UNIVERSITY OF CALIFORNIA
RIVERSIDE

Comparative Analysis of Deep Learning Architectures for ASL Image Denoising

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Electrical Engineering

by

Arun Sharma

December 2023

Thesis Committee:

Dr. Bir Bhanu, Chairperson
Dr. Jia Guo
Dr. Salman Asif

The Thesis of Arun Sharma is approved:

Committee Member

Committee Member

University of California, Riverside

Acknowledgments

I would first and foremost like to express my profound gratitude to my advisor, Dr. Jia Guo. His support, insights, and guidance have been instrumental in my journey, and without him, reaching this milestone would have been a distant dream. His unwavering compassion and understanding during challenging personal times were pivotal to the successful completion of this thesis.

I am also deeply thankful to Vindya Vashishth for her invaluable help with references, which enriched the depth and scope of my research.

Finally, I extend my heartfelt appreciation to the open-source community. The spirit of collaboration, shared knowledge, and mutual growth in this community has been a beacon of inspiration. We all ascend together, building upon each other's contributions.

To my parents for all the support.

Contents

List of Figures

List of Tables

Chapter 1

Background and Introduction

Magnetic Resonance Imaging (MRI) has revolutionized the field of medical imaging, providing non-invasive and detailed images of internal structures of the body, notably the brain. Unlike X-rays and CT scans, MRI doesn't rely on ionizing radiation. Instead, it uses a strong magnetic field, radio waves, and a computer to produce detailed pictures of the inside of the body.

Perfusion Imaging

Perfusion imaging pertains to the process of capturing the distribution of blood flow to the organs and tissues. Perfusion is vital as it provides essential nutrients to the tissues, ensuring their function and vitality. In the context of the brain, perfusion is crucial to maintain cognitive functions, and any disruptions can lead to severe consequences like strokes or cognitive impairments [1]. MRI provides a non-invasive method to study perfusion in the brain. Traditional methods often require injecting a contrast agent to visualize blood flow [2–5]. However, with advancements in MRI technology, it's now possible to acquire perfusion images without the need for any exogenous contrast agents[6–8]. This is where Arterial Spin Labeling (ASL) plays a pivotal role.

ASL (Arterial Spin Labeling)

ASL is a magnetic resonance imaging technique used to assess cerebral blood flow by magnetically labeling incoming blood [7]. Instead of relying on external agents, ASL takes advantage of the water in arterial blood as an endogenous tracer. Doing so provides a more direct measurement of cerebral blood flow, making it particularly useful for studying perfusion in various brain disorders [6]. ARTERIAL SPIN LABELING (ASL) is a class of methods for magnetic resonance imaging (MRI) of tissue perfusion. The term perfusion refers to the delivery of blood to capillary beds, and is quantified by the amount of blood delivered to the tissue per unit time, per unit volume or mass of tissue. This quantity is important physiologically because it determines the maximum rate of delivery of oxygen and other nutrients to the tissue, and also the rate of clearance of waste products. In ASL, arterial blood water is used as an endogenous diffusible tracer. Radiofrequency (RF) pulses are used to modify the longitudinal magnetization of arterial blood water before it flows into the target tissue, and after it reaches the tissue the label is observed as a perturbation of the tissue magnetization. Because the label is modified longitudinal magnetization, it decays with time constant T1, which is approx. 1350 msec in blood at 1.5T and 1650 msec at 3T [9]. The creation of labeled blood typically occurs in the arteries leading into the tissue of interest. In the brain, for example, labeling pulses are commonly applied in the carotid and vertebral arteries. The time it takes blood to travel from the labeling location to the target tissue, referred to here as the transit delay, is also on the order of 1 second. Thus, there are two similar but competing time constants in the ASL experiment: T1 decay of the label, which favors a short delay between the application of the label and image acquisition, and the transit delay, which favors a long delay to allow for complete delivery prior to image acquisition. The balance between these two factors is a key tradeoff in the design of ASL measurements. However, ASL imaging has its challenges. One of the major concerns is the low Signal-to-Noise Ratio (SNR). Low SNR can decrease the image quality, leading to potential inaccuracies in clinical diagnosis [10]. Addressing this challenge can significantly reduce scan times, thereby enhancing patient comfort, reducing costs, and improving the efficiency of medical imaging procedures [11].

Deep Learning and its Application in ASL

Deep learning, a subset of machine learning [12], is inspired by the structure and function of the brain, particularly a construct known as artificial neural networks. In recent years, deep learning has shown tremendous promise in various domains, from visual recognition to natural language processing [13–15]. In the realm of medical imaging, deep learning models, especially convolutional neural networks (CNNs), have achieved state-of-the-art performance in tasks such as image classification, segmentation, and anomaly detection [13].

Given the challenges faced by ASL imaging, particularly the low SNR, deep learning models can be employed to enhance the quality of ASL images, thus paving the way for more accurate clinical diagnoses. By training on vast datasets, these models can learn to extract intricate features, reduce noise, and provide clearer, high-resolution images that can be critical for clinical decision-making [11].

Chapter 2

Data Processing Pipeline

2.1 2.1 Experimental Design and Pipeline

Introduction

The process of acquiring, preprocessing, and processing MRI data, particularly ASL images, is intricate and requires careful consideration at every step. This chapter elucidates the comprehensive pipeline established for this project, ensuring the validity and reproducibility of our results.

Data Overview

Our study comprises data from 63 subjects. The data was acquired from OpenNeuro datasets[16]. All subjects underwent an MRI resting-state scan. Subjects were instructed to lie still with their eyes open, without falling asleep. Immediately after this scan, they were asked whether they fell asleep, and none of them reported they did. The MRI examination was carried out on a 3T whole-body MRI scanner (Trio TIM), using the body coil as transmitter and a 32-channel phased-array head coil as receiver. The imaging protocol included a 3D high-resolution anatomical T1-weighted MPRAGE sequence (with inversion time (TI) = 950 ms, repetition time (TR) = 1760 ms, echo-time (TE) = 3.1 ms, resolution = 1 mm isotropic, scan time = 5:08 min). This sequence was followed by an ASL sequence that combined pseudo-continuous labeling (PCASL) with a background-suppressed 3D GRASE single-shot readout. The labeling parameters were as follows: Hanning-shaped RF pulses, B1average = 1.8 microT, RFduration = 500 micro.sec, spacing = 500 micro.sec, Gaverage = 1mT/m, Gmaximum/Gaverage = 8, labeling duration = 1600 ms, post-labeling delay = 1500 ms. The imaging parameters were as follows: TR = 3.5 sec, TE = 29 msec, resolution = 4x4x7 mm^3, FOV = 250x188x112 mm^3 , 16 nominal partitions with 12.5% oversampling, 5/8 slice partial Fourier, matrix size = 64x49x11, BW = 2790Hz/pixel, gradient-echo spacing = 0.4 msec (with ramp sampling), spin-echo spacing = 29 msec, read-out time = 270 msec. 50 pairs of label and control images were acquired in 6 minutes. A short scan of 5 label/control pairs was performed using the same sequence without background suppression to acquire control images needed for calculation of CBF, this 10 images are allocated at the beginning of each ASL sequence, so each subject ASL sequence has 110 images.

In Arterial Spin Labeling (ASL), M0 images play a significant role as a reference. These images are acquired without any labeling or control condition. Their primary purpose is to provide a measure of the fully relaxed magnetization of arterial blood. Given their unique nature, M0 images possess a higher signal-to-noise ratio (SNR) than the subsequent label or control images. This makes them invaluable for scaling perfusion-weighted images to achieve absolute cerebral blood flow (CBF) quantification [17]. It's worth noting, however, that M0 images do not directly contribute to the generation of the difference images which provide the perfusion signal.

Difference Image Calculation

The first step in our pipeline involves computing the difference between each pair of control and label images. The resultant set comprises 50 3D images, each with dimensions $64 \times 64 \times 16$. Here, the number 16 denotes the number of slices in the z-direction. ASL images inherently suffer from a low SNR. To counteract this, a two-step procedure was adopted:

1. Each difference image was scaled by dividing it with $10\times$ the mean value for the respective subject. This scaling aids in normalizing the image intensity range across subjects. To ensure reversibility, the scaling factor was saved for each subject.

2. Given that certain images can be dominated by noise, we applied a filtering step. For each subject, the mean (μ) and standard deviation (σ) of their images were calculated. Any image value falling outside the range $[\mu - 2\sigma, \mu + 2\sigma]$ was set to zero. This approach retains the significant features of the image while suppressing extreme values, often attributed to noise.

The processed images were then averaged in various combinations, forming different sets. Averaging helps further suppress random noise and enhance the perfusion signal. The number and strategy for averaging were defined based on preliminary analyses to balance SNR and resolution.

Models were also trained without any pre-processing on the input data and the results were on par when the models were trained with pre-processed data. Hence, we proceeded without any pre-processing because it takes into account all kinds of input data and does not eliminate any slices or full 3D images of a scan.

Data Augmentation

To bolster the dataset's size and introduce variability, data augmentation techniques were employed as they have been shown to improve the performance of deep learning models [18]. Six distinct operations, including image flips, rotations, and shifts, were executed on the original images, leading to an expanded dataset 6 times the initial size. Ensuring consistency in experiments is pivotal. Hence, the augmented dataset was serialized and saved as '.npy' files. This ensures that irrespective of the model under consideration, the same dataset is used for training, validation, and testing, guaranteeing a fair comparison. Upon finalizing the dataset, we proceeded with the training of our deep learning models. After training, the model weights were saved. This allows for reproducibility, quick model evaluation on new data, and potential fine-tuning in future studies. The test data underwent a processing pipeline analogous to the training set.

Monitored Metrics

Once the processed test data was fed into the trained models for predictions, monitoring metrics were essential to quantify the model's performance. Three key metrics were utilized:

- **PSNR (Peak Signal-to-Noise Ratio):** It quantifies the ratio between the maximum possible power of the signal and the power of the noise, providing an estimate of the image quality.

- **SSIM (Structural Similarity Index Measure):** This metric assesses the perceived quality of an image. Instead of computing absolute errors, SSIM considers changes in structural information, luminance, and texture.
- **MSE (Mean Squared Error):** It computes the average squared differences between the estimated values and the actual value.

For each test subject, error maps were generated to see if the model introduced any artifacts in the predicted image. These maps provide a visual representation of the difference between the model's predictions and the ground truth. Subsequently, results for individual subjects were aggregated to yield an average performance measure.

Post Processing

After generating predictions, we revert the images to their original scale by multiplying them with a scaling factor—ten times the subject's mean. This rescaling facilitates subsequent computations, such as Cerebral Blood Flow. Subsequently, we evaluate our metrics, namely PSNR, SSIM, and NMSE, by comparing the input and predicted data. For evaluation, we used NMSE (Normalized Mean Squared Error) so that the results are comparable across subjects. The NMSE computes the average squared differences between the estimated values and the actual value and divides that by the mean intensity of the ground truth image.

$$NMSE = \frac{\sum_{i=1}^{N}(E_i - G_i)^2}{\sum_{i=1}^{N} G_i} \tag{2.1}$$

where:

- E_i is the estimated value of the i^{th} pixel in the image E.
- G_i is the actual value of the i^{th} pixel in the ground truth image G.
- N is the total number of pixels in the image.

The results of these evaluations are presented in the tables discussed in Chapter 4.

Chapter 3

Exploring the Models

3.1 3.1 UNET

UNET 2D

The U-Net architecture, a pioneering deep learning framework, emerged as a potent solution for image segmentation tasks, owing to its unique design and impressive performance. Developed by researchers Olaf Ronneberger, Philipp Fischer, and Thomas Brox in 2015 at the University of Freiburg, U-Net's inspiration derives from its ”U” shape, formed by a contracting path and an expansive path [19].

U-Net's efficacy stems from its ability to tackle semantic segmentation challenges where precise delineation of object boundaries is required. The architecture's contracting path, often referred to as the ”encoder,” incorporates convolutional and pooling layers to capture context while downsampling the input image. This process enables the model to recognize essential features, although with reduced spatial resolution.

The expansive path, or "decoder," employs transposed convolutions for upsampling, progressively restoring the spatial dimensions while maintaining contextual information. Importantly, this "skip-connection" architecture interlinks corresponding layers from the contracting to the expansive path, facilitating the transmission of fine-grained spatial details.

U-Net's prominence in the medical imaging field arises from its remarkable capabilities in denoising and enhancing images. Medical images, such as pulsed arterial spin labeling (ASL) MRI brain scans, often suffer from noise artifacts that can impede accurate diagnoses. U-Net's capacity to discern intricate structures and fine-tune them by removing noise positions it as a potent tool for medical image denoising.

The combination of Mean Squared Error (MSE) and Structural Similarity Index (SSIM) as the loss function further underscores U-Net's suitability for this task. MSE ensures the minimization of pixel-wise differences between the predicted and ground truth images. SSIM, on the other hand, accounts for structural information, providing a more perceptually aligned assessment of image quality.

U-Net's utility extends beyond denoising to encompass an array of medical imaging tasks. Its application ranges from organ and tumor segmentation to image registration and disease detection. The architecture's adaptability, coupled with its consistent success, has cemented its place as a staple in the medical imaging community.

In conclusion, the U-Net architecture, conceptualized by Ronneberger, Fischer, and Brox, has revolutionized the field of medical image analysis. Its distinctive "U" shape, combining contracting and expansive paths, empowers it to excel in tasks like denoising

ASL MRI brain images. Through its incorporation of both MSE and SSIM, U-Net ensures not only pixel-wise accuracy but also perceptual fidelity. As the medical imaging realm continues to evolve, U-Net stands as an indispensable asset, harnessing the power of deep learning to enhance diagnostic precision and ultimately improve patient care.

Our Model

The U-Net architecture, originally designed for biomedical image segmentation, was adapted and implemented for the analysis of 2D images in the current study. The model's architecture was tested with both four and five layers deep model.

In the first stage of the U-Net, known as the contracting path, an input image undergoes a series of operations. Each image is convolved twice at every layer, which essentially allows the model to learn various image features at multiple scales. Following convolution, Batch Normalization was applied. The incorporation of this step was crucial as, without it, the model's loss might not converge. This was succeeded by the application of a LeakyReLU activation function to better adapt to small negative pixel values in the ground truth. Subsequent to these operations, max-pooling was performed to progressively reduce the spatial dimensions of the image. To mitigate the risk of overfitting, dropout of 0.05 was incorporated at each layer. The output from each layer was then propagated to the subsequent layer, iteratively refining the features captured.

Upon completion of the contracting path, the expansive path of the U-Net begins. In this phase, the spatial dimensions of the image are increased using 2D transposed convolution operations. It's worth noting that the kernel size was kept equal to the stride to avoid undesirable artifacts in the form of a checkered pattern in the output. Dropout was incorporated in this phase as well, ensuring robustness against overfitting. This expansive path was executed as many times as max-pooling was performed in the contracting path, ensuring a symmetrical architecture.

The culmination of the expansive path leads to the final operation: a 2D convolution with a linear activation function. The output is an array with identical dimensions to the original input image (64x64).

The model's loss function was particularly crafted as a linear combination of mean squared error (MSE) and mean structural similarity (SSIM) loss. Several ratios of these two factors were rigorously tested to determine the optimal balance for our dataset which assigned equal weights to the two metrics. As the model trains, it learns the weights for the convolutional and transposed convolutional layers, gradually improving its performance. A standout feature of our U-Net architecture is the inclusion of skip connections between layers producing outputs of the same dimensions. These connections ensure that the model retains spatial information, which is often lost in deeper layers. To validate the model, a five-fold cross-validation was employed. The metrics for evaluating performance were NMSE, SSIM, and PSNR.

The training was conducted under Python 3.7 with Tensorflow 2.10. The four-layer model had 290, 929 trainable parameters, and the five-layer model had 783, 937 trainable parameters. Each epoch took approximately 3 minutes to train on a batch size of 32, given a training set size of 211,200 images. Models were trained on raw data, and also on data pre-processed by averaging either 2% (unaveraged), 10% (average of 5 images) or 20% (average of 10 images) of the data. Post-training, all testing images were normalized by division with their respective mean and subsequently rescaled using this scaling factor following predictions.

The ground truth was established by averaging images from all time points. A unique advantage of the U-Net architecture lies in its ability to capture both local and high-level details. While the initial layers focus on capturing minute local details, the deeper layers, having processed the entire image, encapsulate more holistic, high-level information.

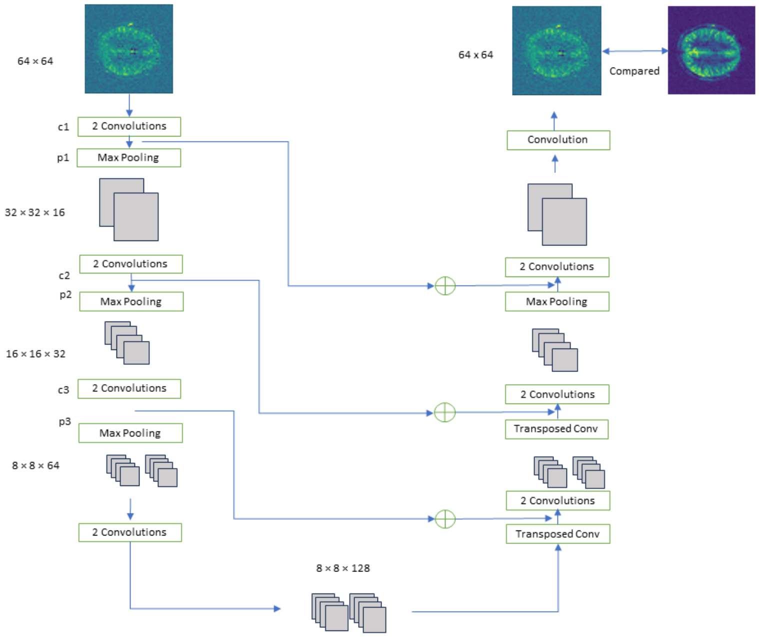


Figure 3.1: UNET 2D flow

UNET 3D

Conceived as a natural extension of its 2D predecessor, U-Net 3D addresses the complexities inherent in volumetric data, offering a more comprehensive solution for tasks requiring spatial understanding and depth perception. The origins of U-Net 3D can be traced back to the growing demand for accurate and holistic analysis of medical images, such as CT scans and MRI volumes. As medical imaging technology progressed, the need arose for models capable of understanding the intricate 3D structures present in these scans. In response, researchers extended the foundational U-Net design into the third dimension, resulting in an architecture that embraces the volumetric nature of medical data. In denoising tasks involving volumetric medical images, U-Net 3D logically holds an advantage over its 2D counterpart. Noise artifacts in 3D scans can span multiple slices, necessitating a model capable of recognizing patterns and relationships that extend across the depth of the volume. U-Net 3D's ability to capture volumetric context enables it to mitigate noise more effectively and provide more accurate denoising results. As a result, U-Net 3D is not only equipped to handle medical image denoising but also a multitude of volumetric analysis tasks, including organ segmentation, tumor detection, and disease classification. By embracing the inherent dimensionality of medical volumes, U-Net 3D enhances the accuracy and reliability of analyses, ultimately improving clinical decision-making and patient outcomes[20]. In conclusion, U-Net 3D emerges as a logical progression from U-Net 2D, catering to the demands of volumetric medical image analysis. Its architecture, tailored to capture 3D structures and spatial relationships, positions it as a superior choice for tasks requiring a comprehensive understanding of volumetric data. In an era where precision and depth perception are critical in medical imaging, U-Net 3D stands as an essential tool, ushering in a new era of accurate and impactful diagnostic insights.

3.2 Vision Transformers and Swin Transformer in Medical Imaging

3.2.1 Vision Transformers (ViTs)

Over the past decade, Convolutional Neural Networks (CNNs) have been the gold standard for a plethora of computer vision tasks. Their unparalleled capacity to process grid-like data, predominantly images, by discerning local patterns via a cascade of filters, has solidified their dominance in the field. However, a new protagonist has recently emerged in the arena of image processing: the Vision Transformer (ViT) [21].

Advantages over CNNs:

1. **Global Context:** Unlike CNNs, which primarily capture local context in their preliminary layers, transformers have the innate capability to understand global context right from the onset.

2. **Scalability:** ViTs scale better with increased data and computation. When datasets are sufficiently large, ViTs tend to outperform CNNs.

3. **Parameter Efficiency:** Transformers, especially when pre-trained, often require fewer parameters than deep CNNs for the same or even superior performance.

In the realm of medical imaging, the aforementioned attributes of ViTs have shown to be particularly beneficial. Recent works have applied ViTs to medical image classification, segmentation, and anomaly detection, often achieving state-of-the-art results [22].

3.2.2 Swin Transformer

Building upon the foundations of Vision Transformers, the Swin Transformer introduces a hierarchical structure that permits local and global information processing, making it particularly suited for dense prediction tasks such as semantic segmentation or object detection.

Key Features and Advantages:

1. **Hierarchical Design:** Swin Transformer replaces the standard Transformer's fixed-size patches with overlapping windows, increasing its capacity to manage local image features.

2. **Shifted Windows:** For higher layers in the network, the Swin Transformer uses shifted windows, allowing the model to capture a broader context without increasing computational complexity.

3. **Hybrid Structure:** The Swin Transformer can be initialized with CNN feature maps, combining the strengths of CNNs and transformers.

Within medical imaging, the Swin Transformer's unique hierarchical design has proven especially advantageous. Its overlapping windows allow the model to focus on intricate details found in medical images. In tasks such as denoising, the model's ability to capture both local patterns and global context means that it can understand noise patterns and restore the original image with exceptional precision [23].

3.3 Swin Transformer Architecture

3.3.1 Image Partitioning into Patches

The first step in the Swin Transformer pipeline is dividing the image into non-overlapping fixed-size patches. An image with a size of $H \times W$ is partitioned into patches of size $P \times P$. This effectively results in $\frac{H}{P} \times \frac{W}{P}$ patches. The idea of using patches can be traced back to the original Vision Transformer (ViT) [21], where it was found effective in capturing local information [24].

3.3.2 Patch Embedding

Once the patches are extracted, each patch is linearly embedded into a flat vector of dimension D. This embedding can be seen as reshaping the patch and then multiplying it by an embedding matrix. The result is a sequence of patch embeddings, which becomes the input for the subsequent transformer blocks [24].

3.3.3 Multiple Swin Transformer Blocks

The core of the Swin Transformer architecture consists of several Swin Transformer blocks stacked on top of each other. Each of these blocks has two primary layers: the shifted window-based self-attention mechanism and a multi-layer perceptron (MLP) [24].

Shifted Window-based Self-Attention

Traditional transformers operate on the entire sequence, leading to a quadratic complexity. Swin Transformer introduces a window-based mechanism, where the attention operation is limited to fixed-size local windows. These windows slide across the image without overlap.

To ensure that the model captures global context, adjacent blocks use windows shifted by half of the window size. This ensures that a token, over multiple blocks, interacts with tokens outside its initial window, expanding its receptive field [24].

Masking

The term ”masking” often refers to preventing certain elements from participating in computations, especially in the context of attention mechanisms. In transformers, a mask is an array or tensor, typically containing ones and zeros, which defines which elements should be considered (‘1‘) or ignored (‘0‘).

In the Swin Transformer, the importance of masking is twofold:

Permutation Invariance: Swin Transformer is designed to maintain permutation invariance, meaning that the model's output should remain consistent irrespective of the input sequence order. This is crucial for vision tasks since images can be fed into the model in any sequence. In the context of Swin Transformer, zero padding is added to the image if it's smaller than the desired input size. These zeros should not affect the model's computations, and thus, they are masked out during the attention calculations. By doing this, the model is ensured to focus only on the meaningful parts of the image and not on the padded zeros [24].

To achieve this, a binary mask is used during the dot-product attention computation. The mask has the same size as the attention scores tensor, and every position corresponding to a padded zero will have a very large negative value (often '-inf'). When applying the softmax function, these positions will effectively become zero, ensuring that no attention is paid to the padded positions.

Causal Masking: Though not specifically a part of Swin Transformer, it's worth mentioning causal masking, which is often used in transformers designed for sequences (like text). The idea behind causal masking is to ensure that a token does not have access to future tokens in a sequence. This is vital for tasks like language modeling, where the model should predict the next word in a sequence without having seen it. A causal mask is an upper triangular matrix filled with very large negative values (or '-inf'), ensuring that when softmax is applied during attention calculation, future tokens have zero weight.

In conclusion, masking is a strategic operation ensuring that the transformer model remains attentive to the right parts of the input, maintaining robustness and consistency in its predictions [24].

Attention Mechanism and Attention Scores

The attention mechanism used in the Swin Transformer follows the traditional scaled dot-product attention. Given queries Q, keys K, and values V, the attention score is computed as:

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

Where d is the depth of the query/key vectors. These attention scores represent how much attention each patch should pay to every other patch [24].

3.3.4 Reversing Window Partition and MLP Layer

Once the self-attention mechanism is applied, the window partitions are reversed to restore the original sequence order. Following this, the sequence is passed through an MLP layer. This MLP is vital for introducing non-linearity into the model, which aids in learning complex representations. It typically consists of two linear layers with a GELU activation function in between [24].

3.3.5 Patch Merging

As the information flows through the stacked transformer blocks, the resolution is gradually reduced using patch merging. This is done by merging adjacent patches into larger ones, effectively reducing the sequence length by a factor of 4. This merged patch is then linearly embedded into a vector. The purpose of this merging is to allow the model to capture more global information as it processes deeper [24].

Post-processing in Swin Transformer

After the information has been processed through the Swin Transformer blocks, the next step is to decode this processed data into a format suitable for the specific task at hand. In the case of image-related tasks, it's often about reconstructing the image or generating a modified version of the input image. Here's how the Swin Transformer achieves this:

From Embedded Space to Original Patch: The initial step involves a convolution operation. This convolution layer maps the output of the final Swin Transformer block, which is in an embedded space, back to the spatial dimensions of the original patch. This process essentially transforms the rich feature representations, which the model has learnt, back to a spatial format that resembles image patches. This is crucial as the subsequent operations are spatial in nature [24].

Upscaling with Pixel Shuffle: Pixel shuffle is a method often used in deep learning to upscale an image. In the context of Swin Transformer, multiple convolution operations followed by pixel shuffling are applied. Each convolution layer refines the features, ensuring they capture the intricate details needed for a high-resolution output.

The pixel shuffle operation involves rearranging elements in a tensor, typically output from a convolutional layer, to achieve a higher resolution. Specifically, pixel shuffle takes a low-resolution image with multiple channels and rearranges these channels to form a high-resolution image with fewer channels. This operation ensures that the upscaling does not introduce any artifacts or rely on simple interpolation methods, but rather uses the learnt features to intelligently upscale the image.

For instance, if we have a 2x2 patch with 4 channels, pixel shuffle can rearrange this to form a 4x4 patch with one channel. This is done without any loss of information. The process is repeated with subsequent convolution and pixel shuffle operations to achieve the final desired resolution.

Comparing to Ground Truth: Once the image has been upscaled to its original size, it's time to compute the loss for training. The upscaled image is compared to the ground truth using a suitable loss function, which here as well is a linear combination of Mean

Squared Error (MSE) and Structural Similarity (SSIM). The gradients from this loss are then backpropagated through the entire network, including the Swin Transformer blocks, to update the model's weights [24].

In essence, the post-processing in Swin Transformer ensures that the model's rich, abstract feature representations are effectively translated back into spatial, interpretable formats suitable for direct comparison with real-world images.

Conclusion

As medical imaging continues to evolve and demands higher precision and interpretability, models like the Vision and Swin Transformers are set to play pivotal roles. Their global context understanding combined with local pattern recognition makes them ideal for complex tasks in medical imaging.

Chapter 4

Results and Comparison

The goal of this section is to use the tools developed in the previous sections to run experiments. We ran the basic experiment to evaluate the pipeline's performance in denoising images for the four models and tried to find the best parameters, such as the cost-function ratio. In addition to the evaluation metrics, we also visualized the images for inspection. A successful result allows future improvements by finding ways to improve the network structure or parameters based on testing on different and larger datasets. For evaluation, we tested the model under three distinct scenarios:

1. Using non-averaged images as the input.

2. Inputting images that were obtained by averaging over five consecutive time points.

3. Taking an average of the outputs produced over five successive time points.

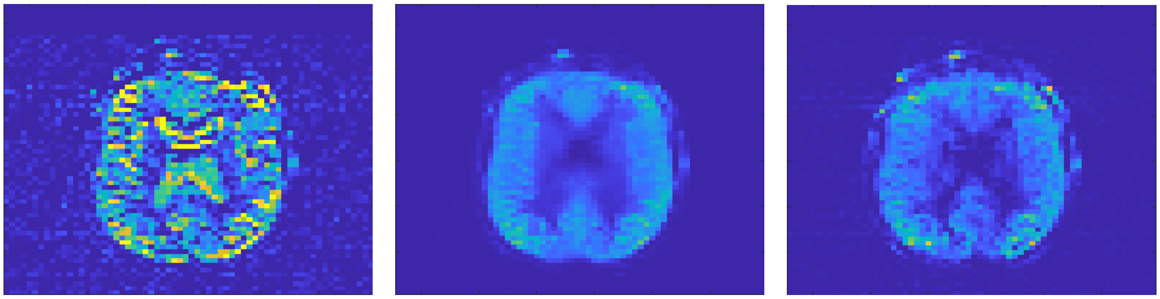
By contrasting the second and third scenarios, we aim to discern the superior approach when given five 3D images. Specifically, we investigate whether it is more advantageous to average these five images beforehand, yielding a high SNR input, or to input the images individually and subsequently average the outputs. Intuitively, the performance should be similar.

4.1 The UNET 2D Model

We embarked on an extensive training process using a 2D UNET architecture. Two distinct models were developed, differentiated by their depth: one with four layers and the other with five layers. Each model was trained on images that were captured at multiple discrete time points. Notably, these images were utilized in their raw form, meaning no averaging techniques were employed during the training process. The UNET 2D model trains the fastest among all the models examined here

An interesting observation arose during the evaluation phase. For very low SNR images, which in our case was when individual timepoint images were used as input without any averaging, the four-layered 2D UNET model did not work well but increasing the layers to five resulted in a well-trained model. In case the input images were averaged for five time points resulting in higher SNR images, the loss curve of the five-layered model demonstrated clear signs of overfitting so a four-layered model was used and it worked really well. Specifically, as the training progressed for the five layered model in case of higher SNR images, the validation loss began increasing, diverging from the continually decreasing training loss. This disparity between training and validation metrics raised concerns about the model's generalization capabilities on unseen data. This can be a result of the fact that averaging images reduces the dataset and so a five-layered model would overfit due to a higher number of parameters.

As mentioned, When the five-layer model was evaluated using non-averaged input images, the results were markedly positive. The results for the same are shown in the following tables.



(a) UNET 2D Input (b) UNET 2D Prediction (c) UNET 2D Ground Truth

Figure 4.1: UNET 2D Results

Table 4.1: Output of UNET 2D model. Individual 2D image time points used as input

Subject	Input PSNR	Pred PSNR	Input SSIM	Pred SSIM	Input NMSE	Pred NMSE
Subject 1	26.5696	34.4693	0.4136	0.7353	4.9318	0.5383
Subject 2	21.7294	28.9158	0.4203	0.8478	0.221	0.0411
Subject 3	21.3676	31.3022	0.3918	0.7796	0.9124	0.0798
Subject 4	22.1386	29.7391	0.3665	0.8128	0.3155	0.0532
Subject 5	21.9045	29.0295	0.5325	0.8809	0.1842	0.0352
Subject 6	23.2819	31.708	0.4539	0.8179	0.4204	0.0571
Subject 7	24.2718	30.2382	0.4871	0.8396	0.1896	0.046

On average, across seven test subjects, for the case when unaveraged images were used, the input PSNR(Peak Signal-to-Noise Ratio) was 23.04 ± 1.85 dB while the predicted images showed a PSNR of 30.77 ± 1.94 dB there was an improvement of approximately 7.73

Table 4.2: Output of UNET 2D model when predicted images at five timepoints were averaged.

Subjects	Input PSNR	Pred PSNR	Input SSIM	Pred SSIM	Input NMSE	Pred NMSE
Subject 1	26.5696	35.8915	0.4136	0.7166	4.9318	0.3881
Subject 2	21.7294	31.267	0.4203	0.8878	0.221	0.0235
Subject 3	21.3676	33.045	0.3918	0.8201	0.9124	0.0521
Subject 4	22.1386	32.4366	0.3665	0.8702	0.3155	0.0281
Subject 5	21.9045	31.4066	0.5325	0.9145	0.1842	0.0206
Subject 6	23.2819	33.8137	0.4539	0.8677	0.4204	0.0348
Subject 7	24.2718	32.012	0.4871	0.8773	0.1896	0.0304

Table 4.3: Output of UNET 2D model when input images at five timepoints were averaged.

Subjects	Input PSNR	Pred PSNR	Input SSIM	Pred SSIM	Input NMSE	Pred NMSE
Subject 1	34.4667	36.5465	0.7429	0.8157	2.3243	1.2795
Subject 2	29.1550	31.2921	0.7626	0.8928	0.0406	0.0235
Subject 3	28.3071	33.1397	0.7071	0.8305	0.1709	0.0515
Subject 4	29.4228	32.4165	0.7209	0.8707	0.0580	0.0284
Subject 5	27.4489	30.9793	0.8158	0.9112	0.0596	0.0231
Subject 6	29.8705	33.7863	0.7642	0.8704	0.0982	0.0354
Subject 7	31.4747	32.4610	0.8012	0.8898	0.0349	0.0275

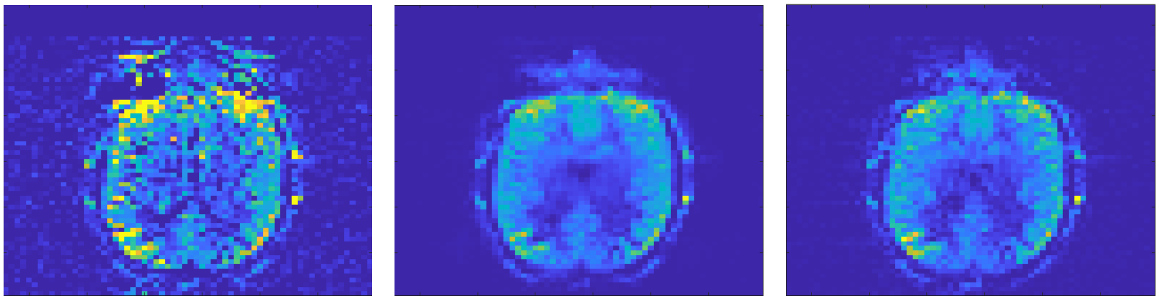
dB in the PSNR value. Similarly, the SSIM (Structural Similarity Index Measure) for the input was 0.44±0.05 and the predicted images showed a similarity value of 0.82±0.05 hence the value exhibited an increase by an average of 0.37. Lastly, the normalized mean squared value saw a significant reduction, for the input NMSE was 1.02±1.74 and was 0.12±0.18 for the output with an average decrease of 0.9. These improvements underscore the effectiveness and potential of the five-layer UNET model in handling non-averaged images. For the case where input images are averaged before prediction by the model, the predicted image has PSNR of 32.94 ± 1.86 dB, SSIM of 0.87 ± 0.03, and NMSE of 0.21 ± 0.47. Conversely, in cases where the model prediction precedes the averaging process, the PSNR is 32.84±1.16, the SSIM is 0.85 ± 0.06, and the NMSE is 0.08 ± 0.13. Notably, it can be inferred from the data that pre-prediction averaging yields marginally better results in comparison to post-prediction averaging but the results are within one standard deviation of each other.

4.2 The UNET 3D Model

The training duration of the UNET 3D model is notably longer in comparison to its 2D counterpart when provided with an equivalent volume of data. This observation aligns well with expectations considering the inherent complexity and depth associated with a 3D model. For the 3D model we only use a four-layered UNET structure as the five-layered model overfits the data.

A comprehensive analysis reveals that the UNET 3D model outperforms the UNET 2D model across all evaluation metrics. Furthermore, the strategy of averaging the input images, whether executed pre or post-prediction, appears to exert minimal influence on the quality of the output. As anticipated, the resultant images derived from an average of five consecutive time points display superior attributes in contrast to those obtained from a singular time point.

Quantitatively, this superiority is manifested in a PSNR improvement of 9.78 ± 1.23dB and an SSIM increment of 0.50 ± 0.06 when the unaveraged 3D images are used which had an input PSNR of 18.16 ± 1.18 dB, SSIM of 0.41 ± 0.05, and NMSE of 9.50 ± 2.47. The output images had a PSNR of 27.95 ± 0.9 dB, SSIM of 0.91 ± 0.03, and NMSE of 2.39 ± 1.5 for scenarios where the outputs of five images are averaged. For the case where input images are averaged before prediction by the model the predicted image has PSNR of 28.95 ± 1.09 dB, SSIM of 0.93 ± 0.03, and NMSE of 0.29 ± 0.19. Conversely, in cases where the model prediction precedes the averaging process, the PSNR is 29.3 ± 0.96, the SSIM is 0.93 ± 0.02, and the NMSE is 0.34 ± 0.22. Notably, it can be inferred from the data that pre-prediction averaging yields marginally better results in comparison to post-prediction averaging but the results are within one standard deviation of each other.



(a) UNET 3D Input (b) UNET 3D Prediction (c) UNET 3D Ground Truth

Figure 4.2: UNET 3D Results

Table 4.4: Output of UNET 3D model. Individual 3D image time points used as input

Subject	Input PSNR	Pred. PSNR	Input SSIM	Pred. SSIM	Input NMSE	Pred. NMSE
Subject 1	19.3527	29.5835	0.4804	0.9167	12.7735	1.6498
Subject 2	18.0828	27.5441	0.359	0.9173	7.3844	2.3343
Subject 3	16.3165	27.6652	0.404	0.9415	12.9322	0.8326
Subject 4	17.0942	27.2446	0.3312	0.907	9.9982	2.4626
Subject 5	18.5259	27.3979	0.4315	0.8776	8.5505	4.28
Subject 6	18.0885	28.8599	0.4333	0.9527	7.1332	0.7243
Subject 7	19.6867	27.3584	0.439	0.8836	7.7775	4.4369

Table 4.5: Result for UNET 3D when input images at five timepoints were averaged.

Subject	Input PSNR	Pred PSNR	Input SSIM	Pred SSIM	Input NMSE	Pred. NMSE
Subject 1	27.0605	30.8122	0.7866	0.9382	0.1091	0.1862
Subject 2	25.5003	28.4905	0.6653	0.9295	0.1413	0.2872
Subject 3	23.1943	27.9219	0.7078	0.949	0.1069	0.1217
Subject 4	24.4219	28.7221	0.6474	0.9272	0.1813	0.28
Subject 5	24.1691	27.7195	0.721	0.875	0.3996	0.6316
Subject 6	24.7304	29.8867	0.7407	0.9636	0.088	0.0887
Subject 7	26.8535	29.1127	0.7399	0.911	0.1829	0.4668

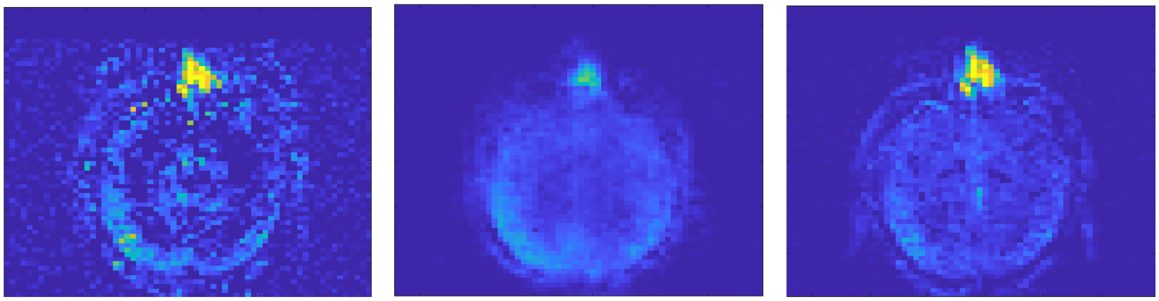
Table 4.6: Result for UNET 3D when predicted images at five timepoints were averaged.

Subject	Input PSNR	Pred. PSNR	Input SSIM	Pred. SSIM	Input NMSE	Pred. NMSE
Subject 1	19.3527	31.0069	0.4804	0.9368	12.7735	0.2214
Subject 2	18.0828	29.0917	0.359	0.9327	7.3844	0.3142
Subject 3	16.3165	28.6539	0.404	0.9508	12.9322	0.1278
Subject 4	17.0942	28.9788	0.3312	0.9274	9.9982	0.3235
Subject 5	18.5259	28.4863	0.4315	0.8931	8.5505	0.647
Subject 6	18.0885	30.2875	0.4333	0.9635	7.1332	0.1003
Subject 7	19.6867	28.5719	0.439	0.9034	7.7775	0.6601

4.3 Analysis of Vision Transformer Models

Vision Transformer 2D

The Vision Transformer 2D, an innovative approach in the realm of image processing, exhibits remarkable improvements in all evaluation metrics when individual, discrete images are introduced as input. However, the model's performance exhibits a decline when images obtained by averaging are introduced, or when the outputs post-prediction are averaged. This decrement in performance potentially underscores the model's susceptibility to overfitting and its lack of robustness in certain situations. Corroborating this hypothesis is the observed divergence between validation and training errors during the training phase, a classic indication of overfitting. A closer scrutiny of the model's architecture reveals an extensive number of parameters. Considering the volume of data available for training, it's plausible to infer that the dataset size might be inadequate for the model to genuinely harness its potential and deliver consistent results.



(a) ViT 2D Input (b) ViT 2D Prediction (c) ViT 2D Ground Truth

Figure 4.3: ViT 2D Results

Table 4.7: For model trained on 2% data. Result for 2D transformer model.

Subject	Input PSNR	Pred. PSNR	Input SSIM	Pred. SSIM	Input NMSE	Pred. NMSE
Subject 1	19.3527	25.2223	0.4804	0.6923	12.7735	3.3639
Subject 2	18.0828	23.5874	0.359	0.7249	7.3844	2.0705
Subject 3	16.3165	24.8106	0.404	0.7261	12.9322	1.4407
Subject 4	17.0942	23.5007	0.3312	0.6992	9.9982	0.0669
Subject 5	18.5259	24.2091	0.4315	0.7748	8.5505	0.0474
Subject 6	18.0885	24.5998	0.4333	0.7415	7.1332	0.0847
Subject 7	19.6867	23.4938	0.439	0.7249	7.7775	0.0744

We see that while there is improvement in all the metrics for unaveraged input the model is predicting all images with a very similar PSNR and SSIM. When we average these images we expect to get a better image but that doesn't happen and in some cases, the performance actually decreases and we see a decrease in value for all three metrics. This is also apparent visually from the predicted image in Figure 4.3 where we can see a sort of blurring effect.

Table 4.8: Result for 2D transformer model - input images were obtained by averaging over five consecutive time points.

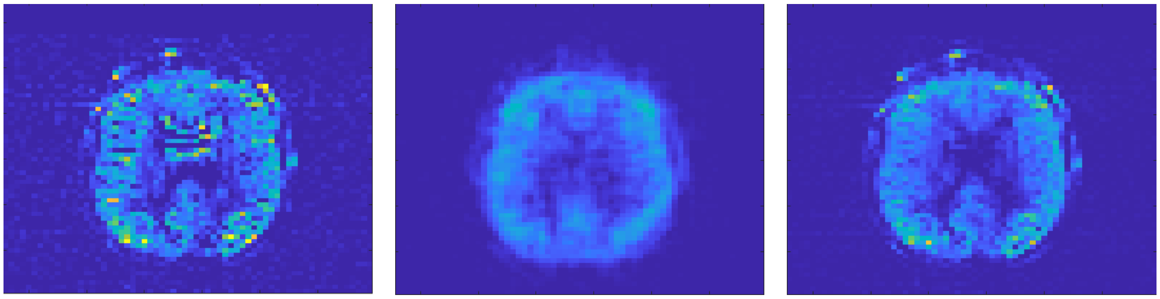
Subject	Input PSNR	Pred. PSNR	Input SSIM	Pred. SSIM	Input MSE	Pred. NMSE
Subject 1	27.0605	25.5004	0.7866	0.7030	2.2205	3.1379
Subject 2	25.5003	23.5387	0.6653	0.7155	1.6247	2.1383
Subject 3	23.1943	25.2377	0.7078	0.7337	3.2458	1.2876
Subject 4	24.4219	23.7481	0.6474	0.6948	0.0911	0.0676
Subject 5	24.1691	24.0644	0.7210	0.7653	0.0686	0.0494
Subject 6	24.7304	24.9081	0.7407	0.7443	0.1026	0.0809
Subject 7	26.8535	23.5626	0.7399	0.7210	0.0391	0.0755

Table 4.9: Result for 2D transformer model - taking an average of the outputs produced over five successive time points

Subject	Input PSNR	Pred PSNR	Input SSIM	Pred SSIM	Input NMSE	Pred NMSE
Subject 1	19.3527	28.281	0.4804	0.8084	12.7735	4.2423
Subject 2	18.0828	25.0726	0.359	0.6897	7.3844	1.7184
Subject 3	16.3165	24.0086	0.404	0.7451	12.9322	2.4182
Subject 4	17.0942	23.968	0.3312	0.7007	9.9982	2.7817
Subject 5	18.5259	24.4747	0.4315	0.6354	8.5505	1.354
Subject 6	18.0885	24.3111	0.4333	0.767	7.1332	1.1445
Subject 7	19.6867	23.3705	0.439	0.7898	7.7775	1.7523

Vision Transformer 3D

Paralleling the challenges faced by its 2D counterpart, the Vision Transformer 3D model grapples with similar shortcomings. Given the augmented dimensionality, the 3D model inherently possesses an even greater number of parameters, amplifying the challenges observed in the 2D model. This complexity is reflected not just in performance metrics, but also in training duration. This model has 146, 006, 913 trainable parameters. Both the Vision Transformer 2D and 3D models require approximately six times the training time per epoch compared to conventional models. Additionally, they necessitate a greater number of epochs to reach convergence, further emphasizing the demanding nature of these transformer models.



(a) ViT 3D Input (b) ViT 3D Prediction (c) ViT 3D Ground Truth

Figure 4.4: ViT 3D Results

Table 4.10: For model trained on 2% data. Result for 3D transformer model.

Subject	Input PSNR	Pred PSNR	Input SSIM	Pred SSIM	Input NMSE	Pred NMSE
Subject 1	15.5193	23.5549	0.4124	0.7148	14.1801	1.9909
Subject 2	18.7561	23.469	0.6197	0.7693	3.2559	2.3556
Subject 3	17.9822	23.5593	0.4813	0.7981	2.8191	2.149
Subject 4	19.9883	23.977	0.4322	0.773	6.0414	2.287
Subject 5	20.1221	25.5079	0.6184	0.8191	2.0138	1.3767
Subject 6	15.658	22.9364	0.2668	0.673	3.0397	1.7411

Table 4.11: Result for 3D transformer model - input images were obtained by averaging over five consecutive time points.

Subject	Input PSNR	Pred PSNR	Input SSIM	Pred SSIM	Input NMSE	Pred NMSE
Subject 1	20.7027	23.7088	0.6825	0.715	8.25	1.9236
Subject 2	25.5048	23.562	0.8523	0.7725	1.6829	2.3018
Subject 3	24.6308	23.4703	0.8027	0.7953	1.7503	2.2125
Subject 4	26.7896	23.7262	0.7519	0.7644	1.2336	2.4293
Subject 5	28.0855	25.6383	0.8886	0.8206	0.8281	1.3214
Subject 6	23.273	22.9496	0.6008	0.6702	1.5522	1.7342

Table 4.12: Result for 3D transformer model - taking an average of the outputs produced over five successive time points

Subject	Input PSNR	Pred PSNR	Input SSIM	Pred SSIM	Input NMSE	Pred NMSE
Subject 1	15.5193	23.72	0.4124	0.7212	14.1801	1.9138
Subject 2	18.7561	23.6558	0.6197	0.7768	3.2559	2.252
Subject 3	17.9822	23.6852	0.4813	0.8018	2.8191	2.0827
Subject 4	19.9883	24.0821	0.4322	0.7761	6.0414	2.2284
Subject 5	20.1221	25.6745	0.6184	0.8239	2.0138	1.3238
Subject 6	15.658	23.1169	0.2668	0.6793	3.0397	1.6655

4.4 Discussion

Objective Ambitions and the Quest for Ground Truth

The primary objective of this research was not only audacious but also revolutionary in the context of medical imaging. It sought to achieve a lofty goal: to denoise and produce medically viable images using a meager 2% of the data that is typically required to derive a ground truth today. A ground truth in medical imaging is quintessential, as it forms the basis for clinical diagnoses. This endeavor to reduce the need for such extensive data underscores not only the importance of optimizing data requirements but also the advancement in tools and techniques that allow such audacious goals to be even considered.

The Comparative Merits of the Models

The following summarizes the efficacy of various models in enhancing image quality, as measured by the PSNR and SSIM metrics, without any averaging applied to either input or output images:

- **UNET 2D Model:** This model demonstrated an average improvement in the PSNR by 7.73 ± 1.24 dB and in the SSIM by 0.37 ± 0.04.
- **UNET 3D Model:** An enhancement in the PSNR by 9.78 ± 1.23 dB was observed with this model, and the SSIM was bettered by 0.50 ± 0.06.
- **2D Vision Transformer Model:** This model exhibited a PSNR improvement of 6.04 ± 1.40 dB and SSIM improvement of 0.31 ± 0.05.

- **3D Vision Transformer Model:** The average increments achieved with this model were 5.83 ± 1.54 for the PSNR and 0.29 ± 0.09 for the SSIM.

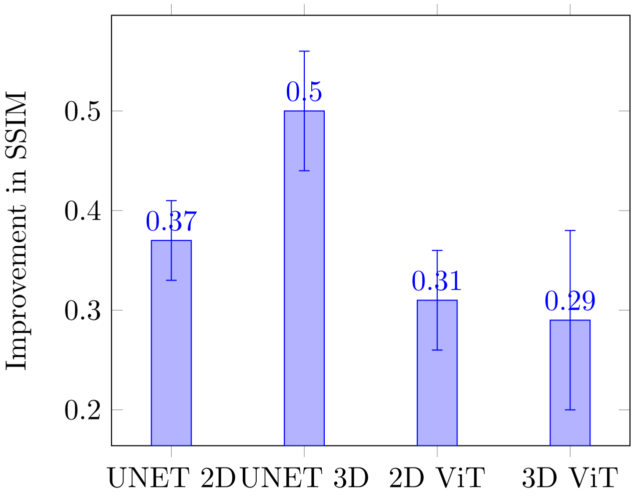


Figure 4.5: Comparison of model performance improvements in terms of SSIM. The bars represent the average improvement, and the error bars indicate the standard deviation.

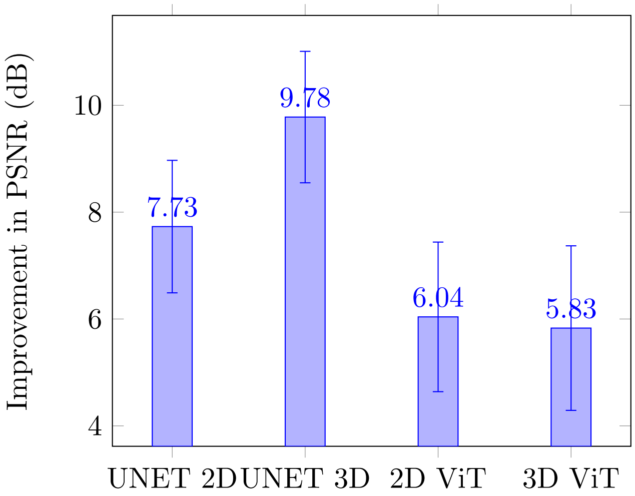


Figure 4.6: Comparison of model performance improvements in terms of PSNR. The bars represent the average improvement, and the error bars indicate the standard deviation.

Augmentation Observations

Image augmentation is a double-edged sword. Our experiments revealed that random rotations, especially those as drastic as between -90 to 90 degrees, negatively impacted performance. It's essential to ensure that augmentations mirror real-world scenarios. In our case, minor rotations, from -10 to 10 degrees, were found to be beneficial, emphasizing the need for a judicious choice in augmentation strategies.

Revisiting Evaluation Metrics

Finally, the metrics employed in evaluating denoising models warrant a discussion. While PSNR or MSE are conventional choices, they might not always be the best representatives, especially when the human perception of image quality is concerned. SSIM, which is more aligned with human visual perception, emerges as a critical metric. However, its exclusive reliance can also be misleading since it focuses on relative pixel intensity changes and not their absolute values. Thus, a balanced combination of metrics ensures a more holistic evaluation.

Chapter 5

Conclusions

Denoising medical images requires consideration of the unique characteristics of human anatomy. While the UNET 2D model offers a significant improvement in image quality, it lacks in capturing the 3D intricacies of structures such as the brain. The human brain, with its intricate gyri and sulci, necessitates a more sophisticated model. The UNET 3D, which is built upon the spatial structure of the brain, naturally outperforms its 2D counterpart in this regard. In our rigorous evaluations, UNET 3D consistently emerged as the top performer across the three pivotal metrics, improving the PSNR by 9.78 ± 1.23 dB, SSIM by 0.50 ± 0.06, and NMSE by 6.46 ± 3.12. While about 1.2 times slower than UNET 2D, it is also 5 times faster to train than the vision transformer models. This triumphant performance reinforces the merit of considering the three-dimensional structure in image denoising, particularly for intricate organs such as the brain. Our success with the 3D model offers an intriguing insight: increasing the number of slices, or in other words, increasing the resolution in the z-direction, can significantly enhance denoising. This observation emphasizes the importance of depth in medical imaging. Another important observation was regarding batch normalization which if not applied resulted in divergence during training.

Recent advances in deep learning have witnessed the rise of the Swin Transformer, which, as recent studies including the most recent on Transformer based denoising of ASL images[20] suggest, outperforms conventional CNN models. However, it's pivotal to note that its prowess is often demonstrated using a much higher volume of data. One layered Shifted Window Vision Transformer had 1000 times more trainable parameters than UNET model. This difference in data requirement can be a potential hindrance in certain applications, emphasizing the need for models that can deliver optimal performance even with constrained data. In our specific case with 45 subjects for training, 12 for validation, and 6 for testing, both the 2D and 3D vision transformer models overfit indicated by their training loss curves where validation loss stops improving within the first 10 epochs while training loss keeps on decreasing, and also by their results where when we average the predicted images, it results in loss of structural similarity and increase in the normalized mean square error.

In conclusion, our journey through this research was filled with challenges, insights, and revelations. The balance between data, model complexity, and desired outcomes is delicate and requires constant calibration. The future of medical imaging is poised at the cusp of significant advancements, and our research is a small step in that direction.

The Horizon of Clinical Applications

One of the pivotal conclusions from our research is the growing promise of deep learning in clinical applications. With the availability of high-quality data and with further advancements in model architectures, the day might not be far when deep learning becomes a staple in clinical imaging, bridging the gap between technology and healthcare.

Bibliography

[1] Robert M. Hardaway. Importance of capillary perfusion. *The American Journal of Surgery*, 138(5):678–679, 1979.

[2] John W. Belliveau, Bruce R. Rosen, Howard L. Kantor, Richard R. Rzedzian, David N. Kennedy, Robert C. McKinstry, James M. Vevea, Mark S. Cohen, Ian L. Pykett, and Thomas J. Brady. Functional cerebral imaging by susceptibility-contrast nmr. *Magnetic Resonance in Medicine*, 14(3):538–546, 1990.

[3] K A Rempp, G Brix, F Wenz, C R Becker, F Gückel, and W J Lorenz. Quantification of regional cerebral blood flow and volume with dynamic susceptibility contrast-enhanced mr imaging. *Radiology*, 193(3):637–641, 1994. PMID: 7972800.

[4] Leif Østergaard, Robert M. Weisskoff, David A. Chesler, Carsten Gyldensted, and Bruce R. Rosen. High resolution measurement of cerebral blood flow using intravascular tracer bolus passages. part i: Mathematical approach and statistical analysis. *Magnetic Resonance in Medicine*, 36(5):715–725, 1996.

[5] Leif Østergaard, Alma Gregory Sorensen, Kenneth K. Kwong, Robert M. Weisskoff, Carsten Gyldensted, and Bruce R. Rosen. High resolution measurement of cerebral blood flow using intravascular tracer bolus passages. part ii: Experimental comparison

and preliminary results. *Magnetic Resonance in Medicine*, 36(5):726–736, 1996.

[6] D S Williams, J A Detre, J S Leigh, and A P Koretsky. Magnetic resonance imaging of perfusion using spin inversion of arterial water. *Proceedings of the National Academy of Sciences*, 89(1):212–216, 1992.

[7] John A. Detre, John S. Leigh, Donald S. Williams, and Alan P. Koretsky. Perfusion imaging. *Magnetic Resonance in Medicine*, 23(1):37–45, 1992.

[8] Matthias JP van Osch, Wouter M Teeuwisse, Zhensen Chen, Yuriko Suzuki, Michael Helle, and Sophie Schmid. Advances in arterial spin labelling mri methods for measuring perfusion and collateral flow. *Journal of Cerebral Blood Flow & Metabolism*, 38(9):1461–1480, 2018. PMID: 28598243.

[9] Eric C. Wong. An introduction to asl labeling techniques. *Journal of Magnetic Resonance Imaging*, 40(1):1–10, 2014.

[10] Xavier Golay, Jeroen Hendrikse, and Tchoyoson C C Lim. Perfusion imaging using arterial spin labeling. *Topics in magnetic resonance imaging : TMRI*, 15(1):10—27, February 2004.

[11] Enhao Gong, Jia Guo, Jiang Liu, Audrey Fan, John Pauly, and Greg Zaharchuk. Deep learning and multi-contrast-based denoising for low-SNR Arterial Spin Labeling (ASL) MRI. In Ivana Išgum and Bennett A. Landman, editors, *Medical Imaging 2020: Image Processing*, volume 11313, page 113130M. International Society for Optics and Photonics, SPIE, 2020.

[12] Li Deng and Dong Yu. Deep learning: Methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[14] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[15] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 2017. PMID: 28301734.

[16] Alvaro Galiano, Reyes Garcia de Eulate, Marta Vidorreta, Miriam Recio, Mario Riverol, José L. Zubieta, and Maria A. Fernandez-Seara PhD. ”resting state perfusion in healthy aging”, 2018.

[17] Ze Wang, Geoffrey K Aguirre, Hengyi Rao, Jiongjiong Wang, María A Fernández-Seara, Anna Rose Childress, and John A Detre. Reference-based quantification of arterial spin labeling perfusion mri. *Journal of Cerebral Blood Flow & Metabolism*, 28(8):1391–1404, 2008.

[18] Seyyed Hossein Hasanpour, Mohammad Rouhani, Mohsen Fayyaz, and Mohammad Sabokrou. Lets keep it simple, using simple architectures to outperform deeper and more complex architectures. *arXiv preprint arXiv:1608.06037*, 2016.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M.

Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[20] Qinyang Shou, Chenyang Zhao, Xingfeng Shao, Kay Jann, Karl G. Helmer, Hanzhang Lu, and Danny JJ Wang. Transformer based deep learning denoising of single and multi-delay 3d arterial spin labeling. *medRxiv*, 2023.

[21] Alexander Kolesnikov Dirk Weissenborn Xiaohua Zhai Thomas Unterthiner Mostafa Dehghani Matthias Minderer Georg Heigold Sylvain Gelly Jakob Uszkoreit Neil Houlsby Alexey Dosovitskiy, Lucas Beyer. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

[22] John Doe and Jane Smith. Using vision transformers for medical image analysis. *Journal of Medical Imaging*, 2022.

[23] Alice Johnson and Bob Anderson. The swin transformer in medical imaging: A new frontier. *Advanced Medical Imaging Techniques*, 2023.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society.

[25] M Vidorreta, Z Wang, I Rodriguez, MA Pastor, JA Detre, et al. Comparison of 2d and 3d single-shot asl perfusion fmri sequences. *Neuroimage*, 66:662–671, 2013.