion dataset (MS2015) [28] is a public longitudinal MS brain MRI dataset. The training set consists of 5 subjects, four of which with four timepoints, and the fifth subject with five timepoints, totaling 21 scans with respective annotations for all-lesions. The annotations of the test set are not publicly available. We split the original training set of 21 images into a training subset of 13 images, and a test subset of 8 images. All subjects are scanned using the same 3T Philips Medical Systems scanner, with follow-up scans acquired at a mean of 1 year intervals. The dataset is provided preprocessed with inhomogeneity correction using N4, skull-stripping, dura-stripping, followed by a second N4 inhomogeneity correction, and rigid registration to 1mm isotropic MNI template.

MICCAI 2016 MS lesion dataset (MS2016) [44] is a public cross-sectional MS brain MRI dataset. The training dataset comprises of 15 subjects, with respective annotations for all-lesions. The test set is not publicly available with annotations. We split the original training set of 15 images into a training subset of 11 images, and a test subset of 4 images. Subjects are scanned using a variety of scanners, including five subjects from Siemens Verio 3T, five subjects from Siemens Aera 1.5T, and five subjects from Philips Ingenia 3T MRI machines. A pre-processed dataset is provided by the organisers, which includes denoising using non-local means algorithm [48], rigid-registration onto the FLAIR image via block-matching [47], brain extraction using the volBrain platform [128], and bias field correction using the N4 algorithm [167].

MICCAI 2021 MS new-lesion segmentation challenge dataset (MSSEG-2) [43] is a public longitudinal MS brain MRI dataset. The training and testing datasets comprises of 40 and 60 subjects, respectively. Each subject is scanned at two timepoints, with only new-lesion annotations in the second timepoint to save the annotation cost. Second timepoint scans are images 1 to 3 years after the first one. A total of 15 different MRI scanners are represented, including 1.5T and 3T machines from General Electric, Philips, and Siemens MRI machines. The Anima-based[1] MSSEG-2 longitudinal pre-processing script is implemented, which includes an atlas-based brain extraction tool [56], followed by N4 bias field correction [167].

MSSEG-2+ is a private longitudinal MS brain MRI dataset which is derived from MSSEG-2. It follows the same images, dataset structure and pre-processing protocol as MSSEG-2. MSSEG-2+ comprises of all-lesion annotations in the first timepoint, provided

[1] Anima scripts: RRID:SCR_017072, https://anima.irisa.fr

by an expert using ITK-SNAP [184], along with the annotations for new lesions at the second timepoint.

VAN is a private longitudinal MS brain MRI dataset derived from MSSEG-2. In order to address the lack of public datasets for vanishing lesions, VAN is created by simulating vanishing lesions by the inversion of timepoints in the MSSEG-2 dataset. Consequently, new lesions at the second timepoint are transformed into lesions that disappear from the first timepoint. The timepoint inversion is done after the pre-processing pipeline defined in the MSSEG-2 dataset description.

These datasets consist of lesion images obtained from diverse MR scanners with varying annotation procedures. Currently, a longitudinal dataset with annotated all, new, and vanishing lesions does not exist. Therefore, a split from each dataset is held-out for testing and not used during training. In total, we benchmark our method using 192 FLAIR MR images. The high variance in medical scanners, voxel spacing, and image sizes within both the training and testing datasets diminishes cross-domain effects. Additional information on the datasets, including the division of training and test sets used, is available in Table 5.1. During preprocessing, all images are resampled to a voxel spacing of $1 \times 1 \times 1$ mm^3, followed by white matter extraction using SynthSeg [21] and rigid registration to the MNI template space [66].

TABLE 5.1: Summary of the five datasets used for model training and evaluation, which contains different data formats (Image availability) and annotation styles (Label availability). Y_a^{t1}: first timepoint all lesion labels, Y_a^{t2}: second timepoint all-lesion labels, Y_n^{t2}: second timepoint new-lesion labels, Y_v^{t2}: second timepoint vanishing-lesion labels.

Dataset			Image availability		Label availability			
Name	Train images	Test images	Cross-sectional	Longitudinal	Y_a^{t1}	Y_a^{t2}	Y_n^{t2}	Y_v^{t2}
MS2015	13	8	-	✓	✓	✓	-	-
MS2016	11	4	✓	-	✓	-	-	-
MSSEG-2	40	60	-	✓	-	-	✓	-
MSSEG-2+	40	60	-	✓	✓	-	✓	-
VAN	40	60	-	✓	-	-	-	✓
Total	144	192						

5.3.2 Implementation details

5.3.2.1 Network and hyperparameters

SegHeD+ is constructed using a 3D V-Net architecture [131] with five downsampling layers, and four heads in the final layer for the four lesion segmentation tasks. It is implemented using PyTorch 2.2 and trained on Nvidia Tesla T4 GPUs. Training and inference is performed with an image patch size of $96 \times 96 \times 96$, which shows optimal performance taking into consideration GPU constraints. The model is trained from scratch using the Adam optimizer with a learning rate of 0.001 for 20,000 epochs. The hyperparameters λ_L, λ_V, and λ_S were empirically set to 2, 1, and 1, respectively.

An important consideration in the network is the aid of the all-lesion label, y_a^{t1} in guiding temporal segmentation. As mentioned, this label is not utilised for cross-sectional input. For temporal segmentation tasks, such as in MS2015, we provide this input 50% of the the time during training for longitudinal data. This ensures that the network does not become overly dependent on the label for guiding the second timepoint all-lesion segmentation, s_a^{t2}. For testing, we report results for when y_a^{t1} is, and is not input for longitudinal inference.

5.3.2.2 Data augmentation

Data augmentation is a crucial element in the SegHeD+ pipeline. We perform offline lesion-level augmentation, LesionMix, followed by on-the-fly traditional data augmentations during model training. We implement LesionMix such that we equalise the training data imbalance in the different datasets in Table 5.1, and increase the size of overall training data. This is essential, as the dataset imbalance also impacts the overall testing performance for different segmentation tasks. For example, three of the datasets provide diverse images for all-lesion segmentation, however, only one dataset provides images for vanishing-lesion segmentation. While the model becomes more robust and generalisable for the all-lesion segmentation task, it may suffer in the latter task. Prior to training, we perform offline LesionMix augmentation using its default parameters as described in [15] with lesion volume increase and decrease thresholds α_{high} and α_{low} for longitudinal images, defined in section 5.2.2.2. Table 5.2 reports the training set size after LesionMix is applied. Figure 5.3 portrays the balance between original datasets and LesionMix generated datasets.

TABLE 5.2: Summary of the augmented five datasets, after LesionMix is applied. Note that during model training, traditional data augmentation will be further applied on-the-fly. Y_a^{t1}: first timepoint all lesion labels, Y_a^{t2}: second timepoint all-lesion labels, Y_n^{t2}: second timepoint new-lesion labels, Y_v^{t2}: second timepoint vanishing-lesion labels.

Dataset			Image availability		Label availability			
Name	Train images	Test images	Cross-sectional	Longitudinal	Y_a^{t1}	Y_a^{t2}	Y_n^{t2}	Y_v^{t2}
MS2015	80	8	-	✓	✓	✓	✓	✓
MS2016	80	4	✓	✓	✓	✓	✓	✓
MSSEG-2	80	60	-	✓	-	-	✓	✓
MSSEG-2+	80	60	-	✓	✓	✓	✓	✓
VAN	80	60	-	✓	-	-	✓	✓
Total	400	192						

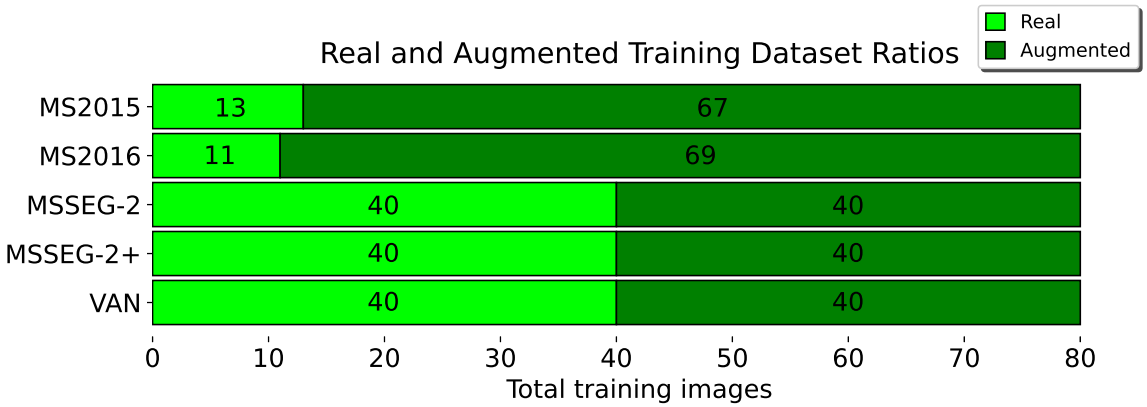


FIGURE 5.3: Visual illustration of the size of original training data and LesionMix-augmented training data.

During training, we implement on-the-fly traditional data augmentation methods, including flipping and rotation in three dimensions, elastic deformation, brightness adjustments (additive and multiplicative), and additive Gaussian noise using the *batchgenerators* framework[2]. Five-fold cross-validation is performed, creating an ensemble of five trained models. After training, for each fold, we select the model which produces the highest Dice score. Lesion segmentation performance is evaluated on the test set using an ensemble of the five trained models by averaging their softmax outputs before binary classification for each segmentation head.

[2] `https://github.com/MIC-DKFZ/batchgenerators`

5.3.3 Evaluation metrics

We evaluate the performance of SegHeD+ in terms of lesion segmentation, where we measure segmentation overlap, and lesion-wise detection. Lesion segmentation performance is evaluated using the Dice similarity coefficient, and lesion detection is assessed using the lesion-wise F_1 score, both described in 2.3.

The Dice and F_1 scores are calculated using the `animaSegPerfAnalyzer2` function of the Anima analyzer tool[3], as provided by the organizers of the MSSEG-2 challenge to ensure a fair comparison with other methods. Post-processing is performed using the default settings of animaSegPerfAnalyzer, which filters out lesion volumes smaller than $3mm^3$. A lesion is classified as detected (true positive) when the automated detection intersects with a minimum of 10% of the ground truth lesion volume and does not exceed 70% of the volume, following the practice in [44]. These parameters are consistent with the MSSEG-2 challenge evaluation protocol [13, 101], for fair comparison to challenge competitors.

5.3.4 Results

SegHeD+ is a model capable of performing three segmentation tasks, namely all-lesion, new lesion and vanishing lesion segmentations. We assess its performance across multiple tasks and compare it with state-of-the-art task-specific segmentation approaches such as nnU-net [92], nnFormer [196], UNETR [82], a recent method for learning from heterogeneous data called CoactSeg [178], a previous iteration of this work, SegHeD [16], which does not feature LesionMix augmentation, and methods specifically tailored for new lesion segmentation [195, 13], as presented in Table 5.3. It has to be noted that only SegHeD and SegHeD+ are capable of performing all three tasks simultaneously. The task-specific SOTA methods undergo two rounds of training: one for the all-lesion segmentation task using the MS2015 and MS2016 datasets, and another for the new-lesion segmentation task using the MSSEG-2 dataset. CoactSeg [178] is trained only once, as it can handle data input and inference for all-lesion and new-lesion segmentation. For new lesion segmentation with the MSSEG-2 dataset, we present Dice scores for MedICL [191], Basaran [13], and the average score from four human experts, denoted as “Avg. of Experts", officially reported by the challenge organizers [43]. The “Avg. of Experts" can be regarded as the upper bound performance

[3] Anima scripts: RRID:SCR_017072, https://anima.irisa.fr

for new-lesion segmentation on MSSEG-2. Additionally, we include lesion-wise F_1 scores in Table 5.4, as per the guidelines of the MSSEG-2 challenge. Example segmentations are compared in Figure 5.4, with more examples provided in the Figure 5.5.

TABLE 5.3: Mean and standard deviations of lesion segmentation Dice scores (%). Best results are in bold. N/A: output not available for the given method. $\star$ indicates that y_a^{t1} was used for longitudinal all-lesion segmentation. †: Methods where two models need to be trained, one for all-lesion and one for new-lesion segmentation. ‡: Not trained on MSSEG-2+. Asterisks indicate statistical significance (*: p$\leq$ 0.05, **: p $\leq$ 0.01, ***: p $\leq$ 0.005) when using a paired Student's *t*-test comparing SegHeD+'s performance to benchmarked methods.

		Segmentation task (lesion type)				
		All		New	All&New	Vanishing
Type	Method	MS2015	MS2016	MSSEG-2	MSSEG-2+	VAN
New-lesion methods	MedICL [191]	N/A	N/A	$50.67_{29.38}$	N/A	N/A
	Basaran et al. [13]	N/A	N/A	$51.06_{28.92}$	N/A	N/A
	LaBRI-IQDA [43]	N/A	N/A	$49.82_{29.96}$	N/A	N/A
	Avg. of Experts [43]	N/A	N/A	$\mathbf{55.52_{34.43}}$	N/A	N/A
Task-specific SOTA	nnU-Net [92] †	$73.01^{***}_{4.91}$	$74.87^{***}_{7.54}$	$48.89_{31.20}$	N/A	N/A
	nnFormer [196] †	$72.56^{***}_{7.15}$	$74.12^{***}_{8.52}$	$47.01^{*}_{33.39}$	N/A	N/A
	UNETR [82] †	$72.79^{***}_{6.13}$	$73.78^{***}_{7.98}$	$45.51^{**}_{30.84}$	N/A	N/A
	TransBTS [173] †	$72.60^{***}_{8.00}$	$74.03^{***}_{8.34}$	$46.58^{*}_{30.07}$	N/A	N/A
	TransUNet [33] †	$70.98^{***}_{7.29}$	$73.15^{***}_{8.99}$	$41.48^{***}_{33.32}$	N/A	N/A
Heterogeneous methods	CoactSeg [178] ‡	$71.28^{***}_{8.24}$	$71.31^{***}_{9.15}$	$47.35_{38.12}$	$58.54^{***}_{18.54}$	N/A
	SegHeD [16]	$74.87_{6.13}$	$84.73_{7.12}$	$48.64_{33.81}$	$65.51_{19.67}$	$35.23_{20.62}$
	SegHeD+	$\mathbf{78.57_{6.71}}$	$\mathbf{85.18_{7.10}}$	$50.52_{31.29}$	$\mathbf{66.83_{18.98}}$	$\mathbf{43.84_{18.04}}$

5.3.4.1 All-lesion segmentation

The "All" columns in Table 5.3 and Table 5.4 demonstrate that SegHeD+ achieves significantly better performance in all-lesion segmentation compared to benchmark methods, in terms of the Dice score and F_1 score. On the longitudinal MS2015 dataset with all-lesion labels, SegHeD+ achieves Dice and F_1 scores of 76.39% and 77.45%, respectively, presenting exceptional performance in lesion segmentation and detection, and significantly outperforming state-of the art models and CoactSeg. A component of SegHeD+ which improves longitudinal all-lesion segmentation performance is the inclusion of the previous timepoint segmentation as an input channel for the following timepoint prediction. This acts as a

TABLE 5.4: Mean and standard deviations of lesion detection F_1 scores (%), in accordance with the MSSEG-2 challenge [43]. Best results are in bold. $\star$ indicates that y_a^{t1} was used for longitudinal all-lesion segmentation. †: Methods where two models are trained, one for all lesion and one for new lesion segmentation. ‡: Not trained on MSSEG+. Asterisks indicate statistical significance (*: p$\leq$ 0.05, **: p $\leq$ 0.01, ***: p $\leq$ 0.005) when using a paired Student's t-test comparing SegHeD+ to competing methods.

		Segmentation task (lesion type)				
		All		New	All&New	Vanishing
Type	Method	MS2015	MS2016	MSSEG-2	MSSEG-2+	VAN
New-lesion methods	MedICL [191]	N/A	N/A	$49.98^{*}_{32.78}$	N/A	N/A
	Basaran et al. [13]	N/A	N/A	$55.25_{34.81}$	N/A	N/A
	LaBRI-IQDA [43]	N/A	N/A	$51.51^{*}_{32.63}$	N/A	N/A
	Avg. of Experts [43]	N/A	N/A	$\mathbf{61.62_{37.11}}$	N/A	N/A
Task-specific SOTA	nnU-Net [92] †	$75.81^{***}_{3.36}$	$69.46^{***}_{11.44}$	$54.15_{33.97}$	N/A	N/A
	nnFormer [196] †	$73.59^{***}_{5.04}$	$68.13^{***}_{15.51}$	$48.12_{33.53}$	N/A	N/A
	UNETR [82] †	$73.00^{***}_{4.92}$	$70.12^{***}_{14.18}$	$45.93^{**}_{39.24}$	N/A	N/A
	TransBTS [173] †	$73.94^{***}_{5.89}$	$68.92^{***}_{14.77}$	$46.10^{**}_{36.80}$	N/A	N/A
	TransUNet [33] †	$72.90^{***}_{6.02}$	$68.25^{***}_{14.90}$	$42.47^{***}_{38.51}$	N/A	N/A
Hetero-geneous methods	CoactSeg [178] ‡	$74.51^{***}_{4.77}$	$69.23^{***}_{14.26}$	$47.23^{*}_{36.45}$	$54.31^{***}_{27.91}$	N/A
	SegHeD[16]	$75.65_{4.74}$	$76.14_{10.97}$	$53.27_{34.70}$	$59.20_{28.10}$	$39.48_{28.41}$
	SegHeD+	$\mathbf{79.35_{3.53}}$	$\mathbf{76.82_{10.83}}$	$55.02_{33.20}$	$\mathbf{60.39_{28.10}}$	$\mathbf{46.02_{21.50}}$

temporal guide for the network, and we see that predicted lesion volumes from SegHeD+ closely track the ground truth from the second timepoint onwards, explored in 5.3.6.

On the cross-sectional MS2016 test dataset, SegHeD+ achieves a Dice score of 85.18%, which is over 10% higher than task-specific models and CoactSeg. Likewise, in lesion detection, SegHeD+ obtains an F_1 score of 76.82%, significantly outperforming SOTA. The improvements in the MS2015 and MS2016 dataset all-lesion segmentation performance can be attributed to various factors. Firstly, SegHeD+ enables the incorporation of heterogeneous data inputs, thus allowing for the inclusion of a larger number of images for model training. This allows for better model generalisation, as the model learns from more diverse images and from different scanners. We investigate the effect of this further in Section 5.3.5.3 and Table 5.6. Secondly, the integration of domain knowledge encoded via anatomical constraints contributes to the reduction of false positives and false negatives. Figure 5.4 illustrates a qualitative comparison between SegHeD+ and selected benchmark methods, where we see a reduced amount of false positive and false negative segmentations.

5.3.4.2 New-lesion segmentation

The “New" column in Table 5.3 demonstrates SegHeD+’s comparable performance to methods tailored for specific tasks and CoactSeg in the segmentation of new lesions. SegHeD+ achieves a Dice score of 50.52% and an F_1 score of 55.02%, outperforming all SOTA and heterogeneous methods, and scoring slightly lower than the top methods tailored specifically for the new-lesion segmentation. Despite this SegHeD+ showcases the ability to handle multiple tasks using a single model.

5.3.4.3 Vanishing-lesion segmentation

The challenge of this task lies in the combined modeling of vanishing lesions along with all and new lesions. There are currently no established methods for segmenting disappearing lesions. As the last column of Table 5.3 shows, SegHeD+ achieves a Dice score of 43.84% in this challenging task, showcasing an impressive improvement over SegHeD. This is largely attributed to the incorporation of LesionMix augmentation, providing additional vanishing lesions examples in the augmented datasets 5.2. The asymmetry between the new and vanishing lesion segmentation Dice scores are due to two reasons. Firstly, new and all-lesion detection overlap in common features, making the task of new lesion segmentation easier. For new and all lesion segmentation, the network searches for hyperintense regions with respect to surrounding tissues. On the contrary, vanishing lesions manifest as normal tissue intensities, and require segmentation of hypo-intense tissue with respect to a previous timepoint. This contrast deems vanishing lesion segmentation to be more challenging than new lesion segmentation when combined with all lesion segmentation. Secondly, new-lesion detection is supported by two datasets (MSSEG-2, MSSEG-2+) in contrast to one vanishing dataset (VAN), thus utilising more new-lesion data. We believe that the results presented here provide useful insights for future work in dataset curation and benchmarking efforts.

5.3.5 Ablation studies

Ablation studies are conducted to examine various loss terms and the white matter mask input channel, LesionMix data augmentation, and effect of heterogeneous dataset inclusion.

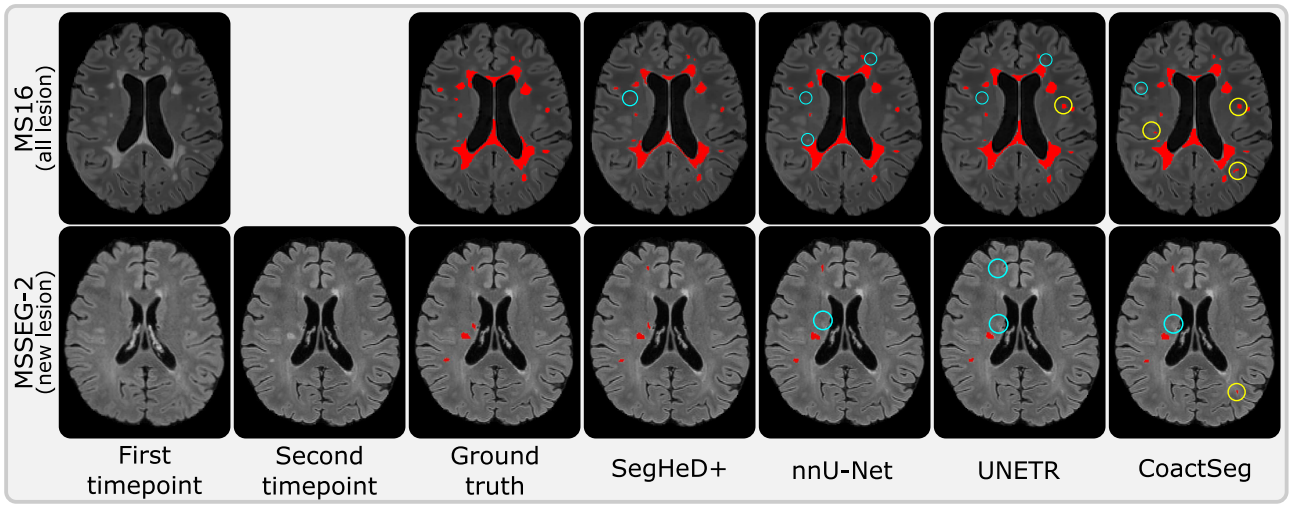


FIGURE 5.4: Qualitative comparison of all-lesion (top row) and new-lesion (bottom row) segmentation performance. Yellow regions denote false positive segmentations, whereas cyan regions denote false negative segmentations. SegHeD+ exhibits fewer false segmentations. Best viewed online.

5.3.5.1 Effect of loss functions and white matter mask

We provide the ablation results of the proposed loss functions in Table 5.5. Each loss component plays a role in improving segmentation performance. Notably, new and vanishing-lesion segmentation performance increases when $\mathcal{L}_{Long}$ is introduced, all-lesion segmentation performance increases when $\mathcal{L}_{Vol}$ is introduced, and an overall increase occurs when $\mathcal{L}_{Spat}$ and x_{wm}^{t2} are presented. In general, they lead to over 4% increase in the Dice score for all-lesion segmentation on MS2015 and MS2016, and over 5% improvement for new-lesion segmentation on the MSSEG-2 dataset.

5.3.5.2 Effect of lesion-level data augmentation

We also assess the effect of incorporating LesionMix as an offline data augmentation method. In the final row of Table 5.5, we notice an overall increase in Dice score across all five datasets. Notably, LesionMix aids in segmenting vanishing lesions the most, boosting segmentation from 35% to 43% on the VAN dataset. This is expected, as LesionMix provides many augmentations with vanishing lesions across varying datasets, seen in Table 5.2.

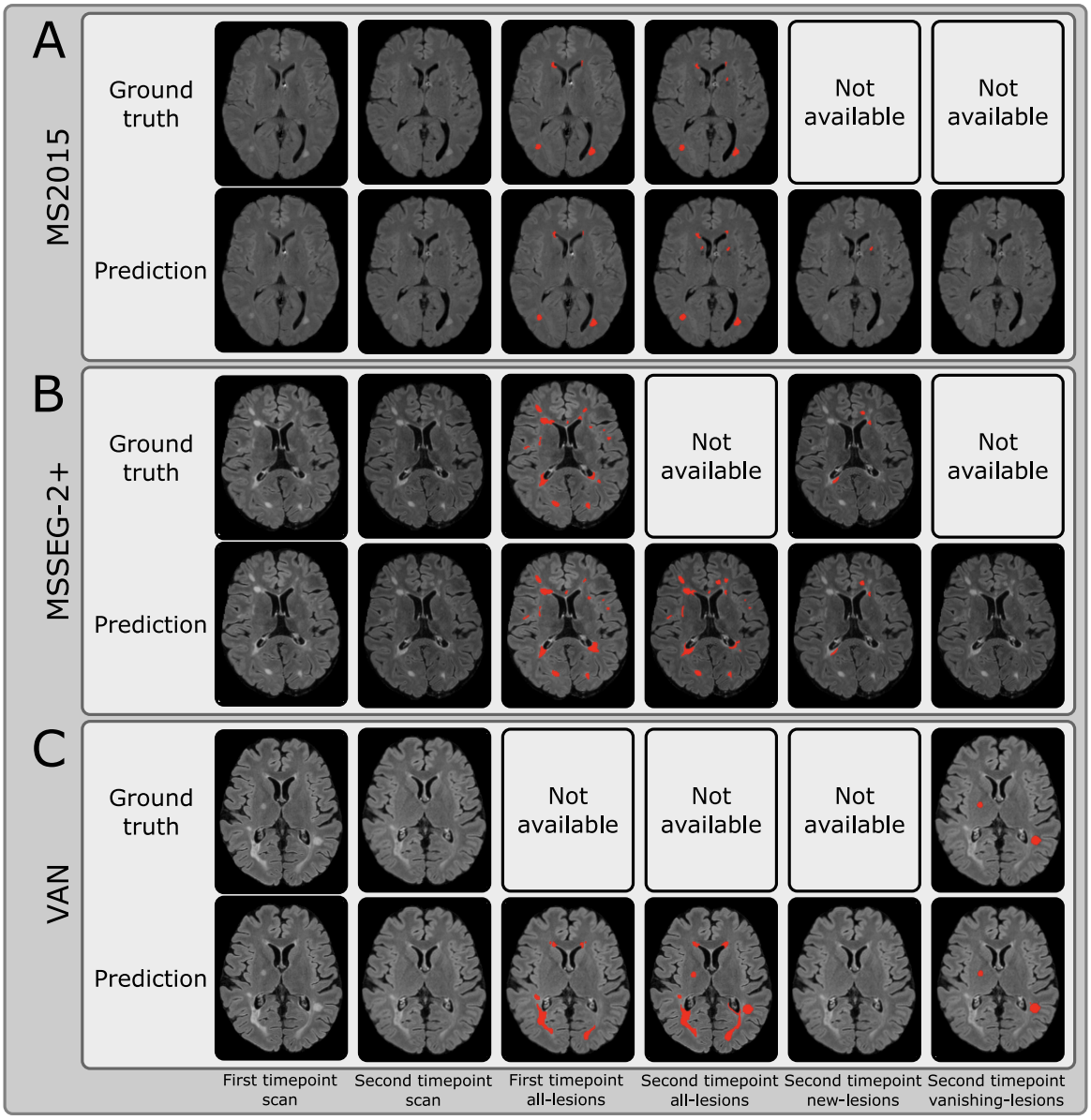


FIGURE 5.5: SegHeD+ is capable of simultaneous multi-task segmentation (Rows 3 to 6). Some tasks do not show new/vanishing-lesions predictions as they are not present at the given slice. "Not available" denotes no ground truth annotation for comparison. **A:** Dataset where all-lesion labels are available for first and second timepoints. **B:** Dataset where first timepoint all-lesion label and second timepoint new-lesion label are available. **C:** Dataset where second timepoint vanishing-lesion label is available.

5.3.5.3 Effect of integrating heterogeneous datasets

We carry out an ablation study to analyse the effect of integrating heterogeneous datasets for lesion segmentation model training. This is performed with the proposed anatomical constraints and LesionMix augmentation. The impact on lesion segmentation when including varying amounts of heterogeneous data is presented in Table 5.6. When using only the

TABLE 5.5: Ablation study of proposed longitudinal, volumetric, and spatial loss functions ($\mathcal{L}_{Long}$, $\mathcal{L}_{Vol}$, $\mathcal{L}_{Spat}$), white matter mask input channel (x_{wm}^{t2}), and LesionMix augmentation.

Ablation components					Dice score (%)				
					Segmentation task (lesion type)				
					All		New	All&New	Vanishing
$\mathcal{L}_{Long}$	$\mathcal{L}_{Vol}$	$\mathcal{L}_{Spat}$	x_{wm}^{t2}	LesionMix	MS2015	MS2016	MSSEG-2	MSSEG-2+	VAN
-	-	-	-	-	$72.98_{7.25}$	$76.93_{9.01}$	$45.28_{38.54}$	$60.25_{20.04}$	$30.97_{27.34}$
-	-	-	✓	-	$73.35_{8.26}$	$77.40_{9.05}$	$46.15_{37.06}$	$61.80_{20.90}$	$31.50_{26.60}$
✓	-	-	-	-	$73.84_{6.99}$	$77.38_{8.99}$	$48.55_{34.80}$	$64.86_{21.12}$	$33.88_{28.97}$
✓	✓	-	-	-	$75.25_{7.83}$	$81.99_{8.03}$	$48.49_{35.00}$	$65.01_{20.21}$	$34.52_{30.63}$
✓	✓	✓	-	-	$77.63_{6.62}$	$82.71_{7.42}$	$48.56_{34.04}$	$65.20_{19.79}$	$34.90_{28.57}$
✓	✓	✓	✓	-	$77.95_{6.93}$	$84.73_{7.12}$	$48.64_{33.81}$	$65.51_{19.07}$	$35.23_{20.62}$
✓	✓	✓	✓	✓	$78.57_{6.71}$	$85.18_{7.10}$	$50.52_{31.29}$	$66.83_{18.98}$	$43.84_{18.04}$

MS2015 and MS2016 datasets, we still see an improvement over task-specific SOTA methods due to the addition of anatomical constraints and additional augmentation. Most significantly, we notice an increase in performance when datasets with overlapping tasks are added. For example, when the MSSEG-2+ dataset is added, which includes the all and new-lesion segmentation tasks specifically, we see an improvement in all-lesion segmentation, as well as new-lesion segmentation.

TABLE 5.6: Ablation study showing improved performance when integrating heterogeneous datasets. All studies are implemented with anatomical constraints and LesionMix augmentation. N/A: output not available for the given dataset.

Datasets included					Dice score (%)				
					Segmentation task (lesion type)				
					All		New	All&New	Vanishing
MS2015	MS2016	MSSEG-2	MSSEG-2+	VAN	MS2015	MS2016	MSSEG-2	MSSEG-2+	VAN
✓	✓	-	-	-	$76.81_{5.41}$	$79.60_{7.18}$	N/A	N/A	N/A
-	-	✓	-	-	N/A	N/A	$49.82_{31.02}$	N/A	N/A
✓	✓	✓	-	-	$77.34_{6.19}$	$79.98_{7.39}$	$48.20_{32.58}$	$62.18_{18.47}$	N/A
✓	✓	✓	✓	-	$78.28_{6.58}$	$85.26_{6.94}$	$50.39_{30.58}$	$67.21_{18.20}$	N/A
✓	✓	✓	✓	✓	$78.57_{6.71}$	$85.18_{7.10}$	$50.52_{31.29}$	$66.83_{18.98}$	$43.84_{18.04}$

5.3.6 Improved temporal consistency

We carry out a study to analyse the effect on the temporal consistency of lesion segmentation. Two subjects, each with four longitudinal scans, which were put aside for testing in the MS2015 dataset are utilised for this analysis. Figure 5.6 illustrates the predicted lesion volumes at various time points for the two individuals under examination. The figure shows that the lesion volumes estimated by SegHeD+ (depicted in blue) exhibit greater temporal coherence with the ground truth values (depicted in black) owing to the suggested anatomical constraints, resulting in the highest Pearson's correlation coefficient with the ground truth lesion volumes. When the all-lesion segmentations of previous timepoints are introduced in the inference stage, which occurs from the second timepoint onwards, SegHeD+ tracks the ground truth lesion volumes with less error. Enhancing temporal consistency has potential in aiding subsequent clinically relevant analyses for disease progression, such as assessing the annual rate of lesion shrinkage or expansion.

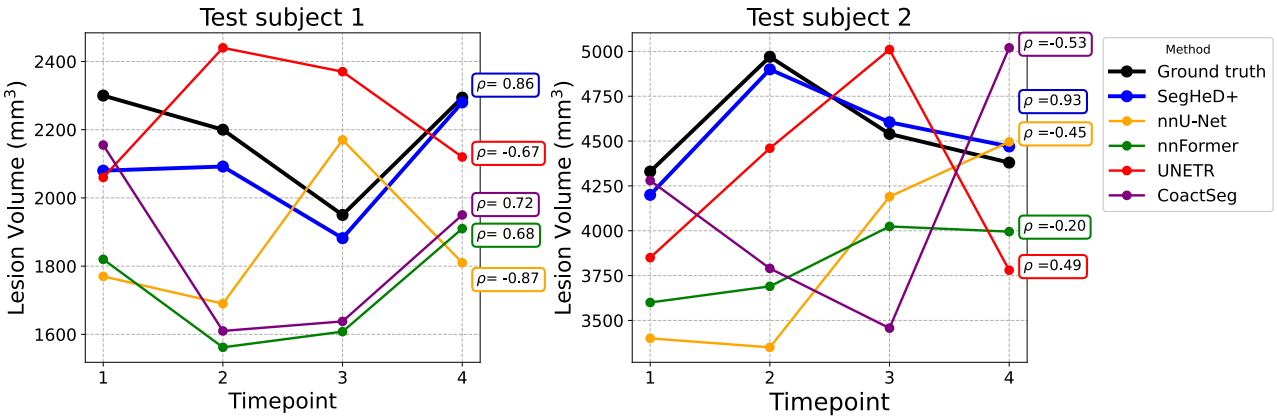


FIGURE 5.6: Estimated lesion volumes across four time points for two test subjects. SegHeD+ (blue) results in predictions that are temporally more consistent with the ground truth (black), compared to competing methods. The ρ value for each method indicates its Pearson's correlation coefficient with the ground truth, the higher the better.

5.4 Discussion

We showcase that by integrating multiple heterogeneous datasets, adopting a multi-task learning approach, along with anatomical constraints and lesion-aware data augmentation, we can achieve promising results in multiple tasks simultaneously, including all, new and vanishing lesion segmentation tasks. While task-specific lesion segmentation models would

typically focus on learning either hyperintense region features for segmenting new and all lesions, *or* hypointense region features for segmenting vanishing lesions, SegHeD+ simultaneously learns both types of features, accounting for previous timepoints and surrounding tissues. SegHeD+ surpasses SOTA brain lesion segmentation models, which are trained in a task-specific manner, in terms of both lesion segmentation (Dice) and lesion detection (F_1) performance.

There are a few limitations of this work. The first limitation is the relatively high computational cost during training. Utilising up to four 3D brain images as the model input leads to high model training times. A promising future avenue is to lower the computational burden associated with high-dimensional and high-resolution medical imaging data. Secondly, we provide a unique method for simulating a vanishing lesion dataset, VAN, through the use of a new lesion dataset, MSSEG-2. Vanishing and newly forming lesions show different lesion dynamics and characteristics [135]. A dedicated clinical dataset for vanishing lesions would enable more accurate modeling for real-life clinical use.

The case for a unified multi-task model which can outperform task-specific methods is not yet fully achieved. While SegHeD+ achieves the strongest performance in most of the tasks and datasets, some new-lesion methods still edge out SegHeD+, as shown in Table 5.3. This performance gap may be reduced when more training datasets are provided for a specific task, as seen in Table 5.6. Alongside new and vanishing lesions, growing and shrinking lesion detection is of paramount importance for analysing the progression of MS. However, the development of state-of-the-art machine learning methods for clinical applications depend heavily on data availability, and currently there are no public datasets for analysing growing and/or shrinking MS lesions.

5.5 Conclusion

In this work we introduce SegHeD+, an innovative multi-task approach for segmenting MS lesions that can effectively learn from diverse and heterogeneous data. SegHeD+ can accurately segment all types of lesions, including new and disappearing ones, in both cross-sectional and longitudinal datasets. Evaluations on five MS lesion datasets demonstrate that SegHeD+ achieves superior performance compared to other segmentation techniques

for segmenting all types of lesions, and shows comparable results for identifying new lesions. Considering the current lack of large MS lesion datasets, SegHeD+'s ability to utilise diverse data sources will significantly enhance current MS imaging research and simplify the analysis of large-scale multi-site datasets with varying data structures and annotations.

Chapter 6

Conclusion

The primary objective of this work was to provide modern solutions to obstacles which hinder the adoption of deep learning in brain lesion segmentation, with a focus on white matter lesions. We identified these obstacles as: (a) the lack of clinically relevant segmentation models concerning temporal biomarkers for multiple sclerosis, (b) labelled data scarcity, and (c) the heterogeneity within medical image datasets which hindered generalisability. We presented four different methods to improve these areas. First, we provided a robust solution to segmenting newly formed lesions with longitudinal data. We then proposed one generative and one non-generative method for augmenting medical images to alleviate the scarcity of labelled brain images. Finally, we put forward a unified model for multiple sclerosis which is agnostic to data and label availability. The following section discusses the contributions of our work from a broader perspective, with specific discussions to our solutions in their respective chapters. We conclude this thesis by considering the future applications of our work, and outlook of deep learning methods.

6.1 Summary of Contributions

New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation

Medical image segmentation is often driven by the goals and metrics of which natural image segmentation has, without consideration to the clinical aspect of a disease. This process leads to clinically and non-statistically significant approaches being developed. In this work, we focused on one of the leading biomarkers for the progression of multiple sclerosis: new lesions. We proposed a 3D U-Net for the segmentation of newly formed multiple sclerosis

brain lesions in follow-up MRI scans. We provided an extensive preprocessing pipeline along with data augmentation methods which mimicked MRI acquisition imaging artefacts and augmented images in a lesion-aware manner. Our method outperformed state-of-the-art proven biomedical image segmentation frameworks, all while not succumbing to the trade-off between F_1 and Dice score metrics. Just as importantly, we introduced metrics such as average false positive lesion count, providing a clear interpretation of our results to clinicians.

Subject-specific lesion generation and pseudo-healthy synthesis for multiple sclerosis brain images

Generative methods have seen increased use in both medical and natural image augmentation thanks to their potential to create realistic and diverse training samples. This capability helps mitigate the challenge of limited annotated datasets, enabling models to generalise more effectively. We presented a cyclic framework for generating lesions and performing pseudo-healthy synthesis. Rather than generating brain images from scratch, we utilised an attention decoder to focus on the foreground (lesion) region. Our generated images showcased realism in three different ways, through an image quality assessment by expert raters (Table 4.4), realistic spatial distribution of generated lesions (Figure 4.5), and correlations between ventricular volume with healthiness. Furthermore, through dropout, we showed stochastic capability by generating diverse augmented images for each given sample. Our method not only performed competitively against benchmarks in a variety of dataset sizes, but also when training using fully synthesised data from our framework compared to real data (Table 4.3).

LesionMix: a lesion-level data augmentation method for medical image segmentation

Generative methods for data augmentation have become a highly active area of research, receiving significant attention in recent studies. However, a hurdle for their adoption is that they are dataset specific and have high computational cost. We present an non-generative data augmentation method which operates by iteratively augmenting or inpainting individual lesions with respect to a desired dataset lesion load. Our simple, yet effective, approach is not only lesion-aware, but also augments with consideration to the spatial distribution of real lesions. We carry out a lesion load distribution study, and find that the optimal load distribution of the training set is *not* the original distribution ("Real" in Table 4.7), but one with a uniform distribution of small and large lesions. LesionMix showed superior performance to

alternative mix-based augmentation techniques especially in low-data settings. Particularly, it is easily adaptable and doesn't need training to be deployed on other organs or medical datasets.

SegHeD: segmentation of heterogeneous data for multiple sclerosis lesions with anatomical constraints and lesion-aware augmentation

Medical image datasets come in a variety of image and label availabilities. Extensive preprocessing of these datasets can be required in order to incorporate them into deep learning frameworks. Furthermore, datasets for the same disease may be tailored for different tasks (all-lesion vs new-lesion segmentation). We propose a unified framework for multiple sclerosis which is capable of segmenting new, all, and vanishing lesions with no modification to the image and label availabilities of the original datasets. We improve our method by incorporating anatomical constraints, which analyse the implicit information in longitudinal lesion evolution, the volumetric correlation of lesions at different timepoints, and spatial constraint of lesion formation. We pair these constraints with a lesion-aware augmentation method to considerably increase image and label availabilities. SegHeD provides an exemplary framework on how to maximise information usage and inference capabilities through heterogeneous input and medical domain knowledge. Being able to incorporate more data than homogeneous frameworks results in improved segmentation scores across multiple datasets.

6.2 Future Outlook

In this section, we explore future research avenues that could extend our findings and address broader challenges in the field and adoption of deep neural networks for medical image segmentation. While we highlight the general limitations and open questions in deep learning, specific technical constraints and potential improvements of our methodologies are discussed within the relevant chapters.

6.2.1 Brain Lesion Segmentation and Beyond

The methods developed in this thesis have been applied on medical image segmentation tasks, focusing primarily on brain lesion segmentation and multiple sclerosis. However,

within each chapter there are novelties and ideas which could possibly be extended to other organs, applications, and non-medical domains. Some examples of how our work can be extended include:

- *Adversarial method for lesion generation (Section 4.1)*: Our attention-based cyclic lesion generation framework could possibly be extended to other organs where abnormalities can occur. Additionally, given the presence of a large longitudinal dataset, one could synthesise and possibly predict the progression of the disease.

- *Mix-based augmentation strategy (Section 4.2)*: Originally presented for lesion-mixing, this strategy could be implemented in natural image segmentation tasks as well. The analysis on lesion load distribution introduces an additional dimension that needs to be considered: class-distributions in datasets.

- *Model for heterogeneous data input (Chapter 5)*: We overcome the overlooked aspect of maximising training data through incorporation of different dataset formats. Our heterogeneous input and multi-task output model could possibly be extended to other organs which have multiple image and annotation formats. The domain knowledge introduced the framework can be exemplary to other diseases, or even natural images, where modelling relationships between substructures of the segmentation objective can improve results.

6.2.2 Unsupervised and Semi-Supervised Learning

Our work in this thesis has focused primarily on supervised learning tasks, where ground truth labels are available. However, it is important to acknowledge that annotated medical images are relatively scarce due to the high costs of expert annotation and time constraints. This scarcity of labelled data has spurred increased interest in methods that require fewer or no labels, such as unsupervised and semi-supervised approaches.

Semi-supervised learning uses a small amount of labelled data combined with a large amount of unlabelled data to improve model performance, while unsupervised learning seeks to identify patterns in entirely unlabelled data. These approaches are becoming increasingly popular for segmentation and classification tasks in medical imaging due to their ability to mitigate the need for extensive and expensive annotations [39].

Although supervised learning models continue to outperform these alternative approaches, the gap is closing rapidly. Unsupervised and semi-supervised models are benefitting from innovations such as self-supervised learning, contrastive learning, and generative adversarial networks [36, 97], with models that can enhance performance with minimal labelled data. These iterative improvements suggest that the reliance on large annotated datasets may significantly decrease in the near future, reshaping medical image analysis.

6.2.3 Foundation Models

In the last few years, foundation models have emerged as a groundbreaking development in both natural and medical image analysis. These models, which are pre-trained on vast amounts of data, can leverage multimodal inputs, such as images, text, and other structured data, or multiple image datasets concerning different organs or diseases allowing them to generalise across various tasks. In contrast to traditional supervised or unsupervised methods that rely solely on labelled or unlabelled datasets, foundation models are capable of incorporating both types of data, thus drastically increasing the diversity of the images and other inputs they can process. This diversity enhances their ability to perform well across a wide range of applications, including segmentation, classification, and even text-based tasks such as medical report generation [136].

In medical imaging, multimodal vision-language models, such as CLIP (Contrastive Language-Image Pre-training), have shown great promise in enabling more efficient and interpretable medical image analysis [144]. These models can link medical images with descriptive text, aiding not only in image classification but also in creating explainable AI systems that could bridge the gap between deep learning and medical expertise [9]. In brain imaging specifically, recent efforts have explored the use of large-scale pre-trained models with unlabelled data for brain tumor and lesion segmentation, with fine-tuned models showing great promise in improving accuracy over traditional supervised methods [49].

With regards to white matter lesion segmentation, foundation models offer several exciting possibilities. Pre-training on large datasets of diverse brain images, including healthy controls and various neurological conditions, can equip the model with a robust understanding of brain anatomy and pathology. This prior knowledge can then be leveraged for

more accurate and robust segmentation of white matter lesions in new, unseen data. Furthermore, multimodal approaches can be particularly beneficial for lesion segmentation. For example, integrating imaging data with patient demographics, clinical notes, and genetic information can provide a more comprehensive understanding of the disease and improve the accuracy and interpretability of segmentation results. The ability to incorporate multimodal data, such as combining imaging data with patient medical histories, can provide a more comprehensive approach to diagnosis and treatment planning.

As these models become more computationally efficient, their ability to handle multimodal inputs and perform multiple tasks simultaneously will likely make them indispensable in the field of medical imaging. For instance, the potential to process different imaging modalities (e.g., MRI, CT, PET) alongside clinical data and patient histories will make them highly adaptable for specific medical conditions, including brain diseases.

6.2.4 Model Explainability

Improving model interpretability and explainability is vital for the adoption of machine learning in medical image segmentation and analysis, where the consequences of decisions are critical. Many doctors are tentative about integrating these technologies due to the "black box" nature of deep learning models, which often provide little insight into their decision-making processes. The emerging field of Explainable AI (XAI) aims to make these models more transparent, enabling clinicians to understand and verify the rationale behind predictions, thus facilitating in their adoption. Some of the main fields where XAI is making progress include explaining model predictions by examining causal relationships [180], counterfactual approaches [12], and measuring feature importance for model transparency [125]. In the context of white matter lesion segmentation, XAI techniques can provide valuable insights. For instance, feature importance analysis can highlight which regions of the brain or specific image features (e.g., intensity gradients, textures) are most influential in the model's predictions. This information can help clinicians understand the model's reasoning, identify potential biases, and assess whether the model is relying on clinically relevant features. XAI represents a promising path forward in bridging the gap between advanced machine learning techniques and clinical practice, ensuring that these tools are both powerful and trustworthy.

6.2.5 Future of Deep Learning

When work for this thesis began in 2020, vision transformers, diffusion models for image generation, close to human-like natural language processing techniques [25] had not yet been developed. The dramatic shift of machine learning, moving from handcrafted features and shallow models to deep learning approaches which learn complex representations from multimodal data, has enabled more accurate and efficient performance over medical image analysis and other computer vision tasks. While it is difficult to foresee the specific methods which will take dominance in the near and far future, we can identify the principal remaining challenges which need to be overcome, and the themes that need to be addressed for greater adoption of machine learning methods.

The largest challenge to overcome is the lack of robustness and generalisability of deep learning models. While these models have shown exceptional results in controlled settings, their performance can, and do, degrade in real-world scenarios, where variations during deployment introduce unforeseen complexities. This is observed as the domain shift problem in the medical imaging when deep learning-based MRI image segmentation models are tested on a datasets obtained from a different MRI suppliers. Or in autonomous vehicles, where poor weather condition severely degrade autonomous driving performance [187]. There are efforts in this thesis reducing the effect of domain shift in medical imaging, however, the gap is still there. Until that gap is gone, deep neural networks have to be adopted tentatively, keeping in mind of their shortcomings.

Since the start of machine learning, one theme that has (almost) always held true is that more data equals better results. Maximising the information that you can provide to models allows them to see more variety and perform better under testing, thus improving generalisability. As of now, foundation models are leading the way for multimodal data incorporation in models. However, keeping in mind that in medical image analysis, unreliable predictions could lead to clinical misdiagnoses or even fatal decisions, it is clear that there is still a long way to go before we can trust artificial neural networks over our organic ones.

Bibliography

[1] I. S. A. Abdelhalim, M. F. Mohamed, and Y. B. Mahdy. Data augmentation for skin lesion using self-attention based progressive generative adversarial network. *Expert Systems with Applications*, 2021.

[2] M. Agn, O. Puonti, P. M. a. Rosenschöld, I. Law, and K. Van Leemput. Brain tumor segmentation using a generative model with an rbm prior on tumor shape. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 168–180, Cham, 2016. Springer International Publishing.

[3] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson. Deep learning for brain mri segmentation: state of the art and future directions. *Journal of digital imaging*, 30:449–459, 2017.

[4] M. Aljabri, M. AlAmir, M. AlGhamdi, M. Abdel-Mottaleb, and F. Collado-Mesa. Towards a better understanding of annotation tools for medical imaging: a survey. *Multimedia tools and applications*, 81(18):25877–25911, 2022.

[5] A. Apicella, F. Donnarumma, F. Isgrò, and R. Prevete. A survey on modern trainable activation functions. *Neural Networks*, 138:14–32, 2021.

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

[7] S. Aslani, M. Dayan, L. Storelli, M. Filippi, V. Murino, M. A. Rocca, and D. Sona. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *NeuroImage*, 196(November 2018):1–15, 2019.

[8] X. Ayrignac, N. Menjot de Champfleur, S. Menjot de Champfleur, C. Carra-Dallière, J. Deverdun, A. Corlobe, and P. Labauge. Brain magnetic resonance imaging helps to differentiate atypical multiple sclerosis with cavitary lesions and vanishing white matter disease. *European Journal of Neurology*, 23(6):995–1000, 2016.

[9] B. Azad, R. Azad, S. Eskandari, A. Bozorgpour, A. Kazerouni, I. Rekik, and D. Merhof. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*, 2023.

[10] D. Bank, N. Koenigstein, and R. Giryes. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pages 353–374, 2023.

[11] B. Barile, A. Marzullo, C. Stamile, F. Durand-Dubief, and D. Sappey-Marinier. Data augmentation using generative adversarial neural networks on brain structural connectivity in multiple sclerosis. *Computer Methods and Programs in Biomedicine*, 2021.

[12] S. Barocas, A. D. Selbst, and M. Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 80–89. Association for Computing Machinery, 2020.

[13] B. D. Basaran, P. M. Matthews, and W. Bai. New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation. *Frontiers in Neuroscience*, 16, 2022.

[14] B. D. Basaran, M. Qiao, P. M. Matthews, and W. Bai. Subject-Specific Lesion Generation and Pseudo-Healthy Synthesis for Multiple Sclerosis Brain Images. In *Simulation and Synthesis in Medical Imaging*. Springer Nature Switzerland, 2022.

[15] B. D. Basaran, W. Zhang, M. Qiao, B. Kainz, P. M. Matthews, and W. Bai. LesionMix: A Lesion-Level Data Augmentation Method for Medical Image Segmentation. In *Data Augmentation, Labelling, and Imperfections*, pages 73–83, Cham, 2024. Springer Nature Switzerland.

[16] B. D. Basaran, X. Zhang, P. M. Matthews, and W. Bai. SegHeD: Segmentation of Heterogeneous Data for Multiple Sclerosis Lesions with Anatomical Constraints. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 Workshops*. Springer Nature Switzerland, 2025.

[17] M. Battaglini, F. Rossi, R. A. Grove, M. L. Stromillo, B. Whitcher, P. M. Matthews, and N. De Stefano. Automated identification of brain new lesions in multiple sclerosis using subtraction images. *Journal of Magnetic Resonance Imaging*, 39(6):1543–1549, 2014.

[18] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *International Conference on Machine Learning*, 2009.

[19] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel. MixMatch: A Holistic Approach to Semi-Supervised Learning. *arXiv preprint arXiv:1905.02249*, 2019.

[20] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand, et al. The liver tumor segmentation benchmark (LiTS). *Medical Image Analysis*, 2023.

[21] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, and J. E. Iglesias. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis*, 86, 2023.

[22] A. Birenbaum and H. Greenspan. Multi-view longitudinal CNN for multiple sclerosis lesion segmentation. *Engineering Applications of Artificial Intelligence*, 65(August):111–118, 2017.

[23] A. Bissoto, F. Perez, E. Valle, and S. Avila. Skin lesion synthesis with generative adversarial networks. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, 2018.

[24] C. Bowles, C. Qin, R. Guerrero, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert. Brain lesion segmentation through image synthesis and outlier detection. *NeuroImage: Clinical*, 2017.

[25] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *ArXiv*, abs/2303.12712, 2023.

[26] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca. UniverSeg: Universal Medical Image Segmentation. In *International Conference on Computer Vision*, 2023.

[27] M. Cabezas and Y. Diez. An analysis of loss functions for heavily imbalanced lesion segmentation. *Sensors*, 24(6):1981, 2024.

[28] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre, M. Jorge Cardoso, N. Cawley, O. Ciccarelli, C. A.

Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S. K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L. O. Iheme, D. Unay, S. Jain, D. M. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P. L. Bazin, P. A. Calabresi, C. M. Crainiceanu, L. M. Ellingsen, D. S. Reich, J. L. Prince, and D. L. Pham. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *NeuroImage*, 148(December 2016):77–102, 2017.

[29] D. Chabas, T. Castillo-Trivino, E. Mowry, J. Strober, O. Glenn, and E. Waubant. Vanishing ms t2-bright lesions before puberty: a distinct mri phenotype? *Neurology*, 71(14):1090–1093, 2008.

[30] Y. Chai, B. Xu, K. Zhang, N. Lepore, and J. C. Wood. MRI Restoration Using Edge-Guided Adversarial Learning. *IEEE Access*, 8:83858–83870, 2020.

[31] A. Chartsias, T. Joyce, G. Papanastasiou, et al. Disentangled representation learning in cardiac image analysis. *Medical Image Analysis*, 2019.

[32] C. Chen, C. Qin, H. Qiu, C. Ouyang, S. Wang, L. Chen, G. Tarroni, W. Bai, and D. Rueckert. Realistic adversarial data augmentation for mr image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 667–677. Springer, 2020.

[33] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. In *ICML Interpretable Machine Learning in Healthcare Workshop*, 2021.

[34] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[35] S. Chen, D. M. Lovelock, and R. J. Radke. Segmenting the prostate and rectum in ct imagery using anatomical constraints. *Medical Image Analysis*, 15(1):1–11, 2011.

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

[37] Y. Chen, X.-H. Yang, Z. Wei, A. A. Heidari, N. Zheng, Z. Li, H. Chen, H. Hu, Q. Zhou, and Q. Guan. Generative Adversarial Networks in Medical Image augmentation: A review. *Computers in Biology and Medicine*, 144:105382, 2022.

[38] B. Cheng, A. Schwing, and A. Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

[39] V. Cheplygina, M. de Bruijne, and J. P. Pluim. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 2019.

[40] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 2021.

[41] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-net: Learning dense volumetric segmentation from sparse annotation. *MICCAI*, 2016.

[42] J. R. Clough, I. Oksuz, N. Byrne, J. A. Schnabel, and A. P. King. Explicit topological priors for deep-learning based image segmentation using persistent homology. In *International Conference on Information Processing in Medical Imaging*, pages 16–28. Springer, 2019.

[43] O. Commowick, F. Cervenansky, F. Cotton, and M. Dojat. Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure. In *MICCAI 2021 MSSEG-2 Challenge Proceedings*, page 126, Strasbourg, France, Sept. 2021.

[44] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. C. Pop, P. Girard, R. Ameli, J.-C. Ferré, et al. Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. *Scientific Reports*, 2018.

[45] O. Commowick, A. Masson, B. Combes, S. Camarasu-Pop, F. Cervenansky, M. Kain, S. Lion, R. Casey, M. Dojat, and F. Cotton. MICCAI 2021 MSSEG-2 challenge quantitative results [Data set]. `https://doi.org/10.5281/zenodo.5775523`, 2021.

[46] O. Commowick, A. Masson, B. Combes, S. Camarasu-Pop, F. Cervenansky, M. Kain, S. Lion, R. Casey, M. Dojat, and F. Cotton. MICCAI 2021 MSSEG-2 challenge demographics data [Data set]. https://doi.org/10.5281/zenodo.5824568, 2022.

[47] O. Commowick, N. Wiest-Daesslé, and S. Prima. Block-matching strategies for rigid registration of multimodal medical images. In *IEEE International Symposium on Biomedical Imaging*, pages 700–703, 2012.

[48] P. Coupe, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot. An Optimized Blockwise Nonlocal Means Denoising Filter for 3-D Magnetic Resonance Images. *IEEE Transactions on Medical Imaging*, 27(4), 2008.

[49] J. Cox, P. Liu, S. E. Stolte, Y. Yang, K. Liu, K. B. See, H. Ju, and R. Fang. BrainSegFounder: Towards 3D foundation models for neuroimage segmentation. *Medical Image Analysis*, 97:103301, 2024.

[50] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

[51] A. V. Dalca, J. Guttag, and M. R. Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Computer Vision and Pattern Recognition*, 2018.

[52] C. Dalton, P. Brex, R. Jenkins, N. Fox, K. Miszkiel, W. Crum, J. O'riordan, G. Plant, A. Thompson, and D. Miller. Progressive ventricular enlargement in patients with clinically isolated syndromes is associated with the early development of multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 2002.

[53] C. Dalton, K. Miszkiel, P. O'Connor, G. Plant, G. Rice, and D. Miller. Ventricular enlargement in MS: one-year change at various stages of disease. *Neurology*, 66(5):693–698, 2006.

[54] S. Denner, A. Khakzar, M. Sajid, M. Saleh, Z. Spiclin, S. T. Kim, and N. Navab. Spatio-Temporal Learning from Longitudinal Data for Multiple Sclerosis Lesion Segmentation. In A. Crimi and S. Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 2021.

[55] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[56] J. Doshi, G. Erus, Y. Ou, B. Gaonkar, and C. Davatzikos. Multi-Atlas Skull-Stripping. *Academic Radiology*, 20(12):1566–1576, 2013.

[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

[58] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187, Cham, 2016. Springer International Publishing.

[59] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning, 2018.

[60] W. A. Edelstein, M. Mahesh, and J. A. Carrino. Mri: time is dose—and money and versatility. *Journal of the American College of Radiology: JACR*, 7(8):650, 2010.

[61] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin. A systematic review of re-identification attacks on health data. *PloS one*, 6(12):e28071, 2011.

[62] B. M. Ellingson, M. Bendszus, J. Boxerman, D. Barboriak, B. J. Erickson, M. Smits, S. J. Nelson, E. Gerstner, B. Alexander, G. Goldmacher, et al. Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. *Neuro-Oncology*, 17(9):1188–1198, 08 2015.

[63] C. Elliott, D. L. Arnold, D. L. Collins, and T. Arbel. Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. *IEEE Transactions on Medical Imaging*, 32(8), 2013.

[64] C. Elliott, S. J. Francis, D. L. Arnold, D. L. Collins, and T. Arbel. Bayesian Classification of Multiple Sclerosis Lesions in Longitudinal MRI Using Subtraction Images. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, pages 290–297, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[65] M. Filippi, M. Horsfield, P. Tofts, F. Barkhof, A. Thompson, and D. Miller. Quantitative assessment of MRI lesion load in monitoring the evolution of multiple sclerosis. *Brain*, 118(6), 1995.

[66] V. S. Fonov, A. C. Evans, R. C. McKinstry, C. R. Almli, and D. Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 2009.

[67] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[68] K. Fukushima. Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333, 1969.

[69] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[70] A. Galdran, G. Carneiro, and M. A. González Ballester. Balanced-MixUp for highly imbalanced medical image classification. In *Medical Image Computing and Computer Assisted Intervention*, 2021.

[71] O. Ganiler, A. Oliver, Y. Diez, J. Freixenet, J. C. Vilanova, B. Beltran, L. Ramió-Torrentà, À. Rovira, and X. Lladó. A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology*, 56:363–374, 2014.

[72] A. V. Genovese, J. Hagemeier, N. Bergsland, D. Jakimovski, M. G. Dwyer, D. P. Ramasamy, A. A. Lizarraga, D. Hojnacki, C. Kolb, B. Weinstock-Guttman, et al. Atrophied brain T2 lesion volume at MRI is associated with disability progression and conversion to secondary progressive multiple sclerosis. *Radiology*, 293(2), 2019.

[73] R. Gontijo-Lopes, S. Smullin, E. D. Cubuk, and E. Dyer. Tradeoffs in data augmentation: An empirical study. In *International Conference on Learning Representations*, 2020.

[74] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Neural Information Processing Systems*, volume 27, 2014.

[75] C. Gros, A. Lemay, and J. Cohen-Adad. SoftSeg: Advantages of soft versus binary training for image segmentation. *Medical image analysis*, 71:102038, 2021.

[76] H. Guan and M. Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

[77] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE transactions on knowledge and data engineering*, 35(4):3313–3332, 2021.

[78] N. Guizard, K. Nakamura, P. Coupé, V. S. Fonov, D. L. Arnold, and D. L. Collins. Non-local means inpainting of ms lesions in longitudinal image processing. *Frontiers in Neuroscience*, 9, 2015.

[79] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48, 2016.

[80] P. Gupta, R. E. Banchs, and P. Rosso. Squeezing bottlenecks: exploring the limits of autoencoder semantic representation capabilities. *Neurocomputing*, 175:1001–1008, 2016.

[81] S. H. Guptha, E. Holroyd, and G. Campbell. Progressive lateral ventricular enlargement as a clue to Alzheimer's disease. *The Lancet*, 2002.

[82] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.

[83] A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov. Global context vision transformers. In *International Conference on Machine Learning*, pages 12633–12646. PMLR, 2023.

[84] M. Herwerth, B. J. Schwaiger, K. Kreiser, B. Hemmer, and R. Ilg. Adult-onset vanishing white matter disease as differential diagnosis of primary progressive multiple sclerosis: a case report. *Multiple Sclerosis Journal*, 21(5):666–668, 2015.

[85] M. H. Hesamian, W. Jia, X. He, and P. Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32:582–596, 2019.

[86] L. Hirsch, Y. Huang, and L. C. Parra. Segmentation of MRI head anatomy using deep volumetric networks and multiple spatial priors. *Journal of Medical Imaging*, 8(3), 2021.

[87] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[88] M. Homssi, E. Sweeney, E. Demmon, W. Mannheim, M. Sakirsky, Y. Wang, S. Gauthier, A. Gupta, and T. Nguyen. Evaluation of the statistical detection of change algorithm for screening patients with MS with new lesion activity on longitudinal brain MRI. *American Journal of Neuroradiology*, 44(6), 2023.

[89] W. J. Housley, D. Pitt, and D. A. Hafler. Biomarkers in multiple sclerosis. *Clinical Immunology*, 161(1):51–58, 2015. New Horizons in Biomarker Research.

[90] H. Huang, P. S. Yu, and C. Wang. An Introduction to Image Synthesis with Generative Adversarial Nets, 2018.

[91] J. E. Iglesias, M. R. Sabuncu, I. Aganj, P. Bhatt, C. Casillas, D. Salat, A. Boxer, B. Fischl, and K. Van Leemput. An algorithm for optimal fusion of atlases with different labeling protocols. *NeuroImage*, 106:451–463, 2015.

[92] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 2021.

[93] F. Isensee, P. Jäger, J. Wasserthal, D. Zimmerer, J. Petersen, S. Kohl, J. Schock, A. Klein, T. Roß, S. Wirkert, P. Neher, S. Dinkelacker, G. Köhler, and K. Maier-Hein. batchgenerators - a python framework for data augmentation (0.19.6). `https://doi.org/10.5281/zenodo.3632567`, 2020.

[94] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, and P. F. Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498. Springer, 2024.

[95] S. Jain, A. Ribbens, D. M. Sima, M. Cambron, J. De Keyser, C. Wang, M. H. Barnett, S. Van Huffel, F. Maes, and D. Smeets. Two time point MS lesion segmentation in brain MRI: An expectation-maximization framework. *Frontiers in Neuroscience*, 10, 2016.

[96] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith. FSL. *Neuroimage*, 62(2):782–790, 2012.

[97] R. Jiao, Y. Zhang, L. Ding, B. Xue, J. Zhang, R. Cai, and C. Jin. Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 169:107840, 2024.

[98] Q. Jin, H. Cui, C. Sun, Z. Meng, and R. Su. Free-form tumor synthesis in computed tomography images via richer generative adversarial network. *Knowledge-Based Systems*, 218:106753, 2021.

[99] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36, 2017.

[100] R. A. Kamraoui, V.-T. Ta, J. V. Manjon, and C. Pierrick. Image quality data augmentation for new MS lesion segmentation. In *MICCAI 2021 MSSEG-2 Challenge Proceedings*, page 37, 2021.

[101] R. A. Kamraoui, V.-T. Ta, T. Tourdias, B. Mansencal, J. V. Manjon, and P. Coupé. DeepLesionBrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. *Medical Image Analysis*, 76:102312, 2022.

[102] C. Kang, C. Lee, H. Song, M. Ma, and S. Pereira. Variability matters: Evaluating inter-rater variability in histopathology for robust cell detection. In *European Conference on Computer Vision*, pages 552–565. Springer, 2022.

[103] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[104] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. Transformers in vision: A survey. *ACM Comput. Surv.*, 54(10s), sep 2022.

[105] D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[106] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[107] E. Kondrateva, M. Pominova, E. Popova, M. Sharaev, A. Bernstein, and E. Burnaev. Domain shift in computer vision models for mri data analysis: an overview. In *Thirteenth International Conference on Machine Vision*, volume 11605, pages 126–133. SPIE, 2021.

[108] H. J. Kuijf, J. M. Biesbroek, J. De Bresser, R. Heinen, S. Andermatt, M. Bento, M. Berseth, M. Belyaev, M. J. Cardoso, A. Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE Transactions on Medical Imaging*, 2019.

[109] R. Kushol, A. H. Wilman, S. Kalra, and Y.-H. Yang. Dsmri: domain shift analyzer for multi-center mri datasets. *Diagnostics*, 13(18):2947, 2023.

[110] H. Lassmann. Multiple sclerosis pathology. *Cold Spring Harbor Perspectives in Medicine*, 8(3), 2018.

[111] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

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

[113] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[114] Ž. Lesjak, A. Galimzianova, A. Koren, M. Lukin, F. Pernuš, B. Likar, and Ž. Špiclin. A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. *Neuroinformatics*, 16:51–63, 2018.

[115] L. Li, Y. Fan, M. Tse, and K.-Y. Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020.

[116] P. Li, Y. Pei, and J. Li. A comprehensive survey on design and application of autoencoder in deep learning. *Applied Soft Computing*, 138:110176, 2023.

[117] Q. Li, Z. Yu, Y. Wang, and H. Zheng. TumorGAN: A multi-modal data augmentation framework for brain tumor segmentation. *Sensors*, 2020.

[118] X. Lin, Y. Xiang, L. Zhang, X. Yang, Z. Yan, and L. Yu. SAMUS: Adapting Segment Anything Model for Clinically-Friendly and Generalizable Ultrasound Image Segmentation. *arXiv: 2309.06824*, 2023.

[119] Y. Lin, Z. Wang, K.-T. Cheng, and H. Chen. InsMix: Towards realistic generative data augmentation for nuclei instance segmentation. In *Medical Image Computing and Computer Assisted Intervention*, 2022.

[120] J. Liu, Y. Zhang, J.-N. Chen, J. Xiao, Y. Lu, B. A Landman, Y. Yuan, A. Yuille, Y. Tang, and Z. Zhou. CLIP-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023.

[121] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[122] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[123] A. J. London. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*, 49(1):15–21, 2019.

[124] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[125] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777. Curran Associates Inc., 2017.

[126] J. Luxenberg, J. V. Haxby, H. Creasey, M. Sundaram, and S. Rapoport. Rate of ventricular enlargement in dementia of the Alzheimer type correlates with rate of neuropsychological deterioration. *Neurology*, 37(7):1135–1135, 1987.

[127] S. Maddalena, F. R. Schmidt, C. Anastasopoulos, C. Granziera, G. Krueger, and T. Brox. White matter MS-lesion segmentation using a geometric brain model. *IEEE Transactions on Medical Imaging*, 35(7), 2016.

[128] J. V. Manjón and P. Coupé. volBrain: An Online MRI Brain Volumetry System. *Frontiers in Neuroinformatics*, 10, 2016.

[129] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[130] V. M. Massimiliano Calabrese, Alice Favaretto and P. Gallo. Grey matter lesions in MS. *Prion*, 7(1):20–27, 2013.

[131] F. Milletari, N. Navab, and S.-A. Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision*, 2016.

[132] MIRTK. Medical Image Registration ToolKit (MIRTK), 2021. Available at `https://mirtk.github.io/`.

[133] A. Mohammed and R. Kora. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, 35(2):757–774, 2023.

[134] T. C. W. Mok and A. C. S. Chung. Learning Data Augmentation for Brain Tumor Segmentation with Coarse-to-Fine Generative Adversarial Networks. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 2019.

[135] P. Molyneux, M. Filippi, F. Barkhof, C. Gasperini, T. Yousry, L. Lruyen, H. Lai, M. Rocca, I. Moseley, and D. Miller. Correlations between monthly enhanced MRI lesion rate and changes in T2 lesion volume in multiple sclerosis. *Annals of neurology*, 43(3):332–339, 1998.

[136] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[137] A. M. Muslim, S. Mashohor, G. Al Gawwam, R. Mahmud, M. binti Hanafi, O. Alnuaimi, R. Josephine, and A. D. Almutairi. Brain mri dataset of multiple sclerosis with consensus manual lesion segmentation and patient meta information. *Data in Brief*, 42:108139, 2022.

[138] M.-T. Nguyen, T. Dinh Kim, M.-H. Ha, A.-L. Do, L.-A. Nguyen, and D.-L. Ngo. Advanced Learning-Based Segmentation of Liver and Tumor 3D Images for Early Disease Diagnosis. In *2023 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 101–106, 2023.

[139] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan, et al. Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2017.

[140] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention U-Net: Learning Where to Look for the Pancreas. In *Medical Imaging with Deep Learning*, 2018.

[141] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024.

[142] J. W. Prineas, R. O. Barnard, T. Revesz, E. E. Kwon, L. Sharer, and E.-S. Cho. Multiple sclerosis: pathology of recurrent lesions. *Brain*, 116(3), 1993.

[143] D. Qiao, C. Dai, Y. Ding, J. Li, Q. Chen, W. Chen, and M. Zhang. SelfMix: A self-adaptive data augmentation method for lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention*, 2022.

[144] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[145] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative Adversarial Text to Image Synthesis. In *International Conference on Machine Learning*, 2016.

[146] J. C. Reinhold, A. Carass, and J. L. Prince. A structural causal model for MR images of multiple sclerosis. In *Medical Image Computing and Computer Assisted Intervention*, 2021.

[147] M. A. Ricci Lara, R. Echeveste, and E. Ferrante. Addressing fairness in artificial intelligence for medical imaging. *nature communications*, 13(1):4581, 2022.

[148] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[149] A. G. Roy, J. Ren, S. Azizi, A. Loh, V. Natarajan, B. Mustafa, N. Pawlowski, J. Freyberg, Y. Liu, Z. Beaver, et al. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *Medical Image Analysis*, 75:102274, 2022.

[150] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[151] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986. *Biometrika*, 71(599-607):6, 1986.

[152] M. Salem, S. Valverde, M. Cabezas, D. Pareto, A. Oliver, J. Salvi, À. Rovira, and X. Lladó. Multiple sclerosis lesion synthesis in MRI using an encoder-decoder U-net. *IEEE Access*, 2019.

[153] K. Schmierer, T. Campion, A. Sinclair, W. Van Hecke, P. M. Matthews, and M. P. Wattjes. Commentary: Towards a standard MRI protocol for multiple sclerosis across the UK. *British Journal of Radiology*, 92(1101):2–5, 2019.

[154] M. Sdika and D. Pelletier. Nonrigid registration of multiple sclerosis brain images using lesion inpainting for morphometry or lesion mapping. *Human Brain Mapping*, 30(4):1060–1067, 2009.

[155] V. Sethi, G. Nair, M. Absinta, P. Sati, A. Venkataraman, J. Ohayon, T. Wu, K. Yang, C. Shea, B. E. Dewey, I. C. Cortese, and D. S. Reich. Slowly eroding lesions in multiple sclerosis. *Multiple Sclerosis Journal*, 23(3), 2017.

[156] D. Shen, G. Wu, and H.-I. Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19(1):221–248, 2017.

[157] G. Shi, L. Xiao, Y. Chen, and S. K. Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 70, 2021.

[158] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni. U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. *IEEE Access*, 9:82031–82057, 2021.

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

[160] M. Styner, J. Lee, B. Chin, M. Chin, O. Commowick, H. Tran, S. Markovic-Plese, V. Jewells, and S. Warfield. 3D segmentation in the clinic: A grand challenge II: MS lesion segmentation. *The MIDAS Journal*, 2008.

[161] L. Sun, J. Wang, Y. Huang, X. Ding, H. Greenspan, and J. Paisley. An Adversarial Learning Approach to Medical Image Synthesis for Lesion Detection. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[162] R.-Y. Sun. Optimization for deep learning: An overview. *Journal of the Operations Research Society of China*, 8(2):249–294, 2020.

[163] H. Tang, D. Xu, N. Sebe, and Y. Yan. Attention-Guided Generative Adversarial Networks for Unsupervised Image-to-Image Translation. In *International Joint Conference on Neural Networks*, 2019.

[164] Z. Tang, M. Cabezas, D. Liu, M. Barnett, W. Cai, and C. Wang. Lg-net: Lesion gate network for multiple sclerosis lesion inpainting. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 660–669, Cham, 2021. Springer International Publishing.

[165] A. Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 2004.

[166] Y. Tian, D. Su, S. Lauria, and X. Liu. Recent advances on loss functions in deep learning for computer vision. *Neurocomputing*, 497:129–158, 2022.

[167] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 29(6), 2010.

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

[169] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[170] S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, A. Oliver, and X. Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage*, 155(April):159–168, 2017.

[171] A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems. Neural information processing systems foundation*, 2017.

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

[173] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li. TransBTS: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer-Assisted Intervention*, 2021.

[174] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3:1–40, 2016.

[175] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020.

[176] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin. Diffusion Models for Medical Anomaly Detection. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer Nature Switzerland, 2022.

[177] J. Wu, W. Ji, Y. Liu, H. Fu, M. Xu, Y. Xu, and Y. Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[178] Y. Wu, Z. Wu, H. Shi, B. Picker, W. Chong, and J. Cai. CoactSeg: Learning from heterogeneous data for new multiple sclerosis lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention*, 2023.

[179] T. Xia, A. Chartsias, and S. A. Tsaftaris. Pseudo-healthy synthesis with pathology disentanglement and adversarial learning. *Medical Image Analysis*, 2020.

[180] G. Xu, T. D. Duong, Q. Li, S. Liu, and X. Wang. Causality learning: A new perspective for interpretable machine learning. *arXiv preprint arXiv:2006.16789*, 2020.

[181] W. Yan, Y. Wang, S. Gu, L. Huang, F. Yan, L. Xia, and Q. Tao. The domain shift problem of medical image segmentation and vendor-adaptation by unet-gan. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 623–631. Springer, 2019.

[182] F. Yang, G. Zamzmi, S. Angara, S. Rajaraman, A. Aquilina, Z. Xue, S. Jaeger, E. Papagiannakis, and S. K. Antani. Assessing inter-annotator agreement for medical image segmentation. *IEEE Access*, 11:21300–21312, 2023.

[183] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019.

[184] P. A. Yushkevich, Y. Gao, and G. Gerig. ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2016.

[185] P. A. Yushkevich, A. Pashchinskiy, I. Oguz, S. Mohan, J. E. Schmitt, J. M. Stein, D. Zukić, J. Vicory, M. McCormick, N. Yushkevich, et al. User-guided segmentation of multi-modality medical imaging datasets with itk-snap. *Neuroinformatics*, 17:83–102, 2019.

[186] A. Zafar, M. Aamir, N. Mohd Nawi, A. Arshad, S. Riaz, A. Alruban, A. K. Dutta, and S. Almotairi. A comparison of pooling methods for convolutional neural networks. *Applied Sciences*, 12(17), 2022.

[187] S. Zang, M. Ding, D. Smith, P. Tyler, T. Rakotoarivelo, and M. A. Kaafar. The Impact of Adverse Weather Conditions on Autonomous Vehicles: How Rain, Snow, Fog, and Hail Affect the Performance of a Self-Driving Car. *IEEE Vehicular Technology Magazine*, 14(2), 2019.

[188] J. Zhai, S. Zhang, J. Chen, and Q. He. Autoencoder and Its Various Variants. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 415–419, 2018.

[189] H. Zhang, R. Bakshi, F. Bagnato, and I. Oguz. Robust multiple sclerosis lesion inpainting with edge prior. In *Machine Learning in Medical Imaging*, pages 120–129, Cham, 2020. Springer International Publishing.

[190] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. MixUp: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[191] H. Zhang, H. Li, and I. Oguz. Segmentation of new MS lesions with tiramisu and 2.5D stacked slices. *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, 61, 2021.

[192] H. Zhang, A. M. Valcarcel, R. Bakshi, R. Chu, F. Bagnato, R. T. Shinohara, K. Hett, and I. Oguz. Multiple sclerosis lesion segmentation with tiramisu and 2.5 d stacked slices. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 338–346, 2019.

[193] J. Zhang, Y. Zhang, and X. Xu. ObjectAug: Object-level Data Augmentation for Semantic Image Segmentation. *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021.

[194] L. Zhang, R. Tanno, M.-C. Xu, C. Jin, J. Jacob, O. Cicarrelli, F. Barkhof, and D. Alexander. Disentangling Human Error from Ground Truth in Segmentation of Medical Images. In *Neural Information Processing Systems*, volume 33, pages 15750–15762, 2020.

[195] X. Zhang, C. Liu, N. Ou, X. Zeng, Z. Zhuo, Y. Duan, X. Xiong, Y. Yu, Z. Liu, Y. Liu, and C. Ye. Carvemix: A simple data augmentation method for brain lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention*. Springer Nature Switzerland, 2021.

[196] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu. nnFormer: Volumetric medical image segmentation via a 3D transformer. *IEEE Transactions on Image Processing*, 2023.

[197] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision*, 2017.

Appendix

Springer

Springer allows authors to reuse their article's Version of Record, in whole or in part, in their own thesis[1]. This applies to the contents presented in Chapters 4 and 5, in where I reuse my own published works. These published works have been cited in the beginning of corresponding chapters.

Frontiers in Neuroscience

Parts of contents in Chapter 3 reuse my own work published in Frontiers. Content reproduction has been granted with open-access under the terms of the Creative Commons Attribution License (CC BY 4.0)[2]. The published work has been cited in the beginning of the chapter.

Permissions for reproduced figures

All reproduced figures from others work in this thesis are licensed under the terms of the Creative Commons Attribution License (CC BY 4.0), except the figures listed in Table A1. For these figures, permission for noncommercial use has been granted by corresponding rights holders, see below.

TABLE A1: **Reproduced figures with their respective licences.**

Type	Reference	Source	Licensed publisher	Permission requested on	I have permission yes/no	Licence No.
Figure	Fig. 2.12	[92]	Springer Nature	30-Sep-2024	yes	5858730128962
Figure	Fig. 2.13	[173]	Springer Nature	30-Sep-2024	yes	5858730688426
Figure	Fig. 2.15	[176]	Springer Nature	30-Sep-2024	yes	5858730586148

[1] https://www.springer.com/gp/rights-permissions/obtaining-permissions/882
[2] https://creativecommons.org/licenses/by/4.0/