MR Contrast Image Synthesis Using Deep Learning

THESIS

Submitted in Partial Fulfillment of

the Requirements for

the Degree of

MASTER OF SCIENCE (Electrical Engineering)

at the

NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING

by

Parisima Abdali

May 2024

MR Contrast Image Synthesis Using Deep Learning

THESIS

Submitted in Partial Fulfillment of

the Requirements for

the Degree of

MASTER OF SCIENCE (Electrical Engineering)

at the
NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING

by

Parisima Abdali

May 2024

Approved: Wang Yao

Advisor Signature

5/11/2024

Date

Ivan Selesnick

Department Chair Signature

Ivan Selesnick

Date

Approved by the Guidance Committee:

Major: Electrical and Computer Engineering

Wang Yao

Yao Wang

Professor
Electrical and Computer Engineering

Date: 5/11/2024

lifeng

Li Feng

Associate Professor
Department of Radiology

Date: 05/13/2024

Chinmay

Chinmay D Hegde

Associate Professor
Electrical and Computer Engineering

Date: 5/10/2024

Microfilm or other copies of this thesis are obtainable from

UMI Dissertation Publishing
ProQuest CSA
789 E. Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Vita

Parisima Abdali was born and raised in the historically and culturally rich city of Tehran, Iran. From a young age, she exhibited the curiosity of an engineer, always keen to understand how things worked and how they could be improved. This innate curiosity laid the foundation for her academic pursuits.

In September 2016, Parisima began her undergraduate degree in Electrical Engineering, specializing in telecommunications, at Imam Khomeini International University. Her dedication and passion for her field were evident as she excelled in her studies, earning the distinction of being ranked first in her graduating class.

Pursuing her passion for signal processing and its applications in deep learning, Parisima embarked on her graduate studies at New York University (NYU) in September 2022. Here, she immersed herself in the study of Machine Learning (ML) and Deep Learning (DL), with a keen interest in their application in the complex domain of medical image analysis.

From September 2023 to May 2024, Parisima dedicated herself to working on her master's thesis. During this period, she joined the NYU Video Lab and Rapid Lab, focusing her research on the innovative use of Generative AI in MR contrast image synthesis.

Acknowledgments

I would like to extend my deepest appreciation to my thesis advisor, Prof. Yao Wang, for her continuous support, insightful feedback, and encouragement throughout my research journey and the development of this thesis. Prof. Wang's mentorship and critical approach to research have been pivotal in honing my ideas and refining my work. Her dedication to excellence and her encouragement to push beyond the limits have been truly inspiring.

My gratitude also goes to Prof. Li Feng and Prof. Chinmay Hegde for their invaluable contributions to my committee. Their expert advice and constructive criticism have played a significant role in guiding the direction of my research and enhancing the caliber of my thesis.

I am particularly thankful to Lavanya Umapathy, whose previous work on this project laid the groundwork for my research. Lavanya's generosity in sharing her expertise and insights has significantly propelled my work forward. I also wish to express my thanks to Nikola Janjušević, a peer whose guidance has been instrumental in my research. Nikola's support and shared wisdom have enriched my academic experience immeasurably.

Lastly, my heartfelt thanks to my parents, whose love, support, and unwavering belief in me have been the backbone of my endeavors. Their endless encouragement has been my source of motivation, and I could not have embarked on and sustained through this journey without them.

Dedicated to my mother, my unwavering light of love and support

ABSTRACT

MR Contrast Synthesis Using Deep Learning

by

Parisima Abdali

Advisor: Prof. Yao Wang, Ph.D.

Submitted in Partial Fulfillment of the Requirements for

the Degree of Master of Science (Electrical Engineering)

May 2024

The synthesis of T1-weighted contrast-enhanced (T1CE) MR images is essential for accurate brain pathology diagnosis while mitigating the health risks from gadolinium-based agents. This thesis introduces a novel two-stage deep learning approach using the BraTS2021 dataset, employing a modified U-Net model for both feature extraction and image synthesis. Our approach, incorporating constrained contrastive learning (CCL) in the full decoder, demonstrated improvements over the baseline. P-values from hypothesis testing with a 5% error rate suggest that the CCL model is significantly better than the baseline in terms of PSNR, SSIM, LPIPS (AlexNet) and LPIPS (VGG). These results highlight the potential of our methodology to significantly enhance diagnostic accuracy and patient safety in clinical settings.

Table of Contents

List of Figures

List of Tables

1. Introduction

In the evolving landscape of medical imaging, enhancing diagnostic accuracy while minimizing patient discomfort and health risks remains a primary objective. Magnetic Resonance Imaging (MRI) plays a pivotal role in modern diagnostics, providing detailed images of tissues, organs, and other internal structures. However, the traditional use of gadolinium-based contrast agents in T1-weighted contrast-enhanced (T1CE) MRI poses challenges, including potential allergic reactions and gadolinium accumulation in the body. Recognizing these challenges, our project aims to revolutionize the field by advancing multi-parametric MRI synthesis, thereby reducing the dependency on multiple scans and contrast agent injections.

The central objective of this project is to diminish the need for gadolinium injections, which are standard for obtaining T1CE images that enhance the visibility of pathological features such as tumors or inflammation. Our novel approach leverages Constrained Contrastive Learning [11], a technique originally devised for segmentation tasks and adapted from existing methodologies in recent studies, to synthesize T1CE images effectively. By employing this innovative technique, we aim to replicate the diagnostic features visible in T1CE images by using the intrinsic tissue properties evident in other MR contrast images.

The hypothesis driving our research posits that understanding and differentiating the local representation of tissue properties in MR-contrast images can significantly aid in synthesizing high-fidelity T1CE images. This approach not only promises to maintain the diagnostic capabilities of traditional contrast-enhanced imaging but also aims to overcome the associated health risks and patient discomfort. Furthermore, the precise reconstruction of tumor regions in synthesized T1CE images showcases the potential of this method in clinical applications.

This thesis explores the challenges associated with traditional T1CE imaging, including managing allergic reactions and artifacts from excessive contrast, and details the development and validation of our synthesis technique. By addressing these challenges, our project aspires to enhance the overall safety and efficacy of MRI procedures, contributing to the broader field of medical imaging.

1.1 Background

Magnetic Resonance Imaging (MRI) [7] is a pivotal diagnostic tool in neurology and neurosurgery, renowned for its superior delineation of anatomical structures of the brain, spinal cord, and vascular systems. Unlike other imaging modalities, MRI excels in providing detailed images across all three anatomical planes: axial, sagittal, and coronal, offering a comprehensive view essential for accurate diagnosis and treatment planning.

1.1.1 Physics of MRI

The underlying technology of MRI revolves around the magnetization properties of atomic nuclei, particularly hydrogen protons found in water molecules within body tissues. The process begins with the application of a strong, uniform external magnetic field, aligning the randomly oriented protons. This alignment is temporarily disrupted by an external Radio Frequency (RF) pulse, causing the protons to absorb energy and deviate from their alignment. As they return to their original state, they emit RF signals, which are captured and analyzed.

The emitted RF signals vary in frequency depending on their location within the body, and these differences are converted into image data through Fourier transformation. This data is then represented as varying shades of gray in a pixelated image. Crucially, by altering the sequence of RF pulses—characterized by variables like Repetition Time (TR) and Time to Echo (TE)—different image contrasts and types can be generated, enhancing the ability to distinguish between different types of tissues and pathologies.

1.1.2 Tissue Characterization

As mentioned, MRI imaging sequences are crucial for obtaining detailed images that highlight different tissue properties and pathologies. The most commonly used sequences in clinical practice are T1-weighted and T2-weighted scans. The four contrast images of interest, which provide the tissue information, are represented in Figure 1.1.

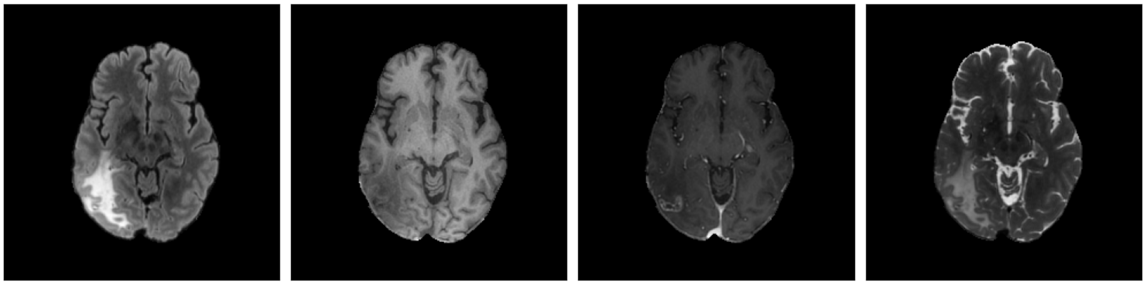


Figure 1.1: Examples of the brain different MR contrast images from BraTS 2021 [2] training dataset. From left to right: T2-FLAIR, T1-weighted, T1-Gd and T2-weighted.

T1-Weighted Images

T1-weighted images are generated using short Time to Echo (TE) and Repetition Time (TR) durations. The resulting images provide high contrast based on the T1 relaxation properties of tissues. In these images, cerebrospinal fluid (CSF) appears dark, while fat and subacute hemorrhage show up as bright, making T1-weighted imaging excellent for assessing the integrity of the blood-brain barrier and for visualizing fatty tissues.

T2-Weighted Images

Conversely, T2-weighted images are produced with longer TE and TR times, which highlight the T2 relaxation properties of tissues. These images make CSF appear bright, which is beneficial for evaluating brain edema, inflammation, and infection. The high contrast for fluid makes T2-weighted scans particularly useful for detecting pathologies filled with fluid, like cysts and tumors.

FLAIR Sequences

Another key sequence is the Fluid Attenuated Inversion Recovery (FLAIR), which is a modification of the T2-weighted scan where TE and TR times are extended to suppress the bright signal of normal CSF. This results in a dark appearance of CSF, enhancing the contrast and visibility of abnormalities such as lesions, which remain bright. FLAIR sequences are indispensable for differentiating between normal fluid spaces and pathological changes.

Gadolinium-Enhanced Imaging

To further enhance diagnostic capability, T1-weighted imaging can be performed with the injection of Gadolinium (Gad), a non-toxic paramagnetic contrast agent, as shown in Figure 1.2.. Gad shortens the T1 relaxation time, causing tissues where Gad accumulates to appear very bright. This property is especially useful for visualizing vascular abnormalities and disruptions in the blood-brain barrier, such as tumors, abscesses, and inflammatory conditions like herpes simplex encephalitis and multiple sclerosis.

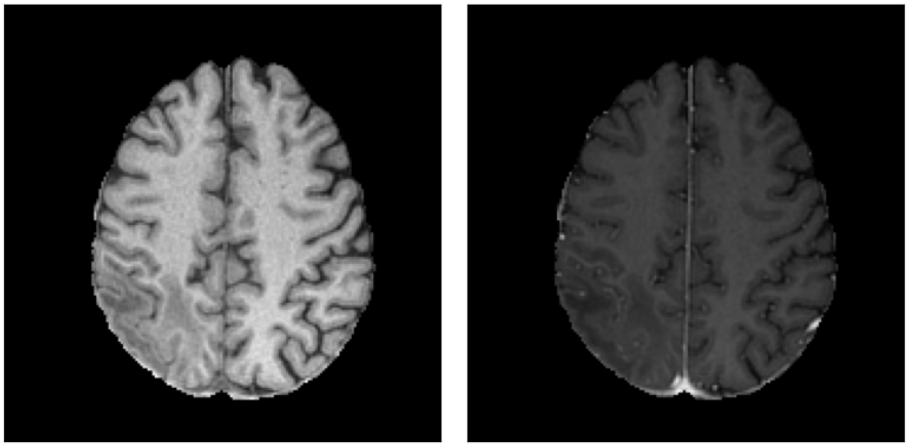


Figure 1.2: Difference between a T1-weighted image (left) and T1 after Gadolinium injection (right)

Table 1.1: Appearance of Brain Tissues in MRI Sequences

Tissue	**T1-Weighted**	**T2-Weighted**	**FLAIR**
CSF	Dark	Bright	Dark
White Matter	Light	Dark Gray	Dark Gray
Gray Matter	Gray	Light Gray	Light Gray
Blood Vessels	Variable	Variable	Variable
Fat (within bone marrow)	Bright	Light	Light
Inflammation (infection, demyelination)	Dark	Bright	Bright

1.2 Training and Evaluation

1.2.1 Dataset Split

To facilitate our synthesis process, we employed a comprehensive and diverse dataset known as Brain Tumor Segmentation (BraTS2021) [2][3][4][5], which comprises a rich collection of MR contrast images suitable for our research. Originally, the dataset included data from 1660 subjects. Following rigorous data cleaning to ensure quality and consistency, we narrowed down the dataset to 838 subjects. The subjects were randomly allocated into different subsets for our deep learning pipeline: 360 for training, 245 for validation, and 70 for testing, with each subset containing 70 slices, viewed as individual samples. The training process was conducted in 2D, treating each slice within the MR scans as an independent image. This approach allowed us to focus on the nuanced differences and similarities between slices, essential for the effective training of our model on local tissue representations and enhancing its capability to synthesize T1CE images accurately.

1.2.2 MR-Contrast Synthesis Using Deep Learning

Our synthesis approach for MR contrast images consists of a two-stage deep learning architecture, primarily based on the U-Net model, which is well-suited for medical image processing due to its ability to handle complex visual data with high efficiency. The first stage of our methodology involves a pretraining phase (6.1) where we employ a constrained contrastive learning methodology. This technique allows the model to effectively extract tissue-specific information from various types of MR images, such as T1-weighted, T2-weighted, and T2-FLAIR. The extracted features are crucial as they form the foundation for the subsequent synthesis stage.

In the synthesis stage (6.2), we leverage the U-Net model, which has been pretrained and whose weights have been refined to enhance their effectiveness for specific tasks. These refined weights are then applied in the synthesis of MR contrast images, particularly T1CE images, as part of the downstream tasks. To optimize the synthesis quality, we experiment with various loss functions, including perceptual loss that incorporates elements of the VGG16 network. This approach ensures that the synthesized images closely resemble the target images, both in terms of quality and diagnostic relevance.

Additionally, we explore two decoder configurations within our U-Net architecture: a full decoder and a partial decoder. For more details about the variations in decoder configurations, please refer to Section (6.1.1) of this document.

1.2.3 Evaluation

We use different metrics to compare three different models—baseline, CL-full, and CL-partial—to assess their performance in MR synthesis. These metrics include mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and convolutional neural network-based metrics such as the Learned Perceptual Image Patch Similarity (LPIPS) [12], employing both AlexNet and VGG architectures. The effectiveness of our full decoder configuration was substantiated through both visual assessments and statistical analyses, particularly using p-values, which confirmed its superior ability in generating T1CE images that closely match the ground truth.

2. BraTS 2021 Dataset

2.1 Background

2.1.1 Overview and Organization

The Brain Tumor Segmentation (BraTS) 2021 Dataset[2][8] is a significant contribution to the field of brain glioma segmentation, organized by the collaborative efforts of the Radiological Society of North America (RSNA), the American Society of Neuroradiology (ASNR), and the Medical Image Computing and Computer Assisted Interventions (MICCAI) society. Its establishment marks a cornerstone in providing a standardized benchmark for the evaluation of brain tumor segmentation algorithms. Renowned for its extensive assortment of multi-institutional, multi-parametric Magnetic Resonance Imaging (mpMRI) data, the dataset facilitates the development of cutting-edge diagnostic tools.

2.1.2 Data Structure

The dataset includes a variety of MRI modalities, co-registered for each subject, with each modality serving a unique purpose in brain tumor analysis [3], [4]. These modalities, which include a mix of pre- and post-therapy brain scans, some displaying resections, were acquired from four distinct centers using MR scanners of varied vendors, field strengths, and imaging sequence implementations. Detailed MRI modality specifications, [8]:

- **T1 (T1-weighted, Native Image):** Acquired in either sagittal or axial 2D planes, these images have a slice thickness ranging from 1 to 6 mm, providing a base view of the brain's anatomy without contrast enhancement (Figure 2.1, second column from left).

- **T1c (T1-weighted, Contrast-Enhanced):** These images are obtained using 3D acquisition techniques, enhanced with Gadolinium contrast to highlight pathological areas. For most patients, these images have a 1 mm isotropic voxel size, offering

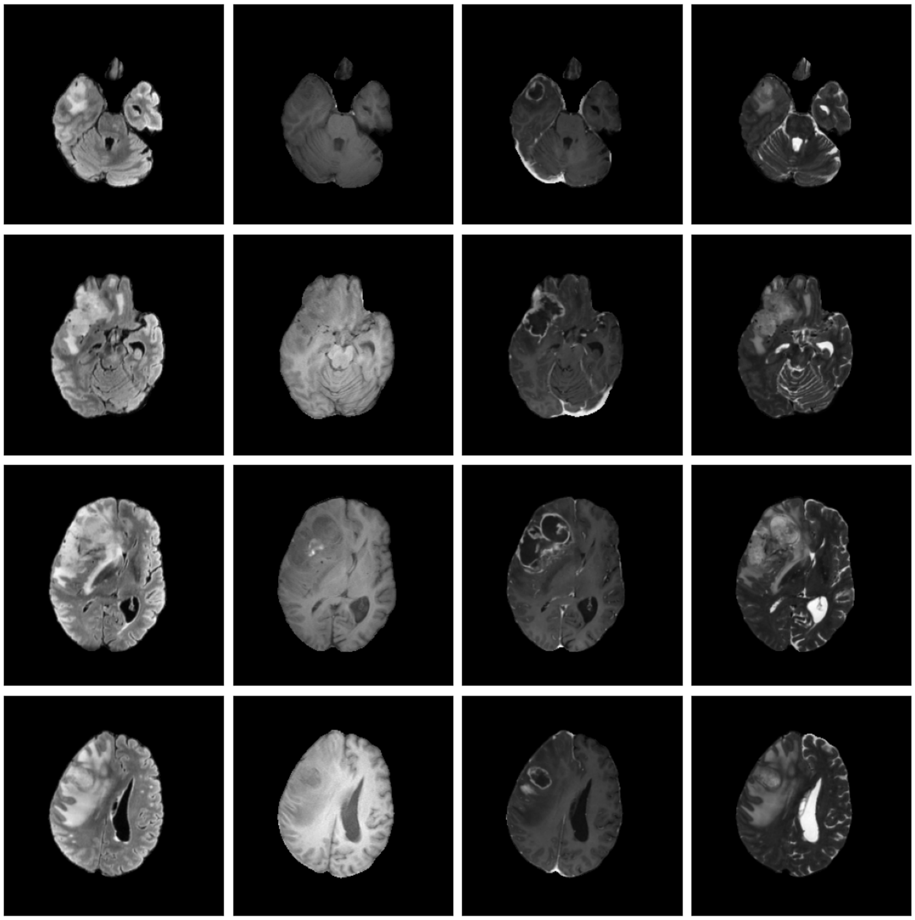


Figure 2.1: Figure 2.1 illustrates four different slices of a brain image from one of the subjects in the BraTS 2021 training dataset. Each row corresponds to a different slice, with columns arranged from left to right showcasing MR contrasts: T2-FLAIR, T1-weighted, T1CE, and T2-weighted. [2].

high-resolution insights into tumor regions (Figure 2.1, third column from left).

- **T2 (T2-weighted Image):** Acquired in axial 2D acquisition, these images have a slice thickness of 2 to 6 mm. They are particularly useful for visualizing the brain's water content and edema surrounding tumors (Figure 2.1, forth column from left).

- **FLAIR (T2-weighted FLAIR Image):** These images can be acquired in axial, coronal, or sagittal 2D planes with a slice thickness of 2 to 6 mm. FLAIR imaging provides valuable information by suppressing the fluid signal, making it easier to detect lesions near the ventricles (Figure 2.1, first column from left).

To ensure uniformity across the dataset, they rigidly aligned each subject's image volumes to the T1c MRI, which typically had the highest spatial resolution. Subsequently, all images were resampled to a 1 mm isotropic resolution in a standardized axial orientation using a linear interpolator. Furthermore, skull stripping was applied to all images to ensure the anonymity of the patients [8].

2.1.3 Reflecting Clinical Diversity

The BraTS dataset encapsulates a retrospective collection of brain tumor mpMRI scans from various institutions, captured under standard clinical conditions but using different equipment and protocols [2]. Specifically, these scans were acquired at four distinct centers—Bern University, Debrecen University, Heidelberg University, and Massachusetts General Hospital—over several years. The use of MR scanners from different vendors, with varying field strengths (1.5T and 3T) and implementations of the imaging sequences (e.g., 2D or 3D), contributes to this diversity. As a result, the dataset exhibits heterogeneous image quality, reflecting the wide range of clinical practices across institutions [8].

2.1.4 Data Preparation

For our synthesis task, we utilized the coregistered multiparametric images (T1w, T1Gd, T2w, T2-FLAIR) to construct a contrast space. This space enabled the generation of a constraint map to learn and discriminate various information types. The selection of images for the constraint map was tailored based on the target image modality. As part of our dataset preparation, we performed preprocessing (details are presented in Chapter 7) to enhance the quality and uniformity of the images. Subsequently, we randomly selected training and validation datasets to ensure a broad representation of the data. These datasets, along with their generated constraint maps, were stored in two separate HDF5 files. This structured approach allowed for efficient access and management during further pretraining and downstream tasks.

3. Constraint Contrastive Learning

3.1 Background

3.1.1 Introduction to Representational Learning

The efficacy of machine learning methodologies is fundamentally linked to the data representation (or features) upon which they operate. A significant portion of the effort in deploying machine learning systems is therefore dedicated to the creation of preprocessing pipelines and data transformations. These transformations are designed to yield a data representation that is conducive to effective machine learning processes. This necessity underscores a critical weakness in current learning algorithms—their limited capacity to autonomously extract and organize discriminative information from the data [6].

Traditionally, overcoming this limitation involves substantial feature engineering, a process where human ingenuity and prior knowledge are leveraged to enhance machine learning models. Although effective, feature engineering is a demanding and labor-intensive process. This process is crucial as it compensates for the shortcomings of current algorithms by generating data representations that aid in the extraction of valuable information, necessary for developing classifiers and other predictive models. Nowadays, various deep learning models are employed to address this limitation and are used as part of pretraining for further downstream tasks. These models automate much of the feature extraction and can learn powerful representations directly from large datasets, thus potentially reducing the need for intensive manual feature engineering.

3.1.2 Contrastive Learning

Building upon the advancements in representational learning, contrastive learning stands as a specialized form of representational learning that leverages unlabeled data to forge powerful feature representations, enhancing performance in a variety of machine learning tasks. This methodology utilizes the availability of semantically similar data pairs and negative samples to refine the quality of data representations [1].

In practice, contrastive learning employs a framework where each data point is paired both with ”positive” samples (similar data points) and ”negative” samples (dissimilar data points). This differential approach is guided by the contrastive loss, which optimizes the weights θ of a deep learning model $\Psi_\theta(.)$ such that the distance between latent representations $(\Psi_\theta(x_a), \Psi_\theta(x_b))$ of pairs of inputs (x_a, x_b)approximates their semantic similarity in input space. Consequently, semantically similar data points (positive samples) are pushed closer together in the representational space, while being distanced from dissimilar data points (negative samples) [11], as an example look at Figure 3.1. This not only amplifies the semantic distinctions between dissimilar data points but also reinforces the similarity among comparable data points.

The theoretical underpinnings of contrastive learning are predicated on the concept of latent classes. It is hypothesized that semantically similar data points are likely drawn from the same latent class. By assuming the existence of these latent classes, the method provides a structured mechanism to learn representations that are not only discriminative but also inherently aligned with the latent semantic structures within the dataset. Such discriminative power is crucial for enhancing the efficacy of classifiers and other predictive models in downstream tasks [1].

3.2 Methodology

In this work, we fully adopt the innovative constrained contrastive learning (CCL) framework introduced by Umapathy et al [11]. in their study on segmentation tasks, applying it to synthesis tasks in MR images. Their original research, detailed in the article ”Reducing annotation burden in MR: A novel MR-contrast guided contrastive learning approach for image segmentation”, highlights how MR imaging offers flexible contrast mechanisms (such as T1, T2, diffusion, etc.) that can be controlled using image acquisition parameters

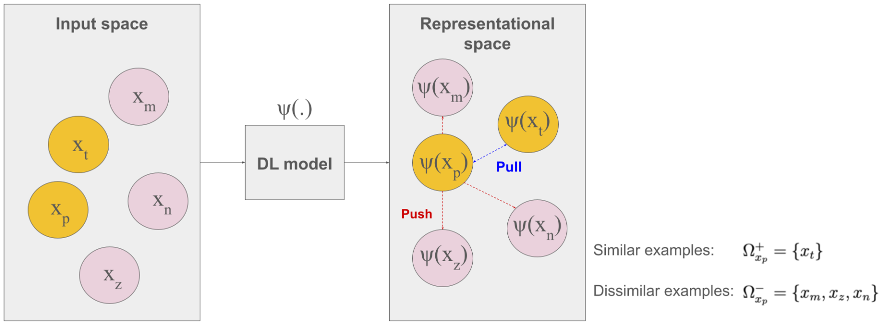

Input space
x_m
x_t
x_p
x_n
x_z
ψ(.)
DL model
Representational space
ψ(x_m)
ψ(x_t)
ψ(x_p)
Pull
Push
ψ(x_n)
ψ(x_z)
Similar examples: $\Omega^+_{x_p} = \{x_t\}$
Dissimilar examples: $\Omega^-_{x_p} = \{x_m, x_z, x_n\}$

Figure 3.1: Transforming data from input space to representational space via a deep learning model, where similar data points are attracted ('Pulled') closer together, while dissimilar data points are repelled ('Pushed') apart, exemplifying the process of contrasting similar and dissimilar examples.

to better characterize underlying tissues. We extend this framework to focus on the synthesis of MR images, leveraging the same MR contrast information via a constraint map to embed tissue-specific details into the representational space. This approach not only maintains the integrity of the original framework but also explores its potential to (1) generate precise tissue characterizations and (2) improve the efficacy of synthesis operations on MR contrast images (such as T1-Gd or T2-FLAIR images).

3.2.1 MR Image contrast space

In our synthesis application for MR imaging, we aim to autonomously detect local areas with semantic consistency by leveraging an associated collection of multi-contrast MR data, termed the contrast space. These regions, characterized by similar underlying tissue attributes, are expected to display analogous signal profiles within this contrast space. The signal contrast in an MR image, denoted as s, is formulated by $f(u, v, \lambda, \phi)$, where (u, v) are the spatial coordinates. Here, λ encapsulates the tissue-specific parameters, while ϕ includes additional imaging parameters such as longitudinal relaxation time $T1$, transverse relaxation time $T2$, or Apparent Diffusion Coefficient, among other MR imaging settings like repetition time (TR), echo time (TE), inversion time (TI), and flip angles [11].

Our hypothesis suggests that local $p \times p$ sections within an MR image, sharing akin λ values, will manifest similar representational features. To pinpoint such $p \times p$ sections within a given image s, we utilize a series of related MR contrast images $S = \{s_i = f(u, v, \lambda, \phi_i)\}_{i=1}^{N}$, where N represents the number of contrasts. The contrast in both the primary MR image and its corresponding contrast space is modulated by the identical tissue-specific parameter λ.

3.2.2 Constraint Maps

Regions within contrast images that share related λ values will naturally reflect similar signal profiles across the contrast space, serving as a proxy for the λ tissue information. We encode this tissue information within the contrast space by conducting a pixelwise principal component analysis on S, retaining fewer than N principal components to reduce dimensionality and filter out noise. Subsequently, an unsupervised clustering technique like K-means, employing K clusters, is applied to these principal component images to create the constraint map C. This map classifies each pixel according to the underlying tissue parameter λ, providing a structured characterization. Please refer to Figure 3.2 to see an example of the generated constraint map from given MR contrast images.

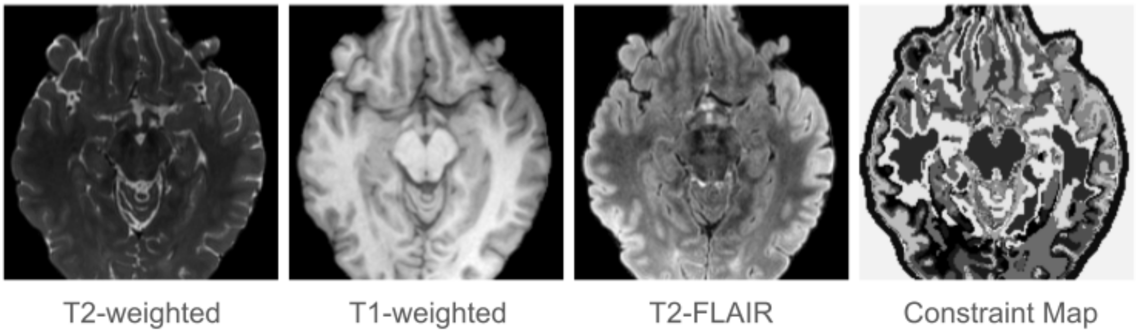


Figure 3.2: Multiparametric images displaying tissue-specific information, accompanied by the generated constraint map.

3.2.3 Strategy and Loss Function

Consider an MR image s and its corresponding constraint map C. The feature representation $\Psi_\theta(s)$ is obtained by processing s through the deep learning model. Define M as the collection of all non-overlapping $p \times p$ patches within s. We select a patch m_a from M at random, which has a feature dimension $\Psi(m_a) \in R^{p \times p \times D}$ where D represents the dimensionality of the space. The following description outlines our contrasting approach to determine the sets of positive ($\Omega^+_{m_a}$) and negative ($\Omega^-_{m_a}$) patches that facilitate the learning of distinct local features.

During training sessions, for each patch m_a, we identify all other patches m_i within s where the similarity in their feature representations surpasses a preset threshold, $d(\Psi(m_a), \Psi(m_i)) \geq s_{\text{thresh}}$. This gives us the set of "representationally similar" embeddings R_{m_a}.

$$R_{m_a} = \{m_i : d(\Psi(m_a), \Psi(m_i)) \geq s_{\text{thresh}}\}$$

(3.1)

We assess similarity in the representational space using the l2-normalized cosine similarity metric. As an alternative approach, the set R_{m_a} can also be constructed by selecting the top-K nearest neighbors in the representational space that are most similar to $\Psi(m_a)$.

Next, we use the constraint map to determine the "parametrically similar" embeddings Λ_{m_a}, those sharing identical signal characteristics and potentially the same tissue parameter λ as m_a:

$$\Lambda_{m_a} = \{m_j : \beta(m_j) = \beta(m_a)\}$$

(3.2)

In this context, $\beta(m_a)$ indicates the predominant class among the pixels within the patch m_a on the constraint map.

The positive and negative neighborhoods are defined through the following constraints for each learning iteration:

$$\Omega^{+}_{m_a} = R_{m_a} \cap \Lambda_{m_a} \tag{3.3}$$

$$\Omega^{-}_{m_a} = R_{m_a} \cap \Lambda^{c}_{m_a} \tag{3.4}$$

Where $\Lambda^{c}_{m_a}$ is the complement of Λ_{m_a}. This strategy, which they refer as constrained contrastive learning [11], ensures that patches that are both representationally and parametrically similar to m_a are grouped closer together, as it is shown in Figure 3.3.

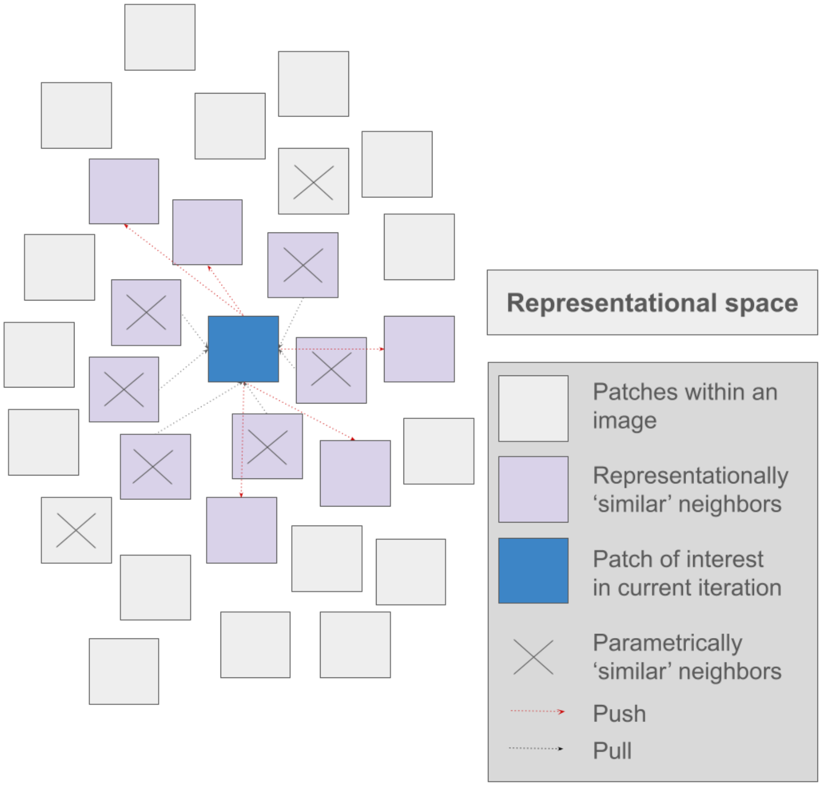

Representational space
Patches within an image
Representationally 'similar' neighbors
Patch of interest in current iteration
Parametrically 'similar' neighbors
Push
Pull

Figure 3.3: Representational space of a selected patch, showing the identification of both representationally and parametrically similar neighbors.

The contrastive loss $L(m_a)$ for a patch m_a is calculated as follows:

$$L(m_a) = \frac{1}{|\Omega^+_{m_a} m_a|} \sum_{m_i \in \Omega^+_{m_a}} l(m_a, m_i) \tag{3.5}$$

$$l(m_a, m_i) = -\log \left(\frac{e^{d(\Psi(m_a), \Psi(m_i))/\tau}}{e^{d(\Psi(m_a), \Psi(m_i))/\tau} + \sum_{x_k \in \Omega^-_{m_a}} e^{d(\Psi(m_a), \Psi(m_k))/\tau}} \right)$$

(3.6)

Here, τ is the temperature coefficient, and $d(a, b) = \frac{a^T b}{\|a\|\|b\|}$ represents the cosine similarity between two l2-normalized feature vectors a and b in the representational space. This loss function is designed to maximize the probability that a selected patch m_a will be recognized as similar to its positively associated local region m_i within $\Omega^+_{m_a}$. By minimizing this loss function, we encourage the representations of patches in $\Omega^+_{m_a}$ to become more similar, while ensuring that those in $\Omega^-_{m_a}$ remain dissimilar. The total contrastive loss for an image is calculated as the sum of losses over T randomly sampled patches in the image:

$$L(s) = \frac{1}{T} \sum_{i=1}^{T} L(m_i) \tag{3.7}$$

4. Unet Architecture

4.1 Introduction

The U-Net model, a significant advancement in the field of computer vision and particularly impactful in the domain of medical image analysis, was proposed by Olaf Ronneberger, Philipp Fischer, and Thomas Brox in their seminal paper presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) in 2015 [9]. Originally developed for biomedical image segmentation, U-Net has demonstrated exceptional versatility and effectiveness, finding application in a broad spectrum of tasks beyond its initial design. At its core, U-Net is a convolutional neural network (CNN) that features a distinctive structure, enabling efficient processing and analysis of visual data, a crucial capability especially in scenarios where data availability is limited.

4.1.1 Architecture

At the heart of U-Net's design is its distinctive ”U” shaped structure, which embodies the encoder-decoder concept:

- **Encoder (Contraction Path):** The encoder or contraction path serves to capture the context within the image. It consists of a sequence of convolutional and max pooling layers. This structure works to progressively reduce the spatial dimensions of the input image while deepening its feature representation, allowing the model to distill and encode high-level features from the visual data.

- **Decoder (Expansion Path):** The decoder or expansion path inversely mirrors the encoder, employing a series of upconvolutions (or transposed convolutions) and concatenations with corresponding feature maps from the encoder through skip connections. This expansion path progressively restores the spatial dimensions while refining the feature depth, enabling precise localization and detailed segmentation by integrating high-level contextual information with low-level spatial details.

- **Skip Connections (Bridging Encoder and Decoder):** A key innovation of the U-Net architecture is the introduction of skip connections that directly link layers in the encoder with corresponding layers in the decoder. These connections are crucial for the transfer of spatial and structural information, which helps in the precise delineation of object boundaries. Skip connections ensure that the decoder has access to both the abstracted features of the deeper layers and the fine-grained details lost during downsampling, facilitating a more nuanced reconstruction of the image.

4.2 Adapting U-Net for Synthesis

While U-Net was initially tailored for segmentation, its encoder-decoder framework, complemented by skip connections, provides a versatile foundation for other complex tasks, such as image synthesis.

In our case, in synthesizing missing MR contrast, T1CE, U-Net can be adapted to handle multi-modal inputs by adjusting the input layer to accept different MRI contrasts as separate channels. This adaptation leverages the encoder to assimilate diverse contextual information from the input contrast images, while the decoder focuses on reconstructing the desired output contrast image with high fidelity. Customizing the output layers and employing specialized loss functions tailored to the synthesis task can further refine the model's output, ensuring that the synthesized images closely match the target characteristics in terms of structure and appearance.

We can customize a U-Net model by keeping some decoder layers unadjusted during pre-training and calculating loss on the feature maps from the lower levels of the decoder. This approach allows the last layers of the decoder to remain as free weights for further adjustments in the downstream tasks, which we will discuss in the Chapter 6. Figure 4.2 illustrates the feature maps extracted from the last and third layers of the decoder.

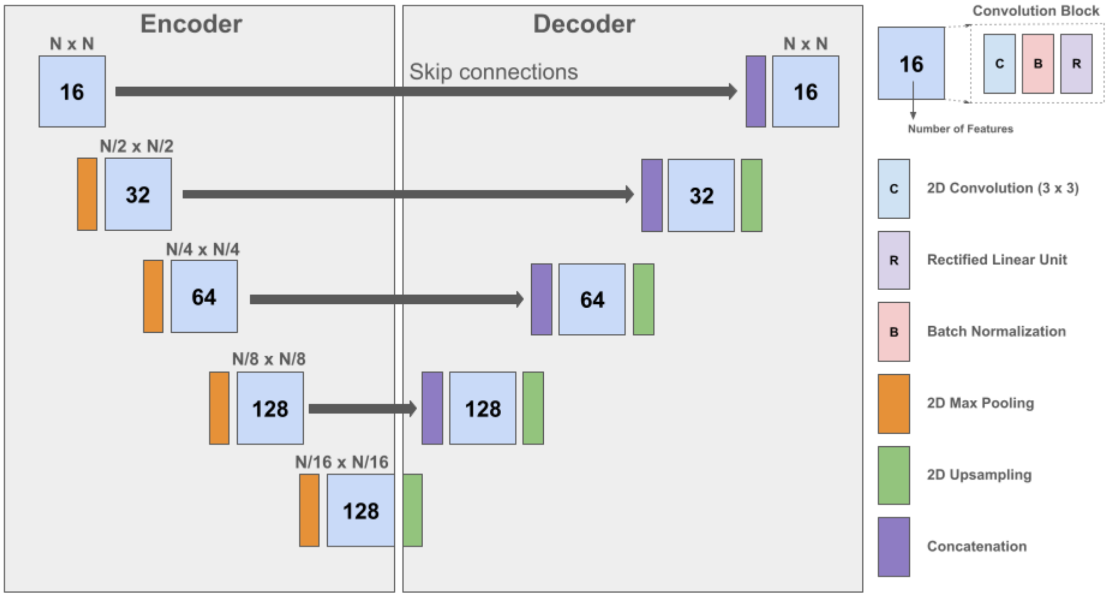
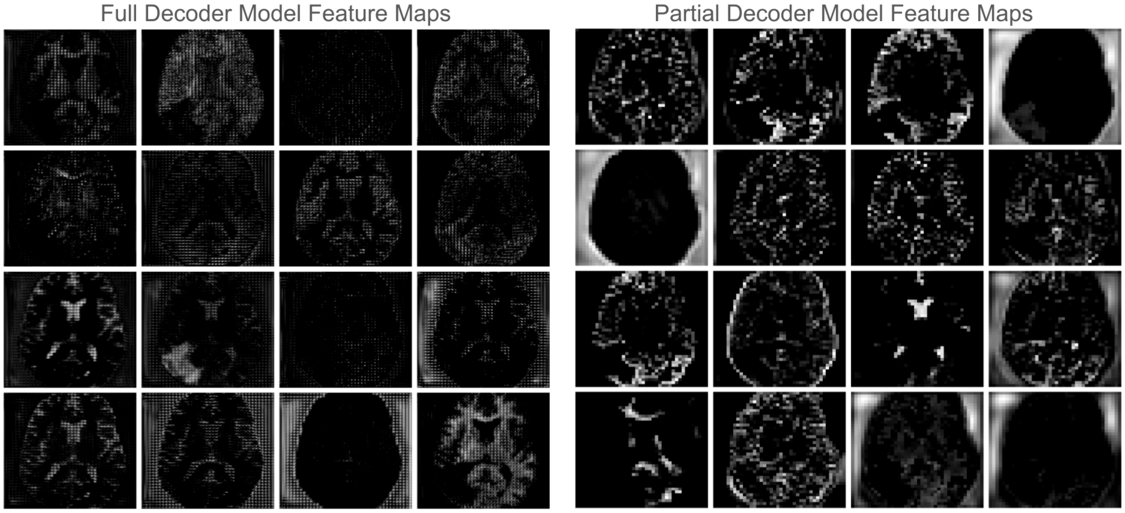


Figure 4.1: We employ a customized version of the UNET model for our training. The 2D Encoder-Decoder architecture (UNET) used in this work is shown here. The encoding path consists of a series of 2D convolutions layers (3x3 kernels) with padded convolutions, batch normalization layers, and rectified linear activation layers. The number of feature maps generated by each convolution layer are noted next to it.


Full Decoder Model Feature Maps

Partial Decoder Model Feature Maps

Figure 4.2: Comparison of feature maps extracted from the last layer of the decoder (full decoder model) and the third layer of the decoder (partial decoder model).

5. Synthesis Loss

In my thesis, the evaluation of loss functions extends beyond traditional pixel-to-pixel approaches like Mean Absolute Error (MAE) and Mean Squared Error (MSE), to include a Perceptual loss framework that integrates the feature-based analysis capabilities of VGG16. Pixel-based losses often lead to over-smoothed results and fail to capture high-frequency details, shortcomings that Perceptual loss mitigates by emphasizing feature similarity over exact spatial alignment. This feature-centric methodology, drawn from the realm of super-resolution, not only enhances the visual sharpness of the generated images but also ensures that the reconstructed output preserves contextual integrity, underscoring the efficacy of the deep learning model employed in my research.

5.1 MAE Loss

Mean Absolute Error (MAE) loss, also known as the L1-norm loss, quantifies the absolute differences between the predicted values and the actual values in an image. In the context of medical image synthesis, MAE loss would measure the absolute pixel-wise discrepancies between the synthetic MRI images I_{synth} and the corresponding ground truth MRI images I_{MR}, thereby providing a direct representation of the average error across the image. This loss function is robust to outliers and often leads to reconstructions with fewer artifacts compared to MSE.

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |I_{\text{MR}_i} - I_{\text{synth}_i}| \tag{5.1}$$

5.2 MSE Loss

Mean Squared Error (MSE) loss, the squared L2-norm, minimizes the sum of the squared differences between the synthesized MRI images I_{synth} and the real MRI images I_{MR}. MSE is highly sensitive to large errors due to its squaring operation, pushing the model to focus on reducing larger errors more aggressively. While this can lead to high-quality reconstructions where minimizing large errors is crucial, it may also result in overly smooth images as it penalizes variance, potentially affecting the sharpness and fine details.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(I_{\mathrm{MR}_i} - I_{\mathrm{synth}_i})^2 \tag{5.2}$$

5.3 Perceptual Loss

The term ”perceptual” implies that the loss function is more aligned with human perception of images. While pixel-wise loss might consider two images to be very different if all pixels are slightly off, perceptual loss might find them to be similar if they share the same structures and textures, much as a human might not notice the pixel differences but would recognize the same objects and shapes in both.

5.3.1 VGG16

VGG16 [10] is a convolutional neural network (CNN) model that has achieved remarkable recognition for its performance in large-scale image recognition challenges. The model is structured as a deep network with 16 weight layers, designed with a philosophy of simplicity and depth. VGG16 employs a series of convolutional layers with small receptive fields of 3×3, which are stacked on top of each other in increasing depth, as it shown in Figure 5.1. Interspersed among these convolutional layers are max pooling layers, which serve to reduce spatial dimensionality and to induce spatial hierarchy in the feature representation. The network concludes with fully connected layers that lead to a final output layer, typically tailored for classification tasks.

The architecture has approximately 138 million trainable parameters, a testament to its depth and capacity for feature extraction. The depth of the network, combined with the small size of the convolution filters, allows VGG16 to learn complex features at various levels of abstraction, which has been a significant factor in its widespread adoption for various computer vision tasks.

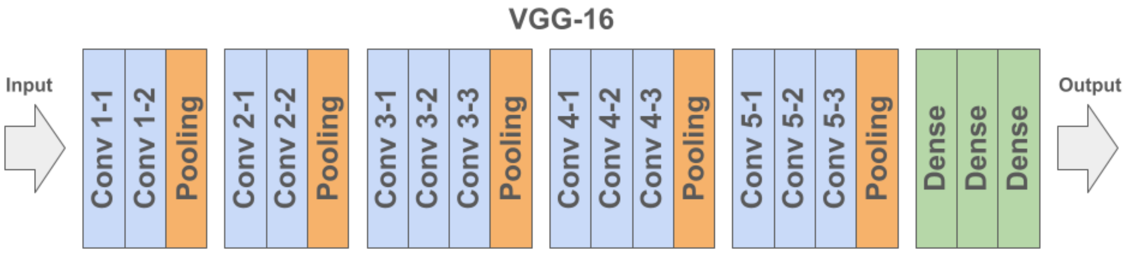


Figure 5.1: In VGG16 there are thirteen convolutional layers, five Max Pooling layers, and three Dense layers which sum up to 21 layers but it has only sixteen weight layers i.e., learnable parameters layer.

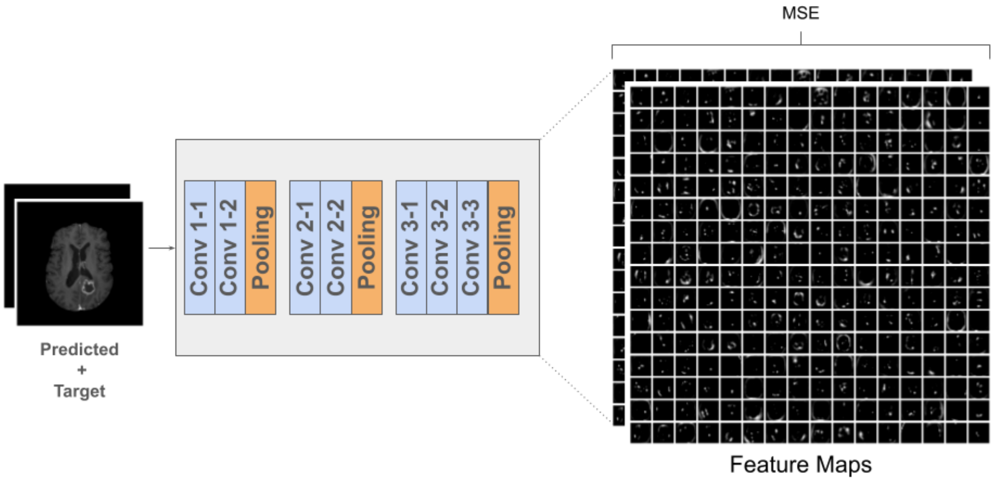


Figure 5.2: To utilize the VGG network as a loss network, we extract the feature maps from the fourth layer of both the input and target in the VGG network and then calculate the mean squared error (MSE).

In the context of image synthesis, pretrained VGG16 can be utilized as a feature extractor where the network is employed in a non-traditional way — not for classifying images, but for comparing different images. When used for synthesis loss calculation, VGG16's intermediate layer activations can serve as a representation of the high-level features learned by the network. By extracting these features from both the target (ground truth) image and the synthesized (generated) image, one can calculate the synthesis loss.

This loss is computed by measuring the similarity between the feature maps of the synthesized image and the feature maps of the target image using a chosen error function, such as the mean squared error (MSE) or the mean absolute error (L1), as visualized in Figure 5.2. This approach is often referred to as perceptual loss or feature reconstruction loss. Since VGG16 is pre-trained on a diverse set of images, its features can capture a wide range of visual patterns and textures, making it particularly effective for such comparisons.

By focusing on feature similarities rather than pixel-level differences, our loss function ensures that the synthesized model prioritizes the replication of perceptually relevant details. This approach involves extracting features from both the predicted and target images using the VGG16 network, and then computing the $L2$ distance or Mean Squared Error (MSE) between these features at corresponding spatial locations, as defined in Equation 5.3 (5.3). In this equation, $\hat{y}_{h,w,c}$ and $y_{h,w,c}$ represent the predicted and target features, respectively. This will enable the generation of images that are visually closer to the target in terms of textures, patterns, and complex structures.

$$L_{\text{Perceptual Loss}} = \frac{1}{H \times W \times C} \sum_{h,w,c} ||\hat{y}_{h,w,c} - y_{h,w,c}||_2^2 \tag{5.3}$$

5.4 Evaluating Synthesis Model Loss Functions

Earlier we outlined the potential for employing various loss functions within our synthesis model—specifically Mean Squared Error (MSE), Mean Absolute Error (MAE), and perceptual loss. To examine their effects, we tested all three models: Baseline, CL-Full, and CL-Partial, against each type of loss function. Our empirical results showcased the superiority of perceptual loss in generating sharp, realistic images that closely mirror the original data.

As evidenced in Figure 5.3, models trained with MSE and MAE tended to produce images with a smoother, more blurred appearance, albeit preserving the general structure of the brain. In stark contrast, models utilizing perceptual loss achieved a higher level of detail and sharpness. This was particularly notable in the models' ability to reconstruct tumor regions with greater clarity and definition. Furthermore, according to the LPIPS values presented in Table 5.1, images generated using Perceptual loss were closer to the original images, indicating its efficacy in preserving textural details that contribute to the perceptual quality of the synthesized images.

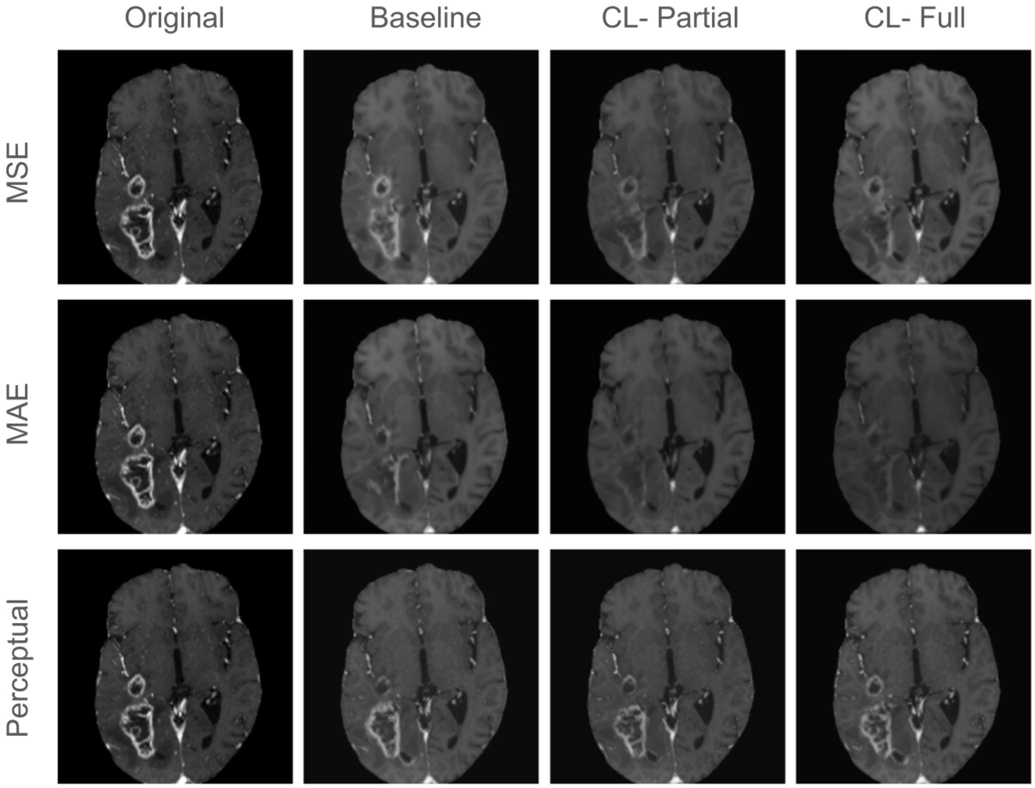


Figure 5.3: Models trained with various loss functions: MSE, MAE, and perceptual loss.

Table 5.1: LPIPS AlexNet values for different models and loss functions

Loss Function	Baseline	CL-Partial	CL-Full
MSE	0.122	0.117	0.123
MAE	0.145	0.155	0.144
Perceptual	0.0867	0.0846	0.0840

6. Model Architecture

The proposed network architecture is designed to generate targeted MR contrast images and consists of two main stages: pretraining and contrast synthesis. As discussed in Section 3.1.2, Contrastive Learning plays a crucial role in utilizing a large amount of unlabeled data for feature representation. In the pretraining stage of our network, this unsupervised learning approach is employed to aggregate tissue-specific information from various MR contrast images. These feature representations are then utilized in downstream task.

In the second stage, the contrast synthesis network leverages the weights developed during pretraining. This stage is specifically tasked with learning the synthesis of targeted MR contrast images. The synthesis network builds on the foundational work of the pretraining stage, further refining all network weights to effectively generate the desired outputs.

6.1 Pretraining

For the pretraining stage of our network, we employ the U-Net architecture, depicted in Figure 4.1. This model processes MR contrast images with spatial dimensions of 160 x 160 pixels. As illustrated in Figure 6.1, in this case our inputs for this model include three types of contrast images: T1-weighted, T2-weighted, and T2-FLAIR. However, the model is flexible and can accept any combination of these images, or even a single image type.

During pretraining, the model generates feature maps from the input images. These maps are then used alongside prepared constraint maps to calculate the contrastive loss (3.5), which guides the learning process.

6.1.1 Full Decoder and Partial Decoder

To explore the impact of different network configurations on learning efficacy, we experimented with modifications to the decoder section of the U-Net architecture. We tested two

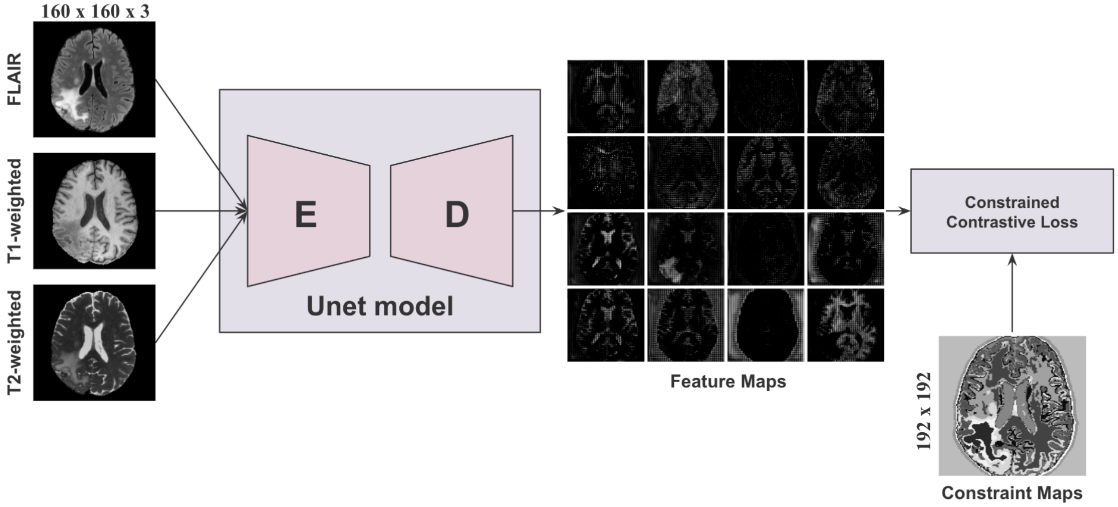


Figure 6.1: Schematic of the Pretraining Network. Three distinct MR contrast images (FLAIR, T1, and T2) are input into the U-Net model, which is composed of an encoder (E) and a decoder (D). The network generates feature maps from these inputs. The contrastive loss is then computed based on these feature maps in conjunction with a corresponding constraint map. The process is applicable for both the complete decoder (CL - Full) and the partial decoder (CL - Partial) scenarios.

scenarios: a full decoder setup (referred to as ”CL - Full”) and a partial decoder setup (”CL - Partial”). In the full decoder scenario, all layers of the decoder are used to reconstruct the data from the latent space back to its original size. The contrastive loss is then calculated based on the feature maps generated from the last decoder layer.

In a typical encoder-decoder architecture, the encoder compresses the input into a latent representation, and the decoder reconstructs the output from this representation. By omitting the final layers of the decoder during the initial training phase — in this case, the last two layers — these layers do not learn from the primary training process. This approach allows the last two layers to retain ”free weights,” which can be adjusted during the downstream task. The output from the earlier layers captures essential data characteristics but is less tailored to the specificities of the input data compared to a full decoder output. The final layers, being free of adjustments during this initial training, are more adaptable for fine-tuning based on the downstream task, which in our case is the synthesis of the T1-CE image. This technique, inspired by prior research, is hypothesized to provide the model with more flexibility in adjusting these weights to enhance performance in other DL tasks. Our experiments aim to evaluate the effectiveness of this strategy compared to the full decoder setup in synthesis task.

6.2 Synthesis Model

The synthesis network represents the second stage of our architecture, wherein we continue to utilize the U-Net model, this time with the objective of generating the target MR contrast images. This stage leverages the pretraining weights acquired in the first stage, applying perceptual loss to guide the synthesis process. The architecture and the initial weights for this stage are illustrated in Figure 6.2.

Input images with dimensions of 192 x 192 pixels are fed into the U-Net model, which then produces a predicted MR image. The fidelity of this prediction is evaluated by comparing it to the target image using a defined loss function. For our network, we employ perceptual loss, utilizing the VGG16 network 5.3.1, which is renowned for capturing high-level content and style from images, to refine our model's predictions.

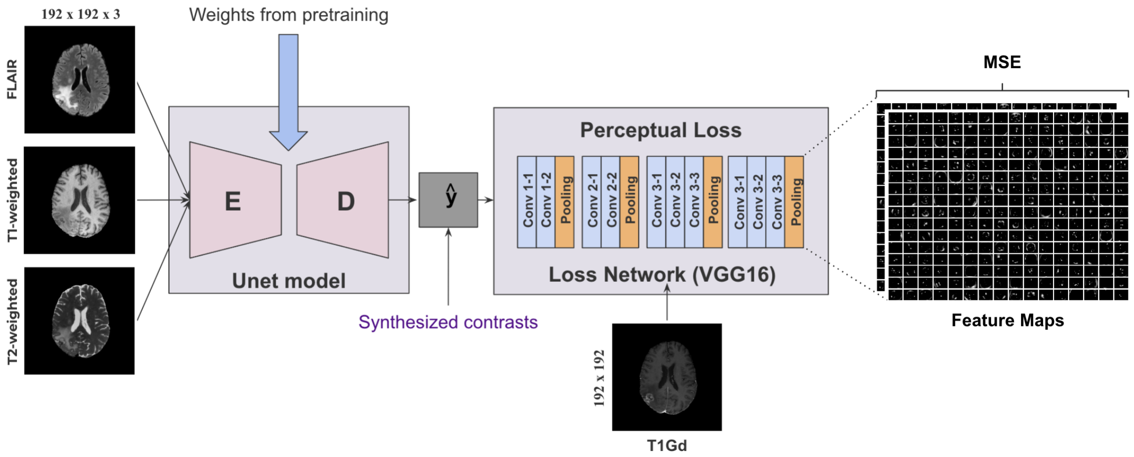


Figure 6.2: Diagram of the Contrast Synthesis Network. The U-Net model, initialized with weights from the pretraining network either CL-partial, CL-full, or baseline (no weights) receives input images of FLAIR, T1, and T2 contrasts with a resolution of 192 x 192 pixels. The model then synthesizes contrast images, which are subsequently evaluated by the loss network, using a VGG-based perceptual loss framework.

6.2.1 Baseline

To assess the benefit of using pretrained weights, we introduce a baseline model for comparison. This baseline model follows the same architectural and training setup; however, it diverges in that it commences training with randomly initialized weights, lacking any pretrained information. The comparison between the pretrained U-Net model and the baseline provides insights into the efficacy of pretraining for the task of MR image synthesis.

7. Data Pre-processing

7.1 Constraint Maps

Constraint maps are crucial for our MR-contrast guided contrastive learning approach. They encapsulate tissue parameter information that assists in emphasizing distinctive features during the pretraining phase. The generation of these maps is an offline preprocessing step.

To create the constraint maps, we first perform Principal Component Analysis (PCA) followed by K-Means clustering on the three available MR contrast images. The PCA step reduces dimensionality and noise, while K-Means clustering, with k set to 20, organizes the MRI data into meaningful clusters. The choice of 20 clusters is based on our analysis presented in Table 1.1, which takes into account the variety of tissue types—such as white and gray matter, necrotic regions, inflammation, water, fat—and the intrinsic contrasts of the images. Opting for more clusters could result in over-segmentation, while fewer clusters might oversimplify the tissue parameters, obscuring critical details needed for effective pretraining.

7.2 Image Processing

The data pre-processing pipeline was carefully designed to standardize the MR images and optimize them for effective network training. The following steps outline the pre-processing applied:

- **Slice Selection:** Each subject's dataset originally comprised 155 slices for each contrast image. We truncated the top and bottom slices to ensure that the remaining images prominently featured visible brain structures.
- **Intensity Clipping:** Intensities were clipped to fall within the 0.01st to 99.9th percentile range, mitigating the effect of extreme values that could skew the analysis.

- **Histogram-Based Contrast Stretching:** We utilized histogram-based contrast stretching for contrast enhancement, rescaling intensities within the brain mask to sharpen the feature distribution.

- **Channel-wise Z-score Normalization:** Channel-wise normalization was implemented by applying zero-mean and unit standard deviation normalization exclusively to the signal within the brain mask. This targeted approach is crucial for addressing the variable acquisition protocols in the BraTS dataset.

- **Cropping for Pretraining:** Originally having dimensions of 240 x 240 pixels, the images were cropped to 160 x 160 pixels during the pretraining phase. This size was selected to balance focus on the brain regions with the computational efficiency of the training process.

- **Cropping for Synthesis Task:** For downstream tasks involving synthesis, we increased the crop size to 192 x 192 pixels to enhance the resolution and the level of detail, which is paramount for generating high-fidelity MR images.

8. Experimental Results

8.1 Experimental settings

The BraTS’21 dataset was randomly divided into three subsets: a training set with 350 subjects, a validation set with 245 subjects, and a test set comprising 70 subjects. We conducted experiments using three models: Baseline, CL-Full, and CL-Partial, across various configurations and settings. The optimal configuration, chosen for its highest PSNR, LPIPS and superior detail preservation in the tumor region, involved synthesizing T1CE images from a combination of T1-weighted, T2-weighted, and T2-Flair images.

During the experimentation phase, several parameters were fine-tuned. For pretraining, a learning rate of 1×10^{-3} was selected, with training extending over 150 epochs. In the synthesis tasks, the learning rate was adjusted to 1×10^{-2} for 20 epochs. We utilized checkpoints throughout the training process to monitor the best loss value on the validation set and adjusted the number of epochs based on these checkpoints, as detailed in the logs shown in Figure 8.1. The temperature setting for the contrastive loss was fixed at 0.1, with a patch size of 4×4.

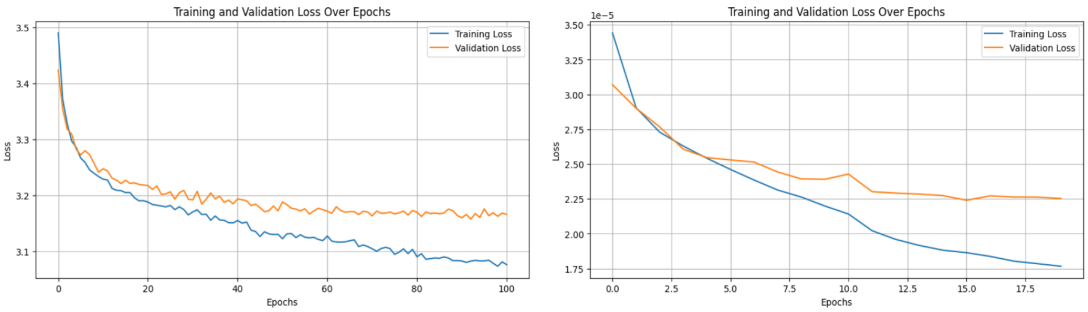


Figure 8.1: The illustrated training and validation curves on the left correspond to the pretraining stage, while those on the right are for the synthesis stage. These plots pertain to the CL-Full model.

In assessing the performance of our models, we employed a comprehensive set of metrics including Mean Squared Error (MSE), Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) [12] with architectures like VGG and AlexNet. While MSE provides a straightforward measure of the average squared difference between the estimated and actual values, SSIM and PSNR offer insights into the perceptual quality of images, focusing on aspects such as contrast, luminance, and noise. LPIPS, particularly when utilized with well-established neural networks like VGG and AlexNet, is especially valuable for image synthesis tasks. This metric assesses the perceptual similarity between synthesized and real images more effectively than traditional pixel-based metrics. By comparing deep features extracted by these networks, LPIPS can capture subtler, more human-perceptible differences, making it a superior choice for ensuring the synthesized images are not only accurate in pixel values but also visually indistinguishable from genuine images in human perception.

8.1.1 Model Comparison

In the synthesis of T1CE images, achieving high accuracy in the representation of tumor and necrotic regions is crucial. For training, the VGG16 network was utilized to calculate the loss, enhancing the model's ability to capture detailed textures and structures. As demonstrated in Figure 8.2, the CL-Full model exhibits superior performance in reconstructing these regions, outperforming both the Baseline and CL-Partial models. As shown in Table 8.1, the CL-Full model has significantly higher values of PSNR and SSIM compared to the Baseline. Additionally, it demonstrates superior LPIPS metrics, evaluated using both AlexNet and VGG, compared to both other models, which underscores its effectiveness. The statistical assessment of these results is detailed in the following section.

8.1.2 Statistical Analysis of Model Performance

To statistically validate the performance enhancement of the CL-Full model over the Baseline in generating T1CE images from a combination of T1-weighted, T2-weighted, and T2-Flair contrast images, a two-tailed paired t-test was conducted on a set of 70 test subjects.

Table 8.1: Comparison of Metrics Across Different Models, where VGG16 was utilized for perceptual loss calculation during the training phase. Metrics used for evaluation include PSNR, MSE, SSIM, and LPIPS metrics evaluated using both 'AlexNet' and 'VGG'.

Model Type	SSIM	MSE	PSNR	LPIPS 'AlexNet'	LPIPS 'VGG'
Baseline	0.752	0.166	27.23	0.0871	0.1494
CL-Partial	0.748	**0.164**	27.23	0.0846	0.1335
CL-Full	**0.757**	0.166	**27.60**	**0.0844**	**0.1240**

Table 8.2 representing key aspects of image quality and fidelity by providing statistical results for different metrics.

Our null hypothesis posited no significant difference between the CL-Full and Baseline models, while the alternative hypothesis suggested a superior performance of the CL-Full model. The significance level was set at 5%, meaning that a p-value lower than 0.05 would indicate a statistically significant difference favoring our alternative hypothesis.

Table 8.2: Statistical comparison of models using t-test under the null hypothesis with a desired confidence of 5%, $p < 0.05$

Metric	Model Comparison	p-value	Conclusion
SSIM	CL-Full vs. Baseline	0.0004	Significant
	CL-Full vs. CL-Partial	0.0000	Significant
MSE	CL-Full vs. Baseline	0.8765	No significant
	CL-Full vs. CL-Partial	0.0966	No significant
PSNR	CL-Full vs. Baseline	0.0000	Significant
	CL-Full vs. CL-Partial	0.0000	Significant
LPIPS 'AlexNet'	CL-Full vs. Baseline	0.0000	Significant
	CL-Full vs. CL-Partial	0.7611	No Significant
LPIPS 'VGG'	CL-Full vs. Baseline	0.0000	Significant
	CL-Full vs. CL-Partial	0.0000	Significant

The statistical analysis in Table 8.1 & 8.2 indicates the following:

- The CL-Full model significantly outperforms the Baseline model across all metrics except MSE, where there was no significant difference.

- For comparisons between CL-Full and CL-Partial, the results are mixed. CL-Full shows significant improvement in SSIM, PSNR, and LPIPS with VGG, but not in MSE and LPIPS with AlexNet where differences were not statistically significant.

- Results suggest that the full implementation of constrained contrastive learning enhances model performance in most tested aspects compared to both a baseline and a partially constrained contrastive model, particularly in terms of image quality metrics like SSIM and PSNR, as well as perceptual similarity as assessed by LPIPS with VGG.

8.2 Dataset and other training details

8.2.1 Data Cleaning

As previously mentioned, our initial dataset comprised approximately 1660 subjects. Upon review, we identified 800 images that were corrupted, blurred, highly noisy, or contained significant artifacts; some were even bisected, as illustrated in Figure 8.3. These issues made them unsuitable for our synthesis tasks and adversely affected the quality of the generated images. While the BraTS'21 dataset includes scans from various clinics using different equipment—a feature we appreciate for its potential to help develop a model that generalizes well across different image qualities—images with excessive corruption can deteriorate model performance. Figure 8.4 demonstrates the notable improvements in the generated images after cleaning the dataset. It is evident that not only the synthesis of the tumor region but also the overall image quality and other details have significantly enhanced.

8.2.2 Normalization

In our code, we provide two normalization options before training: normalizing the entire image to a $[0, 1]$ scale and employing z-mean normalization. Z- score normalization specifically targets the brain tissue, normalizing based on the mean and standard deviation of the brain region, excluding the background. This method has proven especially effective in our experiments (Figure 8.5) for enhancing the synthesis of tumor regions, aligning with our focus on improving tumor visibility in MRI images.

8.2.3 Activation Function

The choice of activation function is pivotal in neural network performance, influencing both the learning dynamics and the quality of the generated output. To determine the most suitable activation function for our model, we conducted a comparative analysis between Tanh and ReLU, as reported in Table 8.3.

Table 8.3: Performance Metrics for Different Activation Functions and Loss Values

Activation	Full			Partial			Base		
	SSIM	MSE	PSNR	SSIM	MSE	PSNR	SSIM	MSE	PSNR
Tanh	0.752	0.169	26.83	0.746	0.175	27.03	0.751	0.170	27.14
ReLU	0.757	0.166	27.60	0.748	0.164	27.23	0.752	0.166	27.23

In comparative performance assessments across different configurations—Full, Partial, and Base—as detailed in Table 8.3, the ReLU activation function consistently outperformed Tanh. This performance contrast with the Tanh activation function is further illustrated in Figure 8.6.

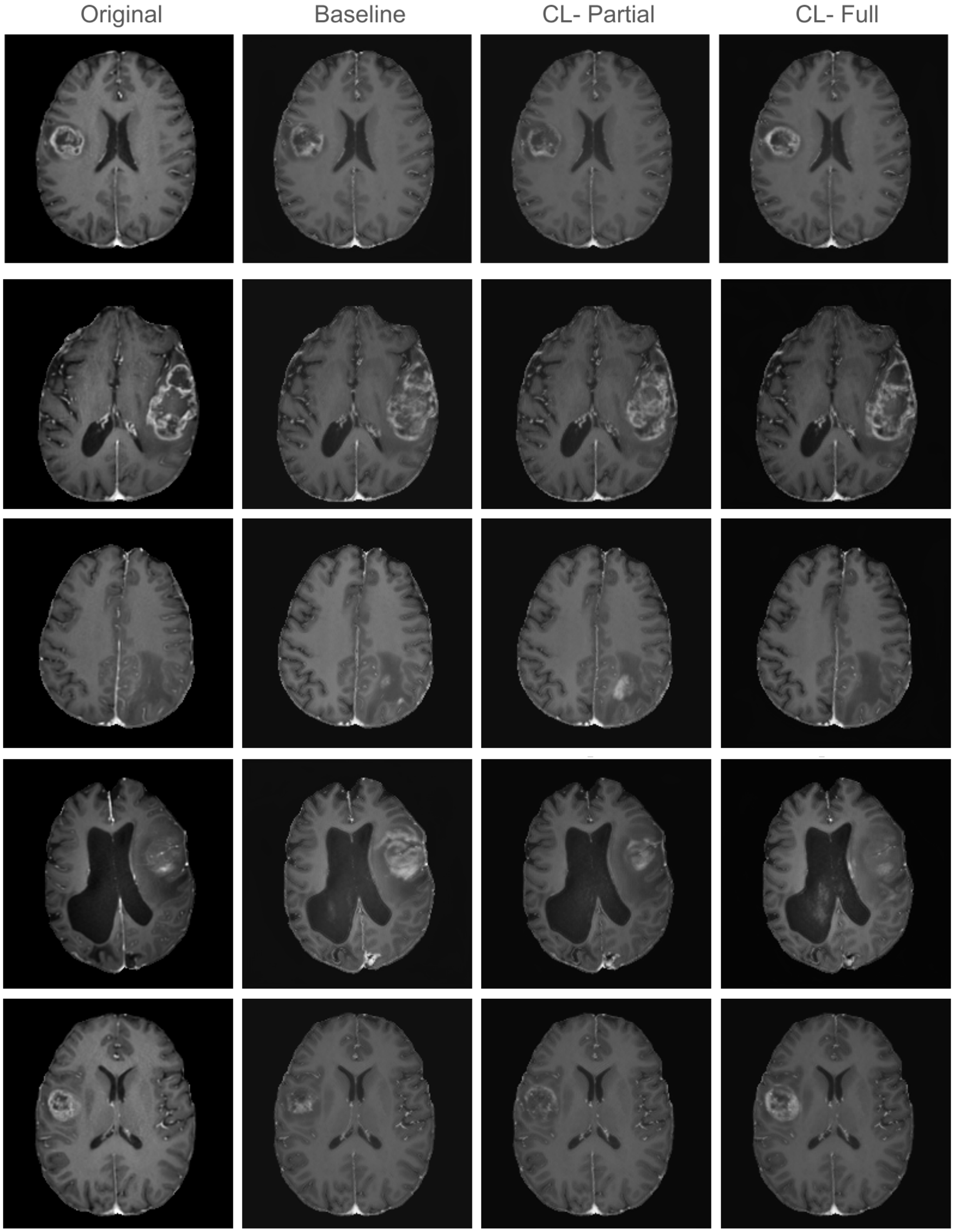


Figure 8.2: Comparison of T1CE synthesis results using a combination of T1-weighted, T2-weighted, and T2-Flair images across different subject samples for three models: Baseline, CL-Full, and CL-Partial.

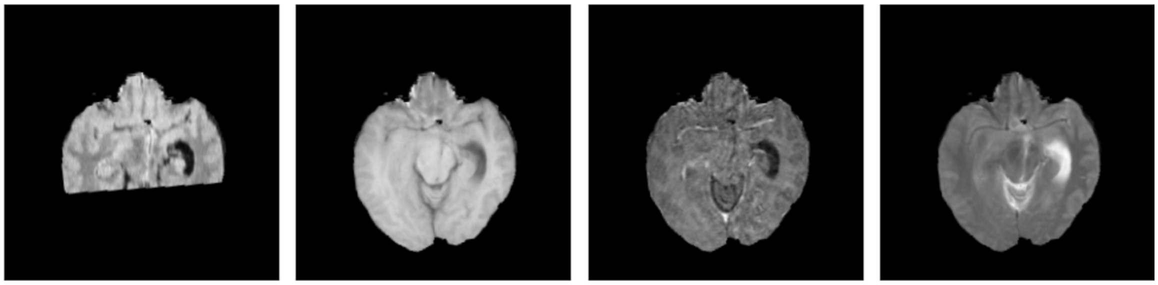


Figure 8.3: Examples of Corrupted MR Images demonstrating the possible challenges by poor-quality data in synthesis task.

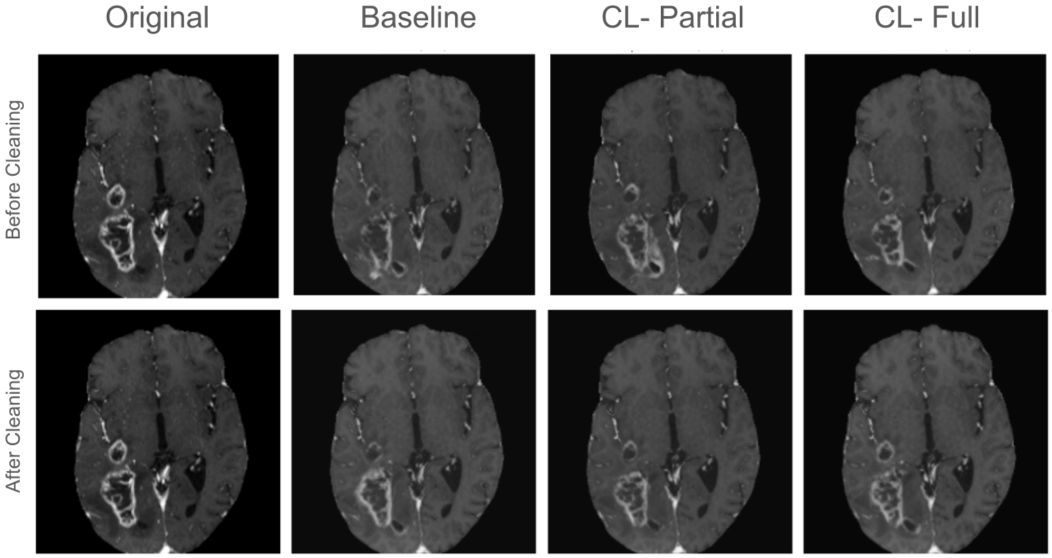


Figure 8.4: An example of synthesized MR contrast image (T1CE) before and after dataset cleaning.

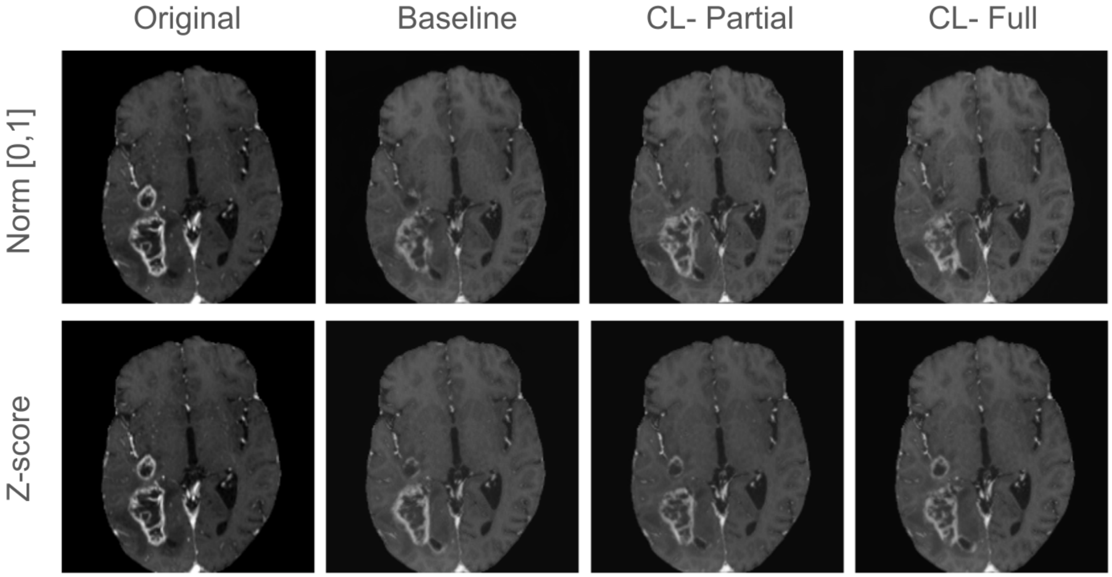


Figure 8.5: Comparison of synthesized results using Z-mean score normalization versus [0, 1] normalization.

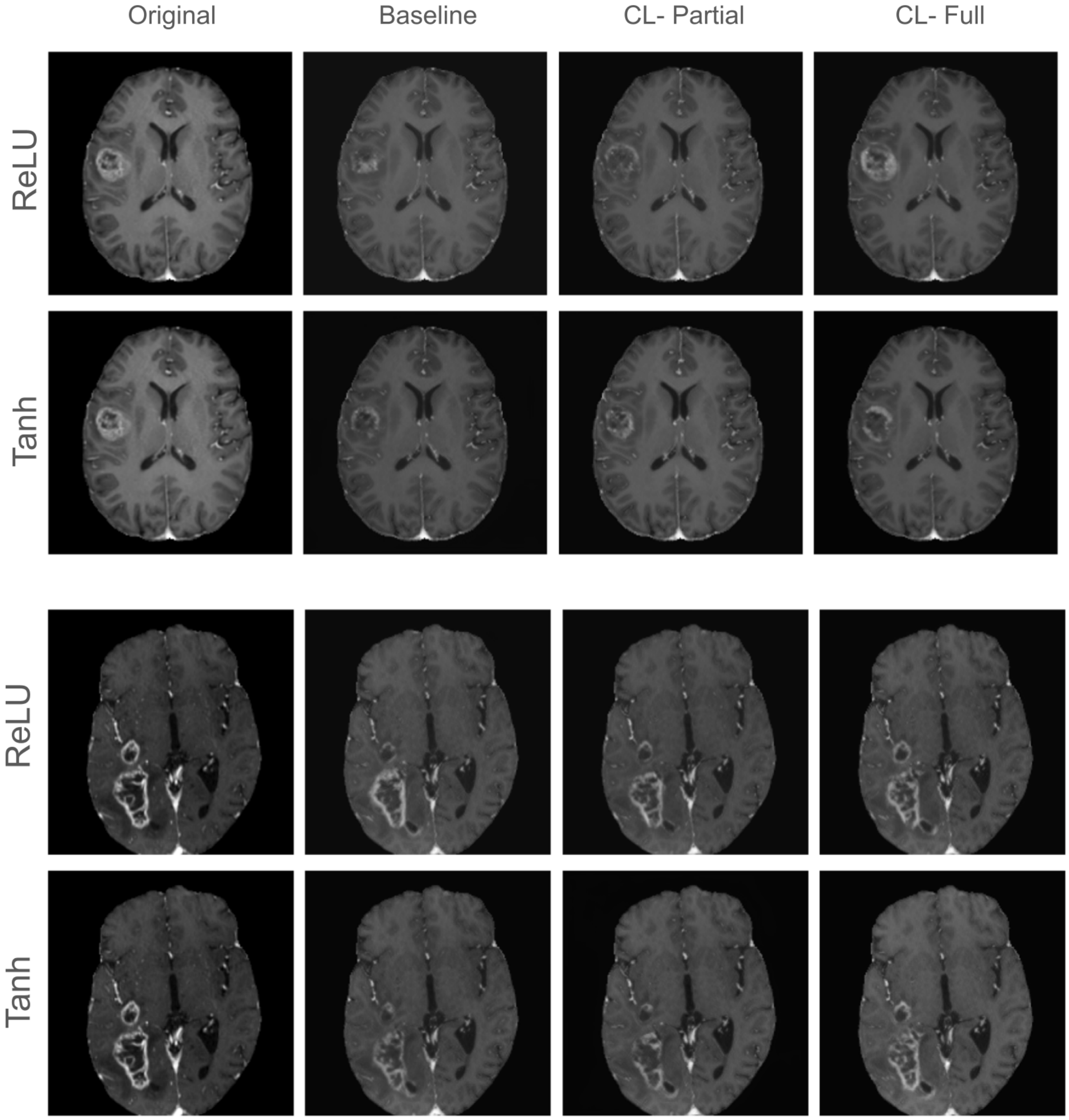


Figure 8.6: Comparison between setting the activation function as ReLU or Tanh for three different models: Baseline, CL-Full, and CL-Partial.

Future Work

This thesis has set the groundwork for several exciting avenues of future research in MR contrast image synthesis using deep learning. As the field progresses, the following strategies could significantly enhance the model's performance and application:

Complex Tasks-Simultaneous Synthesis of Multiple MR Contrasts

Future studies could aim to tackle more complex tasks, such as training models to take only two MR contrast images and simultaneously synthesize two additional contrasts that are missing. This could dramatically improve efficiency in clinical settings, reducing the need for multiple scans. The challenge lies in ensuring the model can capture and relate the nuanced differences and dependencies between various contrast types without losing accuracy.

Incorporate Adversarial Training

Incorporating adversarial training by implementing a discriminator model could refine the realism of synthetic images. Adversarial loss, when applied through such models, pits the generator against the discriminator in a game-theoretic scenario, pushing the generator to produce increasingly realistic images. This approach could help overcome some of the synthetic artifacts and unrealistic texturing that current models might produce, thereby improving the clinical usability of synthesized MR images.

Model Stacking Strategy

Exploring a dual-model approach by stacking CL-Partial and CL-Full models offers a promising direction for enhanced feature extraction and synthesis precision. Stacking models in such a manner could leverage the strengths of each model, potentially leading to better generalization over a wider range of data variations. This strategy might also mitigate overfitting by combining the diverse feature representations learned by each model separately.

Tumor-Centric Loss Weighting

Modifying the synthesis loss function to assign higher weights to tumor masks would prioritize tumor region accuracy, an area of critical importance for diagnostic imaging. By emphasizing tumor areas within the loss function, the model could produce higher fidelity representations of these crucial regions, thus providing radiologists with more reliable images for making accurate diagnoses. This approach necessitates careful calibration to balance the emphasis on tumor regions against the overall image quality.

Bibliography

[1] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning, 2019.

[2] U. Baid et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[3] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, et al. Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. The Cancer Imaging Archive, 2017.

[4] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, et al. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. The Cancer Imaging Archive, 2017.

[5] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Nature Scientific Data*, 4:170117, 2017.

[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014.

[7] Robert W Brown, Yu-Chung Norman Cheng, E. Mark Haacke, Michael R. Thompson, and Ramesh Venkatesan. Magnetic resonance imaging: Physical principles and sequence design. 1999.

[8] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[11] Lavanya Umapathy, Taylor Brown, Raza Mushtaq, Mark Greenhill, J'rick Lu, Diego Martin, Maria Altbach, and Ali Bilgin. Reducing annotation burden in mr: A novel mr-contrast guided contrastive learning approach for image segmentation. *Medical Physics*, 51(4):2707–2720, 2024.

[12] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.