UNIVERSITY OF LEEDS

Deep Learning with Simulated Data for Medical Imaging

Yash Deo

University of Leeds

School of Computing

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

February, 2025

Intellectual Property Statement

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The Thesis is presented in the alternative format as it contains 4 publications (3 Published and 1 in review). The details of the publications and author contributions in the thesis are listed below :-

Learned Local Attention Masks for Synthesizing Vessel Segmentations - Chapter 4

This chapter includes work from the following accepted publication:

> Yash Deo, Rodrigo Bonazzola, Haoran Dou, Yan Xia, Tianyou Wei, Nishant Ravikumar, Alejandro F. Frangi, Toni Lassila. *Learned Local Attention Maps for Synthesising Vessel Segmentations from T2 MRI.* Accepted at the International Workshop on Simulation and Synthesis in Medical Imaging, MICCAI 2023.

The CRediT statement included in the paper submission is as follows:

- **Yash Deo:** Conceptualisation, Investigation, Experimentation, and writing of first draft.
- **Toni Lassila:** Conceptualisation, Supervision, Writing - review and editing.

- **Alejandro F. Frangi, Nishant Ravikumar, Yan Xia:** Supervision.
- **Rodrigo Bonazzola, Haoran Dou:** Conceptualisation and Advice.
- **Tianyou Wei:** Generated segmentation dataset used in this study.

Shape-Guided Conditional Latent Diffusion Models for Synthesizing Brain Vasculature - Chapter 5

This chapter includes work from the following accepted publication:

> Yash Deo, Haoran Dou, Nishant Ravikumar, Alejandro F. Frangi, Toni Lassila. 2023. *Shape-Guided Conditional Latent Diffusion Models for Synthesising Brain Vasculature.* International Conference on Medical Image Computing and Computer-Assisted Intervention, Deep Generative Models Workshop.

My contributions: I wrote the code for the diffusion models, experiments, and the pre-processing of the data.

The CRediT statement included in the paper submission is as follows:

- **Yash Deo:** Conceptualisation, Investigation, Experimentation, and writing of first draft.
- **Toni Lassila:** Conceptualisation, Supervision, Writing - review and editing.
- **Alejandro F. Frangi, Nishant Ravikumar:** Supervision.
- **Haoran Dou:** Conceptualisation and Advice.

Few-Shot Learning in Diffusion Models for Generating Cerebral Aneurysm Geometries - Chapter 6

This chapter includes work from the following publication:

> Y Deo, F Lin, H Dou, N Cheng, N Ravikumar, A.F. Frangi, T Lassila, 2024. *Few-Shot Learning in Diffusion Models for Generating Cerebral Aneurysm Geometries.* International Symposium on Biomedical Imaging, 2024.

My contributions: I wrote the code for the diffusion models, experiments, and the pre-processing of the data.

The CrediT statement included in the paper submission is as follows:

- **Yash Deo:** Conceptualisation, Investigation, Experimentation, and writing of first draft.
- **Toni Lassila:** Conceptualisation, Supervision, Writing - review and editing.
- **Alejandro F. Frangi, Nishant Ravikumar:** Supervision.
- **Haoran Dou:** Conceptualisation and Advice.
- **Fengming Lin, Nina Cheng:** Software.

Anatomy-guided latent diffusion models for generating brain MR angiography images and vessel segmentation masks - Chapter 7

This chapter includes work from the following publication that is under review:

> Y Deo, F Lin, H Dou, N Cheng, N Ravikumar, A.F. Frangi, T Lassila, 2024. *Anatomy-guided latent diffusion models for generating brain MR angiography images and vessel segmentation masks*

My contributions: I wrote the code for the diffusion models, experiments, and the pre-processing of the data.

The CrediT statement included in the paper submission is as follows:

- **Yash Deo:** Conceptualisation, Investigation, Experimentation, and writing of first draft.
- **Toni Lassila:** Conceptualisation, Supervision, Writing - review and editing.
- **Alejandro F. Frangi, Nishant Ravikumar:** Supervision.
- **Haoran Dou:** Conceptualisation and Advice.
- **Fengming Lin, Nina Cheng:** Software.

Acknowledgements

My sincerest thanks go firstly to my supervisors, Dr. Toni Lassila and Dr. Alejandro Frangi, whose invaluable guidance, patience, and steadfast support have been pivotal throughout this journey. Their expertise and insightful critiques have greatly influenced my research, and their encouragement has been a continuous source of motivation. I would also like to express my gratitude to Dr. Nishant Ravikumar for his ongoing input and guidance in my research.

I am deeply appreciative of my colleagues Michael Macraild, Soodeh Kalaie, Ning Bi, Nina Cheng, Fengming Lin, Siyuan Kang, Rodrigo Bonazzola, Haoran Dou, as well as my friends at CISTIB and the University of Leeds. They have contributed to a supportive and enriching environment filled with memorable moments over the past four years.

I also extend their thanks to the School of Computing for their financial support, which made this research feasible.

On a personal note, I am profoundly grateful to my family, Mr. Nitin Deo, Ms. Sawali Deo, and Ms. Sanjna Deo, for their unconditional love and unwavering support not only during my PhD but throughout my life. No words can adequately express my gratitude for their enduring encouragement.

Special thanks go to my father for his continued financial support during my PhD, without which completing this journey would have been impossible. I would also like to thank all family and friends for constantly asking me every other week if I was "done writing the thesis yet."

Finally, I am indebted to all who, directly or indirectly, assisted in bringing this thesis to fruition. Thank you all for your belief in me and my work.

Abstract

This thesis addresses critical challenges in the application of deep learning to medical imaging, particularly focusing on brain vasculature. Deep learning, a specialized branch of machine learning, has shown remarkable potential in tasks such as object classification, detection, and image segmentation. However, its effectiveness is often limited by the scarcity of large, diverse, and labeled datasets in the medical domain. This scarcity stems from the time-consuming and expertise-dependent nature of medical image annotation, making it impractical to manually create extensive datasets for training deep learning models. The research tackles three primary challenges: the disparity in data availability across different imaging modalities, the underrepresentation of certain phenotypes in medical imaging datasets, and the limited availability of data for rare conditions. To address these issues, the thesis explores innovative approaches to data generation and synthesis, aiming to augment existing datasets and create new ones.

The work employs advanced techniques such as cross-modality synthesis using learned local attention masks to generate MRA images from T2-weighted brain MR images, addressing the data availability discrepancy between modalities. Furthermore, the thesis investigates the use of diffusion models for synthetic generation of brain vasculature, particularly focusing on the intricate Circle of Willis to address phenotype underrepresentation. The research also introduces few-shot learning with diffusion models to enable conditional generation of brain vessels with aneurysms, tackling the challenge of limited training data for rare conditions.

By systematically addressing these challenges, this thesis contributes to advancing the field of medical imaging and enhancing the training of deep learning models in healthcare applications.

Contents

List of Figures

List of Tables

Abbreviations

ACA	Anterior Cerebral Artery	ACoA	Anterior Communicating Artery
AVM	Arteriovenous Malformation	CT	Computed Tomography
CSF	Cerebrospinal Fluid	DSA	Digital Subtraction Angiography
MRA	Magnetic Resonance Angiography	MRI	Magnetic Resonance Imaging
MCA	Middle Cerebral Artery	PCA	Posterior Cerebral Artery
PCoA	Posterior Communicating Artery	PET	Positron Emission Tomography
RF	Radio Frequency	SAH	Subarachnoid Hemorrhage
SPECT	Single-Photon Emission Computed Tomography	TE	Time to Echo
TOF	Time of Flight	TR	Repetition Time
3DRA	Three-Dimensional Rotational Angiography	MIP	Maximum Intensity Projection
ANN	Artificial Neural Network	CNN	Convolutional Neural Network
RNN	Recurrent Neural Network	LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit	PCA	Principal Component Analysis
ReLU	Rectified Linear Unit	MSE	Mean Squared Error
GAN	Generative Adversarial Network	VAE	Variational Autoencoder
DDPM	Denoising Diffusion Probabilistic Models	SSIM	Structural Similarity Index Measure
PSNR	Peak Signal-to-Noise Ratio	BPTT	Backpropagation Through Time
SDE	Stochastic Differential Equation	NAC	Non-Attenuation Corrected
AC	Attenuation Corrected	CBCT	Cone-Beam Computed Tomography
UMM-CSGM	Unified Multi-Modal Conditional Score-Based Generative Model	FGDM	Frequency-Guided Diffusion Model
CoW	Circle of Willis	FLAIR	Fluid-Attenuated Inversion Recovery
GANN	Generative Adversarial Neural Network	SGAN	Steerable Generative Adversarial Network
PD	Proton Density	HD95	Hausdorff Distance 95th percentile
MTL	Multi-Task Learning	SSIM	Structural Similarity Index Measure
FID	FrÃ©chet Inception Distance	LDM	Latent Diffusion Model
SDF	Signed Distance Function	ViT	Vision Transformer
MS-SSIM	Multi-Scale Structural Similarity Index		

Chapter 1

Introduction

1.1 Background and motivation

Deep learning is a specialised branch of machine learning recently rising in popularity which involves a set of mathematical operators (neurons) arranged in a layered architecture, designed to perform non-linear operations on the input and converge to a specific output. One of the major branches of deep learning would be supervised learning, which requires a paired set of input and output data. The deep learning network then tries to model a relationship between the input and the corresponding outputs. These techniques can be used to tackle several problems such as object/scene classification, object detection and image segmentation.

An inherent challenge within the domain of deep learning pertains to the susceptibility of its techniques to overfitting. Overfitting denotes a scenario wherein the neural network excessively assimilates the nuances of the training dataset, resulting in great performance on the training data but notably diminished efficacy when applied to unseen or test data. The main contributing factor to overfitting is the inadequacy of the training data. This deficiency arises due to the substantial volume of training data required for deep learning models to achieve optimal performance and increased generalizability, as elucidated in the existing literature [4]. In tasks such as brain tumor segmentation or classification of neurodegenerative diseases, the need for large and diverse datasets is paramount to ensure the model's ability to accurately identify subtle patterns and variations indicative of pathological conditions.

One of the main roadblocks in applying supervised deep learning methods to medical imaging problems is the scarcity of labelled or annotated data. Medical images require expert knowledge and skills to annotate, and this process is often tedious and time-consuming. Therefore, it is not practical to manually create large and diverse datasets for training deep learning models. This motivates the need for exploring alternative ways of generating or simulating data that can augment the existing datasets or create new ones.

Data generation offers a viable solution to this challenge by augmenting existing datasets with synthetic images. These generated images can be used to increase diversity, mitigate class imbalances, and improve model robustness. While data augmentation techniques such as affine transformations, intensity perturbations, and contrast normalization are widely used, they do not create fundamentally new data instances. Deep generative models, such as Generative Adversarial Networks (GANs) and Diffu-

sion Models, can synthesize realistic medical images by learning the underlying data distribution. This is particularly useful for capturing rare cases that are underrepresented in clinical datasets.

It is necessary however to first distinguish between the concepts of simulation and synthesis. Simulation relies on first principles, such as physical laws or mathematical models, to produce images from scratch. Synthesis, on the other hand, uses existing data, such as photographs or videos, to create new images based on some criteria. We also usually assume behind these concepts a natural information processing direction: from data to models with synthesis; and from models to data with simulation [5].

However, if we generate a dataset consisting mainly of simulated images it could lead to poor performance in the network (and overfitting) as important edge cases could be severely under sampled and overlooked by the model. This is a huge problem when the real dataset consists of a very small sample of the edge cases which we want to detect, for e.g., detecting a brain tumour which is found in less than 1% of the population [6], the dataset will have a very low frequency of the edge cases (where brain tumour is present).

When we generate synthetic image, we usually know all the image components such as its location and attributes, which makes synthetic images really good for testing out new algorithms. However, this generation of synthetic images also could shift the mean simulated image too far from the real image distribution. This would make them unsuitable to be used in real world scenarios. Also, for most fields, we require the help of experts while generating / annotating synthetic images and sometimes even experts who know the ground very well can remember things that are not actually present in the image which could lead to a faulty dataset and eventually a faulty model. Therefore, it is important to research the domain shift between real and simulated images and develop algorithms to generate synthetic images in such a way that they can be used for deep learning in medical imaging.

Deep learning methodologies have been effectively utilized to overcome the challenges posed by limited data and class imbalance in medical imaging across various modalities, including Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET) [7]. Despite these advancements, synthesizing virtual vascular structures remains underexplored due to the additional complexity of accurately representing intricate vascular topologies and ensuring the continuity of

vessels. Magnetic Resonance Angiography (MRA) is an imaging modality specifically designed to visualize blood vessels in the brain and other regions. It plays a crucial role in diagnosing vascular conditions such as aneurysms, arteriovenous malformations (AVMs), and stroke. Unlike other imaging modalities, such as Computed Tomography Angiography (CTA), MRA is non-invasive and does not require ionizing radiation or contrast agents in many cases, making it a preferred choice for cerebrovascular assessment.However, acquiring high-quality MRA scans is challenging due to the sensitivity of the imaging process to motion artifacts, long acquisition times, and variations in scanner protocols across institutions. These challenges contribute to inconsistencies in MRA datasets, limiting their utility for deep learning applications. Additionally, small aneurysms and subtle vascular anomalies often go undetected due to limited resolution or contrast variations in standard MRA scans.Given these challenges, synthetic MRA generation has the potential to significantly impact the field of cerebrovascular imaging. Thus, the primary aim of this thesis is to design novel deep learning techniques for generating synthetic data, particularly focusing on cerebral vasculature. This research addresses three key issues: the uneven distribution of data among different imaging modalities, the inadequate representation of certain phenotypes in medical imaging datasets, and the scarcity of data for infrequent conditions. The thesis investigates strategies for data synthesis and generation to enhance existing datasets and develop new ones.

1.2 Contributions

In this thesis, we present several novel methods for generating and synthesizing brain vasculature with a particular focus on synthesizing the Circle of Willis (CoW) and brain aneurysms.

- **Learned local attention maps for cross modality image synthesis [Chapter 4:]** In brain imaging, certain modalities like T1/T2 tend to have a sufficient amount of available data compared to other modalities such as MRA (Magnetic Resonance Angiography) which have limited data available. By focusing on the cross-modality synthesis of T2-weighted brain MR images to MRA using learned local attention mask's. This work aims to enrich population imaging datasets which have a large amount of medical imaging data available for T1/T2 modal-

ities but limited data available for MRA.

- **Shape guided diffusion models to model brain vasculature [Chapter 5]:** The next contribution lies in the generation of different phenotypes of the intricate Circle of Willis (CoW), which serves as the meeting point of major vessels at the anterior portion of the brain. This is particularly motivated by the challenge related to the under-representation of certain phenotypes within available datasets. By addressing this disparity through conditional synthesis of realistic vascular structures, we aim to enrich the existing pool of data, thereby fostering a more comprehensive understanding of brain vasculature. This augmented dataset holds the potential to enhance the accuracy and robustness of neuroimaging analyses, ultimately improving diagnostic capabilities and treatment planning in various cerebrovascular disorders.

- **Few shot diffusion models to generate vessels with aneurysms [Chapter 6]:** The study of aneurysms in different locations in the brain is important to understand variations in their pathophysiology, prognosis, and optimal treatment strategies, contributing to improved patient outcomes and personalized medical interventions. However,given the infrequent occurrence of aneurysms, data availability is severely constrained, particularly for specific anatomical locations.To address this we build upon the previous contribution by further refining the approach by introducing few-shot learning with diffusion models to address the challenge of conditional generation with very limited training data available, specifically targeting the generation of brain vessels with aneurysms.

1.3 Thesis Structure

Chapter 2 provides the clinical background of this study, offering a concise introduction to relevant medical imaging concepts. The clinical background covers brain anatomy and vascular structure. The medical imaging background focuses on brain resonance imaging including its theory, physics, and various sequencing techniques in image acquisition.

Chapter 3 conducts a comprehensive literature review from both technical and methodological standpoints. It delivers a concise introduction to Deep Neural Networks (DNNs) and generative models, elucidating the fundamental components and

theoretical underpinnings. Special attention is given to Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs).

Chapter 4 *[based on publication 'Learned local attention masks for synthesising vessel segmentations']* presents an encoder-decoder model for synthesising segmentations of the main cerebral arteries in the circle of Willis (CoW) from only T2 MRI.We propose a two-phase multi-objective learning approach, which captures both global and local features by using learned local attention maps generated by dilating the segmentation labels, which forces the network to only extract information from the T2 MRI relevant to synthesising the CoW.

Chapter 5 *[based on publication 'Shape-guided conditional latent diffusion models for synthesising brain vasculature']* proposes a novel generative approach utilising a conditional latent diffusion model with shape and anatomical guidance to generate realistic 3D CoW segmentations, including different phenotypical variations. Our conditional latent diffusion model incorporates shape guidance to better preserve vessel continuity and demonstrates superior performance when compared to alternative generative models, including conditional variants of 3D GAN and 3D VAE.

Chapter 6 *[based on publication 'Few-shot learning in diffusion models for generating cerebral aneurysm geometries']* Explores the efficacy of training latent diffusion models (LDMs) for fewshot generation, enabling the generation of detailed vessel segmentations from as few as five images per class. By incorporating set-based vision transformers for class embeddings and leveraging signed distance functions (SDFs) as a novel form of conditioning, our method reduces the need for extensive datasets for training. Comparative studies with established generative models, including variational autoencoders (VAEs) and generative adversarial networks (GANs), highlight the robustness of our approach.

Chapter 7 Proposes an approach which uses conditional latent diffusion models to generate paired images of both the MRA and the corresponding binary mask. By incorporating , multi-task auto-encoders to generate the latent space and using features extracted from Vision transformers we are able to acheive superior results to other established generative models.

Chapter 8 Concluding the thesis, this section provides a summarization of the methodologies and findings presented in the preceding chapters. It critically examines the limitations inherent in the current research and proposes potential avenues for

future investigations.

Chapter 2

Clinical background

2.1 Medical Imaging

Medical imaging serves as a cornerstone in modern healthcare, facilitating the non-invasive visualization and analysis of anatomical structures, physiological functions, and pathological conditions within the human body .In the context of neurological assessment and brain imaging, medical imaging techniques offer invaluable insights into the structure, function, and pathology of the central nervous system. This section provides an introductory overview of medical imaging, encompassing a spectrum of modalities employed in clinical practice and research with a focus on brain imaging.

2.1.1 Introduction to Medical Imaging

Medical imaging encompasses a diverse array of techniques, each leveraging different physical principles and technologies to generate images of internal organs and tissues. These imaging modalities are indispensable tools for healthcare professionals, aiding in the diagnosis, treatment planning, and monitoring of various medical conditions. The evolution of medical imaging has been marked by significant milestones, from the discovery of X-rays by Wilhelm Conrad Rontgen in 1895 to the development of advanced modalities such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Nuclear Medicine Imaging. Over the decades, technological advancements have propelled medical imaging to new heights, enabling improved spatial resolution, faster image acquisition, and enhanced diagnostic accuracy.

At the core of medical imaging lie fundamental concepts that underpin the generation and interpretation of images. These concepts include the interaction of energy with human tissues, principles of image formation, and the manipulation of acquired data to extract diagnostic information. Understanding these fundamental principles is essential for healthcare professionals involved in the interpretation and utilization of medical imaging in clinical practice.

2.1.2 Imaging modalities in medical practice

Medical imaging encompasses a multitude of modalities, each offering unique strengths and applications in clinical practice. A short overview of some of the common imaging modalities used in practice is given below.

X-Ray - X-ray imaging involves the emission of ionizing radiation to produce images primarily of bones and dense tissues. It is commonly used for detecting fractures, assessing joint alignment, and diagnosing conditions such as pneumonia and certain tumors.

Computed Tomography (CT) - Computed Tomography (CT) utilizes X-rays to generate detailed cross-sectional images of the body. CT scanners rotate around the patient, acquiring multiple X-ray projections that are reconstructed into 3D images. CT imaging is valuable for diagnosing acute conditions such as trauma, vascular disorders, and abdominal pathologies.

Magnetic Resonance Imaging (MRI) - Magnetic Resonance Imaging (MRI) utilizes strong magnetic fields and radiofrequency pulses to generate detailed images of soft tissues and organs without ionizing radiation. MRI provides exceptional soft tissue contrast and multi-planar imaging capabilities, making it indispensable for imaging the brain, spine, and musculoskeletal system.

Nuclear Medicine Imaging - Nuclear Medicine Imaging techniques involve the administration of radioactive tracers that emit gamma rays, which are detected by specialized cameras to create images depicting physiological processes at the molecular level. Positron Emission Tomography (PET) and Single-Photon Emission Computed Tomography (SPECT) offer insights into conditions such as cancer, cardiac diseases, and neurological disorders.

Ultrasound Imaging - Ultrasound imaging utilizes high-frequency sound waves to produce real-time images of internal organs and structures. It is widely used in obstetrics for monitoring fetal development and in various clinical scenarios for imaging abdominal organs, the heart, and blood vessels.

2.2 Medical imaging techniques for brain imaging

Brain imaging plays a central role in the diagnosis and treatment planning of neurological conditions. By visualizing structural abnormalities, such as tumors, vascular malformations, and lesions, imaging techniques help clinicians localize pathology, determine its extent, and assess its impact on surrounding brain tissue. Brain imaging techniques have revolutionized the field of neuroscience, allowing for the detection, characterization, and monitoring of various neurological conditions.Brain imaging encompasses a diverse array of techniques, each offering unique insights into different

aspects of brain structure and function. These techniques include Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET), Single-Photon Emission Computed Tomography (SPECT), and Ultrasound, among others. Each imaging modality has its own strengths and limitations, making them suitable for different clinical scenarios and applications. In the context of this thesis we focus on MRI and the 3DRA modalities.

2.2.1 Computed Tomography (CT)

Computed Tomography (CT) and Computed Tomography Angiography (CTA) are among the most widely used techniques in clinical practice for brain imaging, particularly for vascular imaging and emergency diagnostic scenarios. CT is valued for its rapid acquisition time, widespread availability, and ability to detect acute conditions such as hemorrhages, skull fractures, and mass effect due to trauma or tumors. Its angiographic counterpart,Computed Tomography Angiography (CTA), involves the use of iodinated contrast agents to visualize cerebral vasculature with high spatial resolution. CTA is pivotal in diagnosing and characterizing conditions such as intracranial aneurysms, stenosis, arteriovenous malformations (AVMs), and acute ischemic strokes, particularly during the planning of thrombolytic therapy or endovascular interventions. Additionally, CTA is a non-invasive alternative to Digital Subtraction Angiography (DSA), offering a detailed assessment of vessel anatomy with less procedural risk. While CT and CTA have limitations such as ionizing radiation exposure and potential nephrotoxicity from contrast agents, their speed, accuracy, and accessibility make them indispensable tools in both acute and elective brain imaging workflows.While CT is briefly discussed as a general modality, its specialized application as CT Angiography (CTA) for brain imaging warrants deeper exploration due to its clinical significance.

2.2.2 Magnetic Resonance Imaging (MRI)

An MRI is one of the most critical tools used in the field of neurology and neurosurgery as it can generate high resolution images which provide details about the brain, spinal cord, and vasculature. MRI is based on the magnetization properties of atomic nuclei. A powerful, uniform, external magnetic field is employed to align the protons that are normally randomly oriented within the water nuclei of the tissue being examined. This alignment (or magnetization) is next perturbed or disrupted by introduction of

an external Radio Frequency (RF) energy. The nuclei return to their resting alignment through various relaxation processes and in so doing emit RF energy. After a certain period following the initial RF, the emitted signals are measured [8]. A Fourier transform is applied to convert this frequency information into a grey scale image. An MRI machine can generate different types of images by varying the sequence in which the radio waves are applied and collected along with changing parameters such as Repetition Time [TR] (time difference between pulse sequences applied to the same slice) and Time to Echo [TE] (time difference between release and collection of the pulse). [8] . Common MRI sequences are explained below .

T1-weighted MRI -T1-weighted imaging accentuates the differences in tissues' longitudinal relaxation times (T1). The contrast on T1-weighted imaging results from the rate at which excited protons return to their equilibrium, influenced by the tissue environment. For T1-weighted images, a short TR (400-600 ms) and short TE (10-20 ms) are employed. Typically, T1 images depict fat as bright (hyperintense) and water or fluid-filled regions as dark (hypointense). Thus, anatomical structures with high fat content, such as adipose tissue, appear brighter on T1-weighted images, whereas fluid-filled structures, like the brain's ventricles and the spinal canal containing cerebrospinal fluid (CSF), appear dark. [9]

T2-Weighted MRI - T2-weighted imaging accentuates differences in tissues' transverse relaxation times (T2). On T2-weighted images (T2WI), contrast arises from the rate at which excited protons lose phase coherence due to their interactions with the surrounding environment. For T2-weighted images, a long TR (3000-6000 ms) and a long TE (90-110 ms) are employed. In T2WI, fluids appear bright (or hyperintense), making areas filled with cerebrospinal fluid (CSF), such as the brain ventricles or the spinal canal, highly visible. [9]

Magnetic Resonance Angiography (MRA) - Magnetic Resonance Angiography (MRA) is a versatile and non-invasive imaging modality used to visualize blood vessels within the brain. MRA techniques, including Time-of-Flight (TOF), Phase Contrast (PC), and Contrast-Enhanced MRA (CE-MRA), offer distinct advantages in diagnosing vascular conditions such as intracranial aneurysms, arteriovenous malformations (AVMs), and stenoses. TOF-MRA leverages the inflow effect of unsaturated blood into the imaging volume, providing excellent spatial resolution without requiring contrast agents. PC-MRA measures blood flow velocities and directionality, which is particu-

larly useful for characterizing stenotic lesions or abnormal hemodynamics. CE-MRA uses gadolinium-based contrast agents to enhance vessel visibility, offering superior delineation of complex vascular structures.

However, MRA also has significant limitations that influence its clinical application. TOF-MRA can suffer from artifacts due to slow or turbulent blood flow, leading to an underestimation of vessel stenosis or occlusion. Phase contrast techniques are sensitive to motion artifacts and require long acquisition times, making them less suitable for emergency scenarios. Contrast-enhanced MRA, while highly effective, carries the risk of nephrogenic systemic fibrosis (NSF) in patients with renal insufficiency due to gadolinium-based contrast agents. Additionally, MRA generally has lower spatial resolution compared to Computed Tomography Angiography (CTA) or Digital Subtraction Angiography (DSA), limiting its ability to detect small vascular abnormalities such as tiny aneurysms. These limitations underscore the importance of integrating MRA with other imaging modalities to achieve comprehensive vascular assessment.

In summary, the combination of T1-weighted MRI, T2-weighted MRI, and MRA provides comprehensive information for the diagnosis and treatment of neurological disorders, offering detailed anatomical images and vascular mapping capabilities. T1-weighted imaging excels at displaying anatomical details and distinguishing between gray and white matter, while T2-weighted scans are superior for detecting fluid-related abnormalities such as edema, lesions, and certain pathologies. MRA complements these by offering detailed visualization of blood vessels without the need for contrast agents. A visual comparison of the three modalities is given in Figure 2.1.

2.2.3 Three-Dimensional Rotational Angiography (3DRA)

3DRA is a specialized form of angiography that combines X-ray imaging with rotational movements to create three-dimensional reconstructions of blood vessels within the brain. During a 3DRA procedure, a contrast agent is injected into the bloodstream, and a series of X-ray images are acquired as the X-ray source rotates around the patient. These images are then reconstructed using computer algorithms to generate a three-dimensional model of the cerebral vasculature.

3DRA utilizes principles similar to conventional angiography, with the added capability of acquiring images from multiple angles around the patient. This rotational acquisition allows for the visualization of blood vessels from different perspectives,

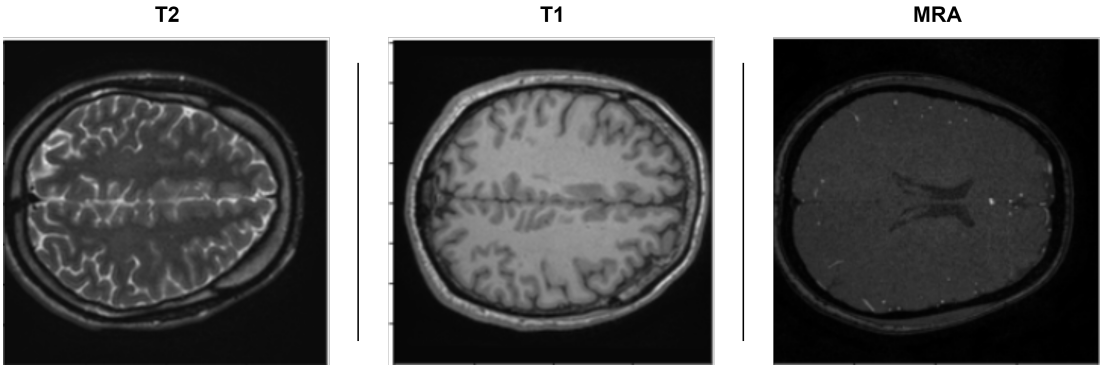

Figure 2.1: This figure shows a visual comparison between a slice of a brain scan taken in T1 , T2 and MRA modality. The T1 weighted scan highoights areas with fat tissues while T2 weighted scan highlight areas with fluids. The MRA on the other hand highlights vascular structures in the brain.

providing a comprehensive view of the vascular anatomy. By reconstructing these images into a three-dimensional model, 3DRA enables precise localization of vascular lesions, assessment of vessel morphology, and planning of endovascular interventions.

3DRA is important for the diagnosis and management of various cerebrovascular disorders, including aneurysms, arteriovenous malformations (AVMs), and stenotic lesions. The three-dimensional reconstructions provided by 3DRA offer detailed information about the size, shape, and spatial relationship of vascular structures, facilitating the identification of abnormal vessels and planning of therapeutic interventions. In addition, 3DRA plays a crucial role in guiding endovascular procedures such as embolization, stent placement, and coiling by providing real-time visualization of catheter navigation and device deployment.

Each MR modality T1-weighted, T2-weighted imaging, MRA, and 3DRA has distinct strengths and limitations, making them suitable for different clinical applications. T1-weighted imaging excels in providing detailed anatomical structures, particularly for fat-containing tissues, but is less effective for detecting pathologies with high water content. T2-weighted imaging, on the other hand, is ideal for identifying edema and inflammation, though it requires longer scan times and may need fat suppression. MRA offers a non-invasive method for vascular imaging, although it has lower spatial resolution and can be prone to artifacts. In contrast, 3DRA provides highly detailed,

real-time vascular images crucial for interventional procedures, but it is invasive, involves radiation, and is less widely available. This makes cross-modality image synthesis important in this context because it can potentially combine the strengths of these different imaging techniques while mitigating their individual limitations. By synthesizing information from multiple modalities, clinicians can obtain a more comprehensive and accurate picture of a patient's condition without subjecting them to additional scans or invasive procedures.

2.3 Introduction to brain vasculature

The brain vasculature consists of a complex network of arteries and veins that supply oxygenated blood to the brain and facilitate the removal of metabolic waste products. This intricate system ensures adequate perfusion of brain tissue and plays a crucial role in maintaining neurological function. The arterial supply to the brain originates from the internal carotid arteries and vertebral arteries, which give rise to a series of interconnected vessels that form the Circle of Willis and its associated branches. Understanding the anatomy and function of the brain vasculature is essential for diagnosing and managing cerebrovascular diseases, including aneurysms.

2.3.1 Circle of Willis

The Circle of Willis is a vital anatomical structure located at the base of the brain that serves as a key collateral pathway for cerebral blood flow. It is formed by the convergence of the anterior cerebral arteries (ACAs), middle cerebral arteries (MCAs), posterior cerebral arteries (PCAs), and communicating arteries (anterior and posterior) that connect the anterior and posterior circulation. The Circle of Willis plays a crucial role in maintaining cerebral perfusion and ensuring adequate blood supply to different regions of the brain. The complete architecture of CoW is shown in Figure 2.2.

While the Circle of Willis is a critical component of cerebral circulation, it's important to note that not all individuals possess a complete or "textbook" version of this arterial network. Studies have shown that a complete Circle of Willis is present in only about 42-52% of the population [10]. Variations in its structure are common and can have significant implications for cerebrovascular health. The most frequent variation is the absence or hypoplasia of one or both posterior communicating arteries (PCoA),

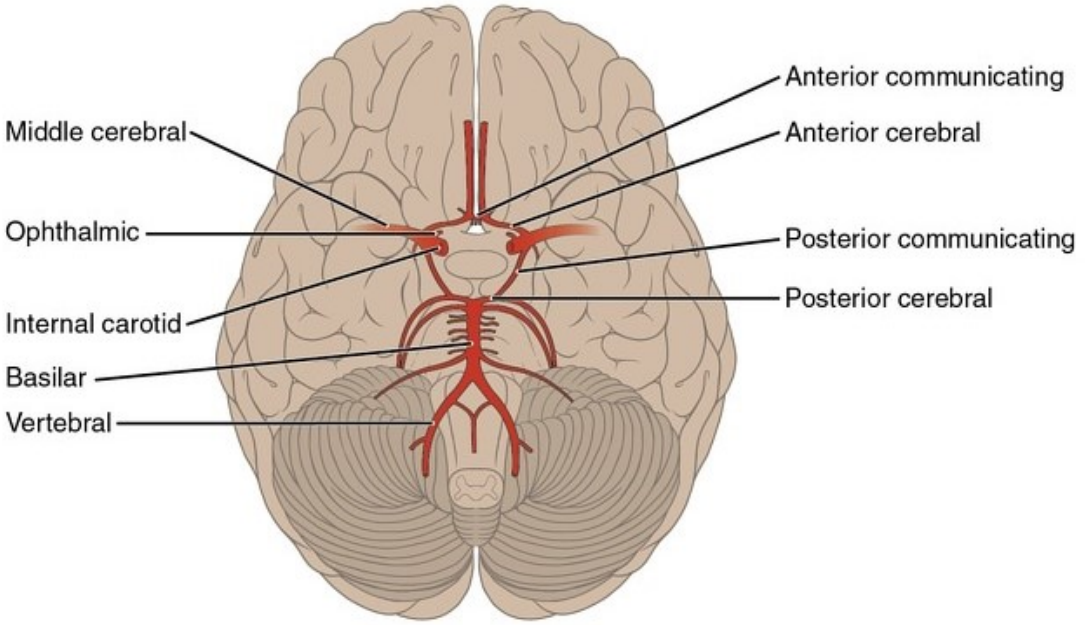

Figure 2.2: Inferior view of the Circle of Willis and major cerebral arteries. This diagram illustrates the critical arterial network at the base of the brain, including the anterior communicating, anterior cerebral, middle cerebral, internal carotid, posterior communicating, posterior cerebral, basilar, and vertebral arteries. The Circle of Willis, formed by these interconnecting vessels, provides crucial collateral circulation to ensure consistent blood supply to different regions of the brain, offering protection against potential ischemic events. The figure was taken from [1].

occurring in up to 30-60% of individuals [11]. Other common variations include absence or hypoplasia of the anterior communicating artery (ACoA) in 5-10% of people, a fetal-type posterior cerebral artery (PCA) in 15-20% of cases, and asymmetry in the A1 segments of the anterior cerebral arteries in about 10% of individuals. [11].These anatomical differences can impact the brain's ability to maintain adequate blood flow in case of occlusion or stenosis in one of the main feeding arteries. For instance, individuals with an incomplete Circle of Willis may be at higher risk for ischemic events during carotid endarterectomy or may have reduced collateral flow capacity in the event of a

stroke [12]. However, it's worth noting that many people with variations in their Circle of Willis remain asymptomatic throughout their lives, as the brain can often adapt to these differences in vascular anatomy [13]. While variations in the Circle of Willis are well-documented, it is important to note that significant vascular variability also exists in other major cerebral vessels, including the MCA and ACA, as well as the posterior circulation.

Variability in the MCA can include differences in branching patterns, such as early bifurcation, accessory MCA branches, or variations in the length of the M1 segment. These variations can influence the presentation and outcomes of cerebrovascular diseases, such as stroke or aneurysms, by altering blood flow distribution or collateral circulation. The ACA similarly exhibits variability in its A1 segment, including hypoplasia or asymmetry, which can affect perfusion in anterior brain regions. Moreover, variations in posterior circulation vessels, such as the basilar artery and posterior cerebral arteries, can impact the brainstem and occipital perfusion, particularly in cases of occlusion or ischemia.

The functional significance of these variations extends beyond the Circle of Willis, as they can influence both the risk of ischemic events and the outcomes of interventions. For example, incomplete or asymmetrical branching patterns may compromise collateral flow, increasing susceptibility to ischemia during vascular events or surgeries. These variations highlight the importance of a comprehensive understanding of cerebral vasculature when diagnosing and managing cerebrovascular diseases, as well as the potential for advanced imaging techniques to map these variations more accurately.

2.3.2 Cerebral Aneurysms

Cerebral aneurysms are abnormal dilations or bulges in the walls of cerebral arteries, often occurring at branching points or bifurcations . These vascular abnormalities pose a significant risk of rupture, leading to potentially life-threatening subarachnoid hemorrhage (SAH) and neurological deficits [14]. Aneurysms can develop as a result of various factors, including genetic predisposition, hemodynamic stress, and underlying vascular diseases such as fibromuscular dysplasia or connective tissue disorders [14]. Common locations for cerebral aneurysms include the anterior communicating artery (ACoA), posterior communicating artery (PCoA), and the bifurcations of the anterior and middle cerebral arteries [15]. The diagnosis and management of cerebral

aneurysms involve a multidisciplinary approach, including advanced imaging studies, neurosurgical interventions, and endovascular treatments. Medical imaging plays a crucial role in the detection, characterization, and monitoring of aneurysms. Computed Tomography Angiography (CTA) serves as an initial screening tool, providing detailed 3D images of cerebral vasculature. Magnetic Resonance Angiography (MRA) offers high-resolution images without radiation exposure, particularly useful for detecting small aneurysms and assessing their relationship with surrounding brain tissue [16]. Digital Subtraction Angiography (DSA) remains the gold standard for aneurysm detection and characterization, offering real-time, high-resolution images valuable for treatment planning and during endovascular procedures [17]. Advanced techniques like 4D-CTA and 4D-Flow MRI allow for the assessment of blood flow dynamics within aneurysms, providing insights into hemodynamic stress and rupture risk . Treatment options include microsurgical clipping, where a metal clip is placed across the aneurysm neck, and endovascular techniques such as coiling, which involves deploying platinum coils into the aneurysm sac to promote thrombosis [17]. Flow diversion devices and stent-assisted coiling represent more recent advancements in endovascular treatment, particularly useful for complex aneurysm morphologies. The choice of treatment modality depends on various factors, including aneurysm characteristics, patient age, and comorbidities. Regular imaging follow-ups are essential for monitoring unruptured aneurysms and assessing treatment effectiveness. Ongoing research focuses on improving treatment outcomes, developing novel endovascular devices, and enhancing our understanding of aneurysm pathophysiology to identify potential targets for pharmacological interventions [17].

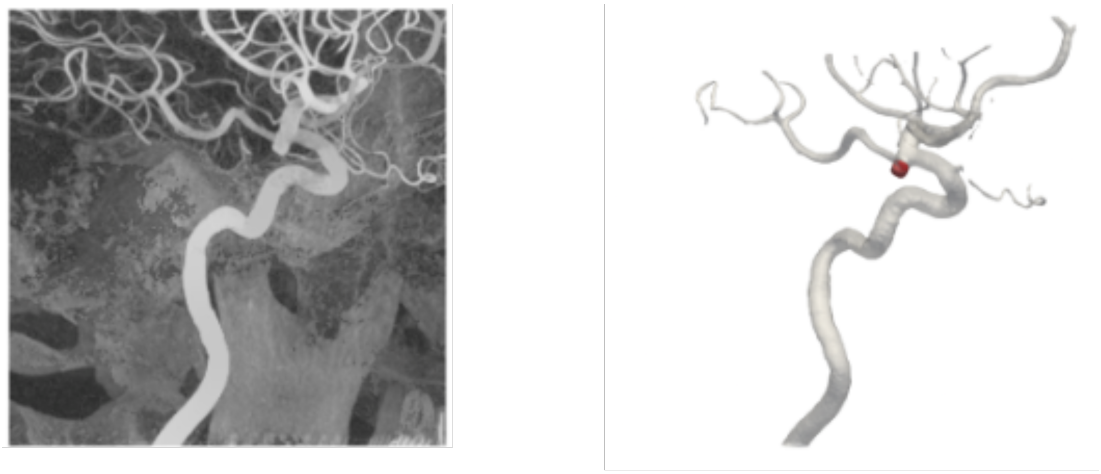

Figure 2.3: This figure shows a 3 Dimensional Rotational Angiography (3DRA) scan (left) which shows contrast enhanced vessels and the segmented vessels (right) and aneurysm (denoted in red) from the scan

Chapter 3

Technical Background

3.1 Deep Learning Background

Deep Learning has emerged as a transformative approach in the field of artificial intelligence, driven by its ability to allow computational models composed of multiple processing layers to learn representations of data with multiple levels of abstraction. This powerful capability enables deep learning models to effectively handle large-scale data and complex tasks, making them indispensable in various domains, including medical imaging.

Before delving into the specifics of artificial neural networks (ANNs), it's important to understand the historical context and foundational concepts that led to the development of deep learning. The journey began in the 1940s, with the pioneering work of Warren McCulloch and Walter Pitts, who proposed the first mathematical model of a neuron in 1943 [18]. Their model laid the groundwork for the development of artificial neural networks, inspired by the biological processes of the human brain. Despite the initial enthusiasm, progress was slow due to the limited computational resources and theoretical understanding available at the time.

Artificial Neural Networks (ANN)

Deep Learning can be described as something which "allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction" [19]. An artificial neural network (ANN) is the backbone of deep learning. Although deep learning has risen in popularity quite recently, the original concept of an artificial neural network (ANN) was initially proposed in 1943 [20]. A neuron (also called Perceptron) is a building block of an ANN and consists of four parts: the input, the output, the weights, and the bias. The way a neuron works is that all inputs are multiplied by their respective weights and a bias is added. This weighted sum is fed to an activation function, which decides if and how a neuron should be activated (for example, the step activation function returns 1 if the weighted sum is greater than a particular value, otherwise 0). The purpose of an activation function is to introduce non-linearity into a neural network. The output is then compared to the ground truth, and the error is back-propagated through the network, updating the weights and biases , this is how a neuron 'learns'. Modern neural networks consist of several neurons arranged in layers where the first layer is called the input layer and the last layer is called the output layer, with all the layers in between called the hidden

layers. The structure of the perceptron and the general architecture of the neural network is show in 7.1

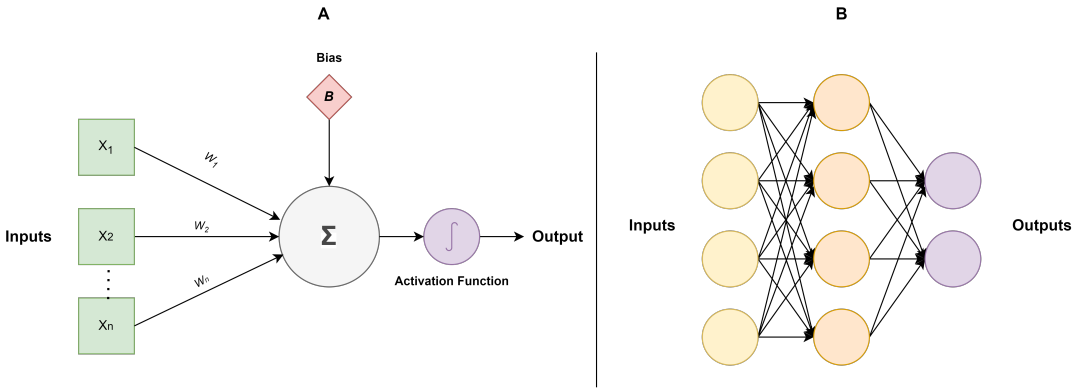

Figure 3.1: Part A of the figure shows the general structure of a perceptron while Part B showcase the general structure of an ANN .

However, traditional neural networks had significant drawbacks, such as being very prone to overfitting and suffering from the vanishing gradient problem. These issues led researchers to favor machine learning techniques such as Support Vector Machines (SVMs) [21]. The resurgence of neural networks was due to a combination of several factors. Firstly, breakthroughs in the field of deep learning brought significant changes to the structure of neurons and introduced algorithms such as Dropout [22], which tackled the overfitting problem. The introduction of activation functions such as the Rectified Linear Unit (ReLU) [23] helped avoid the vanishing gradient problem, making neural networks a viable choice again.

In addition to these algorithmic advancements, improvements in modern computing allowed researchers to use a large number of neurons and hidden layers in their neural architectures. This development enabled them to create far more complex models than those traditional machine learning techniques were capable of. Despite these advancements, the amount of data required to train deep neural networks was still very high and not readily available. However, this problem was soon alleviated with the rapid advancement of the internet, which made massive, labeled datasets readily available for researchers to use.

In conclusion, the convergence of theoretical advancements, innovative algorithms, and enhanced computational power has propelled deep learning to the forefront of

artificial intelligence research. Today, deep learning models are not only feasible but also outperform many traditional approaches, especially in complex tasks such as medical imaging, where they have demonstrated remarkable accuracy and efficiency.

CNN

Traditional artificial neural networks (ANNs) face significant challenges when it comes to handling image data. One major issue is that ANNs do not scale well with the increasing dimensionality of images. For example, a color image of size 256x256 pixels has 196,608 input features (256 * 256 * 3), which results in an overwhelming number of parameters to be learned, leading to increased computational cost and risk of overfitting. Moreover, ANNs lack the ability to capture the spatial hierarchies in images, meaning they cannot effectively recognize patterns such as edges, textures, or shapes that are critical for understanding visual content. These limitations necessitated the development of specialized architectures that could handle the high dimensionality and spatial nature of image data more efficiently.

Convolutional Neural Networks (CNNs) are a type of neural network architecture that has gained tremendous popularity in recent years, especially in the field of computer vision. CNNs consist of a special kind of layer called convolutional layers. Each cell in the convolutional layer applies a convolution operation on the input. A convolution consists of a filter, also known as a kernel, which converts all the pixels in its receptive field into a single value by performing element-wise multiplications followed by a summation.

The neurons in CNNs adopt a ”weight sharing” approach, which allows them to detect complex patterns present in the input data efficiently. This approach significantly reduces the number of parameters in the model, making it more computationally efficient and less prone to overfitting. Owing to this, CNNs have been used very successfully for a wide range of computer vision and image processing tasks.

CNNs use convolutional layers to extract patterns from the input data. In the early layers, simple patterns such as edges and textures are detected. These simple patterns are then combined in the deeper layers to form more complex patterns, such as shapes and objects. After the patterns are extracted, a pooling operation is performed. The function of pooling is to progressively reduce the spatial size of the representation, which helps to reduce the number of parameters and computation in the network, as

well as to control overfitting.

Max pooling, which selects the maximum value from each patch of the feature map, and average pooling, which computes the average value of each patch, are common pooling strategies used in CNNs. These operations help to make the detection of features invariant to small translations of the input, thereby improving the robustness of the model.

Autoencoders

Autoencoders are a type of artificial neural network designed for unsupervised learning tasks, particularly for dimensionality reduction and feature learning. Unlike PCA, which is limited to linear transformations, autoencoders can model non-linear relationships, making them more flexible and powerful for capturing complex data structures.

Structure and Functionality

An autoencoder consists of two main components:

- **Encoder**: The encoder maps the input data x to a latent representation z in a lower-dimensional space. This process captures the essential features of the input data.
- **Decoder**: The decoder reconstructs the original data x from the latent representation z. The goal is to minimize the difference between the input and the reconstructed output, often measured by a loss function such as Mean Squared Error (MSE).

The objective of an autoencoder is to learn a compact and meaningful representation of the data while preserving as much information as possible. The loss function used to train the autoencoder is given by:

$$\mathcal{L}(x, \hat{x}) = \|x - \hat{x}\|^2$$

where $\hat{x}$ is the reconstructed output.

Advantages and Applications

Autoencoders offer several advantages over linear techniques like PCA:

- They can capture non-linear dependencies in the data, making them suitable for complex datasets.
- Autoencoders can be used for various tasks, including dimensionality reduction, anomaly detection, denoising, and feature learning.
- They can automatically learn useful representations from raw data, which can be used for downstream machine learning tasks.

Autoencoders are a foundational concept in deep learning, paving the way for more advanced models like variational autoencoders (VAEs) and generative adversarial networks (GANs). They provide a flexible and powerful tool for reducing dimensionality and uncovering the underlying structure of complex datasets.

RNN

Recurrent Neural Networks (RNNs) [24] represent a powerful class of artificial neural networks designed specifically to handle sequential data. Unlike traditional feedforward neural networks, which assume independence between inputs and outputs, RNNs incorporate feedback connections. These connections allow the network to maintain an internal state, effectively creating a form of memory that captures information from previous inputs. This unique architecture makes RNNs particularly well-suited for tasks involving time series, natural language processing, and other sequence-based problems. The core idea behind RNNs is to leverage sequential information. They are called "recurrent" because they perform the same task for every element of a sequence, with the output being dependent on previous computations. In essence, RNNs have a "memory" which captures information about what has been calculated so far. In its basic form, an RNN consists of an input layer, a hidden layer with recurrent connections, and an output layer. At each time step t, the hidden layer receives input from both the current input x_t and its own output from the previous time step h_{t-1}. This recurrent connection is what allows information to persist. Mathematically, we can describe the operation of a basic RNN as follows:

$$h_t = f(W_{hh}h_{t-1} + W_{xh}x_t + b_h) \tag{3.1}$$

$$y_t = g(W_{hy}h_t + b_y) \tag{3.2}$$

Where $x_t \in \mathbb{R}^d$ is the input vector at time step t, $h_t \in \mathbb{R}^h$ is the hidden state, $y_t \in \mathbb{R}^o$ is the output vector, $W_{hh} \in \mathbb{R}^{h \times h}$, $W_{xh} \in \mathbb{R}^{h \times d}$, and $W_{hy} \in \mathbb{R}^{o \times h}$ are weight matrices, $b_h \in \mathbb{R}^h$ and $b_y \in \mathbb{R}^o$ are bias vectors, and f and g are activation functions. The hidden state h_t serves as the network's "memory", capturing information about what has been seen in previous time steps. Figure 3.2 shows the 'unfolding' process of the hidden state in an RNN.

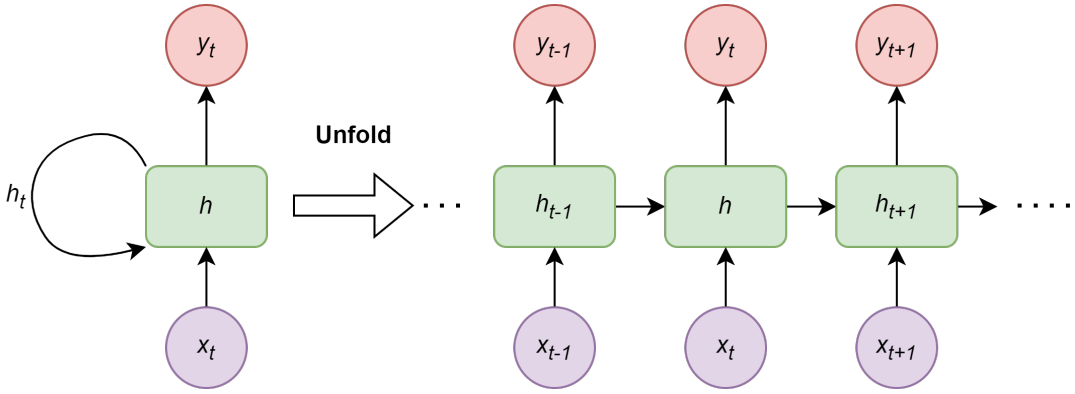

Figure 3.2: This figure demonstrate the working of the hidden state of the RNN .

The output y_t is computed based on this hidden state. Training RNNs involves a process called Backpropagation Through Time (BPTT). This algorithm unfolds the recurrent network into a full network for the entire sequence and then applies standard backpropagation. For a sequence of length T, the total loss L is the sum of the losses at each time step:

$$L = \sum_{t=1}^{T} L_t(y_t, \hat{y}_t) \tag{3.3}$$

Where L_t is the loss function at time step t, y_t is the true output, and $\hat{y}_t$ is the predicted output. The gradients of the loss with respect to the weights are computed by summing over all time steps:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W} \tag{3.4}$$

This training process, while powerful, can be challenging due to the vanishing and exploding gradient problems, especially for long sequences. The vanishing gradient problem occurs when gradients become extremely small as they are propagated back through time, making it difficult for the network to learn long-term dependencies. Con-

versely, the exploding gradient problem occurs when gradients become extremely large, leading to unstable training. To address these challenges, researchers have developed several variants of RNNs, including Long Short-Term Memory (LSTM) [25] networks and Gated Recurrent Units (GRUs) [26]. These architectures introduce gating mechanisms that allow for better control of information flow through the network, mitigating the vanishing and exploding gradient problems.

3.2 Attention Mechanisms in Deep Learning

Attention mechanisms were introduced to enhance the capability of neural networks in processing sequence data, particularly in tasks where different parts of the input sequence have varying levels of importance for generating the output. The concept of attention in neural networks was inspired by the human cognitive process of selectively concentrating on specific aspects of information while ignoring others. In the context of machine learning, attention allows models to focus on the relevant parts of the input when producing each element of the output. The first widely adopted use of an attention mechanism was adopted for neural machine translation [27].

3.2.1 Transformers

Transformers, introduced by Vaswani et al. in their 2017 paper "Attention Is All You Need," represent a significant leap forward in the field of sequence processing. Unlike their predecessors, Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs), Transformers rely entirely on attention mechanisms, dispensing with recurrence and convolutions entirely. This architecture has proven remarkably effective, particularly in natural language processing tasks, and has since become the foundation for many state-of-the-art models. The key innovation of the Transformer model is its use of self-attention, a mechanism that allows the model to weigh the importance of different parts of the input when processing each element. This enables the model to capture long-range dependencies more effectively than RNNs, which process sequences linearly.

The Transformer model is a highly efficient architecture for handling sequential data, particularly in tasks like translation and text generation. It overcomes the limitations of traditional sequence models, such as recurrent neural networks (RNNs), by allowing

parallelization and eliminating the need for sequential processing. The Transformer is composed of two main components: the encoder and the decoder. Let's explore how information flows through these components, step by step.

Encoder Workflow

The encoder processes the input sequence to generate contextualized representations. Tokens are first embedded into continuous vector representations, and positional encodings are added to incorporate sequence information:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

Self-attention then enables each token to attend to all others, computed using query (Q), key (K), and value (V) vectors:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Multi-head attention extends this by applying multiple attention mechanisms in parallel, enhancing the models ability to capture complex dependencies:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

A feed-forward network (FFN) refines each tokens representation, followed by residual connections and layer normalization for stability:

$$\text{LayerNorm}(x + \text{Sublayer}(x))$$

This process repeats across multiple layers, allowing the model to iteratively refine token representations.

Decoder Workflow

The decoder generates the output sequence autoregressively while attending to the encoders output. Target tokens are embedded, and positional encodings are applied. Self-attention in the decoder is masked to prevent future tokens from influencing the current position:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \text{mask}\right)V$$

The encoder-decoder attention mechanism allows the decoder to selectively focus on relevant encoded representations while generating output:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Similar to the encoder, a feed-forward network with residual connections and layer normalization further refines token representations. During training, the decoder uses teacher forcing, while at inference, it generates tokens sequentially, using previous outputs as inputs for the next step.

The combination of the encoders contextualized token representations and the decoders ability to attend to both previous tokens and the encoder output allows the Transformer to generate high-quality sequences for tasks like translation, summarization, and beyond.

The loss is typically computed using cross-entropy, and the model is optimized using variants of stochastic gradient descent, often with adaptive learning rates like Adam.

Transformers have revolutionized natural language processing, forming the basis for models like BERT [28],and GPT [29]. Moreover, the concept of self-attention has proven valuable beyond NLP, finding applications in computer vision [30], speech processing [31], and even protein structure prediction [32].

3.2.2 Vision Transformers (ViTs)

While Transformers were initially designed for natural language processing tasks, their success has inspired adaptations to other domains, including computer vision. The Vision Transformer (ViT), introduced by Dosovitskiy et al. in 2020, successfully applies the Transformer architecture to image recognition tasks, challenging the dominance of convolutional neural networks (CNNs) in this field.

Core Idea

The key insight of Vision Transformers is to treat an image as a sequence of patches, analogous to how a Transformer processes a sequence of words. This approach allows the direct application of the Transformer architecture to image data with minimal modifications. By splitting an image into smaller patches and embedding them as

input tokens, the Transformer can process the image in a similar manner to how it processes text in NLP tasks.

Architecture

The Vision Transformer (ViT) processes images by converting them into sequences of patches and passing them through a Transformer encoder for classification.

The input image of size (H, W, C) is divided into N fixed-size patches of $P \times P$, treated as individual tokens. Each patch is flattened into a vector of size P^2C and projected into an embedding space of dimension D using a trainable linear layer:

$$x_p \in \mathbb{R}^{N \times D}$$

Since Transformers lack inherent spatial awareness, positional embeddings are added to retain spatial information:

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{pos}$$

A learnable class token x_{class} is prepended to the sequence, serving as a global representation used for classification.

The sequence is then processed by a Transformer encoder consisting of L layers, where each layer applies multi-head self-attention (MSA) and feed-forward networks (MLP) with residual connections:

$$z_l' = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \quad z_l = \text{MLP}(\text{LN}(z_l')) + z_l', \quad l = 1, \ldots, L$$

Finally, the class token's output representation undergoes layer normalization and is used for classification:

$$y = \text{LN}(z_L^0)$$

where z_L^0 is the final state of the class token after L Transformer layers. This approach enables ViT to model global dependencies efficiently across image patches.

Training and Performance

Vision Transformers are typically pre-trained on large datasets (e.g., ImageNet-21k, JFT-300M) and then fine-tuned on specific tasks. When pre-trained on sufficiently

large datasets, ViTs have demonstrated the ability to outperform state-of-the-art convolutional neural networks (CNNs) in image classification tasks, while requiring fewer computational resources during training.

In summary, the Vision Transformer architecture successfully adapts the self-attention mechanism to image data by treating images as sequences of patches. This approach, combined with large-scale pre-training, has allowed ViTs to rival or even surpass traditional CNN architectures in many vision tasks.

3.3 Generative models

Generative models are statistical models that learn a joint probability distribution, , over observed data and associated labels (if available). This foundational perspective predates deep learning and has its roots in statistical modeling and probabilistic reasoning. The primary objective of a generative model is to understand the underlying data distribution to allow sampling of new data points that are statistically similar to those in the original dataset. In contrast to discriminative models, which focus on modeling the conditional probability , generative models aim to capture the full joint distribution, offering greater flexibility and broader applications.

Historically, generative modeling has been explored through a variety of approaches such as Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), and Bayesian Networks. These traditional methods provided robust statistical frameworks for capturing complex data distributions and were particularly effective for structured data. However, they often struggled with scalability and handling high-dimensional datasets, limiting their applicability in more intricate domains.

Deep learning has rejuvenated generative modeling by enabling the creation of models capable of learning intricate, high-dimensional data distributions. There are several variations of deep learning based generative models, each with its unique approach and applications. Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) are two of the most widely used generative models. VAEs use a probabilistic framework to learn a latent representation of the data, which can then be used to generate new samples. GANs, on the other hand, employ a game-theoretic approach where two neural networks, a generator and a discriminator, are trained simultaneously. The generator creates new data samples, while the discriminator evaluates their authenticity, leading to the generation of increasingly realistic data. An overview of different

generative model types is show in in Figure 3.3 .

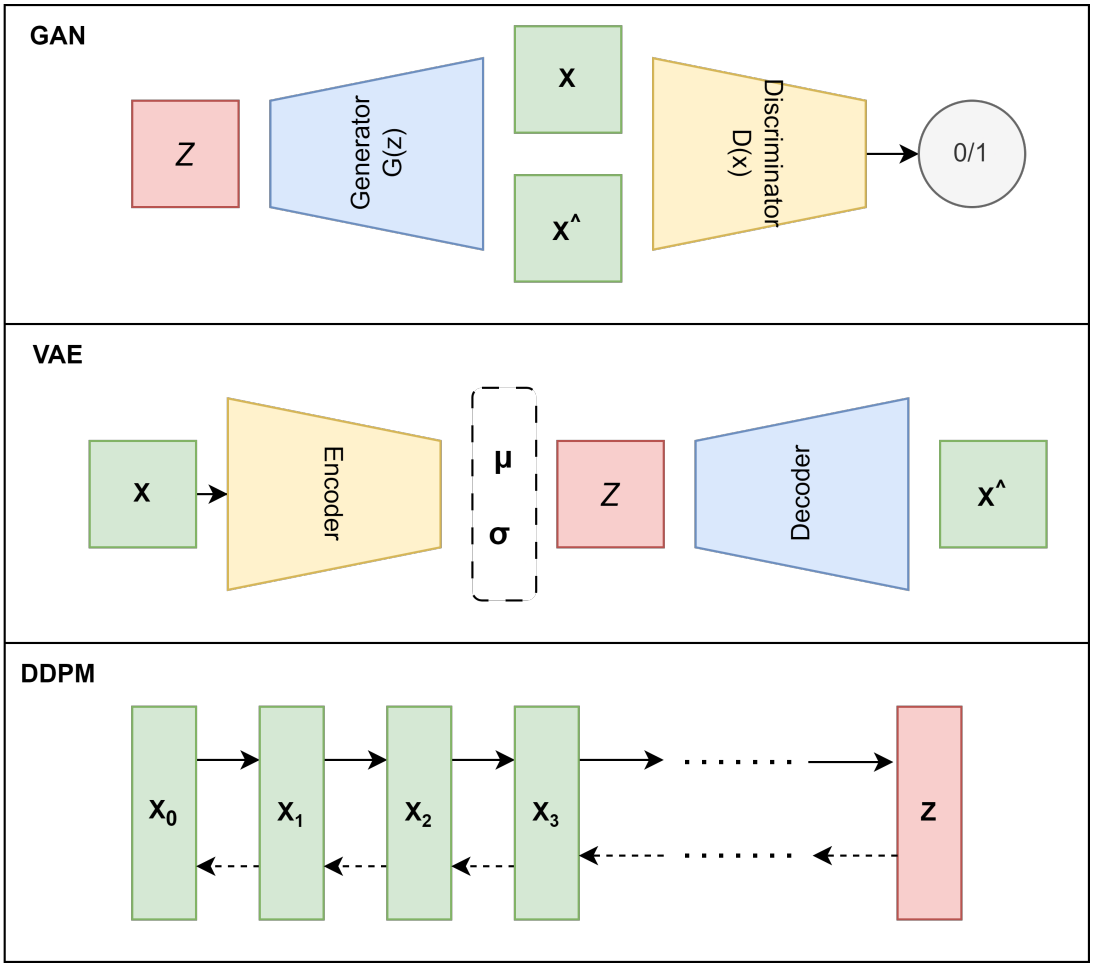

Figure 3.3: Schematic illustration of Variational Auto Encoder (VAE), Generative Adversarial Network (GAN), and Diffusion Models (DDPM). In a GAN, the process begins with a noise vector (Z) that a generator uses to produce an image ($\hat{X}$); this generated image is then compared to a real image (X) by the discriminator, determining the probability of the image being real or fake. Contrastingly, a VAE encodes the input image (X) into distributions within a latent space defined by parameters μ and σ, with the decoder sampling from this distribution to create new images. For a DDPM, noise is progressively added to the input image (X) following a predetermined schedule, and the network is trained to reverse this process to derive an image from the noise.

3.3.1 Variations Auto-encoders

Variational Autoencoders (VAEs) are a class of generative models that combine principles from Bayesian inference and neural networks to learn a probabilistic mapping between data and latent variables. Introduced by Kingma and Welling in 2013 [33], VAEs are particularly powerful for tasks involving high-dimensional data, such as image generation and representation learning.

Architecture of VAEs

The VAE architecture consists of two main components: the encoder (inference network) and the decoder (generative network).

- **Encoder (Inference Network)**: The encoder maps the input data $\mathbf{x}$ to a latent representation $\mathbf{z}$. Instead of mapping directly to a point in the latent space, the encoder learns the parameters of a probability distribution (typically Gaussian), characterized by a mean μ and a standard deviation σ. The encoder network outputs these parameters, which are used to sample $\mathbf{z}$ from the distribution:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu_\phi(\mathbf{x}), \sigma_\phi(\mathbf{x}))$$

  where ϕ denotes the parameters of the encoder network.

- **Decoder (Generative Network)**: The decoder takes the latent representation $\mathbf{z}$ and reconstructs the original data $\mathbf{x}$. The decoder network models the conditional distribution $p_\theta(\mathbf{x}|\mathbf{z})$:

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mu_\theta(\mathbf{z}), \sigma_\theta(\mathbf{z}))$$

  where θ denotes the parameters of the decoder network.

Learning Objective

The objective of training a VAE is to maximize the Evidence Lower Bound (ELBO) on the marginal likelihood of the data. The ELBO can be decomposed into two terms: the reconstruction loss and the regularization term. The reconstruction loss ensures that the decoded samples are close to the original input data, while the regularization term ensures that the learned latent space is smooth and follows a prior distribution, usually a standard normal distribution.

The ELBO is given by:

$$\mathcal{L}(\phi, \theta; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))$$

where $\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))$ is the Kullback-Leibler divergence between the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior $p(\mathbf{z})$.

The first term, the reconstruction loss, can be interpreted as the negative log-likelihood of the data under the decoder's distribution. The second term, the KL divergence, acts as a regularizer that penalizes deviations from the prior.

Reparameterization Trick

To enable backpropagation through the stochastic sampling of $\mathbf{z}$, VAEs use the reparameterization trick. Instead of sampling $\mathbf{z}$ directly from $\mathcal{N}(\mu, \sigma)$, we sample an auxiliary variable ϵ from a standard normal distribution and then transform it:

$$\mathbf{z} = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$$

This reparameterization allows the gradient to propagate through μ and σ during training.

VAEs offer several advantages over traditional autoencoders , namely:

- **Probabilistic Interpretation**: VAEs provide a probabilistic framework for modeling data, allowing for uncertainty quantification and principled Bayesian inference.
- **Continuous and Smooth Latent Space**: The regularization term ensures that the latent space is continuous and smooth, enabling meaningful interpolation between points in the latent space.

3.3.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a class of generative models introduced by Ian Goodfellow and his colleagues in 2014 [34]. GANs consist of two neural networks, a generator and a discriminator, which are trained simultaneously through a process of adversarial competition. This innovative approach has led to significant advancements in generating realistic data across various domains, including images, text, and audio.

Architecture of GANs

The architecture of GANs involves two main components:

- **Generator**: The generator network G takes a random noise vector $\mathbf{z}$ sampled from a prior distribution (typically a standard normal distribution) and transforms it into a data sample $G(\mathbf{z})$. The goal of the generator is to produce data that is indistinguishable from real data.
- **Discriminator**: The discriminator network D takes an input data sample (either from the real dataset or generated by G) and outputs a probability $D(\mathbf{x})$ representing the likelihood that the input data is real. The goal of the discriminator is to correctly classify real and generated data.

Learning Objective

GANs are trained through a minimax game, where the generator and discriminator have opposing objectives. The discriminator aims to maximize the probability of correctly classifying real and generated samples, while the generator aims to minimize the probability that the discriminator correctly identifies generated samples as fake. The objective function for GANs is given by:

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

In this setup:

- $p_{\text{data}}(\mathbf{x})$ is the distribution of the real data.
- $p_{\mathbf{z}}(\mathbf{z})$ is the prior distribution of the noise vector $\mathbf{z}$.

The generator G tries to fool the discriminator D by generating realistic data samples, while D tries to distinguish between real and generated samples. This adversarial process continues until the generator produces samples that are indistinguishable from real data.

Training Process

The training process of GANs involves iteratively updating the parameters of the generator and discriminator. The typical training algorithm is as follows:

1. Sample a batch of noise vectors $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m\}$ from the prior distribution $p_{\mathbf{z}}(\mathbf{z})$.
2. Generate a batch of fake data samples $\{G(\mathbf{z}_1), G(\mathbf{z}_2), \ldots, G(\mathbf{z}_m)\}$ using the generator.
3. Sample a batch of real data samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ from the real data distribution $p_{\text{data}}(\mathbf{x})$.
4. Update the discriminator by maximizing the objective function $\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$.
5. Update the generator by minimizing the objective function $\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$.

This process is repeated for many iterations until the generator produces high-quality data samples that the discriminator cannot distinguish from real data.

3.3.3 Score Based Generative Models

Score-based models are a class of generative models that focus on estimating the gradient (or score) of the data distribution's log density. These models aim to generate new data samples by iteratively refining noisy data points based on the learned score function. The score function provides a direction in which the data can be moved to increase its likelihood under the target distribution. Score-based models have shown great promise in generating high-quality data, particularly in scenarios where traditional generative models might struggle.

Diffusion Models

Diffusion models, a type of score-based generative model, have gained significant attention for their ability to produce realistic and high-quality data samples. These models draw inspiration from non-equilibrium thermodynamics, specifically the process of diffusion, where particles spread from regions of high concentration to low concentration over time.

Diffusion models simulate a forward and reverse diffusion process to generate new data samples. The forward process involves gradually adding noise to the data, transforming it into a noisy distribution, while the reverse process aims to denoise the data, reconstructing the original distribution.

Forward Diffusion Process

The forward diffusion process transforms the data $\mathbf{x}_0$ into a series of progressively noisier versions $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ over T time steps. This process can be described as a Markov chain, where each step adds a small amount of Gaussian noise to the data. Mathematically, the forward process is defined as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

where β_t is a noise schedule that controls the amount of noise added at each step. The initial data $\mathbf{x}_0$ is gradually transformed into pure noise $\mathbf{x}_T$.

Reverse Diffusion Process

The reverse diffusion process aims to recover the original data from the noisy distribution. This process is also described as a Markov chain, where each step denoises the data. The reverse process is defined as:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$$

where μ_θ and Σ_θ are learned functions parameterized by a neural network. The goal is to learn these functions such that the reverse process accurately reconstructs the original data distribution.

Training Objective

The training objective of diffusion models is to learn the parameters of the reverse process by minimizing the Kullback-Leibler (KL) divergence between the true forward process and the learned reverse process. This can be formulated as minimizing the negative log-likelihood of the data under the reverse process:

$$\mathcal{L}(\theta) = \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[\sum_{t=1}^{T} \mathrm{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)\|p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))\right]$$

This objective ensures that the reverse process learns to denoise the data effectively, recovering the original distribution from the noisy data.

Reparameterization Trick and Score Matching

To enable efficient training, diffusion models often use the reparameterization trick, similar to VAEs. Additionally, score-based diffusion models can leverage score matching techniques to directly estimate the score function (gradient of the log density). The score function provides a direction for denoising the data at each step.

Latent Diffusion Models

Latent Diffusion Models (LDMs) [35] present an alternative approach to diffusion models, addressing the computational challenges of standard diffusion models while maintaining their high-quality output. The key innovation of LDMs lies in applying the diffusion process to a compressed latent space rather than the original data space. Mathematically, this can be expressed as a two-step process. First, an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times c}$ is mapped to a latent representation $\mathbf{z} \in \mathbb{R}^{h \times w \times c}$ by an encoder function E. The diffusion process then operates on this latent space over t time steps, following the forward process:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}) \tag{3.5}$$

where β_t is the noise schedule. The reverse process, or denoising, is modeled by a neural network, typically a U-Net, as:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \Sigma_\theta(\mathbf{z}_t, t)) \tag{3.6}$$

where μ_θ and Σ_θ are learned parameters.

The training of LDMs involves two main loss functions. First, the autoencoder (Typically a VAE) is trained using a combination of reconstruction loss and perceptual loss:

$$\mathcal{L}_{\text{AE}} = \mathcal{L}_{\text{rec}} + \lambda\mathcal{L}_{\text{perceptual}} \tag{3.7}$$

where $\mathcal{L}_{\text{rec}}$ is typically the mean squared error between the input and reconstructed image, $\mathcal{L}_{\text{perceptual}}$ is a perceptual similarity metric (often based on VGG features), and λ is a weighting factor.

The diffusion model is then trained in the latent space using a variant of the variational lower bound:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{t,\mathbf{z}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t)\|_2^2\right] \tag{3.8}$$

where t is randomly sampled from the time-steps $\{1, \ldots, T\}$, $\mathbf{z}_0$ is the initial latent representation, ϵ is random Gaussian noise, and ϵ_θ is the noise prediction network (typically implemented as a U-Net). $\mathbf{z}_t$ is obtained by adding noise to $\mathbf{z}_0$ according to the forward process. [35]

The LDM architecture consists of three main components: an autoencoder for data compression and reconstruction, a neural network (typically a U-net) operating in the latent space for the diffusion process, and an optional conditioning network for controlled generation. This design offers several advantages over standard diffusion models, including improved computational efficiency, better scalability for high-resolution image generation, and increased flexibility in model architecture and conditioning mechanisms. The training process occurs in two stages: first, the autoencoder is trained to achieve perceptual compression, preserving semantically relevant information while discarding fine details. Subsequently, the diffusion model is trained in the learned latent space, following a process similar to standard diffusion models but operating on compressed representations.LDMs have demonstrated remarkable success in various applications, including high-resolution image synthesis, text-to-image generation, image inpainting, and domain translation [36–38].

3.4 Deep Learning in medical imaging

In recent years, deep learning has revolutionized various domains, and medical imaging is no exception. The advent of deep learning techniques has led to significant advancements in the analysis, interpretation, and processing of medical images, which are crucial for diagnosis, treatment planning, and patient monitoring. Traditional methods in medical imaging often relied on handcrafted features and statistical models, which, despite their effectiveness, were limited by their ability to generalize across different datasets and imaging modalities.

Deep learning, particularly with the rise of Convolutional Neural Networks (CNNs), has transformed the landscape of medical image analysis. These networks have demonstrated exceptional performance in tasks such as image classification, object detection, and, most notably, image segmentation. The power of deep learning lies in its ability to

automatically learn hierarchical features directly from raw data, eliminating the need for manual feature engineering and allowing models to discover complex patterns in medical images.

In this section, we conduct a comprehensive review of deep learning methodologies as applied to medical imaging, with a particular emphasis on image synthesis and image generation.

In the field of medical imaging, the terms ”image synthesis” and ”image generation” are frequently used as if they were synonymous, although they entail distinct methodologies and applications. Image synthesis involves crafting images based on predefined parameters or input data, ensuring that the produced images closely align with specific anatomical structures or characteristics of the intended modality. This may include tasks such as translating one image type into another (like converting MRI to MRA), enhancing resolution, transferring styles, or image inpainting. Conversely, image generation generally encompasses a more adaptable and frequently probabilistic process, producing new images rooted in learned data patterns. The primary distinction hinges on the degree of control over the resultsâ€”synthesis is typically more deterministic, whereas generation incorporates greater variability, often steered by learned distributions.

3.4.1 Deep Learning in medical image synthesis

Medical imaging plays an indispensable role in the diagnosis and monitoring of treatment in clinical settings by providing detailed and specific information about the human anatomy and physiology. Various imaging modalities, including Computed Tomography (CT), X-Ray, and Magnetic Resonance Imaging (MRI), offer unique structural and functional insights that are essential for making informed clinical decisions. However, acquiring all these scans is not always feasible as certain modalities, such as CT, pose a risk of radiation exposure [39], whereas others, like MRI, can be prohibitively expensive and entail prolonged scanning duration, potentially introducing artifacts [40].

Medical image synthesis is an essential research area in clinical decision-making, aimed at addressing the difficulties in obtaining multiple imaging modalities for an accurate clinical workflow. This method is effective in generating an image of a target modality from an existing source modality among the commonly used medical imaging contrasts, such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI),

and positron emission tomography (PET) [7]. Nevertheless, converting between two imaging modalities is challenging due to the complex and non-linear nature of domain mappings . Medical image synthesis offers a potential solution to these challenges by mapping from a given source image modality to a target modality, enabling the translation of images from one modality to another.

Recent studies have explored various applications of deep learning for image synthesis within medical imaging. Most of these studies utilize additions to two architectures: U-Net or GAN[7]. Examples of modifications made include :-

- **Loss Function Enhancements:** Incorporating perceptual, adversarial, or task-specific loss functions to improve the fidelity of synthesized images.
- **Attention Mechanisms:** Integrating attention modules to focus on critical features or regions within the images.
- **Multi-Modal Fusion:** Utilizing multiple imaging modalities as input to improve synthesis accuracy and robustness.
- **Post-Processing Techniques:** Adding layers or modules for refinement to address specific artifacts or inconsistencies.
- **Domain-Specific Customizations:** Tailoring models to handle unique challenges in specific imaging modalities or clinical applications.

U-net

U-net is a modified variant of a convolutional autoencoder, originally proposed for biomedical image segmentation in 2015 [41]. The architecture of the U-net is illustrated in Figure 3.4. This model refines the traditional autoencoder by incorporating skip connections that span a typically symmetric arrangement of encoder and decoder blocks. Each encoding block transfers its weights and extracted features to the corresponding decoding block via these 'skip connections'. This mechanism enables the initial encoding blocks to more effectively assimilate global information, while subsequent blocks focus on capturing local details. The weights shared across these skip connections ensure the retention of the global information initially captured. The ability of U-net to capture both local and global features makes it particularly effective for tasks where

preserving spatial information is crucial.

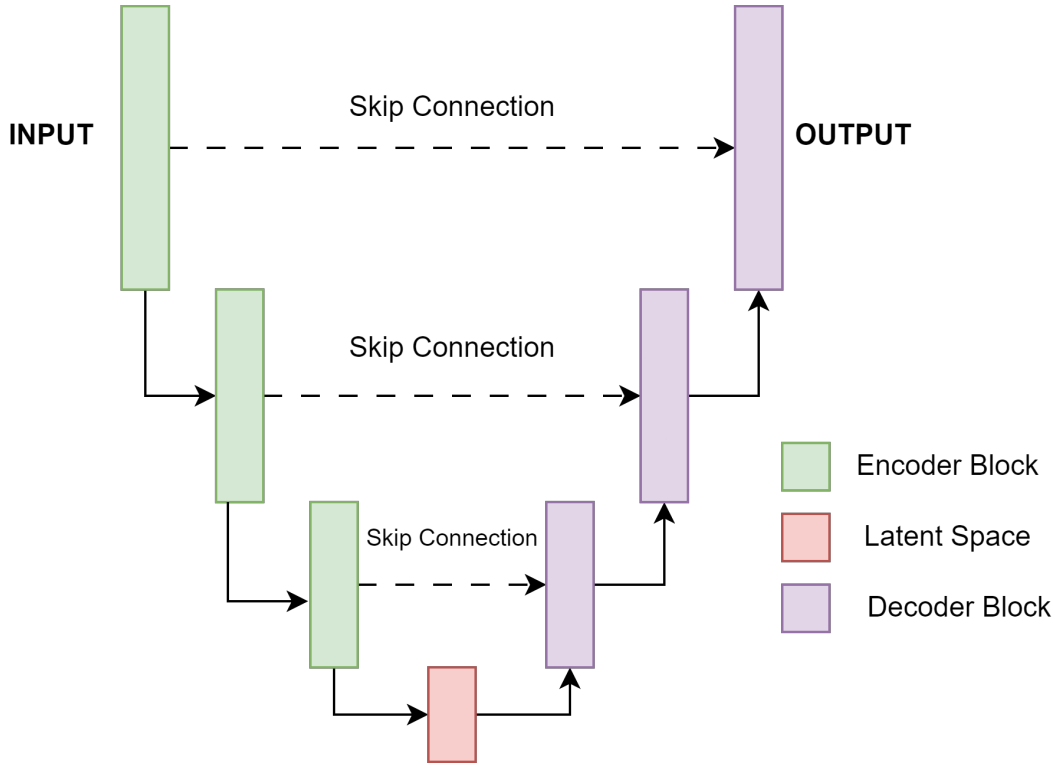

Figure 3.4: This figure showcases the architecture of the U-net.

While originally designed for segmentation tasks, U-nets have been adapted for various tasks, including image synthesis in medical imaging.Various modifications of the U-net architecture have been proposed for different medical imaging tasks, particularly in the realm of image synthesis. Han et .al [42] pioneering work adapted U-net for CT synthesis from MR images by removing fully connected layers and utilizing shortcut connections. Jang et al. and Liu et al. [43, 44] transformed CT synthesis into a segmentation problem by using discretized maps and removing skip connections. Dong et al. [45] addressed the issue of irrelevant high-frequency components in skip connections by introducing a self-attention strategy for generating CT scans from non-attenuation corrected PET images, while Hwang et al. [46] opted to use skip connections only in deeper layers to perform PET attenuation correction. Fu et al. made several improvements over [42], including replacing batch normalization with instance normalization

and using residual shortcuts in order to synthesize CT scans from MRI. Other research has focused on modification of more specific components of the network such as the activation functions and addition of new layers such as dropout, each modification aimed at addressing specific issues or improving performance in particular aspects of image synthesis [47–51].

While U-nets have shown great performance in various medical image synthesis tasks, they have some limitations that led researchers to explore alternative approaches like GANs. U-nets often struggle with generating high-frequency details and can produce overly smooth or blurry outputs, Which is partly due to their reliance on pixel-wise loss functions, which tend to average out fine details [52]. Additionally, U-nets may not capture the full range of image variations present in medical data, leading to less realistic synthesized images. GANs, on the other hand, offer a promising alternative by introducing an adversarial learning framework that can potentially generate more realistic and detailed images [53].

GAN

The medical image analysis (MIA) field historically concentrated on supervised learning, paying less attention to generative tasks. However, this changed drastically with the advent of generative adversarial networks (GANs). Since their inception in 2014 [34], GANs have become the backbone of both medical image synthesis and generation due to their ability to generate realistic and diverse images that are often indistinguishable from real ones. As discussed in the previous section , GANs consist of two neural networks, a generator that creates synthetic images and a discriminator that evaluates their authenticity.

The evolution of GANs in medical image synthesis has seen a diverse array of architectures and modifications, each of which addresses specific challenges in the field. One of the earliest adoption of GANs for a medical image synthesis task utilized a fully convolutional Autoencoder for the generator and a standard Autoencoder for the discrimination, employing binary cross-entropy loss [54]. This approach was further refined by Emami et al. [55] , who introduced conditional GANs (cGANs) for CT synthesis from MR images, allowing both networks to observe input images and thus

improving image-to-image translation.
Developing deep learning models for the synthesis of medical images across different modalities is notably difficult due to the unavoidable misalignment in training datasets, stemming from the challenge of acquiring precisely matched images. This problem was addressed with the advent of CycleGAN, which can handle misalignment and execute unpaired image synthesis tasks [56]. Liang et al.'s [57] application of CycleGAN for CBCT-based synthetic CT utilized a two-generator, two-discriminator architecture capable of managing misaligned training data pairs.

Several components of the GAN architecture have been consistently revised to accommodate various medical imaging modalities. Some researchers have integrated residual blocks into Autoencoders, demonstrating particular efficacy for tasks where the source and target images are similar, such as in CT to CBCT or NAC PET to AC PET conversions [58–60]. These residual connections help the network focus on learning the differences between the image pairs. Conversely, dense blocks, which merge outputs from earlier layers, are preferred for inter-modality synthesis tasks like MR-to-CT and PET-to-CT, capturing multi-frequency data for enhanced modality mapping [61, 62].

Most advancements in GAN architecture target the Generator. Emami et al. [55] altered the autoencoder within the generator network and modified ResNet by eliminating fully connected layers and incorporating transposed convolutional layers to achieve CT synthesis from MRI. Similarly, Kim et al. [63] combined the U-net framework with residual training to enhance MR image resolution. Olberg et al. [64] developed a deep spatial pyramid convolutional structure featuring a spatial pyramid pooling module within a U-net architecture, which facilitates multi-scale feature utilization and supports CT image super-resolution.of Discriminators have stayed relatively straightforward, with innovations mainly improving loss functions. Besides common binary cross-entropy and negative log-likelihood functions, Emami et al. [55] suggested using least-square loss to enhance stability and output quality.

GANs were quite difficult to train however ,and suffered from various issues such as overfitting and mode collapse. The introduction of Wasserstein GAN (WGAN) [65] brought in a novel loss function called Wasserstein loss, as opposed to the Jensen-Shannon divergence loss commonly used in GANs. This change has enabled easier

training with smoother gradient flow and faster convergence. Additionally, to tackle the challenge of vanishing gradients and mode collapse, Ouyang et al. [66] proposed a feature-matching technique. This method encourages the generator to match the expected values of features at the discriminator's intermediate layers, rather than merely focusing on maximizing the discriminator's final output.

These diverse methodologies showcase the swift progression and versatility of GAN architectures in the domain of medical image synthesis. Each alteration is designed to address distinct challenges, ranging from augmentation of image quality and realism to the management of misaligned data and the assurance of stable training paradigms. Nonetheless, despite these advancements, GANs persistently encounter obstacles such as training complexity and, more critically, the generation of 'artifacts' or 'phantoms' in the produced images [67], which poses significant concerns in the context of medical imaging.

Vision Transformers (ViT)

ViTs have recently gained prominence in medical image synthesis due to their ability to model long-range dependencies and capture global contextual information without the inductive biases of CNNs [68]. Unlike CNN-based architectures, which rely on localized feature extraction, ViTs employ self-attention mechanisms that enable more effective feature learning for complex medical image synthesis tasks.

One of the earliest applications of ViTs in medical image synthesis involved their integration with generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). For instance, Choudhury et al. [69] proposed an autoencoder-ViT hybrid for synthesizing cardiac MRI images, demonstrating superior quality and anatomical coherence compared to CNN-based counterparts. Similarly, Hu et al. [70] utilized ViTs for multi-branch attention-based medical image synthesis, showing improved image fidelity in cross-modality transformations such as CT-to-MRI and PET-to-CT synthesis.

A key advantage of ViTs in medical image synthesis lies in their ability to generalize across imaging modalities. Unlike CNNs, which often struggle with domain shifts, ViTs' self-attention mechanisms allow them to retain structural integrity across diverse datasets. Zhao et al. [71] demonstrated this in a study on synthetic CBCT image

generation, where a ViT-based approach significantly outperformed CNN-based models in preserving fine anatomical details. Additionally, ViTs have been employed for multi-view image synthesis in radiotherapy, enhancing the accuracy of sparse-view CT reconstructions [72].

Despite these advancements, ViTs pose several challenges in medical image synthesis. Their high computational cost limits deployment in resource-constrained settings, and their data-hungry nature necessitates large-scale annotated datasets, which are often unavailable in medical imaging. Researchers have attempted to address these issues through hybrid architectures that combine CNNs and ViTs, leveraging the efficiency of CNNs for feature extraction while utilizing ViTs for high-level reasoning [73]. Additionally, model compression techniques such as knowledge distillation and pruning have been explored to make ViTs more computationally viable for clinical applications.

The rapid adoption of Vision Transformers in medical image synthesis highlights their potential to revolutionize synthetic data generation. However, ongoing research is required to optimize their efficiency, reduce computational overhead, and ensure clinical applicability. As ViTs continue to evolve, they are likely to become a cornerstone of medical image synthesis, addressing longstanding challenges in dataset augmentation, cross-modality transformations, and synthetic image realism.

Diffusion models

While GANs have shown remarkable success in image synthesis tasks, including medical imaging applications, they are not without limitations. GANs often struggle with mode collapse, where the generator produces limited varieties of outputs, and training instability, which can lead to unpredictable results [74]. Additionally, GANs may fail to capture the full data distribution, particularly in complex medical imaging scenarios [75, 76]. These drawbacks have prompted researchers to explore alternative approaches, leading to the recent emergence of diffusion models as a promising new direction in image synthesis and translation tasks. Diffusion models, including Denoising Diffusion Probabilistic Models (DDPMs) and score-based diffusion models, have gained traction due to their ability to generate high-quality images with greater stability and diversity compared to GANs [77]. In the context of medical imaging, these models have shown particular promise in addressing the challenges of inter-modality image synthesis [78] . A notable example of this application is the work of Lyu et al. [79], who leveraged

diffusion models to tackle the complex task of translating MRI images to CT scans. This research addresses a significant challenge in medical diagnostics, where the limitations of CT in visualizing soft tissue often require additional MRI scans, leading to increased time, cost and potential image misalignment issues. In a comprehensive study using the Gold Atlas male pelvis dataset [80], diffusion models consistently outperformed CNN and GAN-based methods in terms of Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR) [81].

To further illustrate the versatility and potential of diffusion models in medical imaging, recent research has explored their application in addressing the missing modality problem and improving structure preservation in image translation tasks. Meng et al. [82] introduced a unified multi-modal conditional score-based generative approach (UMM-CSGM) to synthesize missing modalities based on available ones. This conditional SDE model [83] uses a single score-based network to learn various cross-modal conditional distributions. When tested on the BraTS19 dataset [84], which includes four MRI modalities per subject, UMM-CSGM outperformed state-of-the-art methods in generating missing-modality images with higher fidelity and better structural information of brain tissue. Despite the progress in diffusion models, they still encounter difficulties in preserving structural information when translating images, since details from the original domain may be lost during the forward diffusion process [78].To address this, Li et al. [85] developed the Frequency-Guided Diffusion Model (FGDM), which uses frequency-domain filters for structure-preserving image translation. FGDM enables zero-shot learning and can be trained exclusively on target domain data, allowing for direct deployment in source-to-target domain translation without exposure to source domain data during training. An emerging trend in the field of generative medical image synthesis addresses the challenges posed by both 2D and 3D data. For instance, Make-A-Volume [86] offers a diffusion-based framework that circumvents issues such as mode collapse and volumetric inconsistency, by incorporating 2D backbones and fine-tuning volumetric layers for 3D synthesis, this approach offers a computationally efficient solution that maintains coherence across volumes while reducing memory overheads.Additionally , multimodal approaches such as MedSyn [87] are able to generate high feildity 3D CT lung images guided by text information. Apart from image translation tasks , diffusion models have also shown promise in several other image synthesis tasks such as Super resolution [88, 89] image impainting [90]. These advancements

demonstrate the ongoing evolution of diffusion models in medical imaging, addressing critical challenges such as missing modalities and structure preservation, and pushing the boundaries of what is possible in medical image synthesis and translation.

3.4.2 Deep Learning in medical image generation

Up until lately, most research in medical imaging concentrated on supervised learning, with generative tasks receiving less attention [91]. However, this direction took a significant turn with the advent of generative adversarial networks (GANs) and, more recently, denoising probabilistic diffusion models (DDPM).Generative models can potentially alleviate the limitations of data scarcity and class imbalance in medical image analysis by generating realistic-looking images from a learned distribution that follows the real data distribution.

Deep Convolutional Generative Adversarial Networks (DCGAN) [92], designed to tackle the instability in standard GANs by incorporating deeper generator and discriminator architectures, have been extensively utilized in the field of medical image synthesis. Notably, DCGAN has been applied to generate MRI images of prostate lesions [93], X-rays of lungs with cancerous nodules [94], and brain MRIs [95]. More advanced GAN variations, such as Laplacian GAN (LAPGAN) [96], which leverages Laplacian pyramids for enhanced image quality, and progressive GAN [97], have achieved significant success in producing high-resolution images of skin lesions [98]. Utilizing GAN-generated data for training models, rather than relying solely on traditional data augmentation techniques like rotation and shearing, can enhance model performance by up to 16% [91].

Though GANs demonstrate potential in the field of medical image synthesis, they face issues such as instability, mode collapse, and susceptibility to hallucinations [74]. Additionally, recent research indicates that datasets created by GANs do not possess the same depth and variety as actual datasets [99]. On the other hand, diffusion models, despite being relatively new, have been extensively utilized in research for addressing these challenges, offering the promise of generating datasets that are richer and more realistic than those produced by GANs [100].

One notable approach, HistoDiffusion [101], leverages latent diffusion models (LDM) trained on large-scale unlabeled datasets for synthetic augmentation, reducing reliance on expert annotations. This method demonstrated a 6.4% improvement in classifica-

tion accuracy on colorectal cancer histopathology images, highlighting the potential of pre-trained diffusion models in augmenting small labeled datasets.

In the realm of skin disease classification, diffusion models have shown promise in generating diverse images, especially in data-limited environments. Latent diffusion models were used to produce over 450,000 synthetic images, improving classifier performance in underrepresented populations [102].

Furthermore, conditional diffusion probabilistic models (cDPM) have been utilized to generate realistic brain MRIs, providing an alternative to the computationally expensive GAN-based methods. By conditioning on partial slices of MRIs, the cDPM can generate full 3D brain volumes that maintain anatomical consistency, significantly reducing computational requirements while producing high-quality images [103].

Another application of diffusion models is in generating counterfactual images for anomaly detection in brain images. By combining Denoising Diffusion Probabilistic Models (DDPM) with Denoising Diffusion Implicit Models (DDIM), researchers have successfully modified pathological regions while preserving the normal anatomy in surrounding areas [104].

Finally, the use of latent diffusion models in large-scale brain MRI generation has allowed researchers to synthesize realistic high-resolution brain images, with control over variables such as age and sex. The creation of synthetic datasets, such as a publicly available set of 100,000 brain images, demonstrates the scalability and potential of diffusion models in advancing medical imaging research [105].

These developments underscore the increasingly significant role of diffusion models and GANs in the realm of medical imaging, as they confront critical challenges such as data scarcity, modality generation, and the preservation of anatomical accuracy.However, gains in model performance from using generated data for augmentation saturate at higher synthetic-to-real image ratios, underscoring the continued importance of real-world data collection [102].

Chapter 4

Synthesizing Vascular segmentation from T2 Weighted MRI

This Chapter is based on the paper *[Deo, Yash, et al. "Learned Local Attention Maps for Synthesising Vessel Segmentations from T2 MRI." International Workshop on Simulation and Synthesis in Medical Imaging. Cham: Springer Nature Switzerland, 2023.]*

4.1 Introduction

A magnetic resonance angiogram (MRA) contains vital information for visualizing the brain vasculature, including an anastomotic ring of arteries located at the base of the brain called the circle of Willis (CoW). Multiple different topological variants of the CoW exist in the general population [106], and certain variants of the CoW can lead to worse outcomes after stroke [107]. To that end, it would be useful to visualise the main cerebral blood vessels in large imaging datasets and identify them by CoW phenotype to understand their relevance to stroke in the general population. Vessel segmentation from MRA is a well-studied problem with state-of-the-art methods achieving high quality vessel segmentation results [108] with Dice scores as high as 0.91 [109]. However, as MRA acquisition may require the injection of contrast agents and has longer acquisition times, it is not commonly available in population imaging studies. T1- and T2-weighted MRI scans are the most common MR imaging modalities available and are used to study the presence of lesions or other abnormal structures in the brain. While the blood vessels are not explicitly visible in these modalities, they contain latent information that can be used to synthesise the major vessels in the brain.

Generative adversarial neural networks [34] (GANNs) have seen remarkable success in the field of image synthesis, with networks like pix2pix [110] achieving impressive results in paired image-to-image synthesis. GANNs have also been widely used in medical image synthesis in various use cases such as generating T1, T2, and FLAIR images of the brain using Wasserstein-GANNs [65]. Progressively growing GANNs [111] have been used for the generation of retinal fundus and brain images. Previous works on brain MRA synthesis used Steerable GAN (SGAN) [112] to generate MRA from paired T1 and T2 images, or used starGAN [113] to synthesise MRA given T1, T2 and/or a PD-weighted MRI as input. GANN-based approaches such as vox2vox [114] have been used to synthesise segmentations of brain tumour directly from T1, T2, Gadolinium-enhanced T1, and T2 FLAIR modalities. Most GANN based approaches synthesise MRA from multiple other MR modalities, and then require the use of a separate seg-

mentation algorithm, such as U-net (which is popularly accepted as baseline), to segment the brain vascular structures from the synthesised MRA. As the brain vessels form a very small portion of the MRA image, attention mechanisms were introduced to the segmentation algorithms to more accurately capture the small vessels. This has been achieved in networks such as Attention U-Net [115] or more recently transformer based networks such as TransU-Net [116].

In spite of their successes, GANs are notoriously hard to train due to various factors such as training instability and mode collapse [117, 118], while t ransformers, on the other hand, can be extremely computationally expensive to train due to the self-attention mechanism (which has a complexity of $O(n^2)$ where n is the input sequence length) and tend to be very data hungry [119, 120]. On top of that, GANNs tend to produce phantoms (non-existent image features), especially when dealing with very high-resolution images with intrinsic detail arising from medical imaging [121]. To alleviate these issues, we propose multi-task learnable localised attention maps to directly generate vessel segmentations based on a U-Net architecture, which can capture both global and local features from the input domain. Our method requires only the T2 modality as input, which eliminates the need of multiple input modalities. The learned local attention maps enable the trained model to only look for vessels in specific parts of the image, which drastically decreases the number of parameters required to train the synthesis network. Our model consequently synthesises more accurate CoW segmentations with fewer parameters than competing GANN-based approaches.

4.2 Methodology

We propose a deep convolutional encoder-decoder model, which is trained in two phases with multitask learning. At training time, paired T2 images and ground-truth MRA segmentations are available. Our encoder-decoder network captures both global information (by encoding input images into a latent space) and local information (by learning soft attention maps for brain vessels based on MRA segmentations) from the given input images. During training, the model utilizes a two-phase multitask learning approach. In the initial phase, the network functions as a conventional auto-encoder, capturing global features from the input images. In the subsequent phase, an additional output branch is introduced to generate vessel segmentations from the input images. Here, a learned local attention map identifies the most probable locations of vessels

on the T2 images, thereby guiding the network's focus on pertinent local features to enhance the synthesized vessel segmentation masks.A multi-task learning approach is used in the second phase to balance the learning of both the objectives. At inference time, the model proficiently generates brain vessel segmentation masks using only T2 images.

4.2.1 Data and Pre-processing

The model was trained on the IXI dataset [122], which includes imaging data acquired using 3T MRI scanners at Hammersmith Hospital. The dataset comprises paired T2-weighted MRI and MRA scans from 181 patients. The following steps were taken to preprocess the data before feeding it to the model for training :-

Image Registration To ensure that the T2 and MRA images are spatially aligned, we first registered the T2 images with the MRA images using the Rigid registration algorithm provided in the SimpleElastix Python package [123]. This step involves aligning the images based on their geometric properties, correcting any misalignments due to patient movement or differences in image acquisition times and ensures that we have paired T2 and MRA images.

Cropping and Centering The images were centered and cropped from their original resolution of 512×512 pixels to a resolution of 400×400 pixels. This cropping focuses on the region of interest (the brain) and removes extraneous parts of the images, reducing computational load and improving model efficiency.

Intensity Normalization To standardize the intensity values across all images, we used min-max normalization as a pre-processing step. This step adjusts the pixel intensity values to a common scale, ensuring that variations in brightness and contrast do not affect the model's performance. Normalization is crucial for handling the inherent variability in MRI scans.

Ground-Truth Segmentation Generation Ground-truth segmentations of cerebral blood vessels were generated from MRA images using a residual U-Net [2]. This deep learning model was trained to accurately delineate the blood vessels in the MRA

images, providing segmentation masks that serve as the basis for training the synthesis model.

Dilation of Segmentations The fine vessel segmentations obtained from the residual U-Net were expanded by 10 pixels in each direction to create a more inflated / dilated version of the vessel segmentation masks. The optimal dilation width was determined through experimentation.

Creation of Local Attention Maps To generate the local attention map, the dot product is computed between the pairs of the dilated segmentation masks obtained from the previous step and the corresponding T2 slices, as illustrated in Figure. 5.3. These maps identify regions within the T2 images that contain blood vessels and serve as local attention maps during the training phase of our model.

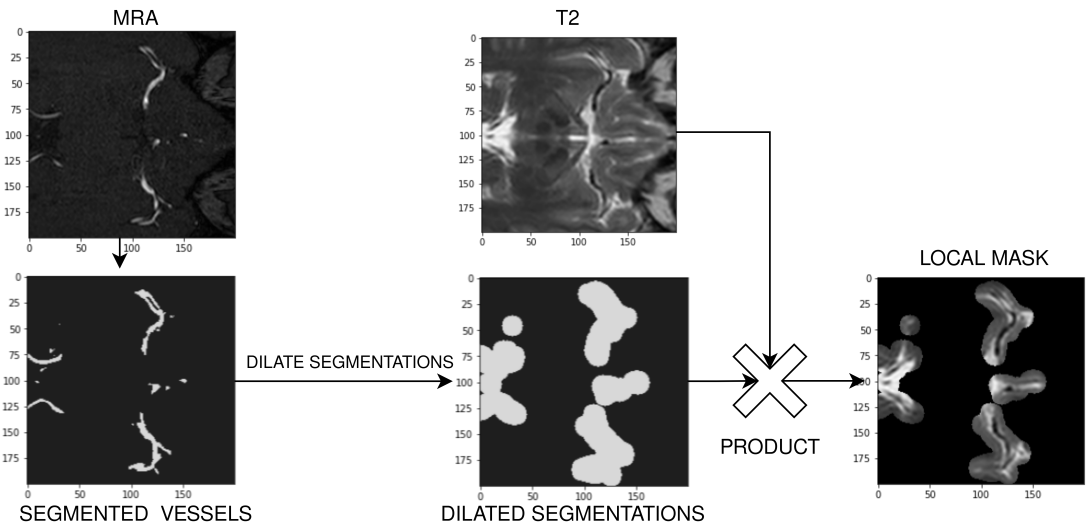

Figure 4.1: This figure demonstrates the method of creating local attention masks. Initially, a binary vessel mask is derived from the MRA using a segmentation algorithm [2], then the segmentations are dilated by extending the binary mask by a specified number of pixels in every direction. Subsequently, a dot product of this dilation with the corresponding T2 slice is calculated to form the local attention mask.

4.2.2 Network Architecture

The proposed model follows the general architecture of the U-Net [41] with one encoder branch and two output branches (Figure. 4.2). The encoder takes the T2 images (400×400) as input and consists of 4 Encoding blocks followed by 3 Residual blocks , similar to the vox2vox-model [114] (The residual blocks are defined the same as in [124]). Each Encoding block consists of three strided convolution layers followed by a max-pooling layer where each convolutional layer is also followed by an instance normalization layer [125].The dimension of the latent space after the encoding branch is 50×50 . The latent space branches out into two output branches: the decoding branch and the synthesis branch (the decoding brach is used to reconstruct the input image and the synthesis branch is used to synthesize the vessel segmentation). Both output branches consists of 4 decoding blocks which follow the same structure as the encoding blocks except the max-pooling layer is swapped with an up-sampling layer . The decoding blocks in the synthesis branch receive skip connections from the corresponding blocks in the encoding branch .

4.2.3 Training and Losses

The network is trained in two phases to effectively capture both the global and local features required to synthesise the vessels from T2 images.

Phase 1:

We pre-train the network on T2 images by first freezing the synthesis branch and only training the decoding branch, effectively training an autoencoder for T2 images. The network is trained with an early stopping criteria based on the loss slope. The only loss calculated in this stage is the T2 reconstruction loss from the decoder branch.The loss function used is L1 and is specified below where X_{T_2} is the ground truth T2 image and $\hat{X}_{T_2}$ is the generated T2 image:

$$\mathcal{L}_{\text{phase 1}} = \text{MAE}(X_{T_2}, \hat{X}_{T_2}) \tag{4.1}$$

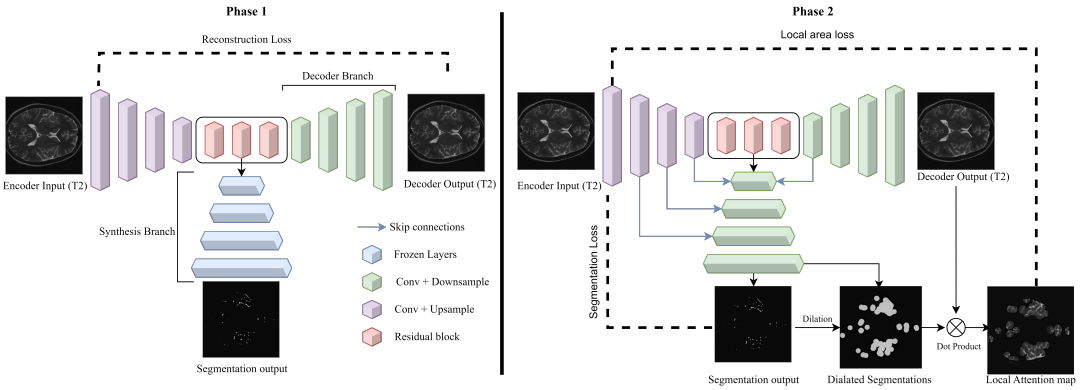

Figure 4.2: Overview of our network architecture and training process. The encoder takes T2-weighted MRI (400×400) as input and compresses it into a smaller latent space (50×50) which then splits into two branches , the synthesis branch (which generates the segmentation) and decoding branch (which reconstructs input). The training takes place in two phases , in the first phase (left) the network is trained as a standard auto-encoder with only the encoding and decoding branch with the reconstruction loss calculated over the decoding branch output. In the second phase (right) both the synthesis and decoding branch are trained simultaneously. The output of the synthesis branch is dilated and multiplied with the output of the decoding branch to generate an attention mask during training . The reconstruction loss is then calculated over this attention mask and the segmentation loss is calculated over the output of the synthesis branch.

Phase 2:

After we complete the pre-training step (Phase 1) , we unfreeze the synthesis branch and train it together with the decoding branch. Although the decoding branch is being trained in this step, the loss calculated for this branch is not the reconstruction loss (as used in Phase 1), but a local loss, which is calculated over the dot product of the output of the decoding branch and the dilated segmentation obtained from the output of the synthesis branch (shown in Figure. 4.2). The loss function used to calculate local loss is still MAE and can be formulated as shown in Equation 4.2 where X_{T_2L} is the ground-truth local attention mask and $\hat{X}_{T_2L}$ is the output local attention mask.

$$\mathcal{L}_{\text{loc}} = \text{MAE}(X_{T_2L}, \hat{X}_{T_2L}) \tag{4.2}$$

For the synthesis branch, on the other hand, we calculate the segmentation loss over the synthesized binary mask segmentation using the DICE score as shown in Equation 4.3 where X_{GT} is the ground truth segmentation and X_{SYN} is the synthesized segmentation.

$$\mathcal{L}_{\text{loc}} = \text{Dice}(X_{GT}, \hat{X}_{SYN}) \tag{4.3}$$

Multi Task Learning for Optimal Training: Developing models with multiple outputs poses substantial challenges. Harmonizing the learning process across various tasks can prove to be difficult, as tasks can compete for model capacity, exhibit different levels of complexity, or require different learning rates, Additionally, the naive amalgamation of loss functions can yield suboptimal solutions, with certain tasks overshadowing others or the model failing to encapsulate critical interrelationships between tasks [126, 127]. These challenges necessitate the exploration of multitask learning (MTL) approaches/algorithms, which offer methods to address these issues and optimize performance across all tasks simultaneously.Hence , To effectively train our model across both tasks (reconstruction and synthesis), we employ a multitask learning algorithm (MTL) to simultaneously converge the outputs of both branches. Multitask learning involves the simultaneous training of a single model to perform multiple tasks, leveraging shared information to enhance efficiency and performance. Research has established that a single model with multiple complementary tasks (outputs) is more efficient and yields performance on par with or superior to that of training separate models for each task, by sharing information across tasks [128, 129]. However, in scenarios where the tasks contrast significantly and one task is substantially harder than the other, deep learning models tend to 'abandon' the difficult task and focus on optimizing the easier one [130]. This is particularly evident in our case, where the decoding branch undertakes a considerably easier task compared to the synthesis branch, causing the model to neglect optimization of the synthesis branch. This is illustrated in Figure. 4.3, which compares the training loss of our model for the two outputs (Reconstruction and Synthesis) with and without the application of MTL.

The landscape of MTL approaches is diverse, ranging from simple to highly sophisticated methods. At the simpler end of the spectrum, we find hard parameter sharing, where a single shared network is used for all tasks with task-specific output layers, and soft parameter sharing, where each task has its own model but parameters are regu-

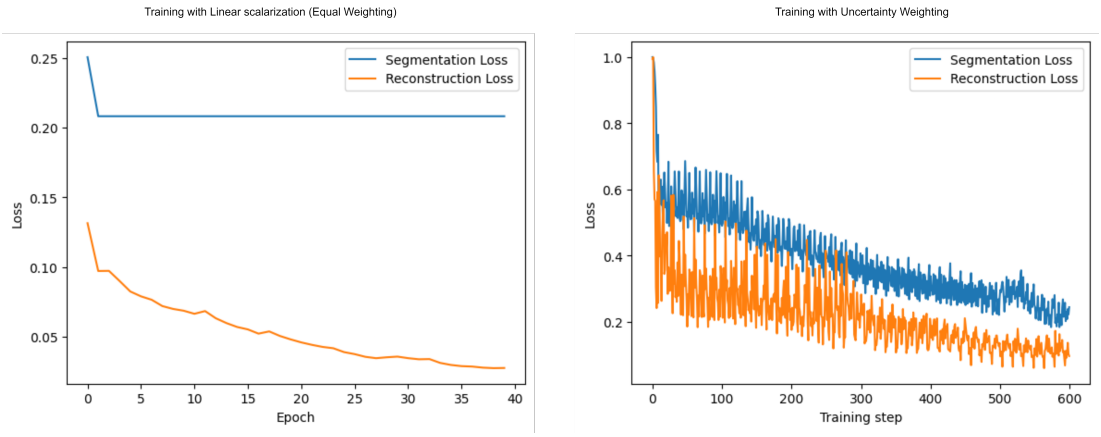

Figure 4.3: Comparison of Model Loss with and without MTL: The left graph shows the loss curves for both segmentation and reconstruction outputs when trained with equal weights, illustrating that the model quickly abandons the more difficult segmentation(Synthesis) task.In contrast, as shown in the graph on the right, applying MTL [3] helps balance the optimization of both tasks.

larized to encourage similarity [131, 132]. These methods, while straightforward to implement, can struggle with conflicting tasks or may not fully capture complex task relationships. More advanced approaches have emerged to address these limitations. NashMTL [133] takes a game-theoretic approach, formulating MTL as a multi-player game and seeking a Nash Equilibrium, which allows it to handle non-convex loss landscapes and adapt to changing task difficulties. CAGrad (Conflict-Averse Gradient descent) [134] goes a step further by explicitly finding a compromise between task-specific gradients, particularly useful when tasks conflict significantly. However, these methods can be computationally intensive and may struggle with very heterogeneous tasks. An additional well-regarded and theoretically grounded approach is uncertainty-based MTL [3], utilizing a Bayesian framework to automatically balance tasks by modeling task-specific uncertainties. We conduct a comparative analysis of NashMTL, CAGrad, and uncertainty-based MTL using our model, as detailed in Table 4.1, to ascertain the most suitable methodology for our current tasks. The results indicate that, although all approaches generally achieve effective task balancing and performance, the uncertainty-based method demonstrates superior performance on the synthesis task, aligning with the primary aim of this paper. As the best performing approach was the

uncertainty-based MTL, where both the losses are weighted based on the assumption of homoscedastic uncertainty for each task , we use this approach to train our multi-task model in phase 2.

Table 4.1: Performance across Synthesis and Reconstruction tasks across different MTL optimization algorithms

MTL Approach	Synthesis Task (Dice score)	Reconstruction Task (SSIM)
NashMTL	0.75	0.93
CAGrad	0.76	0.88
Uncertainty based	0.79	0.89

Local Attention Mask The primary rationale for the implementation of local attention maps is to direct the network's focus to regions within the input T2 image that likely harbor information crucial for the synthesis of vessels. This is achieved by computing the dot product between the synthesized vessel mask and the reconstructed input image during the training process. Nonetheless, given that the output of the segmentation branch encapsulates fine vessel details, the small dimensions of the vessels render the segmentation masks inadequate for the generation of local attention maps. To address this, we dilate these vessel segments by 10 pixels in each direction, thereby forming a local attention mask. The optimal dilation width was determined through empirical experimentation, as depicted in Table 4.3. Subsequently, we conduct pixel-wise multiplication of this local attention mask with the decoder output to derive a local attention map, as illustrated in Figure. 4.1. During the model training phase, the local attention map is derived by performing dilation on the output of the synthesis branch, followed by the computation of the dot product with the output of the reconstruction branch. This generated local attention map is evaluated against ground truth local attention maps to compute the loss.The utilization of a local attention mask compels the network to exclusively learn from a minimal portion of the input image, containing relevant information about the blood vessels, whilst disregarding the extraneous parts. This characteristic significantly reduces the number of parameters required for model training. Moreover, the intrinsic interdependence between these tasks facilitates a synergetic interaction between them, mitigating the otherwise stark contrast between them and enabling Multi-Task Learning (MTL) algorithms to con-

verge more efficiently across both tasks. This assertion is substantiated through an experimental setup wherein the model is trained using the uncertainty-based MTL approach, both with and without the incorporation of the local attention mask. In the initial configuration, the reconstruction loss is computed over the entire image, whereas in the subsequent configuration, the reconstruction loss is evaluated exclusively over the local attention mask. (Figure. 4.4).

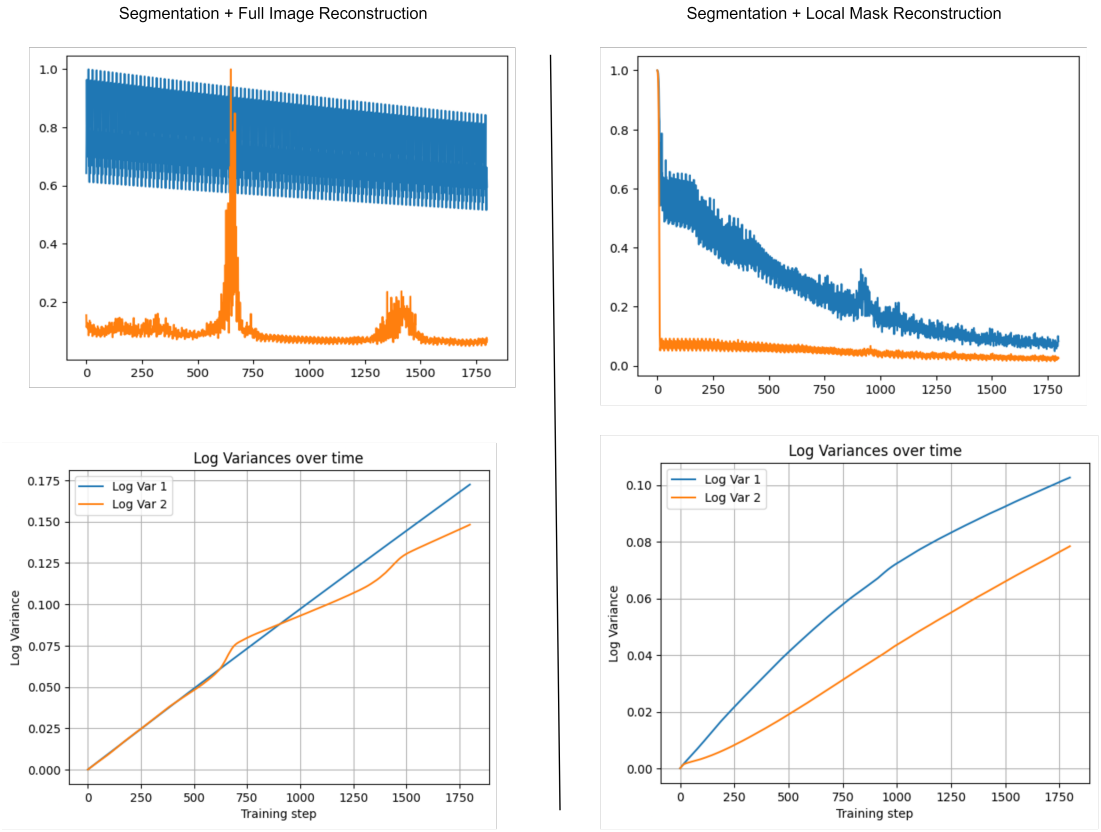

Figure 4.4: A comparative analysis of training losses and logarithmic variances for both tasks, with and without the application of local attention maps. The blue line delineates the segmentation loss, whereas the orange line illustrates the reconstruction loss. The left-hand graphs depict the variations in loss and logarithmic variance when the reconstruction loss is computed over the entire image. In contrast, the right-hand graphs exhibit the variations in loss and logarithmic variance when the reconstruction loss is assessed within the confines of the local attention mask.

In the end , the final loss function for the phase 2 of training of our model is

described in (4.4), where W are the model parameters and we interpret minimising the loss with respect to σ_1 and σ_2 as learning the relative weights for the losses $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{loc}}$ adaptively. We used Dice score as the loss for $\mathcal{L}_{\text{seg}}$ and MAE as the loss for $\mathcal{L}_{\text{loc}}$

$$\mathcal{L}_{\text{phase 2}} = \frac{1}{2\sigma_1^2}\mathcal{L}_{\text{seg}}(\mathbf{W}) + \frac{1}{2\sigma_2^2}\mathcal{L}_{\text{loc}}(\mathbf{W}) + \log \sigma_1\sigma_2 \tag{4.4}$$

4.3 Experiments and results

4.3.1 Implementation Details

All the models were implemented in TensorFlow 2.8 and Pytorch (for nnU-Net) and Python 3. Out of the 181 cases in the dataset we used 150 for training and 31 for testing and validation. All the models were pre-trained on T2 images and grid search was used to optimise the following hyperparameters: (1) batch size, (2) learning rate, (3) number of epochs, and (4) momentum. To train the transformer network, we first used the parameters recommended in [116].

4.3.2 Quantitative Results

To evaluate the results of our model against other methods, we used the segmentation metrics of Dice score and Hausdorff distance (hd95). The results were averaged over the 3D volumes of the 11 leave-out cases and are shown in Table 4.2. Our method clearly outperforms conventional GANN-based synthesis methods, such as vox2vox, and also performs slightly better than state-of-the-art segmentation models like transformer U-Net [116] and nnU-net [135], while also being easier to train with fewer trainable parameters. We experimented with training our model with different input modalities, which showed that using only T1 as an input had the worst performance (average dice 0.64 $\pm$0.04) while the performance of using only T2 (average dice 0.79 $\pm$0.04) and both T1 + T2 (average dice 0.78 $\pm$0.05) was essentially the same, with T1 + T2 requiring additional parameters (33.4 million) compared to using just T2 (26.7 million) as we would need an additional decoding branch for the T1 decoder. A crucial hyperparameter in our model is the dilation width of the segmentations to generate the local attention maps, which was optimised in a separate experiment. (Table 4.3).

Table 4.2: Accuracy of synthesised vessel segmentation masks in a test set of 11 leave-out cases

Model	Model params. ($\times 10^6$)	Dice (95% CI)	HD95 (95% CI)	Model Type
Our model	26.7	0.79 ±0.03	9.1 ±0.5	Segmentation/synthesis
Transformer U-Net [116]	105.8	0.71 ±0.04	10.4 ±0.5	Segmentation
nnU-Net [135]	127.8	0.68 ±0.03	9.3 ±0.4	Segmentation
Vox2vox [114]	78.8	0.67 ±0.05	17.2 ±1.4	Segmentation/synthesis
Pix2pix [110]	36.9	0.55 ±0.04	23.1 ±3.0	Synthesis
U-Net [41] (base)	9.1	0.57 ±0.05	42.6 ±4.2	Segmentation

Table 4.3: Difference in loss with different values of dilation for the local attention mask

Attention mechanism used	Dice (95% CI)	Area covered by mask
No local attention mask	0.62 ±0.04	NA
Mask with no dilation	0.59 ±0.04	1.5%
Mask with dilation by 5 pixels	0.74 ±0.03	8.5%
Mask with dilation by 10 pixels	0.79 ±0.03	18%
Mask with dilation by 15 pixels	0.75 ±0.02	28%
Mask with dilation by 20 pixels	0.75 ±0.03	37%

4.3.3 Qualitative Results

Figure. 4.6 shows a qualitative comparison of our method against pix2pix, vox2vox, U-Net, nnU-net, and transformer U-Net for two samples from the unseen test set. It can be observed that pix2pix and the base U-Net are only able to capture the overall structure of the CoW with a lot of noise. The vox2vox model synthesises the vessels slightly better, but is still unable to capture the finer details and suffers from noise. The nnU-net and transformer U-Net are able to synthesise the vessels with high quality, but struggle to synthesise smaller vessels such as the posterior communicating arteries (PComA) in the first case. An interesting observation can be made in the second case, where the ground truth has faults in the segmentation (especially in the posterior circulation) , the Transformer U-Net, nnU-net, and our model attempt to fix these faults by synthesising a continuous PCA, but our model does better in restoring vessel continuity. Figure. 4.4 shows the CoW synthesis results for the best case, worst case,

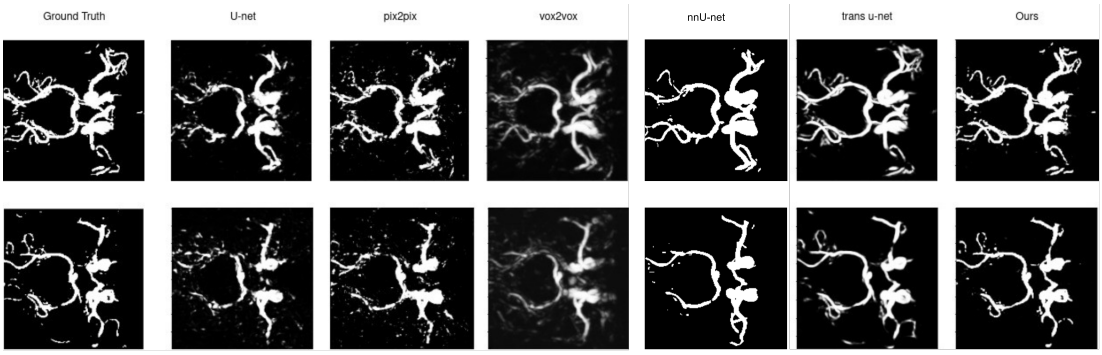

Figure 4.5: CoW synthesis results compared between models. Pix2pix and U-Net are able to capture the overall structure of the Cow but with a lot of noise. Vox2vox performs comparatively better, but still suffers from noise in the outputs. NnU-Net, Transformer U-Net and our method show good results with our method capturing more details and dealing better with noise.

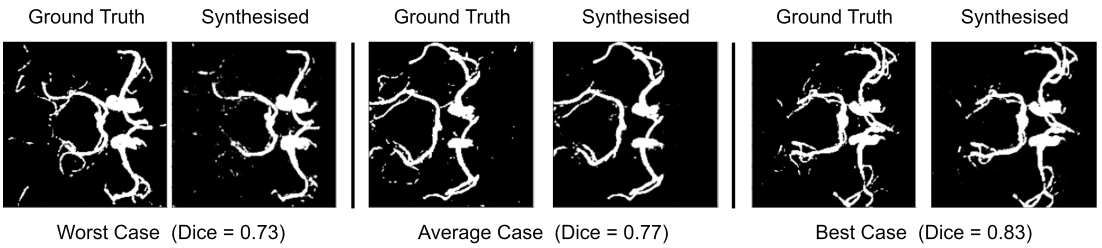

Figure 4.6: CoW synthesis results for the average case, the best case, and the worst case in our unseen test set.

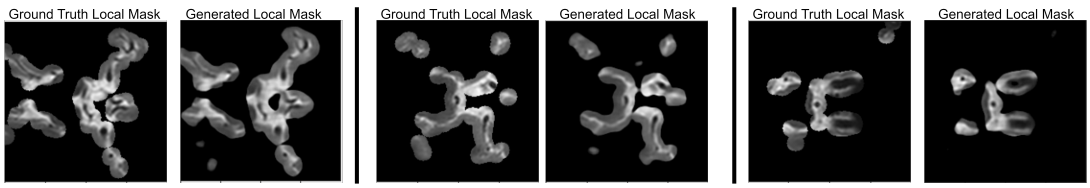

Figure 4.7: Local attention maps learned by the network compared against the ground truth local attention maps.

and median case scenarios. It can be observed that in the worst case the model struggles to synthesise the smaller vessels towards the end of the posterior cerebral circulation, whereas in the median case scenario most of the major vessels are synthesised with only

the small PComA artery missing. The best case is that all the major arteries of the CoW are synthesised while also removing noise from the input image.

4.3.4 Limitations

While our method outperforms state-of-the-art approaches with a much smaller number of trainable parameters and is able to generated the complete structure of the CoW, it can be seen that in some cases the model can struggle to generate some of the finer vessels branching from the main arteries (especially the posterior communicating arteries). This could be either because the input data is of insufficient resolution (T2 images were acquired at 3T) or because the T2 modality does not contain information that could be used to synthesise the anterior circulation. It is possible that additional MR modalities, such as multi-view T1, or a fully-3D neural network architecture could add more information about the posterior and anterior vessels and recover a complete CoW.

4.4 Conclusion

We proposed a multi-output encoder-decoder -based network that learned to effectively synthesize vessels from only T2-weighted MRI using local attention maps and multitask learning. The qualitative and quantitative results show that our method outperformed both the state-of-the-art and conventional segmentation/synthesis algorithms, while at the same time being easier to train with fewer parameters. Future work could invovle additional enhansmenets to the model such as convering the 2D model to a fully 3D synthesis model to achieve even better connectivity of the CoW structure.

Addendum

This addendum addresses specific questions raised about the chapter ”Synthesizing Vascular Segmentation from T2 Weighted MRI.”

Why is GANN used as an acronym, and not GAN? The term ”Generative Adversarial Neural Network (GANN)” is used in this context to emphasize the neural network aspect of the methodology. While ”GAN” is widely recognized, using GANN

provides clarity in distinguishing it from broader GAN-based frameworks and highlights the architectural nuances explored in this work.

Why is the problem called vessel synthesis and not just vessel segmentation? Vessel synthesis involves generating segmentation masks directly from T2-weighted MRI, bypassing the intermediate step of creating an MRA. This broader definition captures the generative nature of the task, which integrates aspects of both synthesis and segmentation, as opposed to traditional segmentation that operates on pre-existing MRAs.

Page 53, Creation of Local Attention Maps: Not really a dot product? Isnt this a Hadamard (elementwise) product? Indeed, the operation described corresponds to a Hadamard (elementwise) product rather than a dot product. The description has been amended in this addendum to ensure terminological precision.

Why does the network in phase 1 not have any shortcut connections? Compressing to 50×50 is not too little? The exclusion of shortcut connections in phase 1 is intentional to ensure that the encoder-decoder framework learns robust feature representations without relying on direct residual connections. The latent space of 50×50 balances the trade-off between dimensionality reduction and preserving sufficient information for subsequent decoding tasks. Empirical results validated that this compression retained the necessary details for accurate synthesis in later phases.

Why do you have a residual connection from the decoder to the segmentation branch? The residual connection from the decoder to the segmentation branch enhances the network's ability to recover fine-grained vessel details by integrating information from both global encoding and local decoding pathways. This connection ensures that the synthesis branch benefits from both reconstructed context and learned attention.

Can you explain the rationale for the dilation a bit better? Dilation expands vessel segmentations to encompass adjacent regions, mitigating the sparsity of vascular features and providing the network with broader spatial context. This dilation was

optimized experimentally to balance vessel prominence and noise suppression, with a width of 10 pixels yielding the best performance.

Multi-task approaches. What are you trying to achieve? The multi-task learning approach aims to jointly optimize reconstruction and segmentation tasks, leveraging shared representations to improve model efficiency and accuracy. If multi-task learning is not utilized , the network typically tends to favor optimizing the easier task while ignoring the hard task.

Why do you think masking for attention makes it better? Masking focuses the network's learning on regions containing vessels, reducing the influence of irrelevant background features. This targeted learning enhances the precision of vessel synthesis while lowering computational overhead by narrowing the models attention to areas of interest.

Chapter 5

Shape-guided conditional latent diffusion models for synthesizing brain vasculature

This Section is based on the paper *[Deo, Yash et.al. "Shape-guided conditional latent diffusion models for synthesising brain vasculature." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 164-173. Cham: Springer Nature Switzerland, 2023.]*

5.1 Introduction

The Circle of Willis (CoW) comprises a complex network of cerebral arteries that plays a critical role in the supply of blood to the brain. The constituent arteries and their branches provide a redundant route for blood flow in the event of occlusion or stenosis of the major vessels, ensuring continuous cerebral perfusion and mitigating the risk of ischaemic events [107]. However, the structure of the CoW is not consistent between individuals and dozens of anatomical variants exist in the general population [136, 137]. Understanding the differences between these variants is essential to study cerebrovascular diseases, predict disease progression, and improve clinical interventions. Previous studies have attempted to classify and describe the anatomical variations of CoW using categorisations such as the Lippert and Pabst system [136, 137]. However, more than 80% of the general population has one of the three most common CoW configurations [138]. The study of anatomical heterogeneity in CoW is limited by the size of available angiographic research data sets, which may only contain a handful of examples of all but the most common phenotypes. The goal of this study is to develop a generative model for CoW segmentations conditioned on anatomical phenotype. Such a model could be used to generate large anatomically realistic virtual cohorts of brain vasculature, and the less common CoW phenotypes can be augmented and explored in greater numbers. Synthesised virtual cohorts of brain vasculature may subsequently be used for training deep learning algorithms on related tasks (e.g. segmenting brain vasculature, classification of CoW phenotype, etc.), or performing in-silico trials.

Generative adversarial networks (GANs) [34] and other generative models have demonstrated success in the synthesis of medical images, including the synthesis of blood vessels and other anatomical structures. However, to the best of our knowledge, no previous study has explored these generative models for synthesising different CoW configurations. Additionally, no previous study has explored the controllable synthesis of different CoW configurations conditioned on desired phenotypes. The synthesis of

narrow tubular structures such as blood vessels using conventional generative models is a challenge. Our study builds upon the foundations of generative models in medical imaging and focusses on utilising a conditional latent diffusion model to generate visually realistic CoW configurations with controlled anatomical variations (i.e., by conditioning relevant anatomical information such as CoW phenotypes). Medical images like brain magnetic resonance angiograms (MRA's) tend to be high-dimensional and as a result are prohibitively memory intensive for generative models. Diffusion models and latent diffusion models (LDM) have recently been used for medical image generation [139] and have been shown to outperform GANs in medical image synthesis [140]. Diffusion models have also been successfully used to generate synthetic MRIs [141–143] but to the best of our knowledge there are no studies that use latent diffusion models or diffusion models to generate synthetic brain vasculature and MRA.

We propose a conditional latent diffusion model that learns latent embeddings of brain vasculature and, during inference, samples from the learnt latent space to synthesise realistic brain vasculature. We incorporate class, shape, and anatomical guidance as conditioning factors in our latent diffusion model, allowing the vessels to retain their shape and allowing precise control over the generated CoW variations. The diffusion model is conditioned to generate different anatomical variants of the posterior cerebral circulation based on the presence or absence of the Peripheral Communiation artery (PComA) [5.1] compares MRAs of two patients with and without PComA]. We evaluate the performance of our model using quantitative metrics such as multiscale structural similarity index (MS-SSIM) [144] and Fr'echet inception distance (FID) [145]. Comparative analyses are conducted against alternative generative architectures, including a 3D GAN and a 3D variational auto-encoder (VAE), to assess the superiority of our proposed method in reproducing CoW variations.

5.2 Methodology

5.2.1 Data and Pre-processing.

We trained our model on the publicly available IXI dataset [122] using the 181 3T MRA scans acquired at the Hammersmith Hospital, London. Images were centred, cropped from $512 \times 512 \times 100$ to $256 \times 256 \times 60$, and the intensity normalised. We then used a Residual U-net [146] to extract vessel segmentations from the MRA. The authors

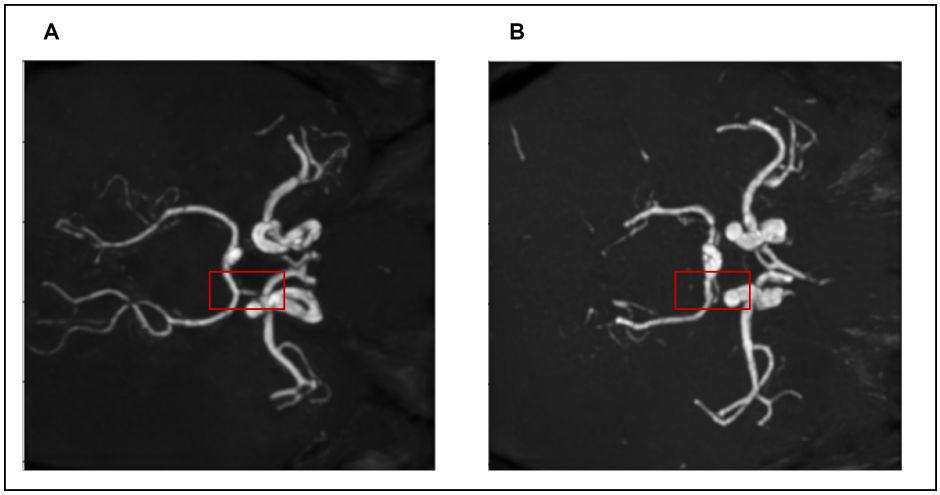

Figure 5.1: Figure showing presence and absence of PComA in the MRA of two patients

manually labelled each case with the presence / absence of one or both peripheral communicating arteries in the CoW. Class 1 includes cases where both the peripheral communication arteries are present (PComA), Class 2 includes cases with only one PComA, while Class 3 includes cases where both PComAs are absent.

5.2.2 Latent Diffusion Model.

Recent advances in diffusion models for medical image generation have achieved remarkable success. Diffusion models define a Markov chain of diffusion steps to add random Gaussian noise to the observed data sequentially and then learn to reverse the diffusion process to construct new samples from the noise. Although effective, vanilla diffusion models can be computationally expensive when the input data is of high dimensionality in image space ($256 \times 256 \times 60$ in our study). Hence, we employ the latent diffusion model (LDM), comprising a pretrained autoencoder and a diffusion model. The autoencoder learns a lower-dimensional latent embedding of the brain vasculature, while the diffusion model focuses on modelling the high-level semantic representations in the latent space efficiently. We employ a depth autoencoder for this objective, facilitating compression of the input image strictly along the channel dimension, thereby eschewing any max pooling or dimensional reduction in the X or Y axes. This process transforms the image from its original dimensionality of $256 \times 256 \times 60$ to $256 \times 256 \times 1$. Despite the challenge presented by such substantial compression, this is effectively achieved

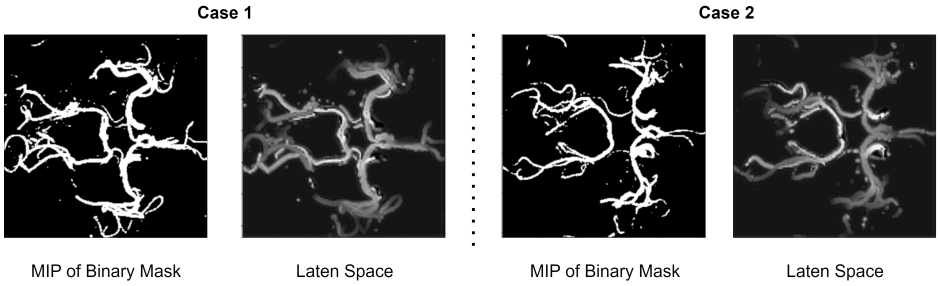

Figure 5.2: This figure shows the compressed latent space for two different input images

due to the nature of our dataset, which consists solely of binary masks of brain vessels. The resultant compressed latent space closely resembles the Maximum Intensity Projection of the image, preserving certain depth information and manifesting almost as a binary mask. Figure 5.2 demonstrates the appearance of the latent space for two distinct input images. Initially, the autoencoder undergoes a pretraining phase. Upon successful training of the compression model, latent representations derived from the training dataset are utilized as inputs to the diffusion model for subsequent analysis and generation.

The diffusion and reverse diffusion process is the same as Following [140] and what is described in the background section of this chapter. The simplified evidence lower bound (ELBO) loss to optimise the diffusion model by Ho *et al.* [140] can be formulated as a score-matching task where the neural network predicts the actual noise ϵ added to the observed data. The resulting loss function is

$$\mathcal{L}_\theta := \mathbb{E}_{\mathbf{x}_0,t,C\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta\left(x_t, t, C_{class}, C_{Ant}, C_{Shape}\right)\|^2\right]$$

where the conditional variables C_{class}, C_{Ant}, and C_{Shape} are employed for conditional generation. Specifically, C_{class} denotes the class variable responsible for determining the class to be generated, while C_{Ant} and C_{Shape} serve as anatomical and morphological guidance conditions, respectively.

We employ a model with a U-net-based architecture as the diffusion model. Our model has 5 encoding blocks and 5 decoding blocks with skip connections between the corresponding encoding and decoding blocks. We replace the simple convolution layers in the encoding and decoding blocks with a residual block followed by a multihead

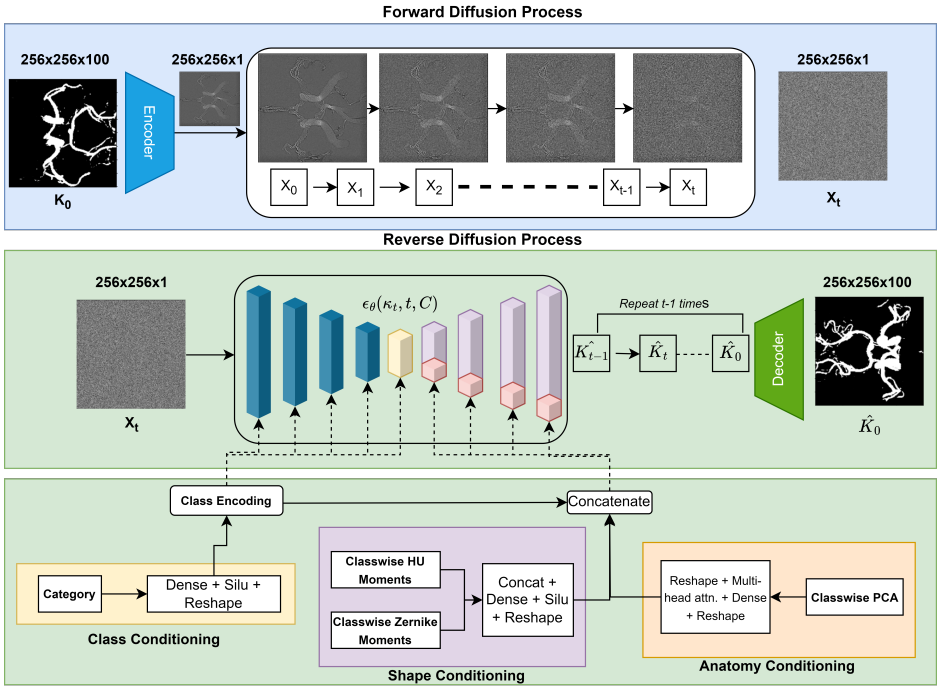

Figure 5.3: Overview of the latent diffusion process. The first panel (in blue) shows the forward diffusion process where noise is gradually added to the image. The second and third panels (in green) show the network architecute/reverse diffusion process and the inclusion of the conditioning variables in the reverse diffusion process.

attention layer to limit information loss in the latent space. Each encoding and decoding block takes the class category (based on CoW phenotypes) as an additional conditional input, while, only the decoding blocks take shape and anatomy features as additional conditional inputs.

5.2.3 Shape and Anatomy Guidance :-

Angiographic medical images exhibit intricate anatomical structures, particularly the small vessels in the peripheral cerebral vasculature. Preserving anatomical integrity becomes crucial in the generation of realistic and accurately depicted vessels. However, diffusion models often face challenges in faithfully representing the anatomical structure, which can be attributed to their learning and sampling processes that are heavily based on probability density functions [147]. Additionally, latent space models are

susceptible to noise and information loss within the latent space [56, 148]. To this end, we incorporate shape and anatomy guidance to improve the performance of our CoW generation.

Shape Guidance

We introduce a shape guiding component to our network to primarily preserve the maintain the shape and continuity of the vessels . Previous studies have demonstrated that the inclusion of geometric and shape priors can improve performance in medical image synthesis [149, 150]. We introduce shape guidance by incorporating class-wise **Hu and Zernike moments** as conditioning variables during model training [151, 152]. This choice stems from the nature of our image dataset, which comprises both vessel and background regions. By including these shape-related moments as conditions, we aim to better preserve vascular structures within the synthesised images.

Hu Moments Hu moments, introduced by Ming-Kuei Hu in 1962 [151], are a set of seven invariant moments derived from the second and third order central moments of an image. These moments are invariant under image transformations such as translation, scaling, and rotation, making them powerful descriptors for shape analysis. The central moments of a 2D image $I(x, y)$ are given by:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q I(x, y)$$

where $\bar{x}$ and $\bar{y}$ are the coordinates of the centroid of the image:

$$\bar{x} = \frac{\sum_x \sum_y x I(x, y)}{\sum_x \sum_y I(x, y)}, \quad \bar{y} = \frac{\sum_x \sum y I(x, y)}{\sum_x \sum_y I(x, y)}$$

Using these central moments, seven invariant moments are defined:

$$
\begin{aligned}
\phi_1 &= \mu_{20} + \mu_{02} \\
\phi_2 &= (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2 \\
\phi_3 &= (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2 \\
\phi_4 &= (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2 \\
\phi_5 &= (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] \\
&\quad + (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \\
\phi_6 &= (\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \\
&\quad + 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \\
\phi_7 &= (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] \\
&\quad - (\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]
\end{aligned}
$$

These moments are particularly effective for binary or grayscale images where shape is the primary characteristic of interest. This makes them well-suited for analyzing vessel masks because they are invariant to rotation, scaling, and translation, making them robust to changes in viewpoint or image orientation.

Zernike Moments Zernike moments [152] are a set of orthogonal moments based on Zernike polynomials, used for image analysis due to their robustness to noise and ability to represent image features compactly and accurately. Zernike polynomials $V_{nm}(x, y)$ are defined on the unit disk (a circular region of radius 1) and are given by:

$$V_{nm}(x, y) = R_{nm}(\rho)e^{im\theta}$$

where $\rho = \sqrt{x^2 + y^2}$ is the radial distance, $\theta = \tan^{-1}(y/x)$ is the angular component, and $R_{nm}(\rho)$ is the radial polynomial defined as:

$$R_{nm}(\rho) = \sum_{s=0}^{(n-|m|)/2} (-1)^s \frac{(n-s)!}{s!((n+|m|)/2 - s)!((n-|m|)/2 - s)!} \rho^{n-2s}$$

The Zernike moments of an image $I(x, y)$ are calculated as:

$$Z_{nm} = \frac{n+1}{\pi} \sum_x \sum_y I(x, y) V_{nm}^*(x, y)$$

where $V_{nm}^*(x, y)$ is the complex conjugate of the Zernike polynomial. Zernike moments are orthogonal, meaning that each moment captures unique information about the image, reducing redundancy. The magnitude of Zernike moments is invariant to rotation,

making them ideal for analyzing objects that may appear in different orientations. Furthermore, Zernike moments are less sensitive to noise compared to other moment-based descriptors, providing stable and reliable shape representation.

Zernike moments are particularly advantageous for analyzing vessel masks due to their ability to capture detailed information about the shape and structure of the vessels. This is crucial for differentiating between various vascular patterns. Their robustness to noise and rotation ensures consistent analysis even in varying imaging conditions. Additionally, Zernike moments provide a compact representation of the image, making it easier to store and process the shape descriptors for large datasets.

Both Hu and Zernike moments are powerful tools for shape analysis in medical imaging. Hu moments offer simplicity and invariance to basic transformations, making them suitable for straightforward shape characterization such as the mostly binary nature of our latent space used for training. Zernike moments, with their orthogonality and robustness, provide a more detailed and noise-resistant representation of complex shapes like vascular structures which is useful as even tough our latent space looks binary there is often noise introduce during the encoding process.

To incorporate the Hu and Zernike moments as conditions, we first calculate the Hu and Zernike moments for each instance of each class, then the class-wise mean is taken and the resulting Hu and Zernike moments for each class are concatenated with each other. An embedding layer comprises a dense layer with a SiLU activation function [153] and a reshape layer to ensure that the data are reshaped into a suitable format for integration as a condition within the decoding branches.

Anatomy Guidance

To further enhance the performance of our model, we incorporate anatomy guidance using principal component analysis (PCA) on images from each class. As the majority branches within the CoW exhibit a consistent configuration with minor variations attributed to the presence or absence of specific branches, the model tends to capture an average or mean representation of the CoW and generates synthetic images with very little variation between them. This characteristic becomes significant due to the limited number of images available per class. To address this, we use PCA components as conditions to enable the model to discern distinctive features specific to each class. We extract seven principal components along with the mean component for

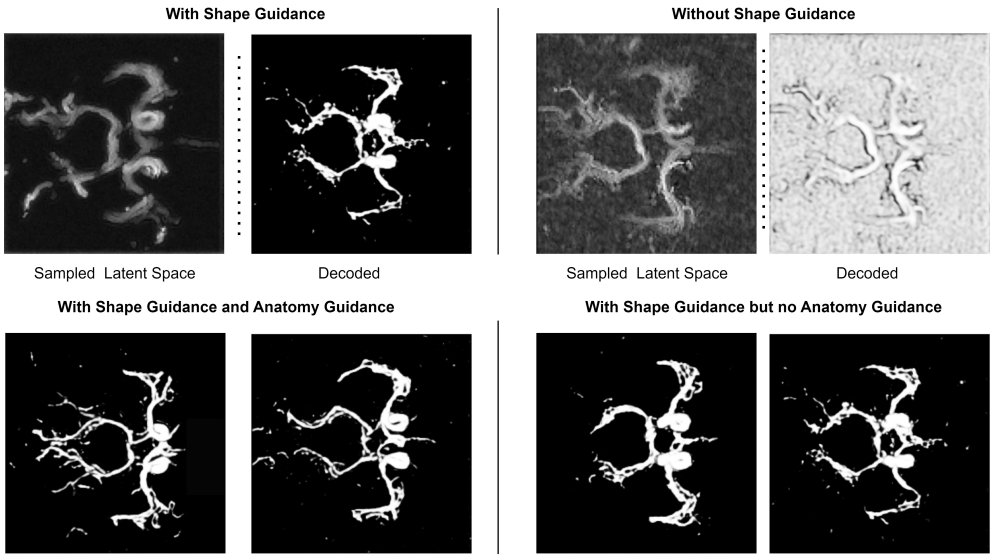

Figure 5.4: Row 1: Comparison of output of the latent diffusion network with and without using shape guidance as conditional input. In each column, the image on the left shows the output of our latent diffusion model and the image on the right shows the result of passing the output through the pretrained decoder and obtaining the Maximum Intensity Projection (MIP); Row 2: compares the output of the network with and without using anatomy guidance as conditional input. The generated images displayed on the right, which are produced without the incorporation of anatomy guidance, consistently exhibit a similar variation of the circle of Willis. Conversely, the images presented on the left, which are generated with the inclusion of anatomy guidance, demonstrate a greater degree of realism and variability in the synthesised circle of Willis variations.

each class, concatenate them, and reshape the data. The resulting features are then passed through a multi-head attention block where each principal feature is treated as a 'token', followed by a dense layer and another reshape operation for integration into the decoding branches.

Figure. 5.4 shows the effect of incorporating shape moments and PCA as conditions in our diffusion process. By incorporating shape and anatomy guidance conditions during the training of our diffusion model, we leverage specific features and knowledge related to the vessel structures and the general anatomy of the images. This approach

promotes the generation of more realistic images, contributing to an improved anatomical fidelity.

5.3 Results and Discussion

5.3.1 Implementation Details.

All models were implemented in TensorFlow 2.8 and Python 3. For the forward diffusion process we use a linear noise schedule with 1000 time steps. The model was trained for 2000 epochs with a learning rate of 0.0005 on a Nvidia Tesla T4 GPU and 38 Gb of RAM with Adam optimiser.

5.3.2 Results and Discussion.

To assess the performance of our model, we compared it against two established conditional generative models: 3D C-VAE [33] and a 3D-α-WGAN [154] along with a vanilla LDM and an LDM with shape guidance. We use the FID score to measure the realism of the generated vasculature. To calculate FID we used a pre-trained InceptionV3 as a feature extractor. A lower FID score indicates higher perceptual image quality. In addition, we used MS-SSIM and 4-G-R SSIM to measure the quality of the generated images [155, 156]. MS-SSIM and 4-G-R SSIM are commonly used to assess the quality of synthesised images. Typically, a higher score is indicative of better image quality, implying a closer resemblance between the synthesised CoW and the ground truth reference. MS-SSIM and 4-G-R SSIM were calculated over 60 synthesised CoW cases for each model. Table 1 presents the evaluation scores achieved by our model, 3D CVAE, and the 3D-α-WGAN and the above metrics. As seen in Table 5.1, our model demonstrates a better FID score, suggesting that the distribution of CoW variants synthesised by our model is closer to that observed in real CoW data, compared to the other models. Additionally, our model achieves higher MS-SSIM and 4-G-R SSIM scores compared to the other methods. These higher scores indicate better image quality, implying that the generated CoW samples resemble the real CoW images more closely. Figure. 5.5 provides a qualitative comparison among the generated samples obtained from the three models to provide additional context to the quantitative results presented in Table 1. As the output of each model is a 3D vascular structure, maximum intensity projections (MIP) over the Z-axis which condense the volumetric representation into a 2D plane

are used to visually compare the synthesised images.

Table 5.1: Quantitative evaluation of Synthetic CoW vasculature

Model	FID ↓	MS-SSIM ↑	4-G-R SSIM ↑
3D CVAE	52.78	0.411	0.24
3D-α-WGAN	12.11	0.53	0.41
LDM	176.41	0.22	0.13
LDM + Shape Guidance	8.86	0.58	0.47
Ours (LDM + Shape & Anatomy Guidance)	5.644	0.61	0.51

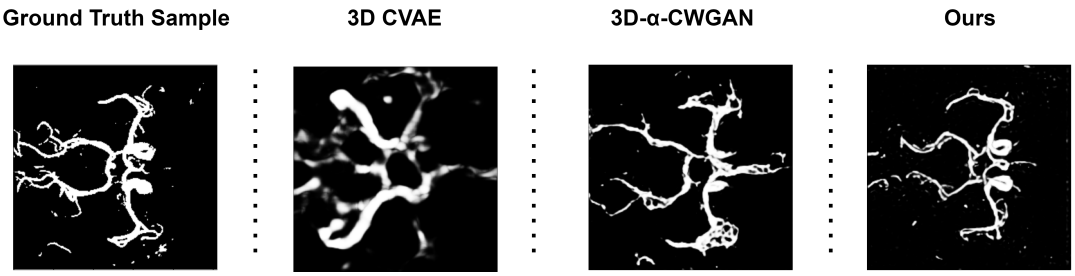

Figure 5.5: Comparison between the maximum intensity projections (MIPs) of a real Circle of Willis(CoW) against those synthesised with 3D CVAE, 3D-α-WGAN, and our model.

Figure. 5.5 reveals that the 3D CVAE model can only generate a limited number of major vessels with limited details. On the other hand, although the 3D-α-WGAN model produces the overall structure of the CoW, it exhibits significant anatomical discrepancies with the presence of numerous phantom vessels. On the contrary, our model demonstrates a faithful synthesis of the majority of CoW, with most vessels identifiable. To generate variations of the CoW based on the presence or absence of the posterior communicating artery, our latent diffusion model uses class-conditional inputs where the classes represent different CoW phenotypes. Consequently, to demonstrate the class-conditional fidelity of the proposed approach, we also evaluate the model's performance in a class-wise manner. The qualitative performance of our model for different classes, compared to real images belonging to those classes, is shown in Figure. 5.6

The results presented in Figure. 5.6 demonstrate the performance of our model in generating realistic variations of the Circle of Willis. Particularly notable is the

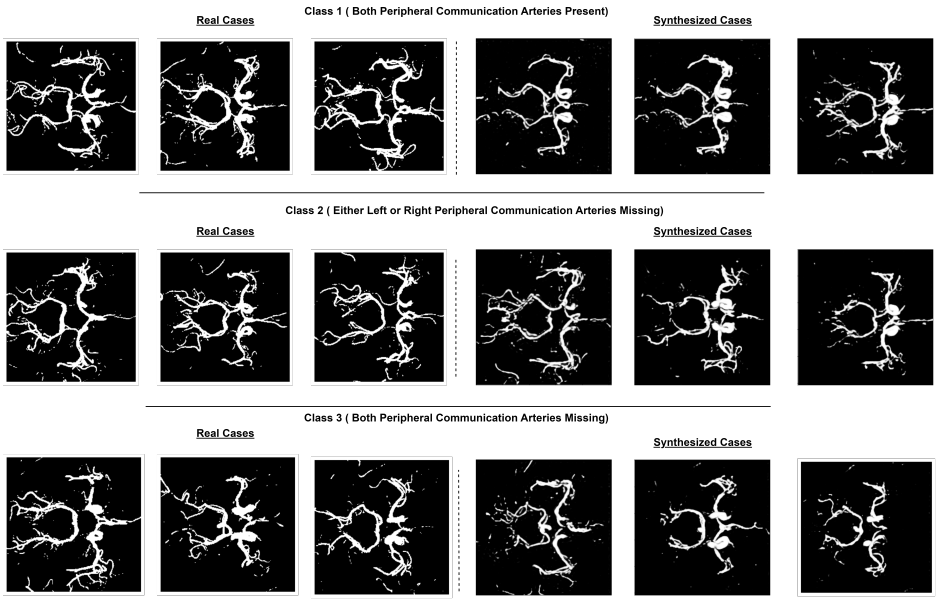

Figure 5.6: Comparison between the real and synthesised maximum intensity projections (MIPs) for each of the three classes

Table 5.2: Quantitative class-wise evaluation of Generated CoW vasculature

Class	FID Score ↓	MS-SSIM ↑	4-G-R SSIM ↑
Class 1	4.41	0.65	0.65
Class 2	3.88	0.52	0.52
Class 3	7.63	0.41	0.41
Overall	5.64	0.61	0.51

model's proficiency in producing accurate representations for classes 1 and 2, surpassing its performance in class 3 due to the limited sample size of the latter. Our model excels in synthesising the posterior circulation and the middle cerebral arteries, showing remarkable fidelity to anatomical structures. However, it faces challenges in effectively generating continuous representations of the anterior circulation. Further investigation and refinement may be required to enhance the model's ability in this specific aspect. In addition to the visual assessment, we also compute class-wise FID scores, along with the MS-SSIM and 4-G-R SSIM scores. These quantitative evaluations serve to provide a more comprehensive understanding of the model performance with respect to each class. The class-wise performance scores shown in Table 5.2 are consistent with our

observations from Figure. 5.6, that the model's performance for class 3 is worse than its performance on classes 1 and 2.

5.4 Conclusion

We proposed a latent diffusion model that used shape and anatomy guidance to generate realistic CoW configurations. Quantitative qualitative results showed that our model outperformed existing generative models based on a conditional 3D GAN and a 3D VAE. Future work will look to enhance the model to capture wider anatomical variability and improve synthetic image quality.

While our method demonstrates potential in conditionally generating CoW phenotypes, it is not always feasible to have sufficient data for every phenotype. If a particular phenotype has limited data, the generative model's performance suffers when generating instances of that phenotype (as illustrated in the class-wise quantitative evaluation in this chapter). This problem is exacerbated in certain datasets where some phenotypes are extremely rare. In the next section, we attempt to address this issue by utilizing a diffusion model on a comparatively simpler dataset of cerebral vessels, which includes aneurysms. We explored a possible approach for enabling the conditional generation of images depicting specific anatomical features, despite the scarcity of data for these features.

Addendum

This addendum addresses some of the specific questions raised about the contents of this chapter and provides some clarifications :-

Why do you want to simulate data? The simulation of data is essential to overcome the limitations posed by the scarcity of large and diverse datasets for rare anatomical variations of the Circle of Willis (CoW). By generating anatomically realistic virtual cohorts, we can augment the representation of underrepresented phenotypes, enabling comprehensive studies of cerebrovascular anatomy and pathology. Simulated

data also supports the training and validation of deep learning models for tasks such as segmentation and classification, ensuring robustness and generalizability.

Why do you want to compress from $256^2 \times 60$ to $256^2 \times 1$? Why work in MIPs? Why not do this in 3D? The compression from $256^2 \times 60$ to $256^2 \times 1$ is achieved to reduce the computational complexity of the diffusion model while preserving critical information. The resultant latent space closely resembles a Maximum Intensity Projection (MIP), which retains depth information effectively in a single 2D representation. Working in MIPs simplifies the dimensionality of the data and aligns with the binary mask nature of the dataset. While a fully 3D approach would provide volumetric fidelity, it introduces significant computational overhead. Our strategy strikes a balance between computational efficiency and anatomical accuracy.

Is the class conditioning fed into the network via concatenation? Yes, class conditioning is incorporated into the network via concatenation. Specifically, class-wise anatomical and morphological guidance features are concatenated with latent representations at specific layers, enabling the model to generate anatomically coherent variations of the CoW based on the desired phenotype.

Validation: Why not compare to 3D VQVAE-Transformer models that were out? While 3D VQVAE-Transformer models have shown promise in generative tasks, the focus of this study is on leveraging latent diffusion models due to their superior performance in preserving fine anatomical structures in high-dimensional data. Additionally, diffusion models offer a robust mechanism for generating synthetic data with conditional guidance, which aligns more closely with the objectives of this research. Future comparisons with transformer-based generative frameworks may further validate the advantages of our approach.

In Table 5.1, can you explain the LDM being 176 of FID? Why so high? Has it converged? The high FID score of 176 for the vanilla Latent Diffusion Model (LDM) reflects its inability to generate anatomically accurate CoW configurations without additional guidance. This model lacks the shape and anatomy guidance mechanisms that significantly improve fidelity and reduce artifacts. While the model

had technically converged during training, its lack of conditioning inputs resulted in suboptimal performance, as evidenced by the poor FID score.

How does the validation show that the classes generated in Figure. 5.6 are correct from an anatomical point of view? The validation demonstrates anatomical correctness by comparing synthesized CoW configurations against real-world examples within the same class. Quantitative metrics such as MS-SSIM and FID scores confirm the visual fidelity of the generated images. Furthermore, the inclusion of class-conditional anatomical features ensures that the generated configurations retain the characteristic patterns of their respective classes. This is visually corroborated in Figure. 5.6, where synthesized CoW images exhibit realistic variations corresponding to their designated phenotypes.

Chapter 6

Few Shot Diffusion Models to Generate Brain Vasculature

6.1 Introduction

Cerebral aneurysms pose significant neurosurgical and neurological concerns and have the potential to lead to life-threatening conditions, like subarachnoid hemorrhage (SAH). Their prevalence in the general population underscores their contribution to morbidity and mortality. A major challenge in this domain arises from the scarcity of comprehensive data, particularly for less common aneurysm phenotypes. This data limitation poses significant obstacles in developing accurate and robust diagnostic models. Generative models, particularly in medical imaging, present a promising solution to this issue. They hold the potential to synthesize high-quality, detailed images of cerebral aneurysms, even in scarcity. However, the effectiveness of traditional generative models is typically constrained by the availability of data, a notable hurdle in the context of aneurysm imaging, where extensive datasets are often lacking.

In the domain of medical imaging, generative models such as generative adversarial networks (GANs) [34] have emerged as a promising solution, offering the potential to create detailed and accurate representations of anatomical structures [157] . More recently, diffusion models (DDPM) have shown great prowess in generating synthetic data [139] and outperforming GANs in image synthesis [140]. Diffusion models have also been successfully used to generate synthetic brain magnetic resonance images (MRIs) [143] and vascular structures [158]. However, the efficacy of these models is often limited by the requirement for extensive training datasets, which are not always available. Even when labeled datasets are available, medical imaging datasets suffer from data imbalance due to certain anatomical phenotypes being underrepresented.

The concept of few-shot learning, a technique for training models with limited data, has become increasingly relevant in medical imaging domains characterized by data scarcity and class imbalance. This is especially true for rare or underrepresented cerebral aneurysm types. While few-shot learning has been explored in diffusion models in prior research [159, 160], our work is the first to our knowledge to apply this concept to the generation of cerebral aneurysms in brain vessel imaging.

Our study addresses this gap by introducing an innovative approach using latent diffusion models (LDMs) with few-shot learning, allowing for the generation of high-fidelity models of brain vessels with aneurysms from a very limited number of samples in each class. Through the integration of transformer-based class embeddings, we reduce the reliance on having a large number of samples from each class to conditionally

generate images. We also leverage signed distance functions (SDF) as a conditioning variable to further enhance the quality of the generated vessels and maintain vessel continuity. We compare the performance of our model against other generative models such as a 3D GAN, 3D variational auto-encoders (VAEs), and also against vanilla diffusion models. To assess the quality of the generated aneurysms, we use metrics such as multi-scale structural similarity (MS-SSIM), Fréchet inception distance (FID), and 4GR SSIM. To our knowledge, this is the first study to use generative/diffusion models to generate synthetic brain vessels with aneurysms.

6.2 Methods

6.2.1 Data and Preprocessing

For training our model, we utilized the @neurIST dataset encompassing 225 3D Rotational Angiography (3DRA) scans of the brain, each with at least one cerebral aneurysm. Out of these, detailed information regarding aneurysm location and other conditional attributes was available for 105 cases. Within these 105 labeled cases, there were more than 15 different classes of aneurysms based on their location with each class having about 7 sample cases on average. In the initial phase of preprocessing, we extracted vessel segmentations from the 3DRA volumes. This extraction was facilitated by the application of VASeg, a segmentation tool designed for vascular imaging [108]. Post-segmentation, the 3DRA volumes underwent a process of centerline cropping, ensuring a focus on the most relevant vascular structures. These cropped segments were then resized to uniform dimensions of $128 \times 128 \times 100$, optimizing them for subsequent processing and analysis. The final step involved the categorization of aneurysms based on their location attributes and saving them as class variables to act as a conditioning vector to the diffusion model. In this study, we mainly focus on basilar tip, medial wall carotid, and ophthalmic segment carotid aneurysms. Each class contains around 5 samples.

6.2.2 Latent Diffusion Model

Diffusion models have demonstrated remarkable success in synthesizing high-quality medical images and vascular structures. Central to the operation of diffusion models is the concept of a Markov chain, which is employed to methodically introduce Gaus-

sian noise into the observed data through a sequence of diffusion steps. The crux of these models lies in their ability to reverse this diffusion process, thereby enabling the generation of new samples from the noise-infused data.

Despite their effectiveness, a notable challenge with conventional diffusion models arises when dealing with high-dimensional data such as the images of size $128 \times 128 \times 100$ used in our study. To circumvent this computational complexity, we have opted to utilize a latent diffusion model (LDM). The architecture of LDM comprises two pivotal components: a pre-trained autoencoder and a diffusion model. The autoencoder is tasked with learning a lower-dimensional latent representation of the brain vasculature from $128 \times 128 \times 100$ to $128 \times 128 \times 1$. This reduction in dimensionality is crucial as it allows for a more manageable and efficient manipulation of data. Concurrently, the diffusion model is designed to focus on modeling the high-level semantic representations within this latent space. By operating in a space of reduced dimension, the LDM alleviates the computational burden but retains the capacity to capture and model the intricate details and nuances of the brain vascular structures.

Like in [140], the diffusion process can be defined through forward and reverse Markov chains, where the forward process iteratively transforms the data x_0 into a standard Gaussian X_T as follows:

$$\begin{aligned} q\left(\mathbf{x}_{1:T}|\mathbf{x}_0\right) &= \prod_{t=1}^{T} q\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right), q\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right) \\ &:= \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right) \end{aligned}$$

where $q\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right)$ is the transition probability at the time step t based on the noise schedule β_t. Therefore, the noisy data $\mathbf{x}_t$ can be formulated as $q\left(\mathbf{x}_t|\mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right)$, where $\alpha_t := 1 - \beta_t, \bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

Consecutively, the reverse process parameterised by θ can then be defined as:

$$\begin{aligned} p_\theta\left(\mathbf{x}_0|\mathbf{x}_T\right) &= p\left(\mathbf{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1}|\mathbf{x}_t\right), p_\theta\left(\mathbf{x}_{t-1}|\mathbf{x}_t\right) \\ &:= \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta\left(\mathbf{x}_t, t\right), \mathbf{\Sigma}_\theta\left(\mathbf{x}_t, t\right)\right) \end{aligned}$$

The simplified evidence lower bound (ELBO) [140] loss can be formulated as a score-matching task, where the neural network predicts the actual noise ϵ added to the observed data:

$$\mathcal{L}_\theta := \mathbb{E}_{\mathbf{x}_0, t, C, \epsilon \sim \mathcal{N}(0,1)}\left[\left\|\epsilon - \epsilon_\theta\left(x_t, t, C\right)\right\|^2\right]$$

where C is the conditioning vector in conditional generation. In our study, the conditioning vector encodes the location of the aneurysm.

The 3D binary masks of the vessels generated from 3DRA volumes in our dataset are passed through the encoder of the pre-trained autoencoder to obtain a dimensionally reduced latent space. This latent space serves as the input for our diffusion model. Consequently, the diffusion model's output is also in this latent space, which is then processed through the decoder of the pre-trained autoencoder to reconstruct the 3D binary masks.

We first train our latent diffusion model unconditionally with no additional condition features on the unlabeled samples in the dataset so that it can learn to generalize the structure of the vessels. After pre training on the unlabeled data, we train the model over limited labeled cases from the three selected classes (basilar tip, medial wall carotid and ophthalmic segment carotid aneurysms) along with class-wise conditioning from a transformer and signed distance fields (SDF) based features.

6.2.3 Transformer based class conditioning

An inherent fault with generative models (especially diffusion) is their intrinsic reliance on substantial data volumes to train effectively and produce convincing outputs. This issue is particularly pronounced in our study, given the limited availability of data, with some classes containing as few as five samples. Such a sparse dataset poses significant difficulties for generative models, as they struggle to accurately approximate the distribution of the data.

To address this challenge, we introduce an innovative approach that integrates transformer [161] -based class features to guide the diffusion process. We employed a set based vision transformers (ViT) [162] model, designed to ingest the entire 3D volume and function as a classifier, determining the specific location of an aneurysm within the brain. Following the successful training of the ViT, we removed its final classification layer. Subsequently, we processed the images from each class through this transformer to extract class-wise encoded features. These features, in conjunction with the class conditioning variables, were then incorporated into the conditioning vector of the diffusion model, enhancing its ability to generate data representative of each class.

6.2.4 Signed Distance Field (SDF) based Conditioning

Although diffusion models show great success in generating medical images, generating vascular structures is challenging as vessels have structural features that need to be maintained, most importantly vessel continuity. Also, aneurysms are small compared to the total size of cerebral vasculature, which makes them hard to track and generate. Studies have shown that adding shape based features to the generative process can improve performance in these tasks [149, 158].

To this end, we incorporate signed distance fields (SDF) as an additional input to the diffusion process. The primary idea behind SDFs is to associate each point in space with a distance value, and the sign of this distance value indicates whether the point is inside or outside of the shape which makes them particularly useful for tasks like shape analysis and 3D rendering. We first convert the segmentation masks for each class into corresponding SDFs. These SDFs act as an input to a 3D ResNet, which similar to the set-based ViT described in the previous subsection is trained to act as a classifier. After successful training, the final output layer is removed and the class-wise features are extracted and incorporated as conditions in the diffusion process. The introduction of these features enhances the quality of the generated vessels by promoting the generation of more continuous vessels as can be seen in Panel B in Figure. 6.2. The overall architecture of the model is shown below in Figure. 6.1

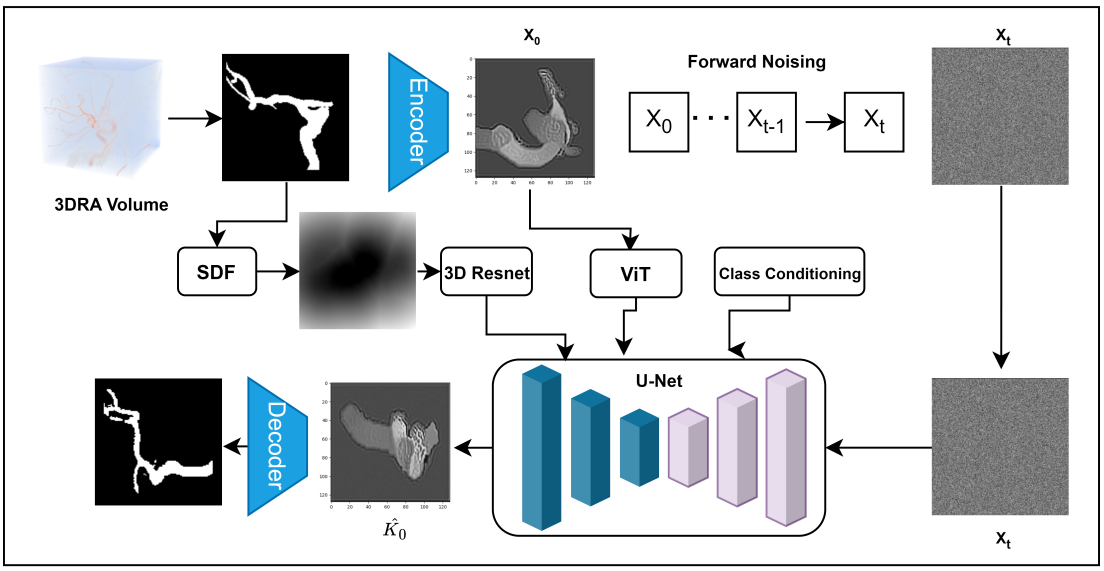

Figure 6.1: Overview of the architecture of the model.

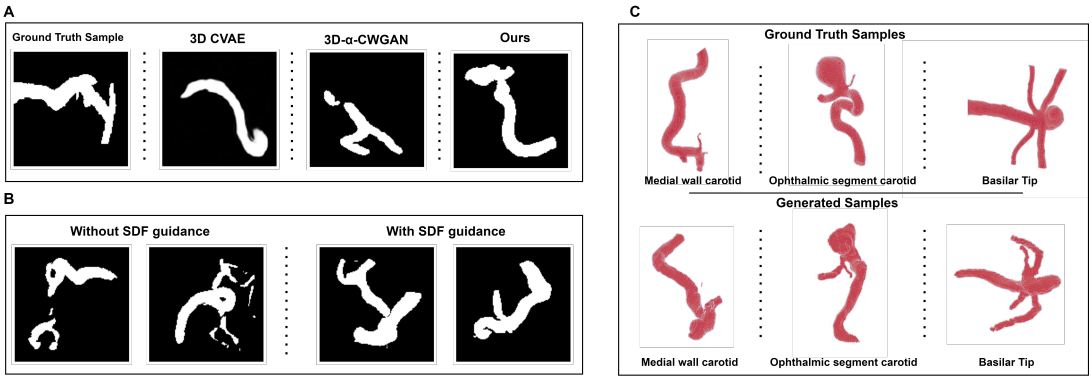

Figure 6.2: Panel A compares the MIPs from the generated cases from different models. Panel B showcases the effect of adding SDF based conditioning to the diffusion process. Panel C compares the volumetric meshes generated from generated and ground truth cases from each class

6.3 Experiments and Results

6.3.1 Implementation Details

All models were implemented in TensorFlow 2.8 and Python 3. For the forward diffusion process we use a linear noise schedule with 1000 time steps. The model was trained for 2000 epochs with a learning rate of 0.0005 on a Nvidia Tesla T4 GPU and 38 Gb of RAM with Adam optimiser. The vision transformer was trained on a Nvidia V100 GPU with 38Gb of RAM.

6.3.2 Results and Discussion

The performance of our proposed model was compared against established generative models serving as baselines. These include a 3D convolutional variational autoencoder (3D C-VAE), a 3D-α-Wasserstein generative adversarial network (3D-α-WGAN)[154], and a conventional diffusion model (Vanilla DDPM). The purpose of this comparison was to ascertain the efficacy of our approach relative to these well-established models in generating high-quality cerebral vascular images. To quantitatively assess the realism of the generated vasculature by each model, we employed the FrÃ©chet inception distance (FID) score. The FID-score was computed using a pre-trained InceptionV3 network

as a feature extractor. It is important to note that a lower FID score is indicative of higher perceptual image quality, reflecting greater realism in the generated images. Additionally, to provide a comprehensive evaluation of image quality, we utilized the multi-scale structural similarity index (MS-SSIM) and 4-G-R SSIM metrics, as outlined in references [155, 156]. These metrics are extensively used in the field to assess the quality of synthesized images. A higher score in both MS-SSIM and 4-G-R SSIM typically signifies superior image quality, implying a closer resemblance to the actual ground truth images. An extremely high score from MS-SSIM and 4-G-R SSIM however could indicate very high levels of similarity between the synthesised cases and the ground truth indicating low variablilty. The MS-SSIM and 4-G-R SSIM scores were calculated over six synthesized cases for each model.

Table 6.1 encapsulates the evaluation scores achieved by our model, 3D C-VAE, 3D-α-WGAN, and Vanilla DDPM, based on the aforementioned metrics. This comparative analysis enables us to elucidate the strengths and limitations of our approach in the context of existing generative models.

Table 6.1: Quantitative evaluation of Synthetic vessels

Model	FID ↓	MS-SSIM ↑	4-G-R SSIM ↑
3D CVAE	8.78	0.36	0.31
3D-α-WGAN	3.55	0.67	0.56
DDPM	4.41	0.69	0.55
Ours	2.56	0.71	0.61

Table 6.1 showcases that our model outperforms the other baselines in terms of FID, indicating that the distribution of the synthesized variants by our model more closely aligns with the real data distribution compared to other evaluated models. Furthermore, our approach outperforms the others in terms of MS-SSIM and 4-G-R SSIM scores, reflecting higher image quality and a closer resemblance of the generated vessels to the real ones.

Figure 6.2 provides a qualitative evaluation through a visual comparison of the synthesized samples from each model. Panels A and B employ maximum intensity projection (MIP) to render 3D binary masks of the vessels onto a 2D plane for analysis. In Panel A, we present the comparisons based on the MIP of the cases generated by each respective model. The convolutional variational autoencoder (VAE) primarily repro-

duces the fundamental structure of the vessels, achieving continuous vessel formation but lacking in variability and branching features. The generative adversarial network (GAN) introduces greater variability and detail in the vessel structures; however, it encounters challenges in maintaining vessel continuity. In contrast, our model excels in generating realistic and continuous vascular structures, closely mirroring the intricacies of actual vessels. Panel B delineates the differential impact of employing SDF-based conditioning in our diffusion model, underscoring its essential role in preserving vessel continuity, a feature that is notably compromised in its absence.

Recognizing the limitations of MIPs in accurately representing the complex three-dimensional nature of vascular structures, we further conducted a comparison using volumetric meshes which are showcased in panel C in Figure 6.2. These meshes were generated from binary masks for each class and compared against their corresponding ground truth samples. This analysis revealed that the cases synthesized by our model not only bear key characteristics akin to the ground truth but also exhibit discernible variability, demonstrating the model's efficacy in replicating both the fidelity and diversity of real-world vascular formations.

While the quality of the generated vessels from our study seems promising, it is important to acknowledge the limitations posed by the lack of extensive training data. This scarcity potentially restricts the variability of the generated cases, as the model's capacity to learn diverse vessel structures is directly tied to the dataset's breadth. Additionally, it is crucial to consider anatomical accuracy in the context of variability. Excessive variability in the generated structures might not accurately reflect the true anatomical complexity of cerebral vessels. Therefore, while our model demonstrates proficiency in replicating realistic vessel structures, the balance between variability and anatomical fidelity remains a key consideration for the authenticity and applicability of the generated outputs.

6.4 Conclusion

This study introduced a novel approach for generating brain vessel segmentations with aneurysms, particularly under the constraint of having classes with limited data.

6.5 Acknowledgement

This research was partially supported by the National Institute for Health and Care Research (NIHR) Leeds Biomedical Research Centre (BRC), UKRI Frontier Research Guarantee INSILICO (EP/Y030494/1) ,EC Sixth Framework Programme @neurIST (FP6-2004-IST-4-027703) and the Royal Academy of Engineering Chair in Emerging Technologies (CiET1919/19).

Addendum

This addendum addresses specific questions raised about the chapter "Few Shot Diffusion Models to Generate Brain Vasculature."

Why use Signed Distance Functions (SDF)? How does it actually work? If it is needed to decode, how do you create them? Signed Distance Functions (SDF) are employed to enhance the anatomical realism and structural continuity of the generated vascular structures. SDF assigns a distance value to each point in space, with the sign indicating whether the point lies inside or outside the structure. This representation ensures that the vascular shape and continuity are preserved during generation. To create SDFs, the binary segmentation masks are transformed by calculating the shortest distance of each voxel to the vessel boundary. Positive distances represent points outside the structure, while negative values correspond to points inside. The resulting SDFs are then used as input features during the diffusion process, promoting more realistic synthesis outcomes.

Why is the segmentation dimension different from the image one? The segmentation dimension differs from the image dimension due to the application of latent space encoding in the diffusion model. By compressing the original $128 \times 128 \times 100$ binary masks into a lower-dimensional latent representation ($128 \times 128 \times 1$), computational efficiency is achieved without compromising the critical structural details necessary for accurate segmentation. This dimensional reduction focuses the model's learning capacity on salient features, optimizing both training and synthesis processes.

Aren't segmentations and images almost the same thing? While segmentation masks and images share similarities, they serve distinct purposes. Images provide

pixel intensity information, whereas segmentation masks are binary representations delineating specific anatomical structures. In this study, segmentation masks explicitly define vascular structures, enabling the model to focus on generating anatomically accurate and continuous vessels. The distinction is crucial as it allows for targeted conditioning and improves the precision of the generated outputs.

What is the advantage of bringing them together? Integrating segmentation masks with corresponding images leverages complementary information. Images provide global context, while segmentation masks offer detailed structural boundaries. Combining these elements enhances the model's ability to synthesize anatomically accurate vascular structures while maintaining continuity and preserving finer details. This synergy ensures that the generated outputs align closely with real-world anatomical variations, improving both the fidelity and applicability of the synthetic data.

Chapter 7

Anatomy-guided latent diffusion models for generating brain MR angiography images and vessel segmentation masks

7.1 Introduction

The Circle of Willis (CoW) is a network of arteries that supplies blood to the brain. The CoW plays a key role in providing a collateral pathway for blood flow in cases where some primary vessels are obstructed, ensuring continuous cerebral perfusion and reducing the risk of ischemic events ([107]). The anatomy of the CoW varies significantly among individuals, with numerous variants observed across the population. Understanding these variations is critical for advancing research into cerebrovascular diseases, predicting disease progression, and enhancing clinical outcomes. Previous research efforts have employed classification systems such as the Lippert and Pabst system to describe these anatomical differences ([136, 137]). Nonetheless, while more than 80% of individuals exhibit one of three most common CoW configurations ([138]), rarer variations have not been as thoroughly investigated.

The exploration of CoW anatomical variations is constrained by the limited size of publicly available angiographic datasets, which often include only a few examples of the less common phenotypes. The aim of this study is to develop a generative deep learning model conditioned on anatomical phenotype for CoW angiography images and their vessel segmentations. Such a model would enable the generation of large, anatomically accurate virtual cohorts of brain vasculature, particularly augmenting the representation of rarer CoW phenotypes. These synthetic cohorts could improve the training of deep learning algorithms for tasks such as brain vessel segmentation and CoW phenotype classification, or could be used to generate virtual patient cohorts for performing in-silico trials for new cerebrovascular treatments ([**?**]).

Generative adversarial networks (GANs) and other generative models have advanced the field of medical imaging synthesis considerably in recent years, showing particular efficacy in creating realistic representations of blood vessels and other complex anatomical structures. More recently, diffusion models and latent diffusion models (LDM) have shown considerable promise with results that outperform GANs ([139, 140]). Diffusion models have also been successfully used to generate synthetic MRIs ([141–143]). While diffusion models in particular have demonstrated encouraging outcomes in generating medical images, there has been limited work done on using them for synthesising vascular structures. This can be attributed to the need to preserve their continuity and structure, particularly in more intricate vascular systems like the CoW or retinal blood vessels. Although the generation of complex vascular structures is

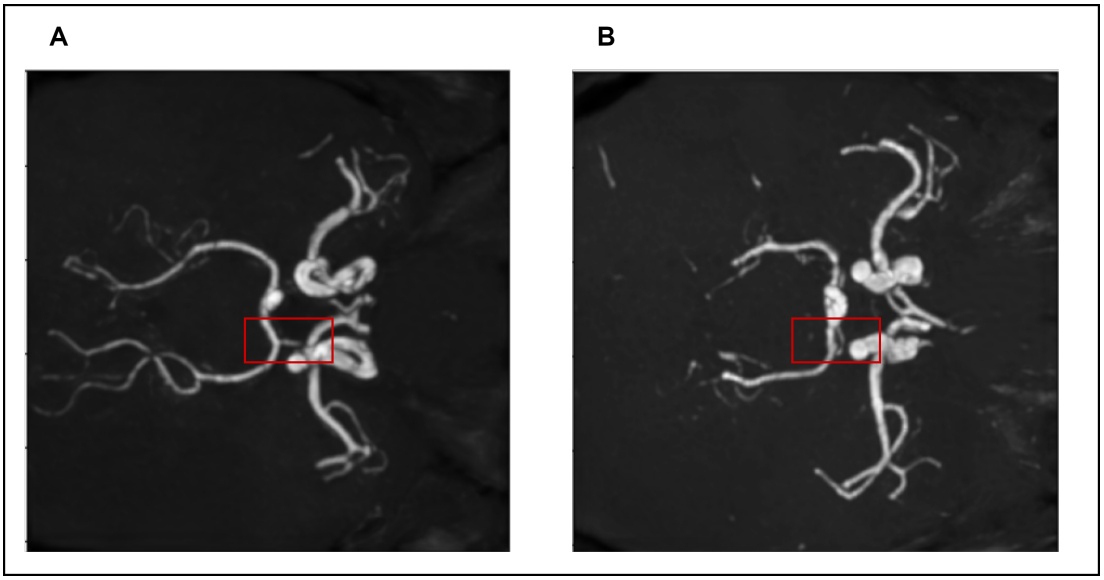

Figure 7.1: Examples of magnetic resonance angiograms from two different patients with and withouth a posterior communicating artery

feasible with certain priors ([163]) or by cross-modality synthesis ([164]), the creation of entirely novel vascular configurations remains challenging. The complexity of this challenge is exacerbated by the high dimensionality inherent in vascular imaging modalities such as Magnetic Resonance Angiography (MRA) and Three-Dimensional Rotational Angiography (3DRA). This high dimensionality necessitates a substantial increase in the volume of data required for training generative models. However, assembling large angiographic datasets is problematic given the limited availability of comprehensive and accessible data in this field. To the best of our knowledge, no previous study has explored these generative models for synthesising different CoW configurations. Additionally, no previous study has explored the controllable synthesis of different CoW configurations conditioned on desired phenotypes.

We introduce a conditional latent diffusion model paired with a multi-channel depth autoencoder, which is engineered to learn latent embeddings of brain vasculature. During inference, this model samples from the learned latent space to synthesise realistic representations of brain vasculature. We enhance the models capabilities by integrating class and anatomical guidance through conditioning factors, including features extracted through VisionTransformer ([165]). This approach facilitates the generation of

continuous vessels and allows precise control over the produced variations of the CoW, even within the constraints of limited data availability. The diffusion model specifically adapts its generation process to produce different anatomical variants of the posterior cerebral circulation, conditioned on the presence or absence of the Peripheral Communicating Artery (PComA) – an artery linking the internal carotid and posterior cerebral arteries (see Fig. 7.1). The performance of the model is quantitatively evaluated using metrics such as the Multiscale Structural Similarity Index Metric (MS-SSIM) and the Fréchet Inception Distance (FID). Additionally, we conduct comparative analyses against other generative architectures, such as a 3D GAN, 3D Variational Auto-Encoder (VAE), and a 3D Denoising Diffusion Probabilistic Models (DDPM), to demonstrate the efficacy and superiority of our proposed method in accurately reproducing CoW variations.

7.2 Related Works

Diffusion models have become a focal point in the advancement of image generative models because of their superior performance in generating realistic and high-quality images. In the study conducted by ([166]), the authors highlight the exceptional capabilities of diffusion models like the Med-DDPM in semantic 3D medical image synthesis. Their work emphasizes how these models excel over traditional GANs by providing enhanced stability and quality in generated images. The study also discusses how Med-DDPM adeptly manages data scarcity and privacy issues, which are critical in the clinical context, thus presenting a substantial improvement in handling the complexities of medical data.

The research conducted by ([167]) investigated the adaptation of conditional diffusion models for applications such as medical image segmentation. Their conclusions reveal that these models adeptly employ image-level annotations to steer both the synthesis and segmentation processes. This method holds particular promise in medical imaging, where obtaining comprehensive annotations can be labor-intensive and costly. The study's results suggest that diffusion models can substantially reduce the dependence on extensive annotations while maintaining competitive accuracy in segmentation tasks. This feature is vital in situations where manual labeling is impractical, thereby enhancing the availability of sophisticated imaging technologies. Conversely, the HistoDiffusion approach [101] takes advantage of latent diffusion models (LDM) trained on

extensive unlabeled datasets for synthetic augmentation, thus decreasing the demand for expert annotations. This technique achieved a 6.4% boost in classification accuracy in colorectal cancer histopathology images, demonstrating the promise of pre-trained diffusion models in bolstering small labeled datasets.

In brain imaging, conditional diffusion probabilistic models (cDPM) have been employed to create lifelike brain MRIs, offering a less compute-intensive alternative to traditional GAN methodologies. By being conditioned on partial MRI slices, cDPMs are capable of generating complete 3D brain volumes, preserving anatomical integrity while notably lowering computational demands, thus ensuring high-quality visuals [103]. Another use of diffusion models in this field includes crafting counterfactual images for anomaly detection in brain scans. By merging DDPMs with Denoising Diffusion Implicit Models (DDIM), researchers have been able to alter pathological areas while keeping the normal structures intact [104]. Additionally, latent diffusion models facilitate large-scale generation of brain MRIs, enabling the synthesis of realistic high-resolution brain images with adjustable features like age and sex. The development of synthetic datasets, such as an openly available collection of 100,000 brain images, highlights the scalability and promise of diffusion models in advancing medical imaging research [105].

Despite the success of diffusion models in generating realistic data in the field of medical imaging, there are significant challenges posed in implementing these models due the lack of high quality training data. Several studies have tackled the problem of training a robust diffusion model with limited data available. In ([168]), the Discriminative Stable Diffusion (DSD) model was developed for few-shot vision and language tasks. DSD utilises pre-trained diffusion models and harnesses cross-attention scores to conduct discriminative tasks, such as image-text matching. This research demonstrated the flexibility of diffusion models to adapt to few-shot learning environments through fine-tuning with attention-based prompt learning. The success of DSD in these tasks underscores the potential of diffusion models to adjust and perform under constrained data conditions. Similarly, ([162]) addressed the over-fitting challenges commonly faced in few-shot image generation. The study proposed that pre-training diffusion models along with feature embeddings as a conditional input from a vision transformer could help the models learn and generalise over limited data and enhanced its generative properties.

7.3 Proposed Method

Standard diffusion models operate by defining a Markov chain of diffusion steps that sequentially introduce random Gaussian noise into the observed data, and then learn to reverse the noise process and reconstruct new samples from the introduced noise using a deep neural network. Although this method is effective, standard diffusion models often become computationally burdensome when handling high-dimensional data, such as the $512 \times 512 \times 100$ image space in our study.

To address this challenge, we develop a latent diffusion model (LDM), which consists of a pretrained autoencoder and a diffusion model. In a typical LDM setup, a variational autoencoder (VAE) is used to learn a compressed, lower-dimensional latent representation of the input data. This allows the diffusion model to focus more efficiently on modeling the high-level semantic features within this reduced latent space, enabling the generation of high-quality medical images with reduced computational demands. This approach not only enhances the efficiency of the image generation process but also maintains the quality of the generated images, crucial for medical applications.

However, VAEs often encounter difficulties in compressing intricate images, such as those obtained from brain angiography, without losing critical details. Vascular structures, which occupy a minimal portion of the overall image, are particularly challenging to reconstruct accurately. To mitigate this issue, we employ a pre-trained multi-task depth autoencoder that selectively reduces the input data across the channel dimension. This approach is designed to preserve essential vascular details, thereby facilitating the generation of more accurate latent spaces for training the diffusion model.

7.3.1 Multitask Depth Autoencoder

In our study, the effectiveness of the latent diffusion model hinges on the quality of the autoencoder utilised during training. Given the direct relationship between the performance of the diffusion model and the autoencoder, we utilise a multi-task depth autoencoder tailored to compress the image stack of MRA scans into a single-channel representation while preserving the spatial dimensions (x- and y-axes). This sort of compression is needed in order to ensure a faithful and complete reconstruction of the vessels, as standard compression using max-pooling layers makes reconstruction of the vessels much more challenging. Given the importance of retaining vessel information that accounts for less than 5% of the MRA image, we introduce a dual-branch ar-

chitecture comprising an output decoding branch and a segmentation branch. The decoding branch facilitates the faithful reconstruction of the input image stack, while the segmentation branch facilitates the delineation of vessel structures within the MRA.

To effectively train this model, we employ multi-task learning (MTL), leveraging the inherent complementarity of the reconstruction and segmentation tasks. Previous research has demonstrated the efficacy of MTL approaches in enhancing model performance across multiple tasks. We tested our model with various multi-task learning (MTL) approaches: Nash-MTL ([133]); average Dice after evaluation 0.76), CAGrad ([134]; average Dice after evaluation 0.74), and uncertainty-based MTL ([3]; average Dice after evaluation 0.79). The best performing version was the uncertainty-based MTL, where both the losses are weighted based on the assumption of homoscedastic uncertainty for each task. The loss function for our multi-output model is described in (7.1), where W are the model parameters and we interpret minimising the loss with respect to σ_1 and σ_2 as learning the relative weights for the losses $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{rec}}$ adaptively. We used Dice score as the loss for $\mathcal{L}_{\text{seg}}$ and MAE as the loss for $\mathcal{L}_{\text{rec}}$.

$$\mathcal{L}_{\text{Total}} = \frac{1}{2\sigma_1^2}\mathcal{L}_{\text{seg}}(\mathbf{W}) + \frac{1}{2\sigma_2^2}\mathcal{L}_{\text{loc}}(\mathbf{W}) + \log \sigma_1\sigma_2 \tag{7.1}$$

Additionally, to ensure that the channel-wise information was not lost during the compression stage we introduce spatial attention followed by squeeze-and-excitation (SE) blocks in every layer. By allowing the network to adaptively recalibrate channel-wise feature responses, SE blocks improve the model's representational power without introducing a significant increase in computational cost or model complexity. An overview of the network architecture is given in Fig. 7.2.

In this work, we use a latent diffusion model (LDM) comprising a pretrained autoencoder and a diffusion model. The autoencoder learns a lower-dimensional latent embedding of the brain vasculature, while the diffusion model focusses on modelling the high-level semantic representations in the latent space efficiently. Following ([140]), the diffusion process can be defined as forward and reverse Markov chains, where the forward process iteratively transforms the data x_0 (i.e. the latent features from the autoencoder in our approach) into a standard Gaussian X_T as following:

$$\begin{aligned} q\left(\mathbf{x}_{1:T}|\mathbf{x}_0\right) &= \prod_{t=1}^{T} q\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right), q\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right) \\ &:= \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right) \end{aligned}$$

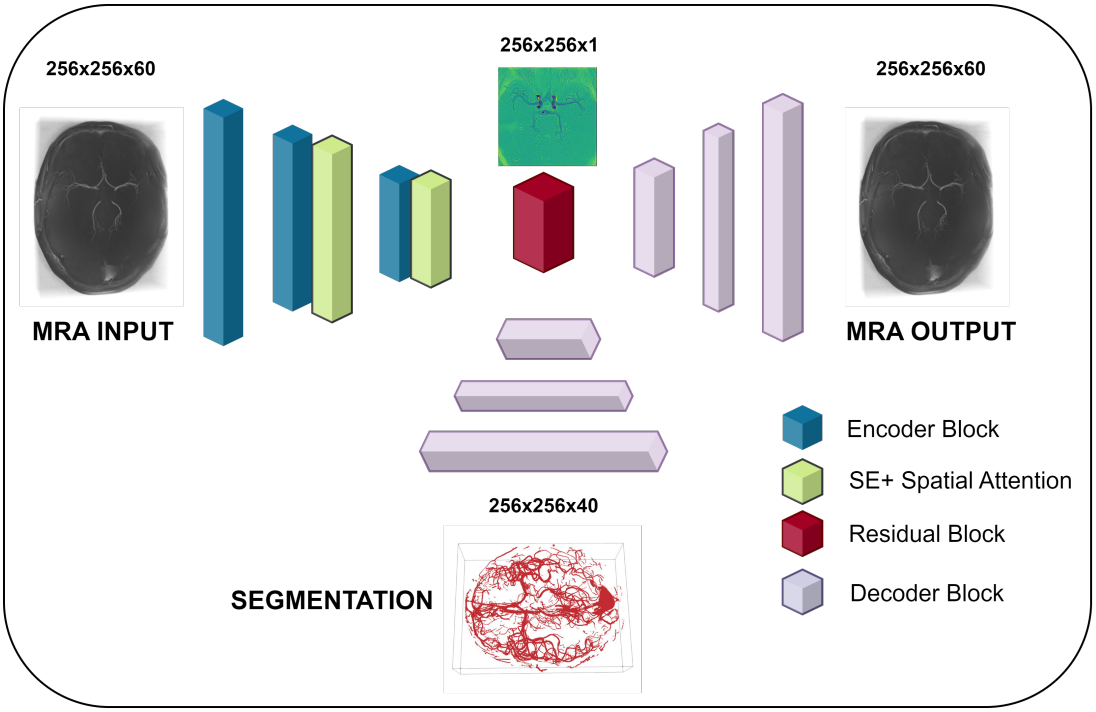

Figure 7.2: Architecture of our multi-task autoencoder

where $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ is the transition probability at the time step t based on the noise schedule β_t. Therefore, the noisy data $\mathbf{x}_t$ can be formulated as

$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$,

where $\alpha_t := 1 - \beta_t, \bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

The reverse process, achieved via a deep neural network parameterised by θ, can then be defined as:

$$p_\theta(\mathbf{x}_0|\mathbf{x}_T) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$
$$:= \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t))$$

The simplified evidence lower bound (ELBO) loss to optimise the diffusion model ([140]) can be formulated as a score-matching task where the neural network predicts the actual noise ϵ added to the observed data. The resulting loss function is

$$\mathcal{L}_\theta := \mathbb{E}_{\mathbf{x}_0,t,C,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(x_t, t, C)\|^2\right]$$

where C is the condition in conditional generation.

The encoder of the pre-trained depth auto-encoder transforms the brain image K_0 into a compact latent representation x_0 with dimensions of $256 \times 256 \times 1$. Once the lower dimensional latent spaces are generated, the latent representations from the training set serve as inputs to the diffusion model for further analysis and generation.

We employ a model with a U-net-based architecture as the diffusion model. Our model has five encoding blocks and five decoding blocks, with skip connections between the corresponding encoding and decoding blocks. We replace the simple convolution layers in the encoding and decoding blocks with a residual block followed by a multi-head attention layer to limit information loss in the latent space. Each encoding and decoding block takes the class category (based on CoW phenotypes) as an additional conditional input, while only the decoding blocks take shape and anatomy features as additional conditional inputs.

7.3.2 Anatomy Conditioning

Angiographic images can capture the small vessels in the peripheral cerebral vasculature. Preserving anatomical integrity is therefore crucial in the generation of realistic and accurately depicted vessels. However, diffusion models often face challenges in faithfully representing the anatomical structure, which can be attributed to their learning and sampling processes that are heavily based on probability density functions ([147]). Previous studies have demonstrated that the inclusion of certain priors, such as geometric and shape priors, can improve their performance in medical image synthesis ([149, 150]). Additionally, generative models based on diffusion techniques inherently depend on large volumes of data to train effectively and yield convincing results. This dependency poses a significant challenge in our study due to the limited availability and high dimensionality of the data. The scarcity of the dataset exacerbates the difficulty for generative models to accurately approximate the data distribution. Moreover, despite phenotypic variations, the CoW generally maintains a consistent structural configuration. This uniformity, coupled with data scarcity, predisposes the model to predominantly learn the average shape of the CoW, thereby diminishing the modelâ€™s variance.

To overcome these limitations, we incorporate transformer-based class features to inform the diffusion process. Specifically, we utilise a set-based Vision Transformer (ViT) model ([162]), which is tailored to process entire 3D volumes and initially serves

as a classifier to distinguish between the three classes. After the ViT is successfully trained, its final classification layer is removed. We then pass the images from each class through this transformer to derive class-specific encoded features. These features, along with class conditioning variables, are integrated into the conditioning vector of the diffusion model, significantly enhancing its capability to produce outputs that accurately represent each class.

7.3.3 Data and Implementation Details

We trained all models on the publicly available IXI data set ([122]) using the 181 3T MRA scans acquired at Hammersmith Hospital, London. The images were centered, cropped from $512 \times 512 \times 100$ to $256 \times 256 \times 100$, and the intensity normalized. We then used a residual U-net [146] to extract vessel segmentations from the MRA. The authors manually labeled each case with the presence / absence of one or both peripheral communication arteries in the CoW. Class 1 includes cases where both PComA's are present, Class 2 includes cases with only one PComA is present, while Class 3 includes cases where both PComA's are absent.

All models were implemented in TensorFlow 2.8 and Python 3. For the forward diffusion process, we used a linear noise schedule with 1000 time steps. The model was trained for 2000 epochs with a learning rate of 0.0005 on a Nvidia Tesla T4 GPU and 38 GB of RAM with Adam optimiser. After the images were generated by our latent diffusion model, a post-processing step of thresholding and finding the largest connected component was implemented to refine the output.

7.4 Experiments and Discussion

To assess the performance of our model, we compared it against three established generative models: 3D C-VAE ([33]), a 3D WGAN ([65]), and a Diffusion model (DDPM) [147]. We use the FID score to measure the perceptual image realism of the generated vasculature. Typically, for medical image models, FID scores are computed using pre-trained features derived from medical datasets [169]. However, recent investigations have revealed that networks like InceptionV3, pre-trained on more general datasets such as Inception, provide a more stable metric for assessing perceptual image realism [170]. Consequently, for our FID calculations, we employed a pre-trained InceptionV3

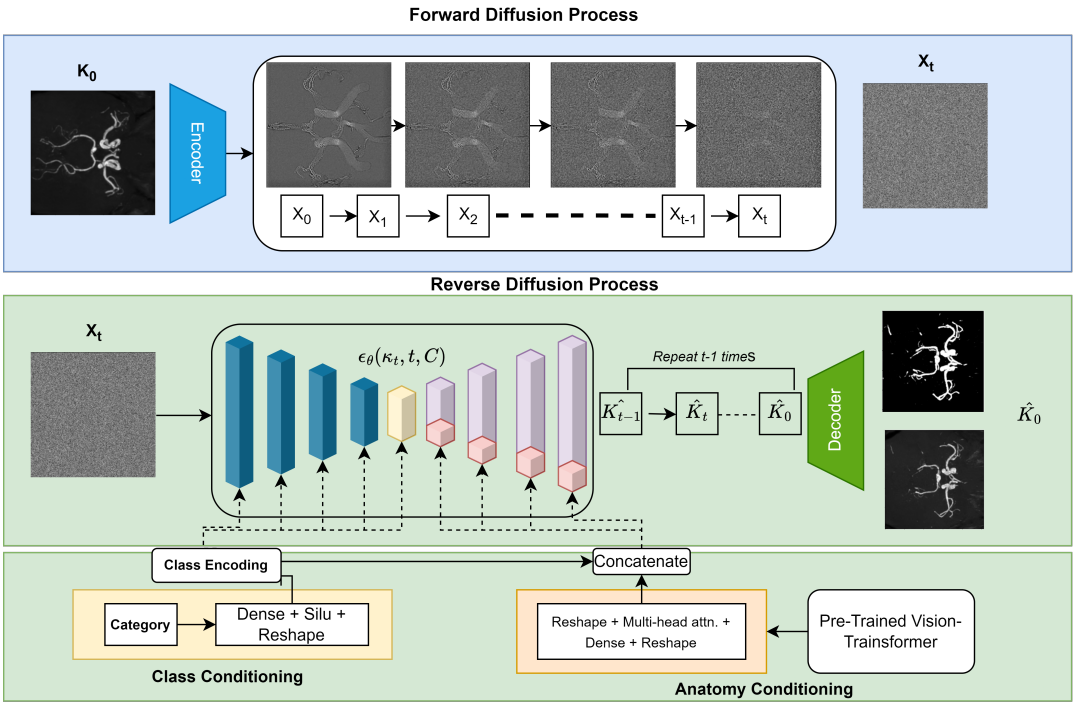

Figure 7.3: The architecture of the latent diffusion model begins with the forward diffusion process, depicted in the first panel, wherein noise is incrementally introduced to the input image. The second panel illustrates the reverse diffusion process, during which the network acquires the ability to predict the noise added at each timestep, thus learning the reverse procedure. Additionally, two conditioning variables are employed: class conditioning and anatomy conditioning. Class conditioning specifies the category of the image that the network is tasked with generating. For anatomy conditioning, a Vision Transformer (ViT) is initially trained to classify different categories of images. Subsequently, the final classification layer is removed to extract class-specific features, which are then input to the network.

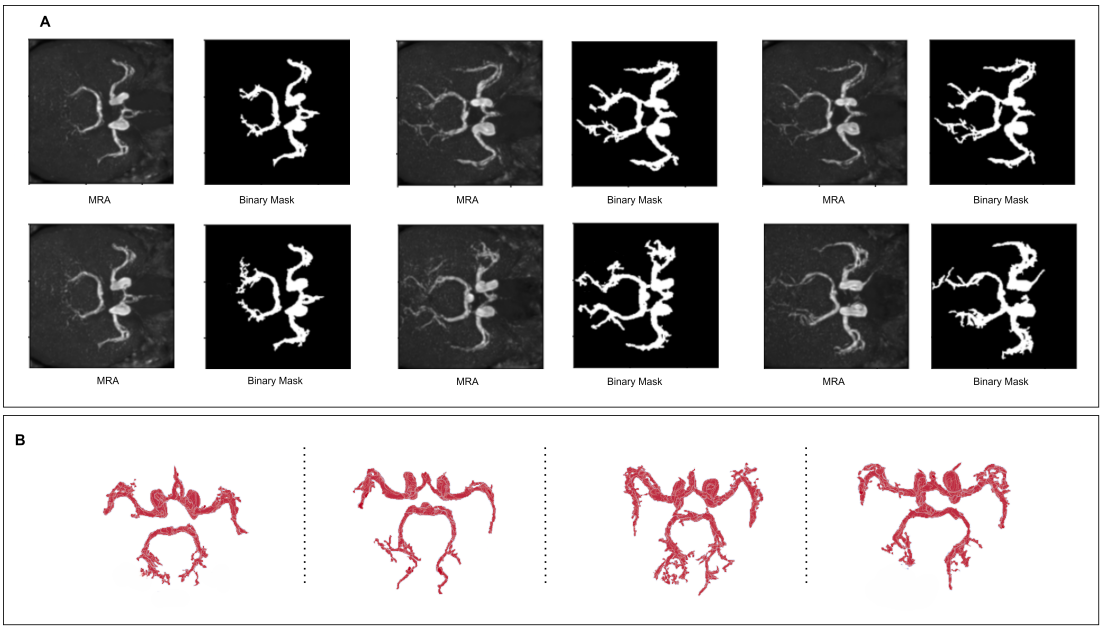

Figure 7.4: [A] Maximum intensity projections of six generated paired MRA and binary segmentation samples from our model. [B] A generated 3D mesh of four additional samples to highlight the continuity and accuracy of the generated vessels.

network as a feature extractor. A lower FID score indicates higher perceptual image quality. In addition, we used MS-SSIM and 4-G-R SSIM to measure the quality of the generated images ([155, 156]). MS-SSIM and 4-G-R SSIM are commonly used to assess the quality of synthesised images. Typically, a higher score is indicative of better image quality, implying a closer resemblance between the synthesised CoW and the ground truth reference, In the realm of image synthesis, a higher score might suggest that the diversity among generated samples is limited. This occurs when the outputs closely align with the actual data distribution, possibly indicating that the model is either producing images that are very close to the dataset's average or replicating images already present within the dataset. In order to compute the MS-SSIM and 4-G-R SSIM scores, we generated 20 synthesized CoW instances for each model. For each synthesized instance, we calculated the MS-SSIM and 4-G-R SSIM scores by comparing against every ground truth data sample. The closest score for each instance was recorded, and finally, we averaged these scores to yield the final score for the model. Table 7.1 presents the evaluation scores achieved by our model, 3D CVAE, WGAN and DDPM in the generation of the MRA. Our model achieved a better FID score than the other models, suggesting that the distribution of CoW variants synthesised by our model is closer to that observed in real CoW data, compared to the other models. Additionally, our model in general achieves higher or comparable MS-SSIM and 4-G-R SSIM scores compared to the other methods. These higher scores indicate better image quality, implying that the generated CoW samples resemble the real CoW images more closely. Similarly 7.2 compares qualitatively the generated binary mask from each of the models across the same metric. These results are similar to results observed from 7.1 apart from the performance of the DDPM which perfrorms much worse in the vessel synthesis task. We also performed a quantitative evaluation of the class-wise results of our model in 7.3, which indicates that the performance of the model in generating cases from class 1 and 2 is better than for generating 3. This could stem from the fact that class 3 has the lowest number of cases in the training data.

Fig. 7.5 provides a qualitative comparison among the samples generated obtained from the three models to provide additional context to the quantitative results presented in 7.1 and 7.2. As the output of each model is a 3D vascular structure, maximum intensity projections (MIPs) over the Z-axis which condense the volumetric representation into a 2D plane are used to visually compare the synthesised images. Panel A of

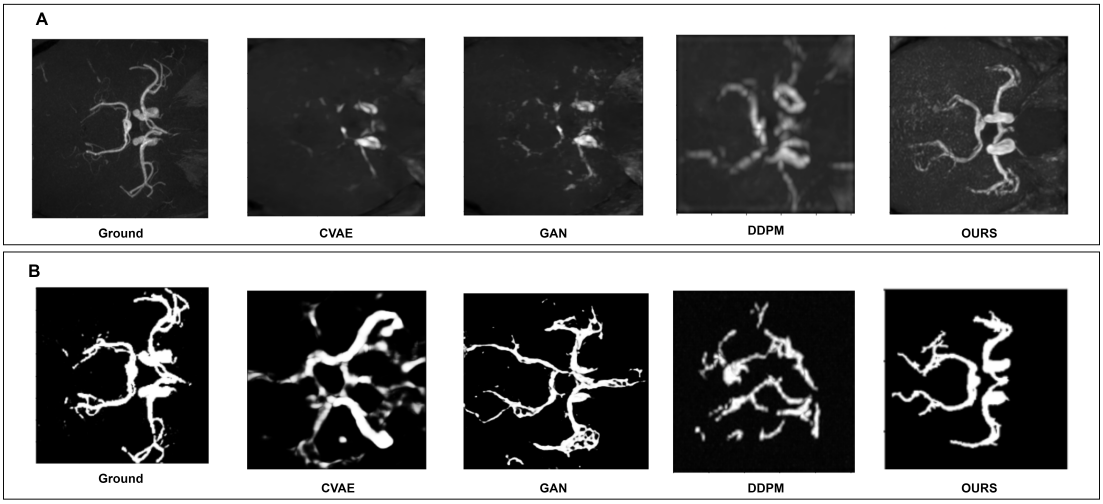

Figure 7.5: [A] Comparison of a MRA generated by our model against established generative models such as VAE, GAN (pix2pix) and a DDPM. The qualitative comparison shows that our mode ourperforms other standard generative models in this task. [B] Comparison of the binary mask generated by our model against established generative models such as VAE, GAN (pix2pix) and a DDPM. The qualitative comparison shows that our mode ourperforms other standard generative models in this task.

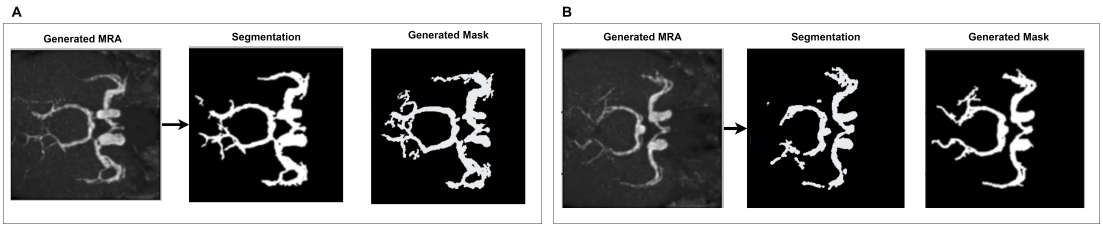

Figure 7.6: Illustration of two scenarios in which the generated MRA image is processed by a pre-trained vessel segmentation network, and the resulting segmentations are compared to the generated binary masks. Panel A depicts a scenario where the segmentation closely approximates the generated binary mask, demonstrating a high degree of accuracy. Panel B presents a scenario where the segmentation deviates slightly from the generated binary mask, evidenced by visible discontinuities in the vessels and some vessels not being segmented. Nevertheless, in both scenarios the major vessels are distinctly captured, and the segmentation outputs remain closely aligned with the generated binary masks.

Fig. 7.5 offers a qualitative evaluation of the MIP derived from the MRA created by our model compared to those from other models like 3D-VAE, pix2pix, and DDPM. Clearly, neither the GAN nor the VAE model can precisely generate the Circle of Willis (CoW) branches, apart from the main arteries like the Middle Cerebral Artery (MCA). This shortcoming is likely due to the lack of an attention mechanism in these models, which is vital for emphasizing the smaller vessels that make up a lesser portion of the MRA images. On the other hand, the DDPM model performs better in reconstructing most CoW branches but still struggles with producing continuous vessels, as evident from the structured discontinuities. Our model, however, excels at rendering a more accurate and high-quality depiction of the CoW. The vessels in our generated images, as seen in the MIP, exhibit markedly better continuity, marking a significant advancement over both conventional and contemporary generative strategies used by the compared models. Panel B of Fig. 7.5 conducts a similar comparison for the generated vessel masks of each model. The VAE can generate some major CoW vessels, but the smaller vessels are inaccurately produced, and the vessels exhibit significant discontinuity. Conversely, the GAN generates a very lifelike vessel mask, capable of synthesizing both larger and smaller vessels, though the smaller vessels'structure may not always appear credible. The Diffusion model, however, fails to produce plausible vessel masks altogether. In contrast, our model successfully generates a realistic and believable Circle of Willis with minimal or no discontinuities in the vessels.

We conduct a more comprehensive qualitative evaluation of the images produced by our model, as illustrated in Fig. 7.4. Panel A of this figure displays the MIP representations of selected samples from both MRA data and the corresponding segmentation obtained through our proposed method. This qualitative assessment reveals that our model effectively recreates the entire CoW, with all primary arteries distinctly visible. Nonetheless, it is essential to mention that the model occasionally generates minor artifacts, especially in smaller vessels located in the posterior communicating area. Recognizing that MIPs are 2D projections of a 3D image and may not always be optimal for visualizing the generated Circle of Willis, we further convert the vessel mask produced by our model into a 3D mesh, represented in Panel B of Fig. 7.4. This 3D mesh demonstrates that our model is capable of generating continuous and realistic variations of the Circle of Willis.

Through qualitative and quantitative experiments, it becomes clear that a disparity

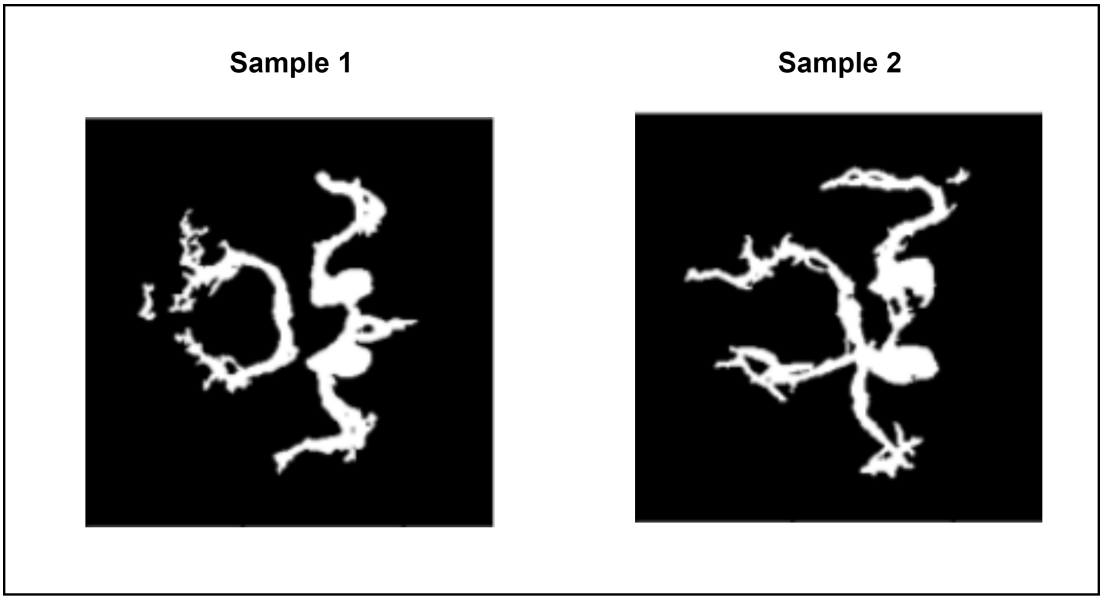

Figure 7.7: Comparison of a generated image that is close to mean of the real MRA image distribution (Sample 1) and a sample image that is furthest away from the mean of the real MRA image distribution (Sample 2)

exists between qualitative and quantitative outcomes. Evaluating diffusion models (or generative models in general) quantitatively remains challenging due to the absence of an objective ground truth. Although this field is active with various proposed evaluation approaches to gauge the visual realism of generated images, there is still a lack of established methods for assessing generative medical models where anatomical accuracy alongside visual realism is crucial ([170–172]). This gap is notably apparent in our case, particularly in the evaluation of the generated MRA. Despite the FID and MS-SSIM scores for DDPM and GAN being nearly identical to those of our model, qualitative assessments clearly reveal that our model significantly outperforms GAN and DDPM. A similar trend is observed in the evaluation of the binary mask – in addition to GAN and DDPM achieving scores comparable to our model, the VAE actually attains a much higher MS-SSIM score than our model, indicating that the binary mask generated by the VAE is closer to the real data compared to our model. However, this result contradicts the qualitative comparison, where our model evidently resembles the real data more closely with higher anatomical accuracy.

While quantitatively assessing the generated data poses difficulties, evaluating the

performance of the binary mask i srelatively straightforward by examining the Maximum Intensity Projection (MIP). This method allows for easy identification of continuity and anatomical correctness of the major vessels (such as MCA and PCA). In contrast, quantitatively evaluating the MRA is more complex due to the presence of soft tissue information in addition to the vessels. This soft tissue information qualitatively resembles noise, making it challenging to determine how closely the generated MRA matches the real one. However, a segmentation algorithm trained to extract vessels from real MRA images considers the latent soft tissue information to produce accurate segmentations. To further assess the quality of the generated MRA, we conducted an additional experiment where the generated MRA was input into a brain MRA segmentation algorithm ([108]) pre-trained on real MRA images to extract vessels. We then compared the segmentation algorithm's output to the paired generated binary mask of the MRA. If the segmentation and the generated binary mask match closely, it indicates that the generated MRA contains similar latent information to that of the real MRA, which enables the segmentation algorithms to work effectively. The results of this experiment are shown qualitatively in Fig. 7.6. In Panel A, the segmentation from the generated MRA closely matches the generated binary mask, with some areas even showing improved vessel connectivity. In Panel B, the segmentation is less accurate than the generated binary mask, displaying evident vessel discontinuities and missing smaller posterior communication arteries. Nonetheless, in both cases, most major arteries are segmented correctly, demonstrating good correlation with the generated binary mask. This provides further evidence that the generated MRA closely resembles the real MRA.

We also conducted an additional experiment to assess the quality of the long-tail generated samples by comparing two distinct images: one generated sample that closely aligns with the mean of the real data distribution and another that is the furthest from the mean. The image near the mean closely resembles the real data, successfully generating all major arteries but exhibiting minimal variance. In contrast, the long-tail image, while generating a plausible-looking CoW, incorrectly positions some arteries in a manner that may not be anatomically plausible. This experiment highlights the challenges in balancing the accuracy and diversity of generated samples in our model.

The results presented demonstrate the performance of our model in generating realistic variations of the CoW. Particularly notable is the model's proficiency in producing

Table 7.1: Quantitative comparison between generated binary mask for the CoW

Model	FID ↓	MS-SSIM ↑	4-G-R SSIM ↑
3D CVAE	8.34	0.71	0.14
Patch-GAN (pix2pix)	2.19	0.31	0.6
DDPM	2.50	0.66	0.11
Ours	1.16	0.64	0.40

Table 7.2: Quantitative evaluation of the generated MRA

Model	FID ↓	MS-SSIM ↑	4-G-R SSIM ↑
3D CVAE	1.47	0.71	0.51
Patch-GAN (pix2pix)	3.417	0.42	0.2
DDPM	11.10	0.31	0.17
Ours	4.58	0.56	0.41

Table 7.3: Quantitative class-wise evaluation of Generated CoW vasculature

Class	FID Score ↓	MS-SSIM ↑	4-G-R SSIM ↑
Class 1	4.41	0.65	0.65
Class 2	3.88	0.52	0.52
Class 3	7.63	0.41	0.41

accurate representations for classes 1 and 2, surpassing its performance in class 3 due to the limited sample size of the latter. Our model excels in synthesising the anterior circulation and the middle cerebral arteries, showing remarkable fidelity to anatomical structures. However, it faces challenges in effectively generating continuous representations of the posterior circulation. Further investigation and refinement may be required to enhance the model's ability in this specific aspect. Furthermore, the addition of further constraints or features may be necessary to prevent the model from occasionally generating vessels that appear implausible. Incorporating these enhancements could help improve the modelâ€™s fidelity and ensure the anatomical accuracy of the generated vascular structures.

7.5 Conclusion

We proposed a latent diffusion model that used shape and anatomy guidance to generate realistic CoW configurations. Quantitative qualitative results showed that our model outperformed existing generative models based on a conditional 3D GAN and a 3D VAE. Future work will look to enhance the model to capture wider anatomical variability and improve synthetic image quality.

Acknowledgments

A.F.F. acknowledges support from the Royal Academy of Engineering under the RAEng Chair in Emerging Technologies (INSILEX CiET1919/19) and ERC Advanced Grant â€“ UKRI Frontier Research Guarantee (INSILICO EP/Y030494/1) schemes. The information contained herein reflects only the authors view, and none of the funders or the European Commission are responsible for any use that may be made of it.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

This thesis explored the development and evaluation of advanced generative models for synthesizing brain vasculature, focusing on the Circle of Willis (CoW) and cerebral aneurysms. We implemented a variety of techniques, such as shape-guided conditional latent diffusion models, multitask learning approaches, few-shot learning for aneurysm generation, and the synthesis of paired magnetic resonance angiography (MRA) and binary segmentation masks.

8.1.1 Key Findings and Contributions

- **Synthesizing Vascular Segmentation from T2-Weighted MRI**: In Chapter 4, we proposed a multitask learning-based encoder-decoder model designed to synthesize CoW vessel segmentation from T2-weighted MRI using localized attention maps. This work introduced an more parameter efficient segmentation approach that eliminates the need for multiple input modalities and optimizes parameter utilization by focusing on specific regions within the MRI images. The model's capability to produce precise CoW segmentations with fewer parameters than its competitors underscores its significance for both clinical and research settings. This is particularly valuable when vascular data from an MRA is crucial but not accessible, yet other non-contrast imaging modalities like T2 are available.

- **Shape-Guided Conditional Latent Diffusion Models**: In Chapter 5, we introduced a shape-guided conditional latent diffusion model, designed to synthesize realistic brain vasculature with emphasis on CoW variations. The inclusion of shape moments and PCA for anatomical guidance was crucial in ensuring the generated vessels exhibited continuity and anatomical accuracy. The use of diffusion models to generate both coarse and fine vascular structures advanced the field of synthetic vascular imaging by allowing for more anatomically realistic outputs.

- **Few-Shot Learning for Aneurysm Generation**: Also in Chapter 5, we presented a novel approach to generating cerebral aneurysm images using limited data. By integrating a transformer-based approach to extract class-specific features and employing signed distance functions (SDF) as shape-based conditions, we were able to synthesize aneurysms with high anatomical fidelity. This

method addressed the challenges posed by the scarcity of labeled data and offered significant improvements in generating rare aneurysm phenotypes.

- **Paired MRA and Segmentation Generation**: A significant advancement in this work was the development of a model capable of generating both MRA and segmentation masks, allowing for the synthesis of complete vascular structures. Utilizing multitask learning, the model demonstrated strong performance in generating high-quality MRAs while maintaining vessel integrity in segmentation outputs. This represents a major step forward in medical imaging synthesis, offering the ability to synthesize both images and corresponding segmentations from a single generative process.

8.1.2 Challenges and Limitations

Several challenges and limitations were encountered during this research. The primary limitation of the segmentation model in Chapter 4 lies in its difficulty in generating finer vessels, especially in the posterior communicating arteries. This may be due to the limited resolution of the input T2 data or the model's inherent limitations in capturing complex vascular features from a single MRI modality.

In the context of Circle of Willis (CoW) generation, although our models demonstrate the capability to produce realistic CoW architectures utilizing conditional anatomical and shape guidance techniques, there remains potential for enhancement. The current models exhibit limitations in accurately capturing the intricate anatomical structures and occasionally yield implausible vascular topologies. This shortcoming is primarily attributed to the absence of a physics-based framework that could assist the model in ascertaining the feasibility of a given vascular structure.

In aneurysm generation, the scarcity of annotated cases remained a challenge. Although the few-shot learning method showed promise, the generated aneurysm images sometimes exhibited artifacts due to the limited data available for rarer aneurysm phenotypes. The use of signed distance functions (SDF) improved vessel continuity, but further refinement is necessary to fully address anatomical inaccuracies in certain cases.

Overall , the biggest challenge that we face with the generative models is to find a good quantitative reference metric to evaluate the outputs of the model. It is naturally more challenging to evaluate the outputs of generative models due to the lack of ground

truth reference , and while there is some promising research in this field with several approachces that try to measure how 'real' the generative images looks [170–172], there is still limited research on how to evaluate generative medical models . The most effective way as of now to evalatute the output of generative medical models still seems to be validations from a clinician or a medical professional.

8.1.3 Future Directions

This work has paved the way for several future research directions:

- **Expanding Data Modalities**: Future research could focus on incorporating additional MRI modalities, such as FLAIR or diffusion-weighted imaging, to capture more detailed vascular structures. Combining multiple modalities could provide richer data for training, allowing for more accurate generation of vascular images.
- **3D Network Architectures**: One natural extension is the implementation of fully 3D generative models to better capture complex vascular structures in three dimensions, improving continuity and anatomical correctness.
- **Clinical Applications and Validation**: Moving forward, it is important to validate the models in clinical settings, focusing on diagnostic applications. Collaborating with clinicians could help refine the models further and ensure that the generated images are relevant and useful in real-world medical scenarios.
- **Refining Anatomical Guidance**: While the transformer-based approach for anatomical guidance proved beneficial, further refinements and the exploration of other feature extraction techniques could yield even more anatomically faithful results.
- **Evaluation Metrics**: As discussed earlier , evalaution of generative medical images is an open challenge and an area of active research. It would be very interesting to explore some registration / physics based evaluation criterion that have some prior information about the anatomy to act as a metric for evaluation of these models.

8.1.4 Final Thoughts

This thesis introduced several novel methods for generating brain vasculature, particularly focusing on the Circle of Willis (CoW) and cerebral aneurysms from MRI data. These contributions address key challenges in medical imaging and open new avenues for clinical and research applications.

A key achievement of this work is the ability to synthesize accurate vascular segmentations from T2-weighted MRI, making vascular analysis accessible even in settings without MRA. The generative models developed, particularly the shape-guided latent diffusion models and few-shot learning techniques, enable the creation of large synthetic datasets, which can improve model training for tasks such as aneurysm detection and vascular segmentation in data-scarce environments. Furthermore, the paired generation of MRA images and segmentation masks facilitates deeper analysis of the relationship between brain anatomy and cerebrovascular conditions, allowing for a more precise study of CoW configurations and their impact on stroke risk and other vascular diseases. The ability to generate aneurysm images for rare phenotypes also supports the development of better diagnostic and intervention strategies.

The research also significantly reduces the burden of manual segmentation by offering accurate, automated solutions that streamline the analysis of large imaging datasets. This innovation enables large-scale population studies and virtual clinical trials, expanding the possibilities for understanding vascular phenotypes and their correlation with neurological conditions. Standardized synthetic datasets created by these models also allow for benchmarking segmentation algorithms, improving diagnostic accuracy and data accessibility.

In conclusion, this thesis represents a step forward in medical image synthesis, providing tools that enhance the accessibility, accuracy, and breadth of vascular research and diagnosis. By bridging the gap between advanced imaging techniques and commonly available MRI modalities, this work has the potential to transform cerebrovascular research, enabling more effective diagnosis and treatment development for vascular diseases.

References

[1] T. N. Page, Circle of willis or circulus arteriosus (2023-09-30)

[2] D. Li, D. A. Dharmawan, B. P. Ng and S. Rahardja, Residual u-net for retinal vessel segmentation, in *2019 IEEE International Conference on Image Processing (ICIP)* (2019), pp. 1425–1429

[3] A. Kendall, Y. Gal and R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in *Proc. IEEE Conf. Comput. Vision Pattern Recog.* (2018), pp. 7482–7491

[4] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao et al., Image data augmentation for deep learning: A survey, *arXiv preprint arXiv:2204.08610* (2022)

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

[6] A. C. Society, Key statistics for brain and spinal cord tumors (2024)

[7] T. Wang, Y. Lei, Y. Fu, J. F. Wynne, W. J. Curran et al., A review on medical imaging synthesis using deep learning and its clinical applications, *Journal of applied clinical medical physics* **22**, 11 (2021)

[8] C. western reserve university, Mri (2023-09-30)

[9] MRI Master, T1 vs t2 mri - understanding the difference, https://mrimaster.com/t1-vs-t2-mri/ (2024)

[10] Z. Guangyu, Q. Yuan, J. Yang and J. Yeo, Experimental study of hemodynamics in the circle of willis, *BioMedical Engineering OnLine* **14 Suppl 1**, S10 (2015)

[11] Radiopaedia contributors, Circle of willis, https://radiopaedia.org/articles/circle-of-willis?lang=gb (2024), accessed: 2024-09-20

[12] A. F. Van Raamt, W. P. Mali, P. J. Van Laar and Y. Van Der Graaf, The fetal variant of the circle of willis and its influence on the cerebral collateral circulation, *Cerebrovascular diseases* **22**, 217 (2006)

[13] E. Fliers, M. Korbonits and J. Romijn, eds., *Handbook of Clinical Neurology Vol.124*, Elsevier (2014)

[14] N. Chalouhi, B. L. Hoh and D. Hasan, Review of cerebral aneurysm formation, growth, and rupture, *Stroke* **44**, 3613 (2013)

[15] J. Van Gijn, R. S. Kerr and G. J. Rinkel, Subarachnoid haemorrhage, *The Lancet* **369**, 306 (2007)

[16] M. H. Vlak, A. Algra, R. Brandenburg and G. J. Rinkel, Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis, *The Lancet Neurology* **10**, 626 (2011)

[17] A. M. Sailer, B. A. Wagemans, P. J. Nelemans, R. de Graaf and W. H. van Zwam, Diagnosing intracranial aneurysms with mr angiography: systematic review and meta-analysis, *Stroke* **45**, 119 (2014)

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

[19] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *nature* **521**, 436 (2015)

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

[21] C. Cortes, Support-vector networks, *Machine Learning* (1995)

[22] 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**, 1929 (2014)

[23] V. Nair and G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in *Proceedings of the 27th international conference on machine learning (ICML-10)* (2010), pp. 807–814

[24] 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**, 6 (1986)

[25] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural computation* **9**, 1735 (1997)

[26] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, *arXiv preprint arXiv:1412.3555* (2014)

[27] D. Bahdanau, K. Cho and Y. Bengio, Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:1409.0473* (2014)

[28] G. Yenduri, R. M, C. S. G, S. Y, G. Srivastava et al., Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions (2023)

[29] G. Yenduri, M. Ramalingam, G. C. Selvi, Y. Supriya, G. Srivastava et al., Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions, *IEEE Access* (2024)

[30] S. Jamil, M. Jalil Piran and O.-J. Kwon, A comprehensive survey of transformers for computer vision, *Drones* **7**, 287 (2023)

[31] S. Latif, A. Zaidi, H. Cuayahuitl, F. Shamshad, M. Shoukat et al., Transformers in speech processing: A survey, *arXiv preprint arXiv:2303.11607* (2023)

[32] Z. Cang and G.-W. Wei, Topologynet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions, *PLoS computational biology* **13**, e1005690 (2017)

[33] D. Kingma and M. Welling, Auto-encoding variational Bayes, *arXiv preprint arXiv:1312.6114* (2013)

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley et al., Generative adversarial networks, *Commun. ACM* **63(11)**, 139 (2020)

[35] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, High-resolution image synthesis with latent diffusion models (2022)

[36] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang et al., Photorealistic text-to-image diffusion models with deep language understanding (2022)

[37] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte et al., Repaint: Inpainting using denoising diffusion probabilistic models (2022)

[38] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz et al., Diffusion models for medical image analysis: A comprehensive survey (2023)

[39] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp et al., Medgan: Medical image translation using gans, *Computerized medical imaging and graphics* **79**, 101684 (2020)

[40] B. Zhan, D. Li, Y. Wang, Z. Ma, X. Wu et al., Lr-cgan: Latent representation based conditional generative adversarial network for multi-modality mri synthesis, *Biomedical Signal Processing and Control* **66**, 102457 (2021)

[41] O. Ronneberger, P. Fischer and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer (2015), pp. 234–241

[42] X. Han, Mr-based synthetic ct generation using a deep convolutional neural network method, *Medical physics* **44**, 1408 (2017)

[43] F. Liu, H. Jang, R. Kijowski, T. Bradshaw and A. B. McMillan, Deep learning mr imaging–based attenuation correction for pet/mr imaging, *Radiology* **286**, 676 (2018)

[44] H. Jang, F. Liu, G. Zhao, T. Bradshaw and A. B. McMillan, Deep learning based mrac using rapid ultrashort echo time imaging, *Medical physics* **45**, 3697 (2018)

[45] X. Dong, T. Wang, Y. Lei, K. Higgins, T. Liu et al., Synthetic ct generation from non-attenuation corrected pet images for whole-body pet imaging, *Physics in Medicine & Biology* **64**, 215016 (2019)

[46] D. Hwang, K. Y. Kim, S. K. Kang, S. Seo, J. C. Paeng et al., Improving the accuracy of simultaneously reconstructed activity and attenuation maps using deep learning, *Journal of Nuclear Medicine* **59**, 1624 (2018)

[47] S. Neppl, G. Landry, C. Kurz, D. C. Hansen, B. Hoyle et al., Evaluation of proton and photon dose distributions recalculated on 2d and 3d unet-generated pseudocts from t1-weighted mr head scans, *Acta Oncologica* **58**, 1429 (2019)

[48] A. Torrado-Carvajal, J. Vera-Olmos, D. Izquierdo-Garcia, O. A. Catalano, M. A. Morales et al., Dixon-vibe deep learning (divide) pseudo-ct synthesis for pelvis pet/mr attenuation correction, *Journal of nuclear medicine* **60**, 429 (2019)

[49] A. P. Leynes, J. Yang, F. Wiesinger, S. S. Kaushik, D. D. Shanbhag et al., Zero-echo-time and dixon deep pseudo-ct (zedd ct): direct generation of pseudo-ct images for pelvic pet/mri attenuation correction using deep convolutional neural networks with multiparametric mri, *Journal of Nuclear Medicine* **59**, 852 (2018)

[50] L. Chen, X. Liang, C. Shen, S. Jiang and J. Wang, Synthetic ct generation from cbct images via deep learning, *Medical physics* **47**, 1115 (2020)

[51] S.-J. Son, B.-y. Park, K. Byeon and H. Park, Synthesizing diffusion tensor imaging from functional mri using fully convolutional networks, *Computers in biology and medicine* **115**, 103528 (2019)

[52] C. Ledig, L. Theis, F. HuszÃ¡r, J. Caballero, A. Cunningham et al., Photo-realistic single image super-resolution using a generative adversarial network, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 105–114

[53] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp et al., Medgan: Medical image translation using gans, *Computerized medical imaging and graphics* **79**, 101684 (2020)

[54] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean et al., Medical image synthesis with deep convolutional adversarial networks, *IEEE Transactions on Biomedical Engineering* **65**, 2720 (2018)

[55] H. Emami, M. Dong, S. P. Nejad-Davarani and C. K. Glide-Hurst, Generating synthetic cts from magnetic resonance images using generative adversarial networks, *Medical physics* **45**, 3627 (2018)

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

[57] X. Liang, L. Chen, D. Nguyen, Z. Zhou, X. Gu et al., Generating synthesized computed tomography (ct) from cone-beam computed tomography (cbct) using cyclegan for adaptive radiation therapy, *Physics in Medicine & Biology* **64**, 125002 (2019)

[58] J. Harms, Y. Lei, T. Wang, R. Zhang, J. Zhou et al., Paired cycle-gan-based image correction for quantitative cone-beam computed tomography, *Medical physics* **46**, 3998 (2019)

[59] X. Dong, Y. Lei, T. Wang, K. Higgins, T. Liu et al., Deep learning-based attenuation correction in the absence of structural information for whole-body positron emission tomography imaging, *Physics in Medicine & Biology* **65**, 055011 (2020)

[60] Y. Liu, Y. Lei, T. Wang, Y. Fu, X. Tang et al., Cbct-based synthetic ct generation using deep-attention cyclegan for pancreatic adaptive radiotherapy, *Medical physics* **47**, 2472 (2020)

[61] X. Dong, Y. Lei, S. Tian, T. Wang, P. Patel et al., Synthetic mri-aided multi-organ segmentation on male pelvic ct using cycle consistent deep attention network, *Radiotherapy and Oncology* **141**, 192 (2019)

[62] Y. Liu, Y. Lei, T. Wang, O. Kayode, S. Tian et al., Mri-based treatment planning for liver stereotactic body radiotherapy: validation of a deep learning-based synthetic ct generation method, *The British Journal of Radiology* **92**, 20190067 (2019)

[63] K. H. Kim, W.-J. Do and S.-H. Park, Improving resolution of mr images with an adversarial network incorporating images with different contrast, *Medical physics* **45**, 3120 (2018)

[64] S. Olberg, H. Zhang, W. R. Kennedy, J. Chun, V. Rodriguez et al., Synthetic ct reconstruction using a deep spatial pyramid convolutional framework for mr-only breast radiotherapy, *Medical physics* **46**, 4135 (2019)

[65] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda et al., GAN-based synthetic brain MR image generation, in *2018 IEEE 15th international Symposium on Biomedical Imaging (ISBI 2018)*, IEEE (2018), pp. 734–738

[66] J. Ouyang, K. T. Chen, E. Gong, J. Pauly and G. Zaharchuk, Ultra-low-dose pet reconstruction using generative adversarial network with feature matching and task-specific perceptual loss, *Medical physics* **46**, 3555 (2019)

[67] X. Zhang, S. Karaman and S.-F. Chang, Detecting and simulating artifacts in gan fake images, in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)* (2019), pp. 1–6

[68] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929* (2020)

[69] Z. Choudhury, B. McCane et al., Medical image synthesis using autoencoder with vision transformer, *IEEE Transactions on Image and Vision Computing* (2024)

[70] Y. Hu, S. Zhang, W. Li, J. Sun and L. Xu, Unsupervised medical image synthesis based on multi-branch attention structure, *Biomedical Signal Processing and Control* (2025)

[71] X. Zhao, Y. Du and Y. Peng, Deep learning-based multi-view projection synthesis approach for improving the quality of sparse-view cbct in image-guided radiotherapy, *Journal of Imaging Informatics in Medicine* (2025)

[72] A. Altalib, S. McGregor and C. Li, Synthetic ct image generation from cbct: A systematic review, *IEEE Transactions on Image and Plasma Medical Applications* (2025)

[73] J. Huang, T. Tan, X. Li, T. Ye and Y. Wu, Multiple attention channels aggregated network for multimodal medical image fusion, *Medical Physics* (2024)

[74] A. Brock, J. Donahue and K. Simonyan, Large scale gan training for high fidelity natural image synthesis. arxiv 2018, *arXiv preprint arXiv:1809.11096* (1809)

[75] M. Arjovsky and L. Bottou, Towards principled methods for training generative adversarial networks, *arXiv preprint arXiv:1701.04862* (2017)

[76] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford et al., Improved techniques for training gans, *Advances in neural information processing systems* **29** (2016)

[77] Y. Song and S. Ermon, Generative modeling by estimating gradients of the data distribution, *Advances in neural information processing systems* **32** (2019)

[78] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz et al., Diffusion models in medical imaging: A comprehensive survey, *Medical Image Analysis* **88**, 102846 (2023)

[79] Q. Lyu and G. Wang, Conversion between ct and mri images using diffusion and score-matching models, *arXiv preprint arXiv:2209.12104* (2022)

[80] T. Nyholm, S. Svensson, S. Andersson, J. Jonsson, M. Sohlin et al., Mr and ct data with multiobserver delineations of organs in the pelvic areaâ€"part of the gold atlas project, *Medical physics* **45**, 1295 (2018)

[81] A. Hore and D. Ziou, Image quality metrics: Psnr vs. ssim, in *2010 20th international conference on pattern recognition*, IEEE (2010), pp. 2366–2369

[82] X. Meng, Y. Gu, Y. Pan, N. Wang, P. Xue et al., A novel unified conditional score-based generative framework for multi-modal medical image completion, *arXiv preprint arXiv:2207.03430* (2022)

[83] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon et al., Score-based generative modeling through stochastic differential equations, *arXiv preprint arXiv:2011.13456* (2020)

[84] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani et al., The multimodal brain tumor image segmentation benchmark (brats), *IEEE transactions on medical imaging* **34**, 1993 (2014)

[85] Y. Li, H.-C. Shao, X. Liang, L. Chen, R. Li et al., Zero-shot medical image translation via frequency-guided diffusion models, *IEEE transactions on medical imaging* (2023)

[86] L. Zhu, Z. Xue, Z. Jin, X. Liu, J. He et al., Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis (2023)

[87] Y. Xu, L. Sun, W. Peng, S. Jia, K. Morrison et al., Medsyn: Text-guided anatomy-aware synthesis of high-fidelity 3d ct images (2024)

[88] K. Zhao, A. L. Y. Hung, K. Pang, H. Zheng and K. Sung, Partdiff: Image super-resolution with partial diffusion models (2023)

[89] J. Wang, J. Levman, W. H. L. Pinaya, P.-D. Tudosiu, M. J. Cardoso et al., Inversesr: 3d brain mri super-resolution using a latent diffusion model (2023)

[90] P. Rouzrokh, B. Khosravi, S. Faghani, M. Moassefi, S. Vahdati et al., Multitask brain tumor inpainting with diffusion models: A methodological report (2023)

[91] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab et al., Gans for medical image analysis, *Artificial Intelligence in Medicine* **109**, 101938 (2020)

[92] A. Radford, Unsupervised representation learning with deep convolutional generative adversarial networks, *arXiv preprint arXiv:1511.06434* (2015)

[93] A. Kitchen and J. Seah, Deep generative adversarial neural networks for realistic prostate lesion mri synthesis, *arXiv preprint arXiv:1708.00129* (2017)

[94] M. J. Chuquicusma, S. Hussein, J. Burt and U. Bagci, How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis, in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, IEEE (2018), pp. 240–244

[95] C. Bermudez, A. J. Plassard, L. T. Davis, A. T. Newton, S. M. Resnick et al., Learning implicit brain mri manifolds with deep learning, in *Medical Imaging 2018: Image Processing*, SPIE (2018), volume 10574, pp. 408–414

[96] E. L. Denton, S. Chintala, R. Fergus et al., Deep generative image models using aï¿¼ laplacian pyramid of adversarial networks, *Advances in neural information processing systems* **28** (2015)

[97] T. Karras, Progressive growing of gans for improved quality, stability, and variation, *arXiv preprint arXiv:1710.10196* (2017)

[98] C. Baur, S. Albarqouni and N. Navab, Generating highly realistic images of skin lesions with gans, in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis: First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018, Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 5*, Springer (2018), pp. 260–267

[99] Y. Skandarani, P.-M. Jodoin and A. Lalande, Gans for medical image synthesis: An empirical study, *Journal of Imaging* **9**, 69 (2023)

[100] P. Dhariwal and A. Nichol, Diffusion models beat gans on image synthesis (2021)

[101] J. Ye, H. Ni, P. Jin, S. X. Huang and Y. Xue, Synthetic augmentation with large-scale unconditional pre-training (2023)

[102] L. W. Sagers, J. A. Diao, L. Melas-Kyriazi, M. Groh, P. Rajpurkar et al., Augmenting medical image classifiers with synthetic data from latent diffusion models (2023)

[103] W. Peng, E. Adeli, T. Bosschieter, S. H. Park, Q. Zhao et al., Generating realistic brain mris via a conditional diffusion probabilistic model (2023)

[104] A. Fontanella, G. Mair, J. Wardlaw, E. Trucco and A. Storkey, Diffusion models for counterfactual generation and anomaly detection in brain images (2024)

[105] W. H. L. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. da Costa, V. Fernandez et al., Brain imaging generation with latent diffusion models (2022)

[106] J. Rosner, V. Reddy and F. Lui, *Neuroanatomy, Circle of Willis*, StatPearls [Internet], StatPearls Publishing, Treasure Island (FL), updated 2023 jul 24 edition (2024), available from: https://www.ncbi.nlm.nih.gov/books/NBK534861/

[107] E. Lin, H. Kamel, A. Gupta, A. RoyChoudhury, P. Girgis et al., Incomplete circle of Willis variants and stroke outcome, *Eur. J. Radiol.* **153**, 110383 (2022)

[108] F. Lin, Y. Xia, S. Song, N. Ravikumar and A. F. Frangi, High-throughput 3dra segmentation of brain vasculature and aneurysms using deep learning, *Computer Methods and Programs in Biomedicine* **230**, 107355 (2023)

[109] R. Xiao, C. Chen, H. Zou, Y. Luo, J. Wang et al., Segmentation of cerebrovascular anatomy from TOF-MRA using length-strained enhancement and random walker, *BioMed Res. Int.* **2020**, 9347215 (2020)

[110] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, Image-to-image translation with conditional adversarial networks, in *Proc. IEEE Conf. Comput. Vision Pattern Recog.* (2017), pp. 1125–1134

[111] A. Beers, J. Brown, K. Chang, J. P. Campbell, S. Ostmo et al., High-resolution medical image synthesis using progressively grown generative adversarial networks, *arXiv preprint arXiv:1805.03144* (2018)

[112] S. Olut, Y. H. Sahin, U. Demir and G. Unal, Generative adversarial training for MRA image synthesis using multi-contrast MRI, in *PRedictive Intelligence in MEdicine: First International Workshop, PRIME 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings Vol.1*, Springer (2018), pp. 147–154

[113] M. Sohail, M. N. Riaz, J. Wu, C. Long and S. Li, Unpaired multi-contrast MR image synthesis using generative adversarial networks, in *Simulation and Synthesis in Medical Imaging: 4th International Workshop, SASHIMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings*, Springer (2019), pp. 22–31

[114] M. D. Cirillo, D. Abramian and A. Eklund, Vox2Vox: 3D-GAN for brain tumour segmentation, in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I.6*, Springer (2021), pp. 274–284

[115] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich et al., Attention U-net: Learning where to look for the pancreas, *arXiv preprint arXiv:1804.03999* (2018)

[116] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli et al., Transunet: Transformers make strong encoders for medical image segmentation, *arXiv preprint arXiv:2102.04306* (2021)

[117] S. Motamed, P. Rogalla and F. Khalvati, Data augmentation using generative adversarial networks (gans) for gan-based detection of pneumonia and covid-19 in chest x-ray images, *Informatics in Medicine Unlocked* **27**, 100779 (2021), epub 2021 Nov 22

[118] N. Kodali, J. Abernethy, J. Hays and Z. Kira, On convergence and stability of gans, *arXiv preprint arXiv:1705.07215* (2017)

[119] Q. Zhang, D. Ram, C. Hawkins, S. Zha and T. Zhao, Efficient long-range transformers: You need to attend more, but not necessarily at every layer, *arXiv preprint arXiv:2310.12442* (2023)

[120] L. Liu, X. Liu, J. Gao, W. Chen and J. Han, Understanding the difficulty of training transformers, *arXiv preprint arXiv:2004.08249* (2020)

[121] B. Yu, Y. Wang, L. Wang, D. Shen and L. Zhou, Medical image synthesis via deep learning, *Deep Learning in Medical Image Analysis: Challenges and Applications* pp. 23–44 (2020)

[122] Information eXtraction from Images Consortium, IXI dataset – brain development, https://brain-development.org/ixi-dataset/, accessed: 2023-02-14

[123] K. Marstal et al., Simpleelastix: A user-friendly, multi-lingual library for medical image registration, https://simpleelastix.github.io/ (2016)

[124] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Comput. Vision Pattern Recog.* (2016), pp. 770–778

[125] D. Ulyanov, Instance normalization: The missing ingredient for fast stylization, *arXiv preprint arXiv:1607.08022* (2016)

[126] S. Ruder, An overview of multi-task learning in deep neural networks, *arXiv preprint arXiv:1706.05098* (2017)

[127] R. Caruana, Multitask learning, *Machine learning* **28**, 41 (1997)

[128] S. Graham, Q. D. Vu, M. Jahanifar, S. E. A. Raza, F. Minhas et al., One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification (2022)

[129] N. Ma, M. Ispir, Y. Li, Y. Yang, Z. Chen et al., An online multi-task learning framework for google feed ads auction models, in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2022), pp. 3477–3485

[130] C. Ding, Z. Lu, S. Wang, R. Cheng and V. N. Boddeti, Mitigating task interference in multi-task learning via explicit task routing with non-learnable primitives (2023)

[131] J. Baxter, A bayesian/information theoretic model of learning to learn via multiple task sampling, *Machine learning* **28**, 7 (1997)

[132] L. Duong, T. Cohn, S. Bird and P. Cook, Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser, in *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)* (2015), pp. 845–850

[133] A. Navon, A. Shamsian, I. Achituve, H. Maron, K. Kawaguchi et al., Multi-task learning as a bargaining game, *arXiv preprint arXiv:2202.01017* (2022)

[134] B. Liu, X. Liu, X. Jin, P. Stone and Q. Liu, Conflict-averse gradient descent for multi-task learning, *Adv. Neural Inf. Process. Syst.* **34**, 18878 (2021)

[135] 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, *Nat. Methods* **18**, 203 (2021)

[136] T.-M. Hoang, T. V. Huynh, A.-V. H. Ly and M.-V. Pham, The variations in the circle of Willis on 64-multislice spiral computed tomography, *Trends Med. Sci.* **2** (2022)

[137] H. Lippert and R. Pabst, *Arterial Variations in Man: Classification and Frequency*, J.F. Bergmann Verlag, Munich (1985)

[138] B. Eftekhar, M. Dadmehr and S. Ansari, Are the distributions of variations of circle of Willis different in different populations?, *BMC Neurol.* **6**, 1 (2006)

[139] F. Khader, G. Mueller-Franzes, S. Arasteh, T. Han, C. Haarburger et al., Medical diffusion–denoising diffusion probabilistic models for 3D medical image generation, *arXiv preprint arXiv:2211.03364* (2022)

[140] G. MÃ¼ller-Franzes, J. Niehues, F. Khader, S. Arasteh, C. Haarburger et al., Diffusion probabilistic models beat GANs on medical images, *arXiv preprint arXiv:2212.07501* (2022)

[141] L. Jiang, Y. Mao, X. Chen, X. Wang and C. Li, CoLa-Diff: Conditional latent diffusion model for multi-modal MRI synthesis, *arXiv preprint arXiv:2303.14081* (2022)

[142] W. Peng, E. Adeli, Q. Zhao and K. Pohl, Generating realistic 3D brain MRIs using a conditional diffusion probabilistic model, *arXiv preprint arXiv:2212.08034* (2022)

[143] W. Pinaya, P. Tudosiu, J. Dafflon, P. Da Costa, V. Fernandez et al., Brain imaging generation with latent diffusion models, *Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022* pp. 117–126 (2022)

[144] Z. Wang, E. P. Simoncelli and A. C. Bovik, Multiscale structural similarity for image quality assessment, in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Ieee (2003), volume 2, pp. 1398–1402

[145] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018)

[146] E. Kerfoot, J. Clough, I. Oksuz, J. Lee, A. P. King et al., Left-ventricle quantification using residual U-Net, in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International*

Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9, Springer (2019), pp. 371–380

[147] J. Ho, A. Jain and P. Abbeel, Denoising diffusion probabilistic models, *Adv. Neural Inf. Process. Syst.* **33**, 6840 (2020)

[148] R. Shu, H. H. Bui, H. Narui and S. Ermon, A dirt-t approach to unsupervised domain adaptation (2018)

[149] B. Brooksby, H. Dehghani, B. Pogue and K. Paulsen, Near-infrared (NIR) tomography breast image reconstruction with a priori structural information from MRI: algorithm development for reconstructing heterogeneities, *IEEE J. Sel. Top. Quantum. Electron.* **9**, 199 (2003)

[150] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp et al., Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis, *IEEE Trans. Med. Imaging* **38**, 1750 (2019)

[151] M. Hu, Visual pattern recognition by moment invariants, *IRE Trans. Inf. Theory* **8**, 179 (1962)

[152] A. Khotanzad and Y. Hong, Invariant image recognition by Zernike moments., *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 199 (1990)

[153] S. Elfwing, E. Uchibef and K. Doya, Sigmoid-weighted linear units for neural network function approximation in reinforcement learning., *Neural Netw.* **107**, 3 (2018)

[154] G. Kwon, C. Han and D. Kim, Generation of 3D brain MRI using auto-encoding generative adversarial networks, *Medical Image Computing and Computer Assisted Interventionâ€“MICCAI 2019* **22**, 118 (2019)

[155] C. Li and A. Bovik, Content-partitioned structural similarity index for image quality assessment., *Signal Processing: Image Communication* (2010)

[156] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and Ommer, Analyzing the role of visual structure in the recognition of natural image content with multi-scale

SSIM, *Human Vision and Electronic Imaging XIII. vol. 6806, pp. 410â€“423. SPIE* (2008)

[157] Y. Deo, R. Bonazzola, H. Dou, Y. Xia, T. Wei et al., Learned local attention maps for synthesising vessel segmentations from t2 mri, in *International Workshop on Simulation and Synthesis in Medical Imaging*, Springer (2023), pp. 32–41

[158] Y. Deo, H. Dou, N. Ravikumar, A. F. Frangi and T. Lassila, Shape-guided conditional latent diffusion models for synthesising brain vasculature, *arXiv preprint arXiv:2308.06781* (2023)

[159] A. Sinha, J. Song, C. Meng and S. Ermon, D2c: Diffusion-decoding models for few-shot conditional generation, *Advances in Neural Information Processing Systems* **34**, 12533 (2021)

[160] G. Giannone, D. Nielsen and O. Winther, Few-shot diffusion models, *arXiv preprint arXiv:2205.15463* (2022)

[161] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones et al., Attention is all you need, *Advances in neural information processing systems* **30** (2017)

[162] S. H. Lee, S. Lee and B. C. Song, Vision transformer for small-size datasets, *arXiv preprint arXiv:2112.13492* (2021)

[163] S. Go, Y. Ji, S. J. Park and S. Lee, Generation of structurally realistic retinal fundus images with diffusion models, *arXiv.org* **abs/2305.06813** (2023)

[164] Y. Xia, N. Ravikumar, T. Lassila and A. F. Frangi, Virtual high-resolution mr angiography from non-angiographic multi-contrast MRIs: synthetic vascular model populations for in-silico trials, *Med. Image Anal.* **87**, 102814 (2023)

[165] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai et al., An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv.org* **abs/2010.11929** (2020)

[166] Z. Dorjsembe, H.-K. Pao and F. Xiao, Conditional diffusion models for semantic 3D medical image synthesis, *arXiv.org* **abs/2305.18453** (2023)

[167] X. Hu, Y.-J. Chen, T.-Y. Ho and Y. Shi, Conditional diffusion models for weakly supervised medical image segmentation, *arXiv.org* **abs/2306.03878** (2023)

[168] X. He, W. Feng, T.-J. Fu, V. Jampani, A. R. Akula et al., Discriminative diffusion models as few-shot vision and language learners, *arXiv.org* **abs/2305.10722** (2023)

[169] S. Chen, K. Ma and Y. Zheng, Med3d: Transfer learning for 3d medical image analysis (2019)

[170] M. Woodland, A. Castelo, M. A. Taie, J. A. M. Silva, M. Eltaher et al., Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend (2024)

[171] Q. Xu, G. Huang, Y. Yuan, C. Guo, Y. Sun et al., An empirical study on evaluation metrics of generative adversarial networks (2018)

[172] B. Wang, Y. Zhu, L. Chen, J. Liu, L. Sun et al., A study of the evaluation metrics for generative images containing combinational creativity, *AI EDAM* **37**, e11 (2023)