POLITECNICO DI TORINO

MASTER's Degree in Biomedical Engineering

MASTER's Degree Thesis

Optimization of Generative Models using No-Reference Metrics: Application to MRI to CT Translation

Supervisors

Prof. Filippo MOLINARI

Prof. Massimo SALVI

Tutors

Annalisa LETIZIA

Vincenza TUFANO

Candidate

Marika ALECCI

December 2024

Summary

In the field of medical imaging, the application of generative artificial intelligence techniques for cross-domain translation represents a research area of significant interest. Nevertheless, a comprehensive review of the literature has revealed a lack of established metrics that can guarantee the reliability and quality of generated images.

This thesis investigates the use of the pix2pix image translation model for the transformation of medical images from Magnetic Resonance Imaging (MRI) to Computed Tomography (CT). In the initial phase of the study, a series of experiments were conducted to evaluate the efficacy of different combinations of loss functions and transfer learning techniques.

Subsequently, in order to identify objective metrics for an accurate evaluation of the produced images, in addition to the utilisation of conventional Full-Reference metrics, including MAE, MSE, PSNR and SSIM, No-Reference metrics such as NIQE, ILNIQE and PIQE were also examined. Moreover, NIQE and ILNIQE were introduced as loss functions during the model training process, with the objective of improve the quality of the generated images.

The best results, according to FR metrics, were obtained from the trail using BCE+L1 as loss function and without transfer learning, with MAE of 0.0746 ± 0.0598, MSE of 0.0239 ± 0.0397, PSNR of 18.6500 ± 3.9336 and SSIM of 0.7040 ± 0.1086. For this same trial, the NIQE value is 12.4894 ± 1.7046, ILNIQE is 45.7019 ± 3.5649 and PIQE is 43.3655 ± 5.1703.

The results demonstrated the challenge of using both Full-Reference and No-Reference metrics to assess the quality of the synthesised images, although ILNIQE showed particular promise. However, the application of ILNIQE as a loss function exhibited limitations due to the high computational time required, whereas NIQE, in combination with other traditional loss functions, produced satisfactory outcomes.

It can be concluded that further developments are required in order to validate and improve the reliability of the translated images.

Acknowledgements

Firstly, I want to thank Professor Molinari, without whom this thesis would not have been possible. He was one of the first to arouse my interest in artificial intelligence applied to the biomedical field. It is thanks to professors like him, who are able to transmit passion for what you do, that I have been able to approach this university journey with enthusiasm and dedication.

A special thanks to Professor Salvi for his availability, for giving me interesting ideas for this thesis work and for following me through to the end, which is why I particularly wanted his name to appear among the supervisors.

Finally, but most importantly, I would like to thank my tutors at Teoresi, Annalisa and Vincenza, who have constantly followed and encouraged me. Their guidance and support have been invaluable and I feel that I have learnt a lot thanks to them, even though I still have a lot to learn. I am thankful for the chance to grow and to engage in constructive comparison.

It is perfectly acceptable to fall sometimes.

Table of Contents

List of Tables

List of Figures

Acronyms

AI Artificial Inteligence

BCE Binary Cross Entropy

BIQA Blind Image Quality Assessment

CNN Convolutional Neural Network

CT Computed Tomography

DCGAN Deep Convolutional GAN

DL Deep Learning

FID Fréchet Inception Distance

FR Full-Reference

GAN Generative Adversarial Network

ILNIQE Integrated Local Natural Image Quality Evaluator

LSGAN Least Squares GAN

MAE Mean Absolute Error

ML Machine Learning

MRI Magnetic Resonance Imaging

MSE Mean Squared Error

MVG Multivariate Gaussian

NIQE Natural Image Quality Evaluator

NR No-Reference

PIQE Perception-based Image Quality Evaluator

PSNR Peak Signal-to-Noise Ratio

SSIM Structural Similarity Index

TL Transfer Learning

WGAN Wasserstein GAN

WGANGP Wasserstein GAN with Gradient Penalty

Chapter 1

Introduction

In recent years, Artificial Inteligence (AI) has grown so exponentially that it has completely revolutionised everyday life. Its influence extends into many fields, including medicine. Recently, however, most AI research has focused on generative AI. Generative models, which represent one of the most promising technologies to date, are a class of algorithms designed to create new and realistic contents. These models are trained on a specific dataset, such as images, text or sounds, from which they are able to extract features and patterns to create new examples that match the observed data.

In contrast to discriminative models, which are employed for the purpose of classification or prediction by learning the decision boundary between classes, generative models are designed to capture the distribution of data in order to replicate it and generate new data.

An example of a popular generative model is the Generative Adversarial Network (GAN). A GAN consists of two neural networks, the generator and the discriminator, which compete with one another during the training process (Fig. 1.1). The generator has the task of generating data, while the discriminator evaluates its authenticity by distinguishing between real and false data. The training process is adversarial because the generator tries to create increasingly realistic data that can fool the discriminator, which has to train more and more to be able to recognise this data as false.

Generative models are gaining ground in many applications, including the medical field, where they offer several advantages: they can help improve diagnosis, create new therapies and better understand certain diseases. However, their integration into clinical practice needs to be carefully evaluated from both a clinical and ethical perspective to ensure their efficacy and safety.

For example, these models can be used to improve the quality of medical images, generate new data to fill in gaps, create realistic simulations of medical procedures to help doctors train without using patients, and synthesise new drugs.

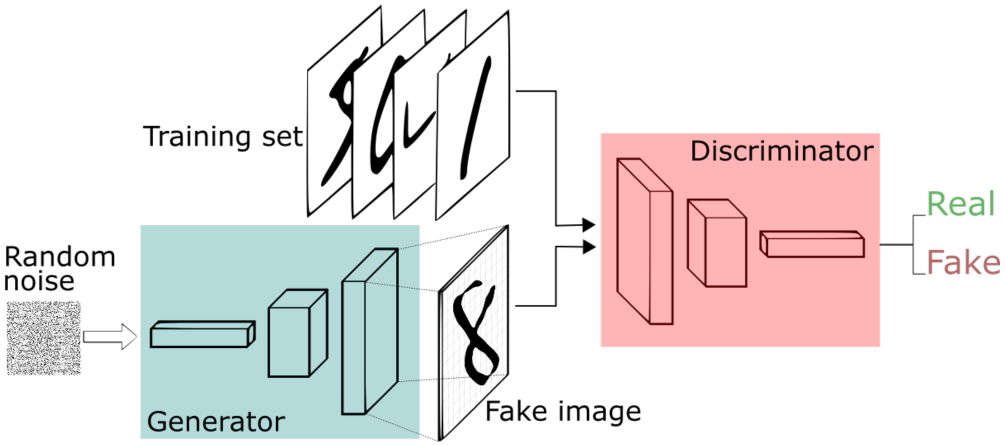


Figure 1.1: Representation of a Generative Adversarial Netwotk.

In particular, one of the applications of generative models in medical imaging is image translation, a process in which one type of scan - e.g. Computed Tomography (CT) - is translated into a scan of another type - e.g. Magnetic Resonance Imaging (MRI). There are several reasons why this translation process can be useful: synthesising one scan from another can certainly provide additional information for diagnosis, while saving time and money.

In the field of medical imaging, MRI and CT are diagnostic techniques widely used in various fields of study that provide detailed images of the human body. MRI uses magnetic fields to create a three-dimensional image of soft and hard tissue [1]. It can be used to visualise internal organs, the skeleton and joints. CT uses ionising radiation to produce three-dimensional images of internal organs, bones, blood vessels and lymph nodes. It is useful to visualize bone structures and detect acute conditions, such as hemorrhages or fractures, but also for planning radiotherapy for a tumour.

Both scans are often needed to make an accurate diagnosis and to plan treatment. However, CT uses higher doses of radiation than a normal X-ray [2], so it is avoided whenever possible.

Also for this reason, i.e. to reduce the radiation dose to the patient, image translation from MRI to CT can offer significant advantages.

There are several studies in the literature on the translation of medical images using GANs in particular. As explained in Section 1.1, different models have been proposed to obtain images that are more realistic and consistent with the original ones. However, the generated images need to be validated in order to be used in clinical practice for correct diagnosis, and to date there is no consensus on the type of metrics to be used to validate the images or the level of accuracy required. Clear validation of the images is essential if these technologies are to be integrated into clinical practice.

Traditional metrics used to evaluate synthesised images are Full-Reference (FR), i.e. they require a reference image against which the generated image is compared. These metrics aim to assess the differences between the target and generated images, but do not always reflect the visual fidelity and quality of the generated images. Another type of metric is the No-Reference (NR) metric, which does not require a reference image and aims to assess the quality of an image. These metrics could be a valid alternative to the metrics used so far.

In this work, after analysing the state of the art and the applications of image translation in the medical field, a GAN conditional model, the *pix2pix*, has been used to generate CT images from MRI images. This model has shown remarkable success in various image translation tasks (Fig. 1.2).

During the training phase, the model is presented with pairs of MRI and CT images of the same subject and body region. The model is trained to identify statistical correspondences between the features of the MRI and CT images. It should be noted that the model does not directly "extract" physical information from the MRI images, as these images do not contain the same basic data: an MRI does not contain the electron density information needed to generate a CT image. Instead, the model learns a mapping between CT and MRI images. Once trained, by acquiring an MRI image, the model can generate a CT image that matches the images observed during training.

The generated images may not capture all the nuances or abnormalities present in real CT images. Therefore, a critical aspect of the research was to carefully evaluate the generated images, particularly using NR metrics.

Finally, an attempt was made to optimise the pix2pix model using NR metrics as Loss Function.

The following chapters present a comprehensive account of the methodology

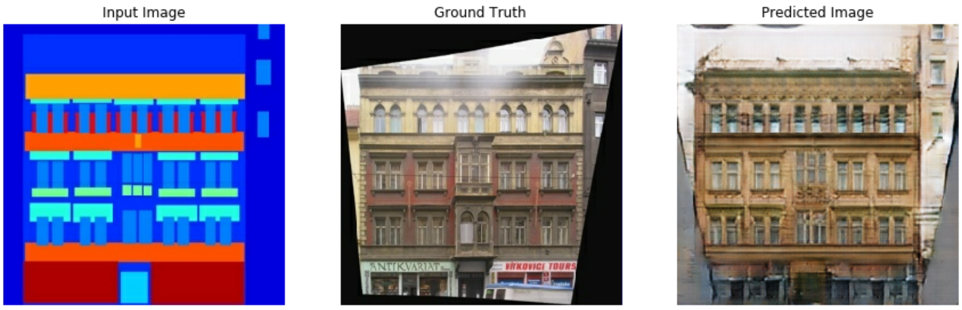


Figure 1.2: An example of image-to-image translation using a pix2pix model on the Facades dataset: from a sketch to an actual image of a building facade.

employed in this study, the findings derived from the various experiments, a critical analysis of the results, and finally, potential future developments or enhancements.

The following section presents a review of the existing literature on the use of AI in the field of medical image translation.

1.1 Medical Image Translation: Review

The generation of synthetic medical images through Machine Learning (ML), particularly using GANs, represents a notable innovation in the field of medical diagnostics. This section presents an overview of the current state of research, an outline of the adopted methodologies and an indication of the challenges in this field of study.

A comprehensive systematic literature review conducted in 2023 [3] identified 689 articles using specific keywords related to medical image translation using Deep Learning (DL) techniques. Through a series of screenings based on title, abstract and full content, the authors selected 99 relevant articles, published between 2017 and 2023.

These studies focused primarily on the application of GAN architectures for translation between different types of medical images, highlighting the significance of these techniques in enhancing diagnoses, treatment planning and clinical research.

The review shows that the most prevalent synthesis among the articles (76 of them) is from MRI to CT, as shown in Figure 1.3. The majority of studies that performed this synthesis are motivated by MRI-only radiation therapy, which circumvents the necessity for CT radiation exposure while simultaneously reducing costs and time.

Other motivations include the transformation of MRI datasets into MRI/sCT paired datasets and the completion of datasets through the synthesis of missing images.

Despite the potential clinical advantages in neurology due to the superior tissue contrast provided by MRI, few studies have been conducted on the synthesis of MRI images from CT. Indeed, the synthesis of MRI from CT could enhance the efficacy and precision of treatment for patients suffering from stroke.

The majority of studies employed both GANs and Convolutional Neural Network (CNN), but no clear consensus has emerged on which framework is more suitable for medical image synthesis. A significant number of articles employed GANs, with the majority introducing modifications and novel contributions to the fundamental models with the aim of enhancing image synthesis.

Few articles used CNNs, in particular variants of U-Net, showing that this architecture works well for image synthesis despite its main use in segmentation.

A significant challenge is the limited availability of large medical datasets,

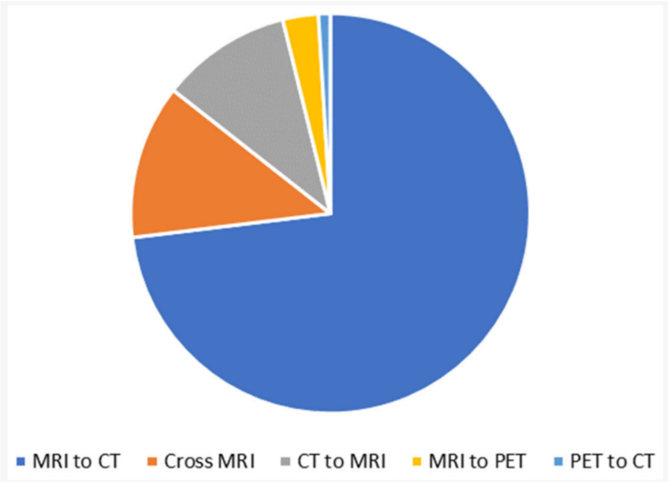


Figure 1.3: Type of synthesis from the review.

particularly those that are paired. The lack of paired and aligned images hinders the use of supervised learning for synthesis between different modalities.

In terms of evaluation methods, 36 different methods for evaluating model performance were identified, as summarised in Figure 1.4.

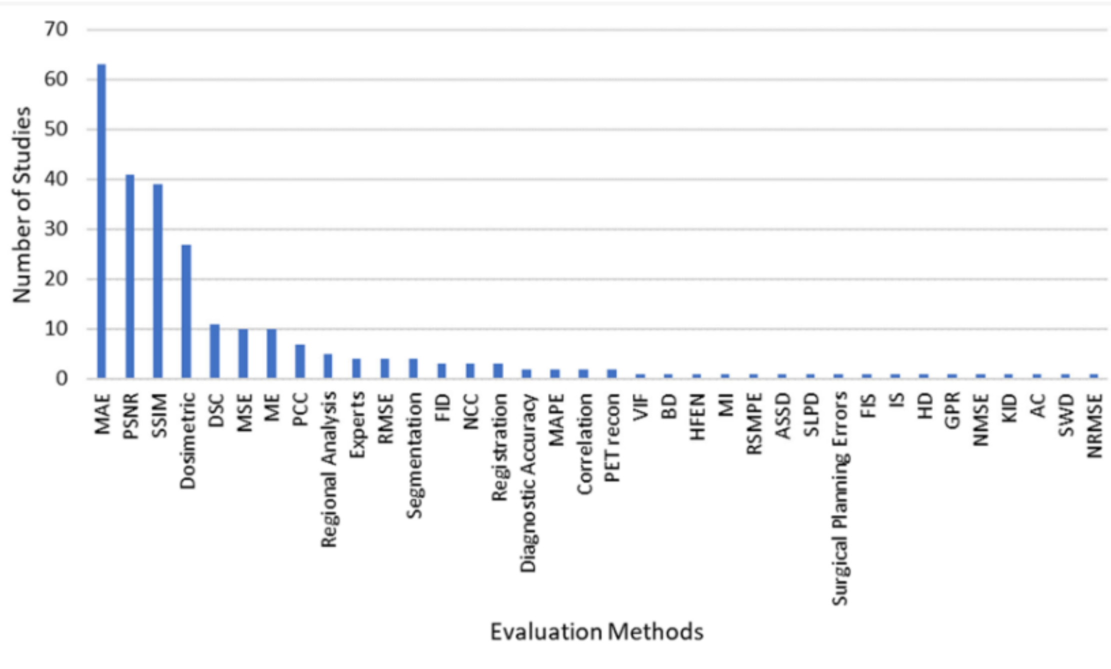


Figure 1.4: Methods for evaluating the synthetic images based on the studies seen in the review.

The most common metrics (used in more than 30 articles) include Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Thus, there seems to be no standard on which metrics to use, which would be useful for comparing performance between different studies. In addition, there does not seem to be agreement on the level of accuracy required for synthetic medical images.

Some models produce blurred images that mask the details of features on a smaller scale, despite the fact that the metrics give good results. It is therefore necessary to establish a benchmark for image quality in order to ensure the suitability of models for use in a clinical setting.

The review concludes that further research is needed to determine which deep learning methods are most effective and accurate in synthesising medical images for use in a clinical context.

1.1.1 An in-depth analysis of the literature

With the aim of specifically identifying the methods and materials for medical image translation, six articles were selected from the aforementioned review ([4], [5], [6], [7], [8], [9]), while another four articles were identified externally ([10], [11], [12], [13]). The ten articles selected for analysis were chosen based on the following criteria: the year of publication, selecting only articles from 2021 onwards; the type of architecture, limited to GANs; the presence of a public dataset composed of paired images.

The objective of all studies is to enhance the performance of GANs by incorporating additional networks, thereby combining different methodologies to generate superior images. Typically, these additions involve Transformer modules, as in the articles [4], [11] and [13]. The Transformer is a deep learning model that adopts the self-attention mechanism by differentially weighting the importance of each part of the input data [14].

Other articles ([8], [9], [10]) have instead used a CycleGAN, a GAN used to translate images from one class to another without the need for direct correspondence between input and output images. Ideally, however, they would like to have a correspondence between these two images in order to improve the training of generative models. In fact, it has been demonstrated that even CycleGANs give better results when the images are paired.

Finally, two papers ([6] and [7]) compare CycleGANs with U-Net, some of them modified with attention blocks, and show that U-Net performs better.

All reviewed articles agree on the usefulness and necessity of generating images of different domains from a single scan. In the context of radiotherapy based only on MRI scans, the generation of synthetic CT (sCT) images is crucial to obtain electron density information [8]. In fact, MRI images are widely used in radiotherapy because they provide superior soft tissue contrast compared to CT, allowing better differentiation of tumors. However, CT images are useful for dose calculation in treatment [9]. It should be noted that image acquisition in these two modalities introduces spatial uncertainties because the images are not aligned. These problems can be solved by synthesizing CT images from MRI images.

One problem with GANs is that they may fail to identify crucial relationships between the input and the target, as illustrated in the article [7]. The distribution of statistical data can have a significant impact on the output, with the potential to influence the inclusion or exclusion of crucial structures, such as lesions, in the generated data. GANs are particularly effective at tasks where there is no single correct answer, such as generating images from text or the translation of photographs into different artistic styles. In medical contexts, where the focus is on contrast and the presence of specific structures, it is critical to ensure that the structures of interest are accurately preserved in the generated image. Coupled GANs, such as *pix2pix*, attempt to address these issues by comparing the generated image to the true target. However, coupled GANs remain susceptible to challenges related to the tendency of GANs to adapt to the distribution of the training data. In fact, training GANs is notoriously difficult.

In the article [11], the model used is pre-trained on paired data and then retrained on unpaired data, incorporating knowledge from a pre-trained non-medical model. The application of Transfer Learning (TL), which is the technique of reusing a model developed for one task as a starting point for a model on another related task, demonstrated superior efficacy compared to data augmentation, a technique used to artificially increase the available training data. This is due to the fact that transfer learning requires less time to achieve comparable results. This approach, which exploits the knowledge acquired from a pre-trained model, has the potential to enhance efficiency and performance.

As previously stated, some articles were selected on the basis of the utilisation of open-source datasets. In fact, as reported in the review, the lack of medically paired datasets represents a significant challenge. Four datasets were identified from the different articles. Among the paired datasets, however, only one had already been aligned [15] using a software called Elastix [16]. The Section 2.2 explains the difference between pairing and alignment.

A further point of particular concern is the quantitative analysis of medical image synthesis. Evaluation metrics such as MAE, PSNR and SSIM are most commonly used, which was also evident in the review. In addition to these metrics, Mean Squared Error (MSE) and Fréchet Inception Distance (FID) are widely used as evaluation metrics in various studies.

Making quantitative comparisons between different articles is not always feasible, either because they do not all use the same metrics or because the image normalizations are different, resulting in incomparable values. In addition, different models treat different parts of the body and the direction of synthesis is not always the same. These factors should be taken into account in order to make effective comparisons.

Qualitative assessments are also important, including expert judgement, dose calculation, image brightness assessment or tissue contrast assessment. Some datasets contain images with tumor lesions, thus it is necessary to assess whether the lesion is propagated in the generated image using segmentation algorithms.

For istance, the article [6] reveals that qualitative analysis performed by experienced radiologists showed significant variability in the scores assigned. This highlights the challenges associated with subjective evaluation of synthetic images. In addition, it was discovered that the quantitative metrics do not fully reflect the visual sharpness of the images, underscoring the complexity of assessing the realistic quality of the generated images.

The article [5] presents a test, the *Visual Turing Test*, designed to assess image quality by experienced radiologists. This test was performed on 59 non-training patients to compare the similarity of axial synthetic MRI lumbar images generated from axial CT lumbar images and true axial MRI lumbar images. A scoring sheet was created containing 600 axial images, 150 of which were true MRI images and 450 of which were synthetic images generated from various studies. These images were randomly distributed on the card. Four participants participated in the study: two board-certified radiologists and two radiology residents. Each participant was presented with five images on a screen: one CT scan, one true MRI and three synthetic MRI images generated by three different algorithms. The participants had to select the two MRI images they thought were more accurate than the CT image, sorting them by accuracy.

The results demonstrate that images generated by supervised learning, i.e. by pairing and aligning images, exhibit the highest level of accuracy. These images were selected as the most accurate, although they were rarely selected as the first choice.

To conclude, since GANs may miss important details, making them less reliable for generating accurate medical images, it is essential to have robust evaluation methods that can accurately assess and quantify the quality of the generated images to ensure their safety and reliability for clinical use.

Chapter 2

Materials and Methods

This chapter describes the materials and methods used to implement a model for translating MRI into CT images. The training process of the model is illustrated, with a particular focus on the tuning of different loss functions and the evaluation of the generated images.

For the model architecture, the simple and well-established pix2pix architecture [17] was selected. The dataset used is the Gold Atlas, which provides paired and aligned MRI and CT images.

The following section provides an overview of GANs, with a particular emphasis on the pix2pix model and the concept of loss function. The dataset used is presented, the model architecture is described and the training process is outlined, including the selection of hyperparameters and the different loss functions employed. Finally, the selected evaluation methods are presented.

2.1 Fundamentals of GANs

Generative Adversarial Networks (GANs), introduced for the first time in 2014 by Ian Goodfellow and his collaborators [18], represent a novel deep learning architecture. The innovative idea of GANs is based on making use of two different neural networks that are trained in a competitive manner. The first network, called *Generator*, is supposed to generate data while the second network, called *Discriminator*, has to distinguish between real data and data created by the generator.

The training process of GANs is therefore a competition between the generator and the discriminator. The generator tries to produce data that are increasingly realistic, with the aim of deceiving the discriminator. In response, the discriminator keeps improving its ability to recognize synthetic data. This dynamic leads to an iterative enhancement of both networks.

Several modifications have been proposed for GANs in later years with the objective of improving their architecture. Some important examples include Deep Convolutional GAN (DCGAN), which introduces the use of CNN in GANs in order to systematically stabilize the training, and Wasserstein GAN (WGAN), using *Wasserstein Distance* as a loss function to avoid stability issues.

In 2016, the first GAN network for image-to-image translation was developed, the pix2pix. Furthermore, CycleGANs have been developed as an evolution of the pix2pix model, enabling translations between domains without the necessity of a direct correspondence.

Since their introduction, GANs have been employed for various applications in numerous domains due to their capability to generate outputs that seem to be of higher quality and more credible.

However, despite their impressive capabilities, GANs present challenges and limitations. Training a GAN is known to be highly difficult: *Mode Collapse*, in which the generator produces only a limited diversity of samples, and *Non-Convergence*, where the generator and discriminator never actually stabilize, are problems that require careful consideration of the architectural and hyperparameter choices.

2.1.1 Training of a GAN

The training process of a GAN involves an adversarial process, in which the generator and the discriminator compete against each other in a zero-sum game.

The discriminator is a simple classifier that tries to distinguish real data from generated data. During its training, the discriminator is presented with real data, used as positive examples, and fake data generated by the generator, used as negative examples. The generator is a neural network that accepts a random noise sample as input and produces an output that is close to the original distribution.

The discriminator classifies both real and false data by penalizing misclassifications using a loss function. Then, the discriminator updates its weights through backpropagation in order to maximize its loss function. During this phase, the generator is not trained and its weights remain constant while it produces data to be provided to the discriminator.

The next step consists of freezing the weights of the discriminator, meaning that they are no longer modified. The generator is then trained by updating its weights in order to minimize its loss function. In this case, the role of the discriminator is to classify the image generated by the generator. Based on this classification, the generator's weights are updated.

The process will then repeat, as the generator would have improved in generating the images. The figure 2.1 shows the training process of a GAN.

Both generator and discriminator are neural networks with specific parameters

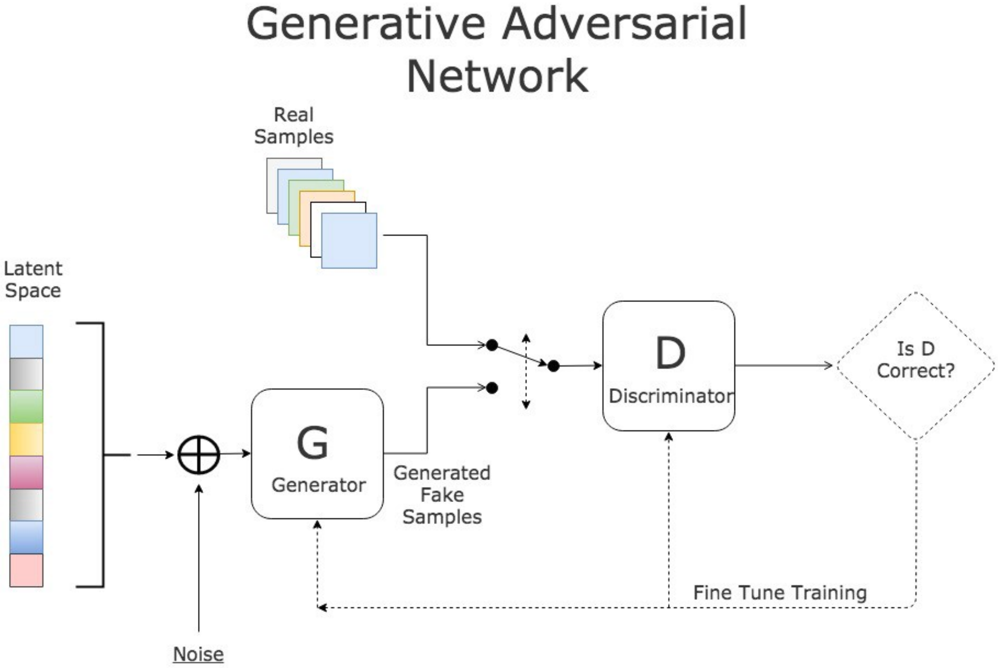


Figure 2.1: Training process of a GAN: the generator is given a noisy input from which it produces samples labeled as false. The training samples are instead labeled as real. All these samples, along with their corresponding labels, are shown to the discriminator which is trained to recognize real data from false data. During the training process of the generator, it generates data that are classified by the discriminator. Based on the result obtained, the generator modifies its weights.

that represent the characteristics of both networks. These parameters include layer size, number of neurons, activation functions, etc. In particular, the generator has parameters θ_g, and the discriminator has parameters θ_d.

A prior distribution p_z is defined from which the latent space z originates, which represents the random input. The generator represents a differentiable function G that maps the input data z, which come from distribution p_z, to a new distribution p_g of pseudo-samples x. The discriminator D produces a single scalar output, representing the probability that x comes from the training data - the real distribution - rather than from p_g.

The first step is to calculate the distribution of the training samples p_{data}. This distribution is difficult to determine. Traditional methods assume that the distribution $p_{data}(x)$ follows a mixed Gaussian distribution and use it as a solution through maximum likelihood. However, when the model is complicated, it is often not possible to calculate this, and the resulting performance is limited due to the limited expressive capacity of the Gaussian distribution itself.

The discriminator is trained to maximize the probability of correctly assigning the label to both the training samples and the samples from G. Simultaneously, the generator is trained to minimize its loss function with the goal of making the generated distribution p_g as similar as possible to the real distribution of images p_{data}.

The learning process can thus be conceptualised as a minmax optimisation problem of a two-player game between D and G:

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

2.1.2 pix2pix

The pix2pix model is an approach to image translation that employs Conditional GANs (cGANs), which represent an advancement over traditional GANs. A cGAN uses labels to generate new data with similar characteristics to the training data. In contrast to conventional GANs that use random noise to generate data, the network is provided with a condition or information regarding the desired output. This allows a more targeted and predictable outcome, rather than a random one.

Proposed by Isola et al. in 2017 [17], the pix2pix model addresses various image transformation problems where the goal is to learn a mapping from input images to output images through the use of paired training data.

In a pix2pix, the generator takes a target image x as input and generates an output image $G(x)$. The discriminator takes a pair of images as input, the generated image and the target image. The discriminator is tasked with identifying whether the pair of images is false or real. The goal of the discriminator's training is to ensure that each pair $\{x, G(x)\}$ is identified as false, while the training pairs $\{x, y\}$ are correctly classified as real. In contrast, the generator is trained by trying to make the pairs $\{x, G(x)\}$ be recognised as real, with the aim of fooling the discriminator.

The generator in pix2pix is typically a *U-Net*, a CNN developed specifically for biomedical image segmentation problems [19]. The architecture of a U-Net consists of an encoder that down-samples the input image into a feature map, and a decoder that up-samples it back to the original resolution, using deconvolution layers.

The distinctive feature of a U-Net is the skip connections between corresponding layers of the encoder and decoder, which facilitate the transfer of detailed information from the initial stages of the encoder to the late stages of the decoder.

The discriminator is a *PatchGAN*, a type of discriminative network that classifies not the entire image but small portions of the image, called patches. In this way, the discriminator can capture more detailed information about the image, thereby enhancing the network's efficiency in detecting local artefacts that might otherwise be missed if attention were focused on the entire image.

This approach enables the discriminator to provide more precise feedback to the generator, which can result in the generation of higher-quality images. Additionally, it reduces the computational complexity compared to analysing the entire image.

Each image patch consists of 70x70 pixels, and the discriminator will have an output of size NxN, where each N value classifies an individual image patch.

2.1.3 Loss Functions

A loss function, also known as a cost function or objective function, quantifies the difference between the output of a model and the actual values. The loss function, therefore, represents a measurement of the model's efficacy in predicting the expected outcome.

In the context of GANs, the loss function plays a pivotal role in guiding the generator and discriminator training process. The selection and design of the loss function can significantly influence the performance and behaviour of GANs.

There are different types of loss functions, some of which are explained below for the purposes of this thesis.

- *Adversarial Loss.*

  The adversarial loss is used in the context of adversarial training, particularly in GANs. Each GAN comprises two loss functions, one pertaining to the generator and the other to the discriminator, that operate in conjunction with one another within the context of adversarial learning. The generator loss encourages the generator to produce images that can fool the discriminator into classifying them as real. The discriminator loss encourages the discriminator to correctly distinguish between real and generated images.

  The standard adversarial loss for the generator G and the discriminator D in a GAN can be expressed as:

  $$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[\log(1 - D(G(\mathbf{z})))]$$

  where p_{data} is the distribution of real data and $p_{\mathbf{z}}$ is the distribution of the input noise vector.

  There are other types of adversarial loss functions, which are essentially modifications of the original GAN. The cGAN, for instance, employs a specific conditional loss function, defined as follows:

  $$\mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log D(\mathbf{x}, \mathbf{y})] + \mathbb{E}_{\mathbf{x},\mathbf{z}}[\log(1 - D(x, G(\mathbf{x}, \mathbf{z})))]$$

  where G is the generator, D the discriminator, x the input image, y is the ground truth image and z the random noise vector.

The Wasserstein distance, a loss function used in WGANs, evaluates the distance between two probability distributions: the actual data distribution and the one generated by the model.

Finally, the Least Squares GAN (LSGAN) model uses the least squares loss function for the discriminator.

Both WGAN and LSGAN will be explained in Section 2.3.1.

- *Loss Functions for Regression: L1 and L2.*

  Regression is a statistical problem concerned with the prediction of real variables called target from a set of independent variables, also known as features. Regression algorithms are designed to predict a continuous value based on a set of input variables.

  A variety of loss functions are available for the resolution of regression problems, including L1 and L2.

  L1 Loss, also know as Mean Absolute Error (MAE), measures the mean absolute difference between true and predicted value.

  $$\mathcal{L}_{L1} = |y - \hat{y}|$$

  where y is the real value of the target, $\hat{y}$ is the value predicted by the model and $|y - \hat{y}|$ is the absolute value of the error.

  L2 Loss, also known as Mean Squared Error (MSE), measures the mean squared difference between the actual value of the target y and the value predicted by the model $\hat{y}$:

  $$\mathcal{L}_{L2} = (y - \hat{y})^2$$

  where $(y - \hat{y})^2$ is the quadratic error.

  The L1 loss function is less susceptible to the influence of outliers than the L2 loss function, due to the fact that it does not square the error.

- *Loss Functions for Classification.*

  Classification is a type of problem in which the objective is to assign each piece of data to a predefined category or class. This is employed in scenarios where a discrete target input variable is present and the goal is to map it to discrete output variables.

  A widely used loss function for classification problems is the Cross-Entropy Loss, also called Log Loss. The cross-entropy loss quantifies the discrepancy between the probability distribution predicted by the model and the actual class distribution. In the context of a binary classification, the Binary Cross Entropy (BCE) is often used. In mathematical terms, it is defined as:

$$\mathcal{L}_{BCE}(y, f(x)) = -\left[y \cdot \log(f(x)) + (1-y) \cdot \log(1-f(x))\right]$$

where y is the true binary label (0 or 1) and $f(x)$ is the predicted probability that the observation belongs to class 1 (between 0 and 1).

- *Perceptual Loss.*

  The Perceptual loss function is a loss function used in ML to compare high-level features between images using pre-trained network, such as VGG or CNN. Unlike traditional pixel-based loss functions, a perceptual loss function is designed to capture perceptual and semantic differences, thereby resulting in images that are more visually natural. It is particularly useful in applications such as style transfer, super-resolution and image synthesis. However, it requires a pre-trained network, which can be computationally expensive, and the choice of comparison levels can affect the results.

The choice of loss function can profoundly influence the behavior and performance of GANs. Some loss functions can lead to more stable training of GANs. For instance, the WGAN uses a different loss formulation that improves stability and convergence compared to the standard adversarial loss.

Loss functions like perceptual loss can lead to higher quality and more realistic images because they focus on high-level features rather than just pixel-wise differences.

Specific loss functions can help mitigate mode collapse, a common problem in GANs where the generator produces limited diversity in the generated images.

In the case of pix2pix, the generator's adversarial loss function is modified by the addition of the L1 loss function, resulting in more powerful outcomes. The combined loss function is:

$$\mathcal{L}_{\text{pix2pix}} = \mathcal{L}_{\text{GAN}}(G, D) + \lambda \mathcal{L}_{L1}(G)$$

where λ is a weighting factor that balances the contribution of the L1 loss and the adversarial loss.

By carefully selecting and tuning the loss function, it is possible to improve the stability, convergence and overall quality of the generated images, leading to better performance.

2.1.4 Biomedical image translation challenges for GANs

The application of GANs in the context of biomedical image translation presents a multitude of advantages, but also presents significant challenges that need to be addressed to ensure effectiveness and safety, especially in a clinical setting.

Generative models are capable of producing images of remarkable realism. However, they can also introduce unintended alterations by adding, removing, or modifying details not present in the source image, a phenomenon known as hallucination.

This characteristic presents a significant risk in a clinical context, where even minor inaccuracies could potentially compromise a patient's diagnosis or treatment plan. Therefore, the reliability of GANs is contingent upon the capacity to develop systems that mitigate these errors, and ensuring the reliability of the generated images remains an open challenge.

A further limitation is the requirement for a substantial dataset for the training of GANs. Although GANs are capable of learning rapidly with a relatively small amount of data, the use of limited datasets can result in overfitting issues, reducing the model's ability to generalise correctly to new data.

Training GANs is a notoriously challenging process, with issues pertaining to model convergence and stability. There is a risk of mode collapse, where the generator produces nearly identical images despite different inputs. However, there is evidence that adversarial networks that perform translation between image domains show greater stability than traditional GANs because feature maps are strongly conditioned by an existing reference image [20].

2.2 Dataset

Several options were evaluated regarding the choice of dataset. For the pix2pix model, it is advisable to use a paired and aligned dataset. Among the datasets identified in the literature ([21], [22], [23], [24], [25]), only four are paired. However, the pairing of images does not necessarily include their alignment: two biomedical images are considered paired if they represent the same anatomical area or tissue of the same subject, but were acquired using different imaging modalities or at different times.

Image alignment is a specific technical process that involves bringing two or more images into a common spatial reference frame. This involves transforming one or more of the images to precisely match the same anatomical region of the reference image. The primary goal is to accurately overlay the images to allow for direct comparison. Alignment can be performed by automatic or manual techniques using transformation algorithms to minimize differences between images.

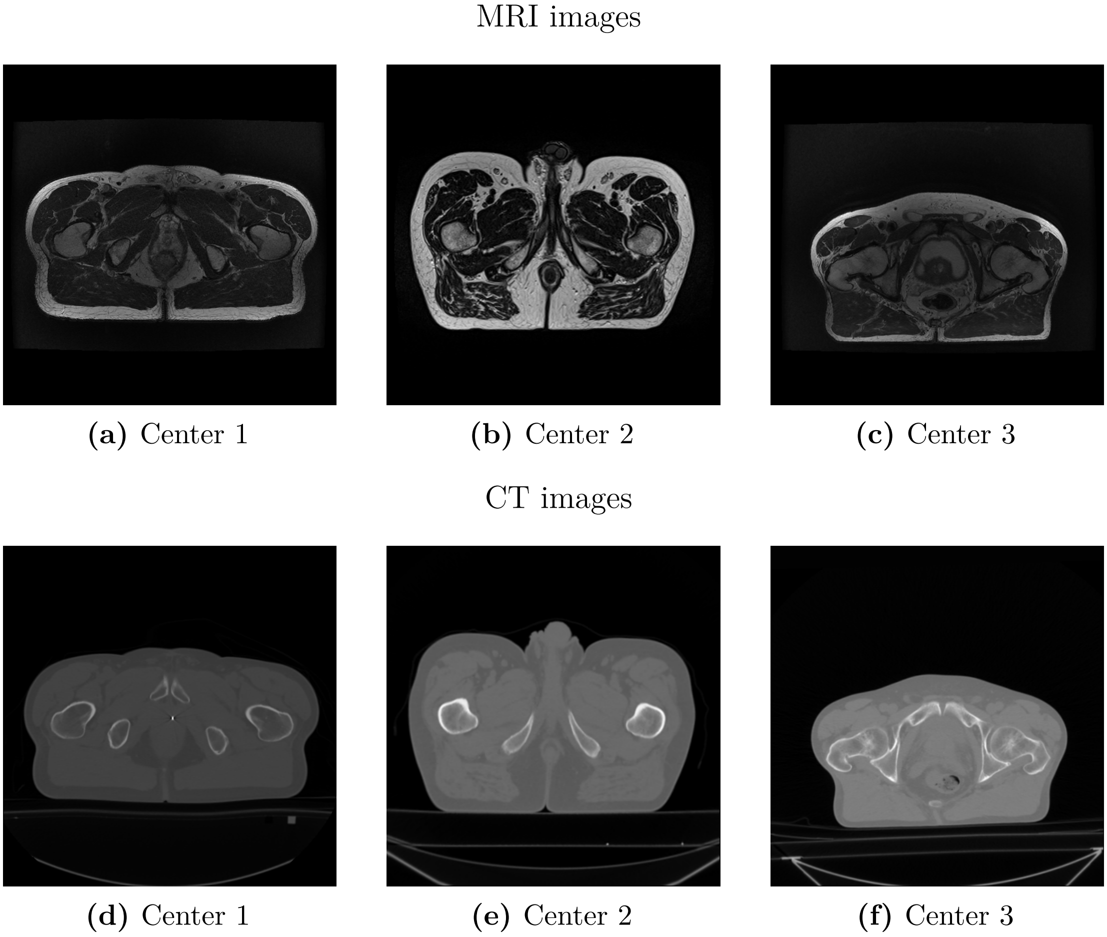
Software tools are available to achieve alignment of paired images, such as the Elastix software [16] used to align the Gold Atlas dataset. These software tools require the input of several parameters, which are not easy to define. The choice of these parameters can significantly affect the final alignment result and, consequently, the quality of the images generated by the translation model.

In order to achieve the objectives of this study, the Gold Atlas dataset was selected as it comprises paired and aligned MRI and CT scans.

2.2.1 Gold Atlas

The Gold Atlas dataset [23] contains pelvic scans from 19 male patients from 3 different diagnostic centers. In total, there are 1925 T1-weighted MR scans, 1788 T2-weighted MR scans, 2975 CT scans, and 1788 CT images aligned from the T2-weighted MR scans. The images are in *dicom* format.

The dataset can be downloaded from [15]. Some examples of images are shown in Figure 2.2.



MRI images

(a) Center 1 **(b)** Center 2 **(c)** Center 3

CT images

(d) Center 1 **(e)** Center 2 **(f)** Center 3

Figure 2.2: MRI and CT images aligned from the 3 different acquisition centers.

2.2.2 Data Preparation

The training and test sets were constructed by randomly sampling images from each centre, with 80% of the images allocated to the training set and 20% to the test set. The training set comprises 1429 images, while the test set consists of 359 images.

During image registration, especially in the case of aligned CT images, it is common to obtain some images that are completely black. These images contain no useful information and can interfere with model training. Therefore, it is important to identify and remove such images from the dataset.

After removing the black images, the training set contains 1402 paired images, while the test set contains 354 paired images.

Subsequently, the images were normalised. Prior to this, they were converted into *PyTorch* tensors with pixel values between 0 and 1. Normalisation was achieved by subtracting the mean from each pixel and dividing by the standard deviation, both of which were 0.5. This resulted in the values being mapped in the range [-1, 1].

This type of normalisation is common for deep learning models because it helps to stabilise and optimise the networks.

2.3 Model Architecture

The model used to translate MRI scans into CT images is based on the pix2pix architecture with a *U-Net256* generator and a *70x70 PatchGAN* discriminator [26].

The MRI image is used as input to the generator that generates a synthetic CT image. PatchGAN evaluates the pair generated CT image and input MRI image and determines whether it is real or not.

The training goal is to minimize the generator loss function and maximize the discriminator loss function, thus progressively improving the quality of the generated images.

The training process of the model is schematised in the Figure 2.3.

2.3.1 Loss Functions

Three classic loss functions were employed in this study: Binary Cross Entropy (BCE), Least Squares GAN (LSGAN) and Wasserstein GAN with Gradient Penalty (WGANGP). The $L1$ loss function was added for BCE and LSGAN.

In addition, two no-reference metrics, Integrated Local Natural Image Quality Evaluator (ILNIQE) and Natural Image Quality Evaluator (NIQE), which are discussed in detail in the Section 2.5, were used as loss functions.

The different loss functions are discussed in detail below.

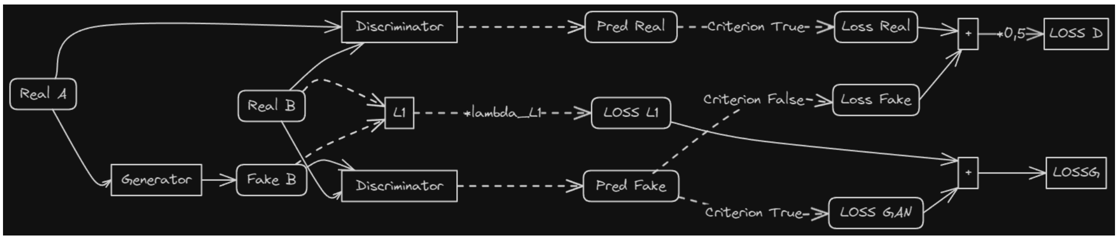


Figure 2.3: Training scheme and loss functions of pix2pix. Real image A is given as input to the generator, which produces a false image B. The corresponding real image B, or target image, and the generated image B are given to the discriminator, which classifies the pair as real or false, thus giving a prediction for the false image. By applying the True criterion, according to which the image should be classified as real, the first Loss Function, Loss GAN, is calculated. In addition, the false image produced is compared with the reference image and the L1 norm is calculated, resulting in L1 Loss. Loss GAN and L1 Loss together give the overall loss function for generator training. In regard to discriminator training, its loss function is the average of two loss functions. The first, Loss Fake, is derived by applying the criterion that the prediction made by the discriminator on a fake image must be False. The second, Loss Real, is obtained by providing the discriminator with the real input image A and the target image B, and applying the criterion that its prediction must be True.

Binary Cross Entropy (BCE)

BCE measures the error between predicted probabilities and actual binary labels.

The combination of BCE and L1 Loss is commonly used in binary regression and classification applications that require an additional penalty for prediction errors. This loss function is particularly useful in models that must balance classification accuracy with consistency in predicting continuous values.

The combined BCE+L1 loss function is given by the weighted sum of these two components:

$$\mathcal{L}_{\text{BCE+L1}} = \text{BCE} + \lambda \cdot \text{L1}$$

where λ is a hyperparameter that balances the contribution of L1 loss against BCE. Usually, λ is equal to 100.

Least Squares GAN (LSGAN)

LSGAN is a variant of GANs that uses a loss function based on the squares of the residuals, instead of the traditional binary loss function, to provide feedback that is more stable and less susceptible to explosive gradients.

The loss function for the discriminator in LSGAN is:

$$\mathcal{L}_{\text{D}} = \frac{1}{2}\left(\mathbb{E}_{x\sim p_{\text{data}}}[(D(x)-1)^2] + \mathbb{E}_{z\sim p_{\text{z}}}[(D(G(z)))^2]\right)$$

where $D(x)$ is the probability that the input x is real, and $G(z)$ is the generated image.

The combination of LSGAN and L1 Loss is used to improve the quality of the generated images.

The loss function for the generator in LSGAN is:

$$\mathcal{L}_{\text{G}} = \frac{1}{2}\mathbb{E}_{z\sim p_{\text{z}}}[(D(G(z))-1)^2]$$

The combined LSGAN+L1 loss function is given by:

$$\mathcal{L}_{\text{LSGAN+L1}} = \mathcal{L}_{\text{G}} + \lambda \cdot \text{L1}$$

where λ is a hyperparameter controlling the relative importance of L1 Loss with respect to the generative loss of LSGAN. As before, λ is usually equal to 100.

Wasserstein GAN with Gradient Penalty (WGANGP)

The WGANGP is an advanced variant of GANs that uses the Wasserstein distance as a measure of loss. This approach provides greater stability and convergence than traditional GANs by using the Wasserstein distance to calculate the divergence between the generated and true image distribution.

The loss function for WGANGP is based on the Wasserstein distance between the distributions and can be expressed as

$$\mathcal{L}_{\text{D}} = \mathbb{E}_{x\sim p_{\text{data}}}[D(x)] - \mathbb{E}_{z\sim p_{\text{z}}}[D(G(z))]$$

where $D(x)$ is the score of the discriminator for the real image, and $D(G(z))$ is the discriminator score for the generated image.

In addition, WGANGP includes a gradient penalty term to help keep the discriminator value function in a stable region. The gradient penalty is given by

$$\mathcal{L}_{\text{GP}} = \mathbb{E}_{\hat{x}\sim\hat{p}}\left[(\nabla_{\hat{x}}D(\hat{x})\|_2 - 1)^2\right]$$

where $\hat{x}$ is an interpolated sample between the real and generated images.

Thus, the total loss function for WGANGP is

$$\mathcal{L}_{\text{WGANGP}} = \mathcal{L}_{\text{D}} + \lambda \cdot \mathcal{L}_{\text{GP}}$$

where λ is a hyperparameter that offsets the gradient penalty term. λ was chosen equal to 10.

Natural Image Quality Evaluator (NIQE) and Integrated Local Natural Image Quality Evaluator (ILNIQE)

NIQE and ILNIQE are metrics used to assess image quality, particularly without the need for a reference image. These metrics, which will be detailed in Section 2.5, evaluate the quality of an image by analysing the natural statistical properties of high quality images.

Both these metrics are used as perceptual loss functions: the metric is applied to the generated image and the reference image, after which the difference between the two values is calculated and the result is used in backpropagation.

In order to use these metrics as loss functions, the implementation in Python was employed, [27] for NIQE and [28] for ILNIQE.

The calculation of both these metrics requires some parameters obtained after the training process that is performed in MatLab, as explained in Section 2.5.

2.4 Training details

On the choice of hyperparameters there has been the problem of overfitting. In comparison to conventional neural networks, GANs appear to exhibit a reduced propensity for overfitting. This is primarily due to the fact that the generator does not receive the anticipated output directly, but rather receives feedback from the discriminator. The generator does not have access to the training set: the information about the training data comes from the discriminator. In other words, the generator is unable to directly replicate examples from the training set [29].

Furthermore, the adversarial process serves as a form of regularisation, as the generator is required to continuously enhance its performance in order to remain aligned with the evolving discriminator. The generator's capacity for constant improvement and adaptation renders it more challenging for it to overfit.

For these reasons, there was minimal attention devoted to the tuning of hyperparameters, with the following parameters maintained at a fixed value: learning rate of 0.0002, batch size of 1, Adam as optimizer with λ_1 equal to 0.5 and λ_2 equal to 0.999. The total number of epochs is typically set at 500.

2.4.1 Transfer Learning

TL is a machine learning technique whereby a model that has been trained on a specific task is reused as the basis for another related task. In contrast to training a model from the outset, a pre-trained model based on a substantial dataset is employed and adapted to a novel task, frequently utilising a smaller dataset. The benefit of TL is that the pre-trained model has already acquired pertinent features from the initial data, which can be reused and adapted to the new task, accelerating the training process and enhancing performance, particularly when the data available for the new task is limited.

In the context of this study, a number of pre-trained models [30] were employed and subsequently subjected to a fine-tuning process on the specific medical dataset.

In particular, only the generator weights were extracted from the pre-trained model. This approach permitted the generator training process to start from a point where the model had already acquired useful knowledge regarding the translation of images from one domain to another, thus obviating the necessity to start from scratch. Subsequently, the aforementioned weights were fine-tuned using the specific medical dataset.

The pre-trained models employed in this study comprise those designed for day-to-night translation (*day2night*), label-to-photo image translation (*label2photo*), map-to-satellite image translation (*map2sat*) and its inverse (*sat2map*).

2.5 Evaluation methods

In the field of medical imaging, most evaluation metrics for such images are full reference [31], meaning they require an original reference image to assess quality. These metrics compare the generated image to the reference image to determine the fidelity and quality of reproduction. The most common full reference metrics used in the biomedical field, and also in this study, are Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Mean Absolute Error (MAE) and Mean Squared Error (MSE).

There are also metrics called No-Reference (NR) metrics, or Blind Image Quality Assessment (BIQA), that do not require a reference image to evaluate image quality. These metrics are especially useful when a reference image is not available.

NR metrics fall into two categories: opinion aware and opinion unaware metrics. Opinion aware metrics are designed to be aware of the subjective opinions of users. These metrics are trained on datasets containing distorted images accompanied by subjective scores assigned by human observers. The training process is supervised, which means that the algorithm learns to assess the quality of the images based on the collected human judgements.

However, in order to obtain an opinion aware metric, it is necessary to have many image samples and their subjective scores, which can be expensive and laborious to collect. In addition, these metrics tend to have a relatively weak generalising ability, as they are closely linked to the data on which they are trained. Despite these limitations, opinion aware metrics generally perform better than other metrics precisely because they more closely reflect human perceptions of image quality.

Opinion unaware metrics do not require the use of subjective scores for their training. This means that there is no need to collect human ratings for images, making the development process less costly and less complex. Due to this feature, opinion unaware metrics tend to have a superior generalisation capability, adapting better to a variety of scenarios and data not seen during training.

For the purposes of this research, it would not have been easy to obtain subjective scores from experienced radiologists, so only opinion unaware metrics were considered. Some examples of these metrics, which were used in this research, include Natural Image Quality Evaluator (NIQE), Integrated Local Natural Image Quality Evaluator (ILNIQE) and Perception-based Image Quality Evaluator (PIQE).

Mean Absolute Error (MAE)

MAE is a linear error measure that represents the distance, in absolute value, between the predicted value and the actual value. It is particularly useful when a simple and easily interpretable measure is desired, and it is resistant to outliers.

Given a set of n samples, MAE is calculated by averaging the absolute error between the predicted value $y_{\text{pred},i}$ and the real value $y_{\text{true},i}$ of each i-th sample:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_{\text{pred},i} - y_{\text{true},i}|$$

Lower MAE values indicate greater similarity between the two images.

Mean Squared Error (MSE)

MSE is a metric that measures the mean of the squares of the differences between predicted and real values. The MSE is a more robust measure of error than the MAE, as it penalises larger errors more severely by squaring the differences. Consequently, the MSE is a preferred metric in contexts where large errors are particularly costly, such as in the medical field.

The formula for calculating the MSE is as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_{\text{pred},i} - y_{\text{true},i})^2$$

where $y_{\text{pred},i}$ is the predicted value for the sample i, $y_{\text{true},i}$ is the true value for the sample i and n is the total number of samples.

Lower MSE values indicate greater similarity between the two images.

Peak Signal-to-Noise Ratio (PSNR)

PSNR is a metric that is primarily employed for the assessment of the quality of compressed or reconstructed images and videos. The PSNR metric compares the maximum power of a signal with the power of noise that affects the fidelity of the representation. The value is expressed in terms of the logarithmic decibel scale (dB), with higher values indicating greater similarity to the original image and thus superior quality.

The formula for calculating PSNR is as follows:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{\text{MSE}}\right)$$

where MAX_I is the maximum possible pixel value (e.g., 255 for an 8-bit image) and MSE is the Mean Squared Error between the original image and the reconstructed image.

Structural Similarity Index (SSIM)

SSIM is an advanced metric used to measure the perceived similarity between two images. Unlike MAE, MSE and PSNR, which are based on pixel-per-pixel differences, SSIM evaluates images in terms of luminance, contrast and texture, providing a more accurate assessment of visual quality.

The formula for calculating SSIM is as follows:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where:

- x and y are the images to be compared;
- μ_x and μ_y are the averages of the images x and y;
- σ_x^2 and σ_y^2 are the variances of the images x and y;
- σ_{xy} is the covariance between x and y;
- C_1 and C_2 are two constants to stabilize the division in cases where $\mu_x^2 + \mu_y^2$ and $\sigma_x^2 + \sigma_y^2$ are very small.

SSIM ranges from -1 to 1, where 1 indicates identical images.

Natural Image Quality Evaluator (NIQE)

NIQE evaluates image quality based on statistical models of natural image features [32], without the need for a reference image. It builds a simple space domain natural scene statistic model and then extracts features from it to construct a "quality aware" collection of statistical features.

The Multivariate Gaussian (MVG) model is used to estimate the global distribution of natural high quality images, which is compared with the corresponding MVG model of the distorted image to predict the final quality score.

The formula can be expressed as follows:

$$\text{NIQE}(I) = \sqrt{(\mu_I - \mu_r)^T \left(\frac{\Sigma_I + \Sigma_r}{2}\right)^{-1} (\mu_I - \mu_r)}$$

where:

- I is the image to be evaluated;
- μ_I and Σ_I are respectively the mean and the covariance matrix of the statistical features of the image I;
- μ_r and Σ_r are respectively the mean and the covariance matrix of the statistical features of a natural image model.

A lower NIQE score indicates better image quality.

In order to estimate the global distribution of high-quality images, a natural image model was constructed comprising all CT images in the Gold Atlas dataset, including both acquired and aligned images. The dataset was employed for the purpose of training the NIQE metric.

The training of this metric and its calculation were both performed in the *MatLab* environment using the *niqe* function [33].

Integrated Local Natural Image Quality Evaluator (ILNIQE)

ILNIQE is an advanced metric that combines multiple local and global image features to provide a reference-free quality assessment [34]. ILNIQE is an extension of NIQE that improves the quality score by taking into account local variations within the image. Unlike NIQE, which uses a global model, ILNIQE performs a local evaluation, analysing different regions of the image to provide a more detailed assessment. In addition, colour, gradient and frequency features are included in the feature extraction step. As with NIQE, a lower ILNIQE score indicates better image quality.

Once more, the metric was initially trained using the identical dataset of CT images employed for the NIQE metric.

The training and the calculation of this metric were both performed using the *MatLab* implementation provided by the authors in the original paper [34].

Perception-based Image Quality Evaluator (PIQE)

PIQE is a metric that evaluates image quality based on human perception by analyzing non-overlapping blocks of the image to estimate local distortion [35].

The following steps are involved in the PIQE algorithm:

1. Calculate the Mean Subtracted Contrast Normalised (MSCN) coefficient for each pixel in the image.
2. Divide the input image into non-overlapping blocks and identify the spatially highly active blocks based on the variance of the MSCN coefficients.
3. In each block, assign a distortion score due to block artefacts and noise using the MSCN coefficients.
4. Aggregate the distortion scores to calculate the overall PIQE score.

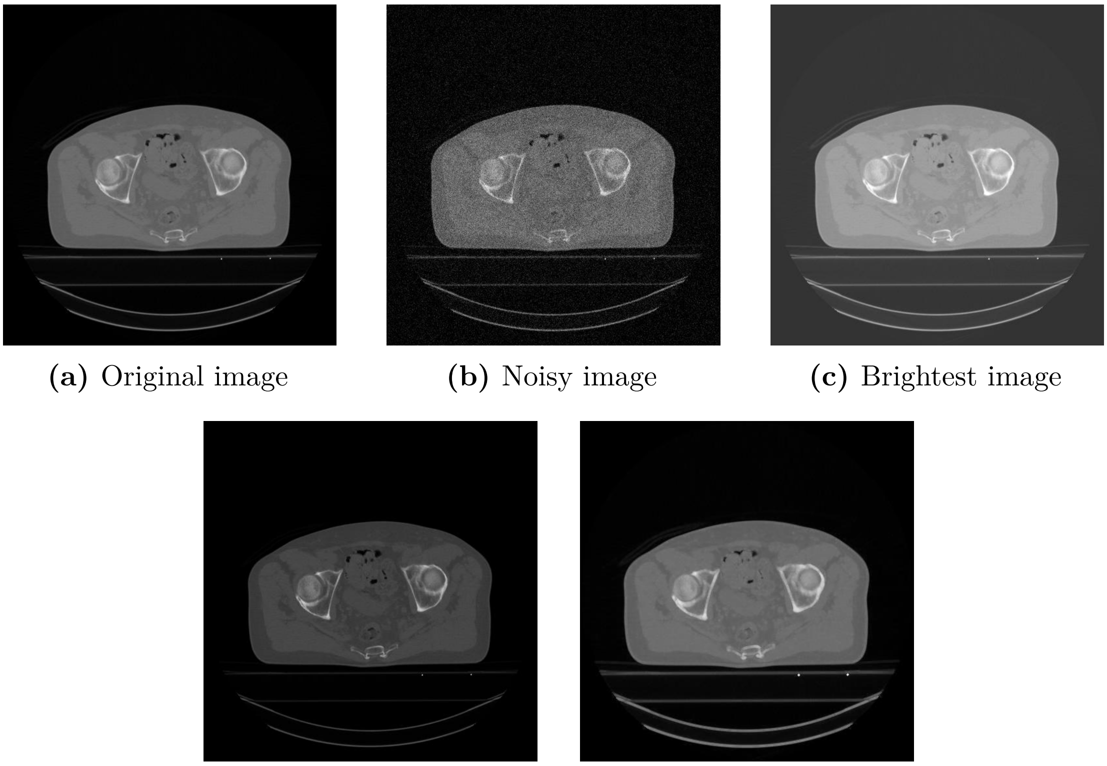
The final result is a score representing the quality of the image, where lower values indicate superior quality and higher values indicate higher perceived distortion.

In contrast to NIQE and ILNIQE, PIQE does not necessitate training. The function developed by Matlab has been used [36].

2.5.1 Addition of artificial distortions

In order to assess the ability and efficacy of NR metrics to evaluate CT images, a series of distortions were added to the original CT images (Figure 2.4) to simulate different levels of image degradation. The original CT images include both those acquired and those subsequently aligned, amounting to 4731 images in total. The introduced distortions include:

- **Noise**: White gaussian noise, with a mean of 0 and variance of 0.005, was applied. This type of distortion introduces random variations in pixel intensity, similar to the noise that can occur during image acquisition.
- **Brightness**: The brightness of the image was increased by adding a value of 0.2 to each pixel. This results in an overall brighter image.
- **Contrast**: A contrast factor of 1.5 was applied, increasing the difference between light and dark pixels. This distortion makes the bright areas of the image brighter and the dark areas darker, enhancing the distinction between different regions of the image.
- **Dilation**: A disc-shaped structural element with a radius of 1 pixel was employed for the purpose of dilation. This results in the structures appearing larger, causing a blurring effect.



(a) Original image **(b)** Noisy image **(c)** Brightest image

(d) Increased contrast **(e)** Dilated image

Figure 2.4: The images were modified by introducing artificial distortions that were not present in the original image (a). White Gaussian noise was added (b), the brightness was increased (c), the contrast was increased (d) and dilation was applied (e).

Chapter 3

Results

This chapter presents the findings of the experiments conducted, illustrating the methodology employed for each macro-experiment and the results obtained.

The initial phase of the study involves the tuning of loss functions and the utilisation of several pre-trained models from which TL was carried out. Once the optimal combinations for the model had been identified based on the classic Full-Reference (FR) metrics, the generated images were evaluated using NR metrics. This was done to ascertain how they were evaluated differently compared to the previous metrics and to determine the reliability of their evaluation. Ultimately, two of these NR metrics were employed as loss functions.

3.1 MRI to CT translation: tuning of Loss Functions and Transfer Learning

In the initial trials, a number of pre-trained models, renowned for their efficacy in processing large datasets, were selected. The weights of these models were loaded from the pre-trained checkpoints, thus providing a robust foundation for fine-tuning on the experiment-specific dataset. Furthermore, a number of loss functions were evaluated for each pre-trained model.

Following the training phase, the models were evaluated on the test set in order to ascertain their capacity to generalise to data that had not been seen during training. The results of the various combinations were then compared in order to ascertain which loss function yielded the most optimal results for different pre-trained model.

Three distinct loss functions were evaluated at this stage: BCE, LSGAN and WGANGP. For each loss function, different pre-trained models were used to perform TL, as shown in the Table 3.1.

Loss Function	**Transfer Learning (TL)**
BCE+L1	no TL
BCE+L1	label2photo
BCE+L1	sat2map
BCE+L1	map2sat
BCE+L1	day2night
LSGAN+L1	no TL
LSGAN+L1	map2sat
LSGAN+L1	day2night
WGANGP	no TL

Table 3.1: Combination of loss functions and pre-trained models employed for transfer learning.

The outcomes for the BCE and LSGAN loss functions are presented in Table 3.2, where the best values for each metric are highlighted. In both cases, the optimal results are obtained when TL is not employed.

BCE + L1				
TL	**MAE**	**MSE**	**PSNR**	**SSIM**
None	**0.0746** ± 0.0598	**0.0239** ± 0.0397	**18.6500** ± 3.9336	**0.7040** ± 0.1086
sat2map	0.0797 ± 0.0617	0.0266 ± 0.0428	18.0038 ± 3.6949	0.6739 ± 0.0984
label2photo	0.0792 ± 0.0691	0.0279 ± 0.0488	18.308 ± 4.0935	0.6827 ± 0.1136
day2night	0.0811 ± 0.0612	0.0276 ± 0.0427	17.8112 ± 3.7579	0.6605 ± 0.1082
map2sat	0.0812 ± 0.0574	0.0249 ± 0.0376	17.8966 ± 3.3251	0.6052 ± 0.1049

LSGAN + L1				
TL	**MAE**	**MSE**	**PSNR**	**SSIM**
None	**0.0744** ± 0.0637	**0.0245** ± 0.0441	**18.5832** ± 3.7475	**0.6992** ± 0.1056
map2sat	0.0809 ± 0.0600	0.0277 ± 0.0406	17.6582 ± 3.6754	0.6819 ± 0.1071
day2night	0.1086 ± 0.0957	0.046 ± 0.0716	16.4509 ± 4.6892	0.637 ± 0.1682

Table 3.2: A comparative analysis of the performance of various transfer learning models using the Binary Cross-Entropy (BCE) loss function with L1 regularisation and the Least Square GAN (LSGAN) loss function with L1 regularisation. The mean value and standard deviation are given for each metric.

With regard to WGANGP, the images generated at an early stage of training are characterised by a high degree of noise, and there is no discernible improvement as the training progresses, regardless of whether TL is employed.

For this reason, it was decided not to continue with the use of WGANGP as loss function, which proved to be unsuitable for the purposes of this study. Figure 3.1 shows examples of images generated using WGANGP as loss function.

Figure 3.2 illustrates the results of some test images generated based on the best proof, which employs a BCE loss function and does not utilise TL. It is evident that the first generated image (b) exhibits a high degree of visual disturbance, although this may be attributed to the fact that it was derived from a particularly dark input image. Indeed, most of the images generated that are disturbed have dark images as input.

Conversely, in the second case (e) the generated image displays a notable degree of visual realism when compared to the target image.

Further examples of images generated by the other tests are provided in the Appendix A.

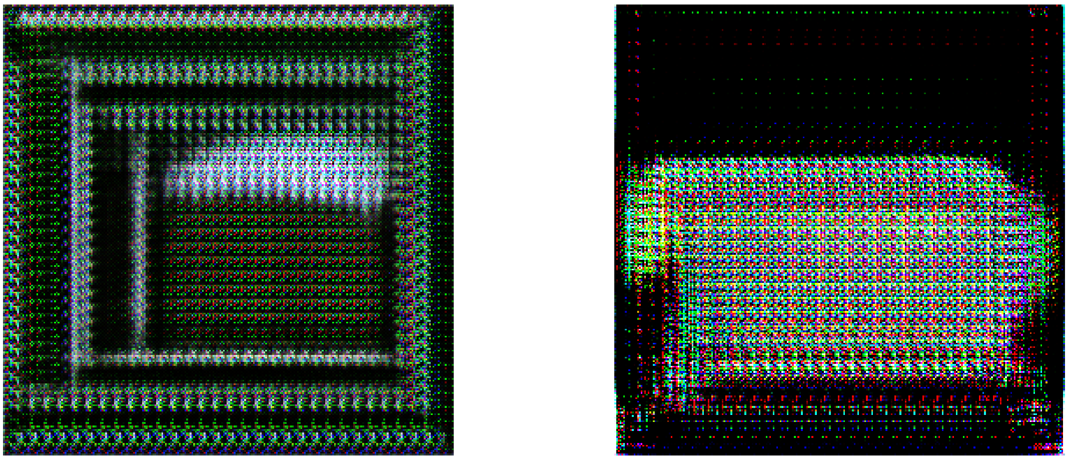
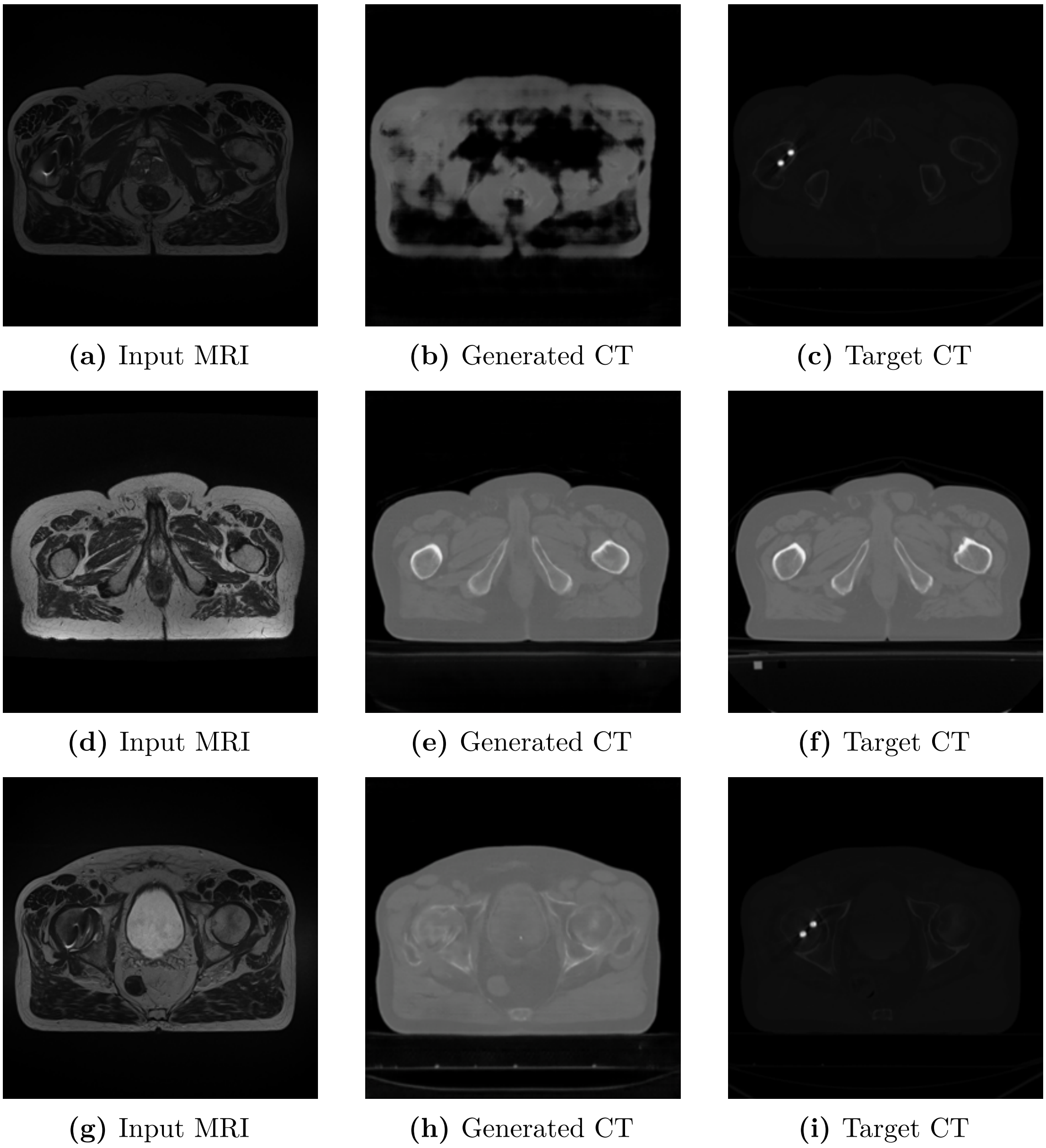


Figure 3.1: Images generated at different epochs of the training process using the WGANGP loss function.



(a) Input MRI

(b) Generated CT

(c) Target CT

(d) Input MRI

(e) Generated CT

(f) Target CT

(g) Input MRI

(h) Generated CT

(i) Target CT

Figure 3.2: Images generated using the optimal test parameters, specifically a BCE loss function and no transfer learning. In instances where the input image is particularly dark (a), the model produces an image (b) that is markedly disturbed.

3.2 Evaluation of images with No-Reference metrics

The following experiment is designed to evaluate the previously generated images using NR metrics, to highlight any differences in evaluation compared to the FR metrics. The table 3.3 shows the results obtained and the best values for each metric are highlighted.

In contrast to the findings of FR metrics, the results of NR metrics indicate a divergent trend across different tests, suggesting that there is no consensus on the optimal test based on the obtained values.

BCE + L1			
TL	**NIQE**	**ILNIQE**	**PIQE**
None	12.4894 ± 1.7046	**45.7019** ± 3.5649	43.3655 ± 5.1703
sat2map	12.4772 ± 1.6389	47.1191 ± 4.1651	43.3568 ± 5.1836
label2photo	**11.7013** ± 1.6330	51.7022 ± 5.9915	43.5101 ± 5.2716
day2night	13.5151 ± 1.8433	52.7218 ± 6.0553	41.9557 ± 5.8597
map2sat	15.5571 ± 1.9670	58.7143 ± 7.3563	**28.3280** ± 6.2557

Table 3.3: A comparison of No-Reference metrics for different types of transfer learning on the test using BCE+L1 as loss function. The mean value and standard deviation are given for each metric.

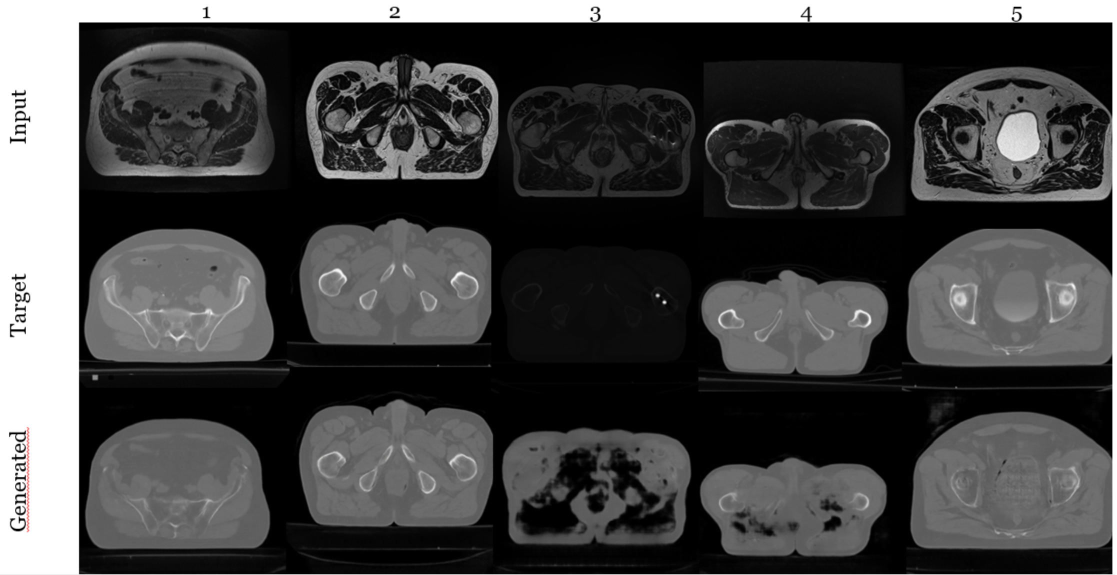
In order to compare the evaluation of the different metrics, examples of generated images with the results of the used metrics are shown in the Figure 3.3. The images were generated using BCE+L1 as loss function and without TL, which represents the optimal test result at least according to traditional metrics.

The FR metrics appear to indicate a consensus regarding the optimal quality image (the second one), whereas the NR metrics do not. However, it is possible to conduct multiple comparisons between the values of the metrics and the quality of the images. For instance, the SSIM metric assigns the same value (0.57) to both the first and third images, despite the significant differences in quality.

3.2.1 Addiction of distortions

At this point, distortions were added to the original CT images to see how the NR metrics evaluate these distorted images. Table 3.4 shows the results of the metrics for different types of distortion. For better visualisation, the metric values have been graphed in the Figure 3.4.

As expected, all values are better in the case of original images. Noise corrupted images are rated very poorly by the NIQE and ILNIQE metrics, where the deviation



MAE	0.0919	**0.0351**	0.2040	0.0671	0.0523
MSE	0.0190	**0.0062**	0.1121	0.0209	0.0106
PSNR	17.2146	**22.1002**	9.5030	16.8077	19.7485
SSIM	0.5723	**0.7798**	0.5717	0.7034	0.7356
NIQE	14.1028	11.7025	13.3494	13.6943	**10.8760**
ILNIQE	47.1147	49.8288	51.8468	**45.0584**	45.9387
PIQE	46.0285	46.0285	56.4836	**30.2080**	36.6761

Figure 3.3: The generated images are presented alongside the input and target images, along with the corresponding metrics.

from the other types of distortion is significant, while PIQE rates the quality as better.

It is possible to compare NR metric values on distorted images with those on generated images. According to NIQE, the performances on generated images are comparable to images with added noise. According to ILNIQE, generated images are more comparable to images modified by adding brightness, contrast and dilation. According to PIQE, generated images have even better quality than original images.

DISTORTIONS	NIQE	IL-NIQE	PIQE
Original	3.3354 ± 0.9296	10.9848 ± 1.2006	48.9643 ± 13.3857
Noise	13.1102 ± 1.5888	87.5214 ± 11.1443	60.2569 ± 1.6544
Brightness	9.4724 ± 1.0862	39.6495 ± 7.1086	61.7990 ± 5.1047
Contrast	9.3102 ± 1.8813	36.8218 ± 7.5531	63.7549 ± 6.0060
Dilation	10.9482 ± 1.1422	37.8523 ± 4.5924	64.2619 ± 4.5485

Table 3.4: No-Reference metric values under different types of distortion compared to the original images. The mean value and standard deviation are given for each metric.

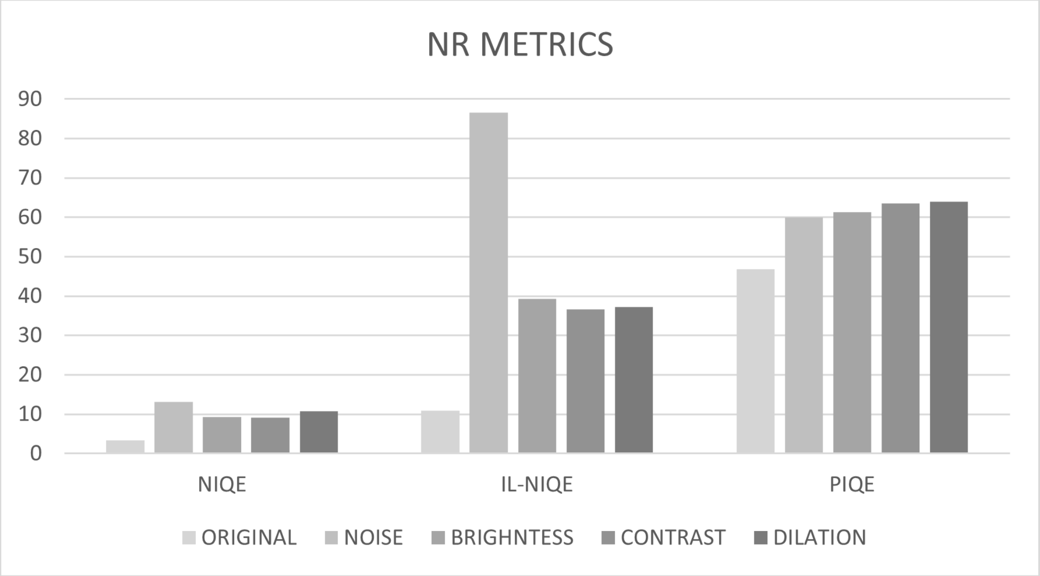


Figure 3.4: NR metrics for the different distortions applied to the original images.

3.3 ILNIQE as Loss Function

Among the NR metrics employed for the performance evaluation, ILNIQE proved to be the most effective in assessing the realistic quality of the generated images. Therefore, it was chosen as loss function. In particular, finetuning was performed on the pre-trained model using BCE+L1 as the loss function and without TL, which results as the optimal approach from the previous iterations. Training was then continued using ILNIQE as the new loss function for the generator.

The primary issue at this point is that the metric ILNIQE requires a processing time that is excessively long. The function, implemented in Python, requires over 40 seconds to calculate the quality of a single image. An initial test was conducted by training the model for a single epoch, which took 38 hours.

3.4 NIQE as Loss Function

In light of the above-mentioned limitations of ILNIQE, NIQE was employed as loss function for the generator, and several experiments were conducted. Initially, an attempt was made to commence training from scratch using solely NIQE as loss function. However, these attempts proved unsuccessful, as the model was unable to generate realistic images (Figure 3.5).

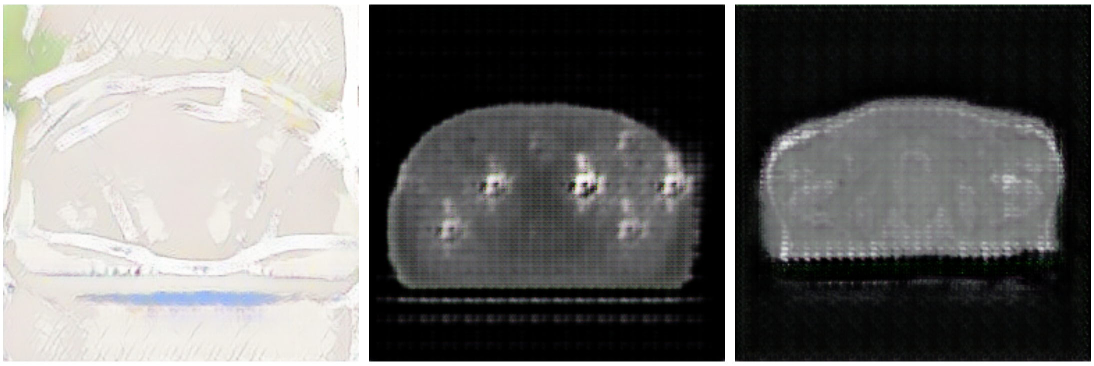


Figure 3.5: Images generated by the model trained using NIQE as the only loss function for the generator.

Subsequently, NIQE was combined with the BCE function and an attempt was made to train the model for 100 epochs and then continue the training by fine-tuning the model using only BCE as loss function. In this case, the images generated during the first 100 epochs with NIQE loss exhibited poor quality (Figure 3.6) which, however, improves when training was continued using only BCE loss.

The final experiment employs a pre-trained model with a BCE+L1 loss function

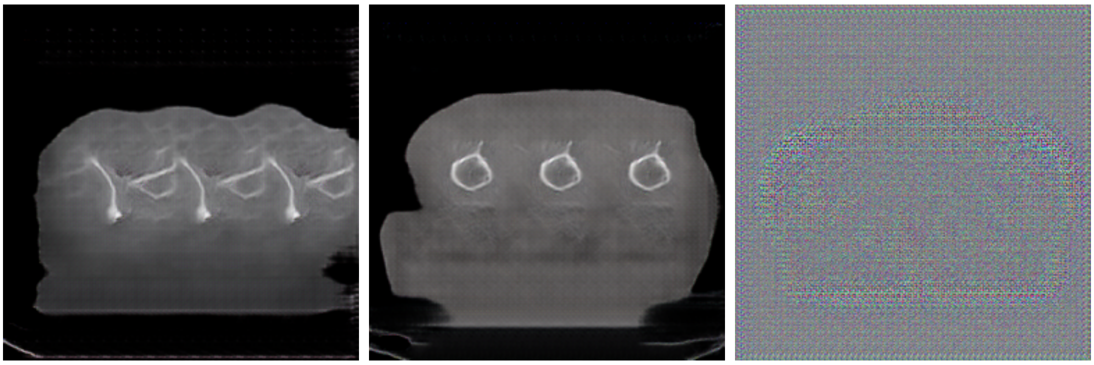


Figure 3.6: Images generated during the initial 100 training epochs of the model with NIQE associated with BCE as the loss function for the generator.

and no TL, and fine tuning using NIQE in combination with BCE as loss function for the generator. Several tests were carried out by changing the weights of NIQE within the loss function. The new loss function, therefore, can be expressed as follows:

$$\mathcal{L} = \lambda_{NIQE} * NIQE + BCE$$

where λ_{NIQE} represents the weight associated with NIQE, and varies from 0 to 1 in increments of 0.1.

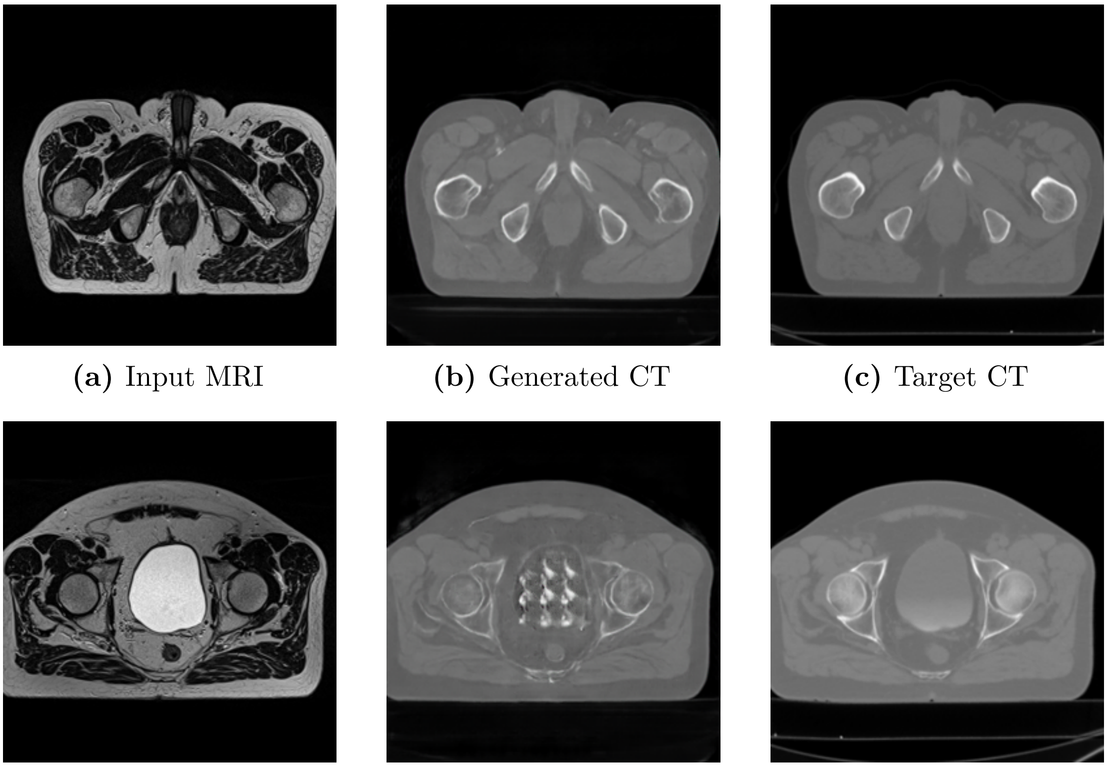
The Table 3.5 presents the results of the final test in terms of performance.

Figure 3.7 shows the images obtained from the optimal test, wherein the λ_{NIQE} value was set to 0.3. In some cases, the model went into Mode Collapse and all the images of the generated test set were identical.

The images of the other tests are shown in the Appendix A.

λ_{NIQE}	**MAE**	**MSE**	**PSNR**	**SSIM**
1	0.0803 ± 0.0544	0.0249 ± 0.0355	17.8964 ± 3.4390	0.6781 ± 0.0993
0.9	0.2841 ± 0.0830	0.2202 ± 0.0812	6.8362 ± 1.4969	0.2959 ± 0.0416
0.8	0.2013 ± 0.0441	0.1008 ± 0.0336	10.1872 ± 1.3733	0.4020 ± 0.0544
0.7	0.3309 ± 0.0495	0.2134 ± 0.0452	6.8088 ± 0.9588	0.2565 ± 0.0418
0.6	0.0795 ± 0.0539	0.0260 ± 0.0357	17.8076 ± 3.5859	0.6783 ± 0.0959
0.5	0.0821 ± 0.0623	0.0279 ± 0.0417	17.7476 ± 3.7530	0.6676 ± 0.1175
0.4	0.0807 ± 0.0602	0.0267 ± 0.0413	17.8269 ± 3.5778	0.6672 ± 0.1075
0.3	**0.0734** ± 0.0578	**0.0249** ± 0.0399	**18.2946** ± 3.6704	**0.6939** ± 0.0938
0.2	0.1994 ± 0.0543	0.0858 ± 0.0361	11.1794 ± 2.4498	0.5366 ± 0.0969
0.1	0.0842 ± 0.0577	0.0268 ± 0.0388	17.7122 ± 3.5605	0.6698 ± 0.0985
0	**0.0746** ± 0.0598	**0.0239** ± 0.0397	**18.6500** ± 3.9336	**0.7040** ± 0.1086

Table 3.5: A metric comparison conducted by varying the weight λ_{NIQE} of the loss function NIQE during the final training phase. The overall loss function is given by the sum of BCE and NIQE multiplied by λ_{NIQE}.



(a) Input MRI **(b)** Generated CT **(c)** Target CT

(d) Input MRI **(e)** Generated CT **(f)** Target CT

Figure 3.7: Generated images (b)(e) compared with respective input (a)(d) and target (c)(f) images, when training with a pre-trained model and using $\lambda_{NIQE} * NIQE + BCE$ as the loss function, with λ_{NIQE} sets to 0.3.

Chapter 4

Discussions and Conclusions

After a careful review of the literature regarding techniques already implemented for the translation of medical images from one domain to another, several challenges were identified, including the limited availability of aligned datasets and the lack of consensus on which metrics to use or the required level of accuracy.

This thesis thus has two principal objectives. Firstly, to assess the feasibility of translating biomedical images using the simple generative model pix2pix, with modifications to the loss function employed by the networks and the incorporation of transfer learning. Secondly, it attempts to ascertain whether No-Reference metrics can be reliable for evaluating the generated images, with a view to potentially utilising them for optimising the model.

Initially, therefore, the impact of three different loss functions - BCE, LSGAN and WGANGP - was evaluated in combination with the use of different pre-trained models used to perform transfer learning. The results demonstrated that both BCE and LSGAN, combined with L1, appear to be suitable for the translation task, with BCE+L1 loss function performing better. In both cases, however, the use of transfer learning did not lead to performance improvements. It seems probably that the pre-trained models used have a correspondence between one domain to another that is too different from that between MRI and CT. A more accurate assessment of the type of correspondence is required in order to identify more suitable models.

WGANGP, on the other hand, is rather unsuitable, although its implementation should be investigated further as it is widely used to generate high-quality images and to avoid stability problems.

The CT images generated up to this point have achieved a satisfactory level of visual quality when compared to the reference images. However, some images are severely compromised, with large dark areas within them. This may be attributed to the very dark input MRI images. In order to achieve the desired result, it may be necessary to implement a pre-processing stage involving more appropriate normalisation or brightness enhancement, specifically tailored for these images.

Without the input of a clinical expert, it is difficult to say whether or not the synthesised images can be considered realistic. Moreover, although FR metrics appear to remain the most effective for assessing the quality of generated images, they require a reference point for comparison. However, this reference is not always readily available.

Therefore, the research concentrated on the utilisation of metrics that have not been previously employed in the scientific literature for the assessment of generated medical images: the NR metrics, which do not require a reference image for their calculation. Of these, NIQE, ILNIQE and PIQE were chosen because their training does not require subjective scores on the quality of the images associated with the training images. In fact, PIQE requires no training at all. Nevertheless, obtaining these scores would be appropriate in view of further investigation of NR metrics, in order to make more accurate evaluations of medical images. This would require crowdsourcing involving experienced radiologists who would be asked to rate the images shown according to specific parameters.

Among the NR metrics used, there is no consensus on which test is the best. The only one that agrees with the results previously obtained is ILNIQE.

To further evaluate these metrics, a series of distortions were introduced to the original CT images, including noise, increased brightness, increased contrast and pixel dilation. Subsequently, the performance obtained on the distorted images was compared to that obtained on the generated images.

It is expected that the generated images are at least of lower quality than the original ones. This is true for NIQE and ILNIQE, but not for PIQE, which can therefore be considered unsuitable. In fact, according to PIQE, even noisy images are considered to be of higher quality than images with added contrast, even if only slightly.

ILNIQE proves to be the most accurate, assessing the quality of generated images as superior to noisy images but worse than brighter images or those with added contrast or dilation. Further considerations would be needed on the applicability of this metric, perhaps with the assistance of clinical experts.

Given the promising results obtained with ILNIQE, an attempt was made to use this metric as a loss function with the goal of optimising the model. Unfortunately, it was not possible to train the network for a sufficient number of epochs, as the computational time required is too high, despite the high-performance capabilities of the available computers. This factor represents a significant obstacle to the practical applicability of the method. As a result, the focus was redirected to NIQE as the loss function.

After several attempts and combinations of NIQE and BCE, it became evident that it is always preferable to start training with BCE as the loss function and then introduce NIQE for fine-tuning. Indeed, in cases where training began directly with NIQE, performance was unacceptable, and the images were very degraded.

The results of the various tests performed indicate that the translation of medical images from MRI to CT is feasible even with a simple generative model, with which satisfactory results can be obtained.

However, these images still have to be critically analysed before they can be used in a clinical setting. It remains uncertain whether the NR metrics employed offer a realistic assessment, although there appear to be promising foundations for optimising generative models through these metrics.

Chapter 5

Future Works

The findings presented in this thesis establish a foundation for future research in the domain of medical image translation through the utilisation of generative models. Nevertheless, there are a number of topics that warrant further investigation, as well as possibilities for future research. This section will present some potential future insights.

Dataset Limitations

One ongoing challenge is the limited availability of paired medical datasets. While several datasets have been developed through research, they have been isolated to specific anatomical regions. It may be feasible to integrate these datasets in order to train models on different body parts at the same time, thereby obtaining a generalised model. Nevertheless, it is also evident that this would necessitate the acquisition of a considerably larger number of images, thereby underscoring the continued necessity for the creation of a large dataset.

An additional method for increasing the amount of images is through the utilisation of Data Augmentation, which remains a viable option in the event of a scarcity of data. This technique involves the generation of new images from the original ones through the application of various forms of distortion, rotation, translation, and other geometric transformations. Additionally, modifications can be made to aspects such as brightness and contrast. It is recommended to evaluate this approach for future implementations, with due care taken to maintain coupling and alignment between images.

Optimization of the Model

The pix2pix model remains a good, simple and cost-efficient approach to image translation. However, this approach can be enhanced by incorporating more sophisticated modules or by identifying novel, more appropriate loss functions.

Furthermore, although research has not demonstrated significant enhancements resulting from the use of transfer learning, it is possible that more specialised pre-trained models could give better results, thus further reducing the computational cost for training.

Image Quality Assessment

A future development would certainly be to explore other metrics for the evaluation of synthesised images, particularly for medical applications. This is because the majority of existing metrics have been developed for natural images, and may not fully align with the specific requirements of medical images. It would therefore be beneficial to identify more appropriate metrics or to adapt existing ones to the clinical context.

Given that numerous NR metrics necessitate the input of subjective scores for their training, it would be appropriate to obtain these scores from experienced radiologists.

Moreover, it may be worthwhile to consider the implementation of new metrics that are specifically tailored to the evaluation of medical images.

Clinical Validation

The input of radiologists and other medical professionals in this field is pivotal.

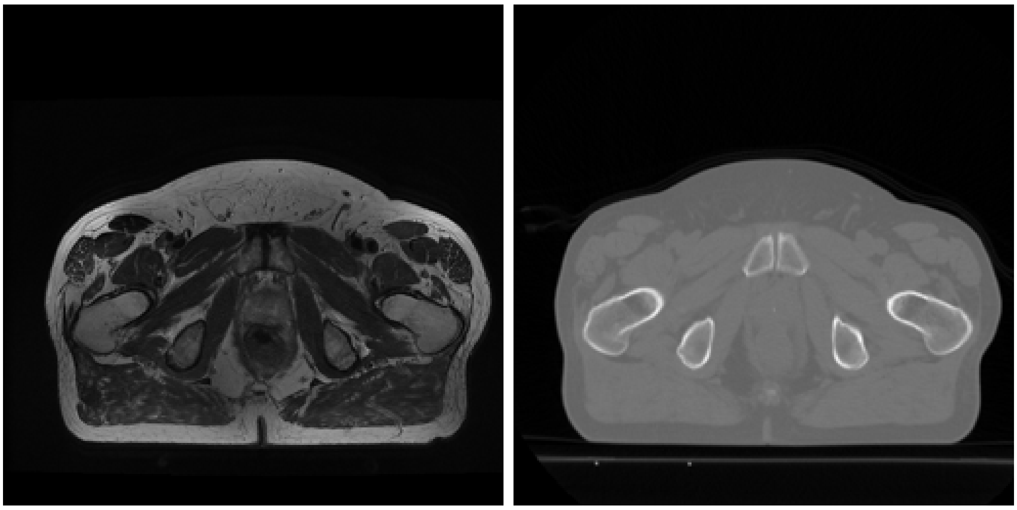
The subsequent phase will entail transitioning into a clinical setting, where the models will be validated through collaboration with healthcare institutions and radiologists. This is a crucial step to guarantee that the generated images are accurate and useful in practice.

In this regard, an ambitious attempt is the incorporation of explainability into generative models for medical image translation. It is imperative that medical professionals have confidence in and comprehend how these models make decisions. The creation of interpretability tools that can provide information regarding the rationale behind the generation of specific images or the manner in which particular features are translated will enhance the clinical adoption and reliability of these models.

Appendix A

Generated Images

The following appendix will present the images generated by the various tests performed. In all cases, reference is made to the same MRI input image and target CT image shown in the Figure A.1.

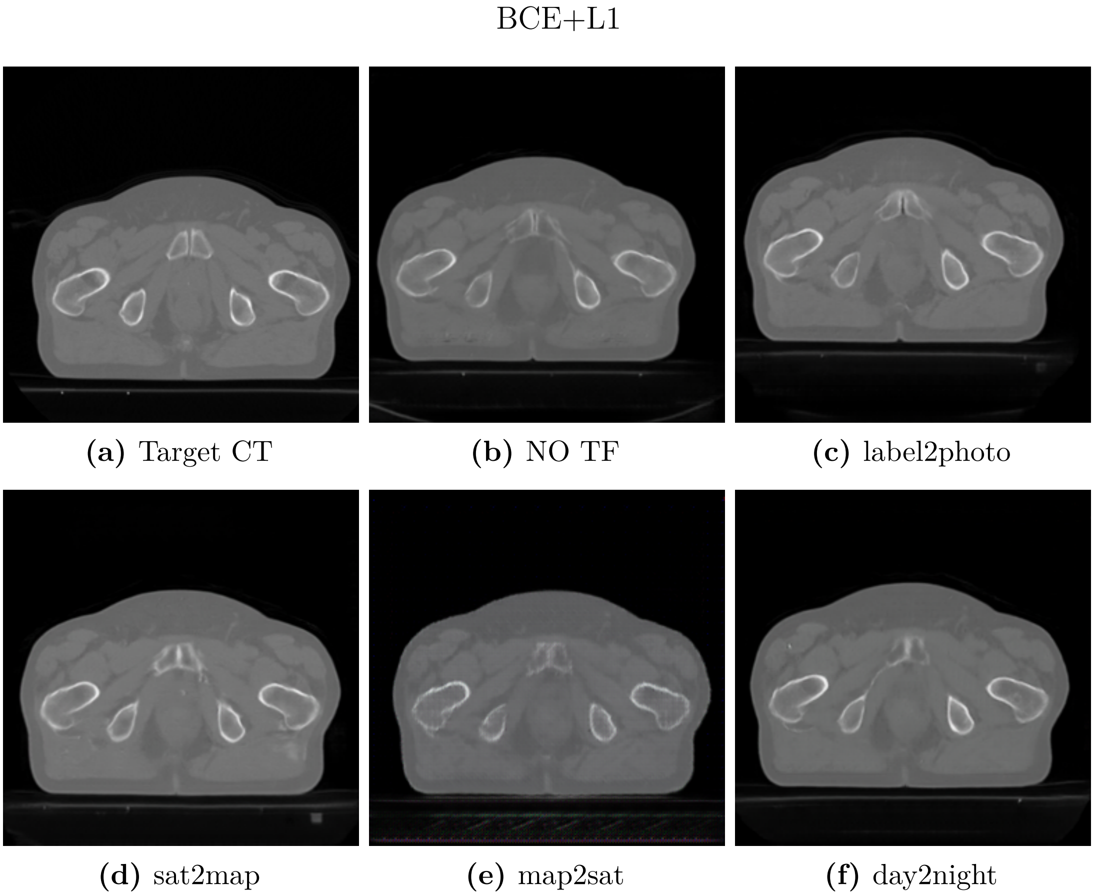


(a) Input MRI **(b)** Target CT

Figure A.1: The input (a) and reference (b) images are provided for comparison purposes, with the objective of facilitating an evaluation of the results of the various tests.

A.1 Tuning of Loss Functions and Transfer Learning

The images presented in Figures A.2 and A.3 illustrate the generated test images of the models when trained with different parameters, specifically when the loss function is modified and applying or not transfer learning from different pre-trained models.

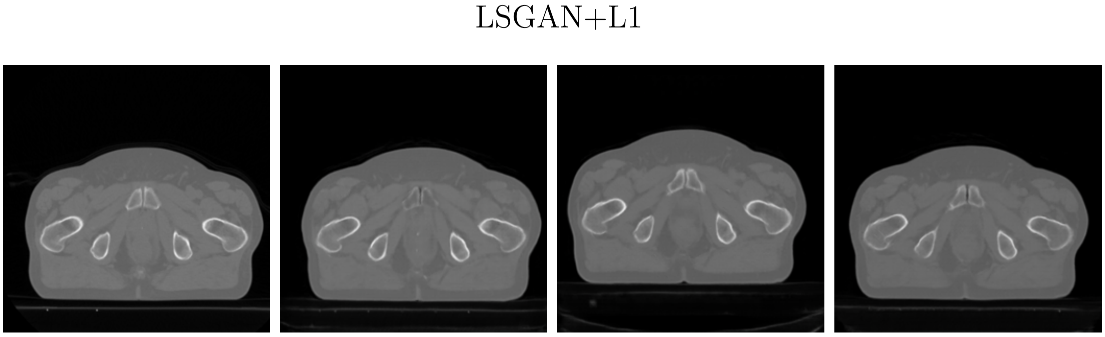


BCE+L1

(a) Target CT **(b)** NO TF **(c)** label2photo

(d) sat2map **(e)** map2sat **(f)** day2night

Figure A.2: Images generated by the various tests utilising BCE as the loss function. The CT image (a) is the reference. In trial (b), no transfer learning was employed. The pre-trained models used are: label2photo (c), sat2map (d), map2sat (e) and day2night (f).



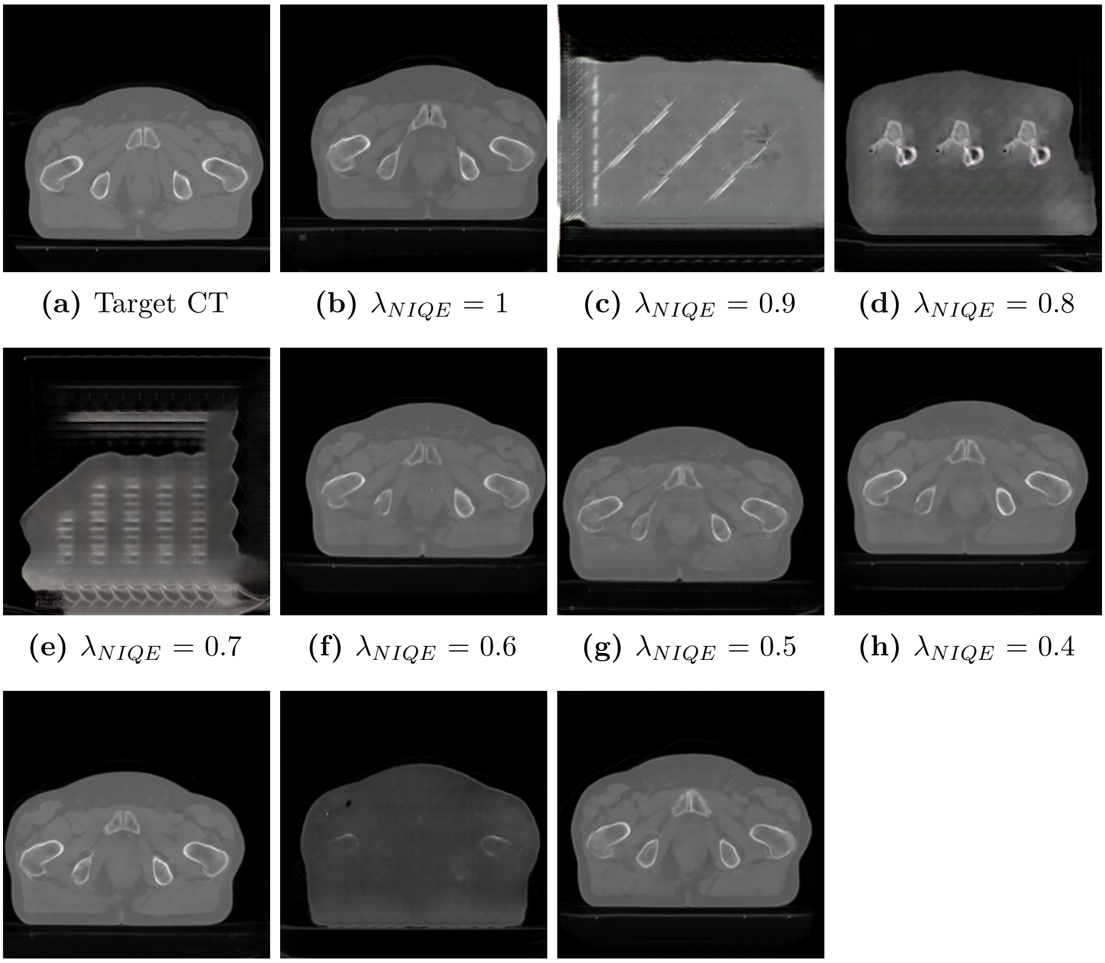
LSGAN+L1

(a) Target CT **(b)** NO TF **(c)** map2sat **(d)** day2night

Figure A.3: Images generated by the various tests utilising LSGAN as the loss function. The CT image (a) is the reference. Two pre-trained models were employed for the various tests: map2sat (c) and day2night (d). In trial (b), no transfer learning was employed.

A.2 NIQE as Loss Function

The Figure A.4 illustrates the images generated by varying the λ_{NIQE} value in the test in which the loss function NIQE is used in combination with BCE, as detailed in Section 3.4.



(a) Target CT **(b)** $\lambda_{NIQE} = 1$ **(c)** $\lambda_{NIQE} = 0.9$ **(d)** $\lambda_{NIQE} = 0.8$

(e) $\lambda_{NIQE} = 0.7$ **(f)** $\lambda_{NIQE} = 0.6$ **(g)** $\lambda_{NIQE} = 0.5$ **(h)** $\lambda_{NIQE} = 0.4$

(i) $\lambda_{NIQE} = 0.3$ **(j)** $\lambda_{NIQE} = 0.2$ **(k)** $\lambda_{NIQE} = 0.1$

Figure A.4: Images generated by varying the weight λ_{NIQE} associated with NIQE when the loss function NIQE+BCE is used.

Bibliography

[1] Vijay P.B. Grover, Joshua M. Tognarelli, Mary M.E. Crossey, I. Jane Cox, Simon D. Taylor-Robinson, and Mark J.W. McPhail. «Magnetic Resonance Imaging: Principles and Techniques: Lessons for Clinicians». In: *Journal of Clinical and Experimental Hepatology* 5.3 (2015), pp. 246–255. ISSN: 0973-6883. DOI: `https://doi.org/10.1016/j.jceh.2015.08.001`. URL: `https://www.sciencedirect.com/science/article/pii/S0973688315004156` (cit. on p. 2).

[2] Mettler FA Jr, Huda W, Yoshizumi TT, and Mahesh M. «Effective doses in radiology and diagnostic nuclear medicine: a catalog». In: *Radiology* (2008). DOI: `10.1148/radiol.2481071451` (cit. on p. 2).

[3] Jake McNaughton, Justin Fernandez, Samantha Holdsworth, Benjamin Chong, Vickie Shim, and Alan Wang. «Machine Learning for Medical Image Translation: A Systematic Review». In: *Bioengineering* 10.9 (2023). ISSN: 2306-5354. DOI: `10.3390/bioengineering10091078`. URL: `https://www.mdpi.com/2306-5354/10/9/1078` (cit. on p. 4).

[4] Jitao Li, Zongjin Qu, Yue Yang, Fuchun Zhang, Meng Li, and Shunbo Hu. «TCGAN: a transformer-enhanced GAN for PET synthetic CT». In: *Biomed. Opt. Express* 13.11 (Nov. 2022), pp. 6003–6018. DOI: `10.1364/BOE.467683`. URL: `https://opg.optica.org/boe/abstract.cfm?URI=boe-13-11-6003` (cit. on p. 6).

[5] Ki-Taek Hong, Yongwon Cho, Chang Ho Kang, Kyung-Sik Ahn, Heegon Lee, Joohui Kim, Suk Joo Hong, Baek Hyun Kim, and Euddeum Shim. «Lumbar Spine Computed Tomography to Magnetic Resonance Imaging Synthesis Using Generative Adversarial Network: Visual Turing Test». In: *Diagnostics* 12.2 (2022). ISSN: 2075-4418. DOI: `10.3390/diagnostics12020530`. URL: `https://www.mdpi.com/2075-4418/12/2/530` (cit. on pp. 6, 8).

[6] Kexin Wei, Weipeng Kong, Liheng Liu, Jian Wang, Baosheng Li, Bo Zhao, Zhenjiang Li, Jian Zhu, and Gang Yu. «CT synthesis from MR images using frequency attention conditional generative adversarial network». In: *Computers in Biology and Medicine* 170 (2024), p. 107983. ISSN: 0010-4825.

DOI: https://doi.org/10.1016/j.compbiomed.2024.107983. URL: https://www.sciencedirect.com/science/article/pii/S0010482524000672 (cit. on pp. 6, 8).

[7] Jake McNaughton, Samantha Holdsworth, Benjamin Chong, Justin Fernandez, Vickie Shim, and Alan Wang. «Synthetic MRI Generation from CT Scans for Stroke Patients». In: *BioMedInformatics* 3.3 (2023), pp. 791–816. ISSN: 2673-7426. DOI: 10.3390/biomedinformatics3030050. URL: https://www.mdpi.com/2673-7426/3/3/50 (cit. on pp. 6, 7).

[8] J. Wang, B. Yan, X. Wu, X. Jiang, Y. Zuo, and Y. Yang. «Development of an unsupervised cycle contrastive unpaired translation network for MRI-to-CT synthesis». In: *Journal of Applied Clinical Medical Physics* 23.11 (Nov. 2022), e13775. DOI: 10.1002/acm2.13775. eprint: 2022Sep28 (cit. on p. 6).

[9] Kévin Brou Boni, John Klein, Akos Gulyban, Nick Reynaert, and David Pasquier. «Improving generalization in MR-to-CT synthesis in radiotherapy by using an augmented cycle generative adversarial network with unpaired data». In: *Medical Physics* 48.6 (Mar. 2021), pp. 3003–3010. DOI: 10.1002/mp.14866. URL: https://hal.science/hal-03204110 (cit. on pp. 6, 7).

[10] Saba Nikbakhsh, Lachin Naghashyar, Morteza Valizadeh, and Mehdi Chehel Amirani. *Enhanced Synthetic MRI Generation from CT Scans Using CycleGAN with Feature Extraction*. 2023. arXiv: 2310.20604 [eess.IV]. URL: https://arxiv.org/abs/2310.20604 (cit. on p. 6).

[11] Alaa Abu-Srhan, Israa Almallahi, Mohammad A.M. Abushariah, Waleed Mahafza, and Omar S. Al-Kadi. «Paired-unpaired Unsupervised Attention Guided GAN with transfer learning for bidirectional brain MR-CT synthesis». In: *Computers in Biology and Medicine* 136 (2021), p. 104763. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2021.104763. URL: https://www.sciencedirect.com/science/article/pii/S0010482521005576 (cit. on pp. 6, 7).

[12] Aolin Yang, Tiejun Yang, Xiang Zhao, Xin Zhang, Yanghui Yan, and Chunxia Jiao. «DTR-GAN: An Unsupervised Bidirectional Translation Generative Adversarial Network for MRI-CT Registration». In: *Applied Sciences* 14.1 (2024). ISSN: 2076-3417. DOI: 10.3390/app14010095. URL: https://www.mdpi.com/2076-3417/14/1/95 (cit. on p. 6).

[13] Shouang Yan, Chengyan Wang, Weibo Chen, and Jun Lyu. «Swin transformer-based GAN for multi-modal medical image translation». In: *Frontiers in Oncology* 12 (2022). ISSN: 2234-943X. DOI: 10.3389/fonc.2022.942511. URL: https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2022.942511 (cit. on p. 6).

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need.* 2023. arXiv: `1706.03762 [cs.CL]`. URL: `https://arxiv.org/abs/1706.03762` (cit. on p. 6).

[15] Tufve Nyholm et al. *Gold Atlas - Male Pelvis - Gentle Radiotherapy.* Zenodo, June 2017. DOI: `10.5281/zenodo.583096`. URL: `https://doi.org/10.5281/zenodo.583096` (cit. on pp. 7, 19).

[16] Stefan Klein, Marius Staring, Keelin Murphy, Max Viergever, and Josien Pluim. «Elastix: A Toolbox for Intensity-Based Medical Image Registration». In: *IEEE transactions on medical imaging* 29 (Nov. 2009), pp. 196–205. DOI: `10.1109/TMI.2009.2035616` (cit. on pp. 7, 18).

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. *Image-to-Image Translation with Conditional Adversarial Networks.* 2018. arXiv: `1611.07004 [cs.CV]`. URL: `https://arxiv.org/abs/1611.07004` (cit. on pp. 11, 14).

[18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. *Generative Adversarial Networks.* 2014. arXiv: `1406.2661 [stat.ML]`. URL: `https://arxiv.org/abs/1406.2661` (cit. on p. 11).

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation.* 2015. arXiv: `1505.04597 [cs.CV]`. URL: `https://arxiv.org/abs/1505.04597` (cit. on p. 14).

[20] Jelmer M. Wolterink, Konstantinos Kamnitsas, Christian Ledig, and Ivana Išgum. *Generative adversarial networks and adversarial methods in biomedical image analysis.* 2018. arXiv: `1810.10352 [cs.CV]`. URL: `https://arxiv.org/abs/1810.10352` (cit. on p. 18).

[21] Omar Al-Kadi, Israa Almallahi, Alaa Abu-Srhan, Mohammad Abushariah, and Waleed Mahafza. «Unpaired MR-CT Brain Dataset for Unsupervised Image Translation». In: *Data in Brief* (2022). DOI: `10.17632/z4wc364g79.1` (cit. on p. 18).

[22] URL: `https://learn2reg.grand-challenge.org/Datasets/` (cit. on p. 18).

[23] T. Nyholm, S. Svensson, S. Andersson, J. Jonsson, M. Sohlin, C. Gustafsson, and A. Gunnlaugsson. «MR and CT data with multiobserver delineations of organs in the pelvic area: Part of the Gold Atlas project.» In: *Medical Physics (Lancaster), 45(3), 1295–1300* (2018). URL: `https://doi.org/10.1002/mp.12748` (cit. on pp. 18, 19).

[24] URL: `https://brain-development.org/ixi-dataset/` (cit. on p. 18).

[25] URL: `https://www.kaggle.com/datasets/awsaf49/brats2020-training-data` (cit. on p. 18).

[26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. «Image-to-image translation with conditional adversarial networks». In: *IEEE Conference on Computer Vision and Pattern Recognition.* 2017 (cit. on p. 20).

[27] URL: `https://github.com/guptapraful/niqe/tree/master` (cit. on p. 23).

[28] URL: `https://github.com/IceClear/IL-NIQE` (cit. on p. 23).

[29] Ian Goodfellow. *NIPS 2016 Tutorial: Generative Adversarial Networks.* 2017. arXiv: `1701.00160 [cs.LG]`. URL: `https://arxiv.org/abs/1701.00160` (cit. on p. 23).

[30] URL: `http://efrosgans.eecs.berkeley.edu/pix2pix/models-pytorch/` (cit. on p. 24).

[31] Li Sze Chow and Raveendran Paramesran. «Review of medical image quality assessment». In: *Biomedical Signal Processing and Control* 27 (2016), pp. 145–154. ISSN: 1746-8094. DOI: `https://doi.org/10.1016/j.bspc.2016.02.006`. URL: `https://www.sciencedirect.com/science/article/pii/S1746809416300180` (cit. on p. 24).

[32] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. «Making a "Completely Blind" Image Quality Analyzer». In: *IEEE Signal Processing Letters* 20.3 (2013), pp. 209–212. DOI: `10.1109/LSP.2012.2227726` (cit. on p. 26).

[33] URL: `https://it.mathworks.com/help/images/ref/niqe.html` (cit. on p. 27).

[34] Lin Zhang, Lei Zhang, and Alan C. Bovik. «A Feature-Enriched Completely Blind Image Quality Evaluator». In: *IEEE Transactions on Image Processing* 24.8 (2015), pp. 2579–2591. DOI: `10.1109/TIP.2015.2426416` (cit. on p. 27).

[35] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, Sumohana S. Channappayya, and Swarup S. Medasani. «Blind image quality evaluation using perception based features». In: *2015 Twenty First National Conference on Communications (NCC).* 2015, pp. 1–6. DOI: `10.1109/NCC.2015.7084843` (cit. on p. 28).

[36] URL: `https://it.mathworks.com/help/images/ref/piqe.html` (cit. on p. 28).