POLITÉCNICO DE LEIRIA | ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO

Politécnico de Leiria
Escola Superior de Tecnologia e Gestão
Departamento de Engenharia Eletrotécnica
Mestrado em Eng.ª Eletrotécnica

JOINT CODING OF MULTIMODAL BIOMEDICAL IMAGES USING CONVOLUTIONAL NEURAL NETWORKS

JOÃO OLIVEIRA PARRACHO

Leiria, October 2021

POLITÉCNICO DE LEIRIA | ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO

Politécnico de Leiria
Escola Superior de Tecnologia e Gestão
Departamento de Engenharia Eletrotécnica
Mestrado em Eng.ª Eletrotécnica

JOINT CODING OF MULTIMODAL BIOMEDICAL IMAGES USING CONVOLUTIONAL NEURAL NETWORKS

JOÃO OLIVEIRA PARRACHO
Number: 2192600

Dissertation performed under the supervision of Professor Pedro Antonio Amado Assunção, Professor Lucas Arrabal Thomaz, and Professor Luís Miguel de Oliveira Pegado de Noronha e Távora

Leiria, October 2021

ACKNOWLEDGMENTS

I would like to start by thanking to my advisers Dr. Lucas Arrabal Thomaz, Dr. Luís Miguel de Oliveira Pegado de Noronha e Távora, and Dr. Pedro Antonio Amado Assunção for their guidance and support that were crucial for this research. Also, I would like to thank Dr. Sérgio Manuel Maciel de Faria for his tremendous support and knowledge share during my research journey.

This work was supported by Instituto de Telecomunicações, by the Programa Operacional Regional do Centro, project PlenoISLA POCI-01-0145-FEDER-028325, and by FCT/MCTES through national funds and when applicable co-funded by EU funds under the project UIDB/EEA/50008/2020, and by INCD funded by FCT and FEDER under the project CPCA/A2/6986/2020. I would like to thank to the members of these projects.

I would like to thank to Instituto de Telecomunicações for hosting the work, providing the necessaries resources for its success and for the fundamental research environments. A special thanks to all my colleagues at the Multimedia Signal Processing group from Leiria Delegation of Instituto de Telecomunicações, for all the debates, knowledge shared, and support given. I would like to thank to the School of Technology and Management of Polytechnic of Leiria for providing great facilities and resources that were crucial to the success of my research.

A special thanks to all my friends for the unconditional support and patient during this journey. Finally I would like to thank to my parents, Rui Miguel Lourenço Parracho and Maria José Freitas de Oliveira and to my sister Susana Margarida Oliveira Parracho, for their support.

RESUMO

Atualmente, o grande volume de dados gerados por tecnologias suportadas em imagens médicas com diferentes modalidades, coloca problemas na gestão desse tipo de informação, tanto ao nível do armazenamento como de partilha e transferência através de redes de comunicação. Uma das formas de minimizar esses problemas consiste em usar métodos de compressão eficientes, que devem ser implementados de forma a reduzir a quantidade de recursos de armazenamento e transmissão exigidos por tal quantidade de dados.

Neste âmbito, esta dissertação apresenta o trabalho de pesquisa sobre um metodo de compressão sem perdas desenvolvido para codificar imagens de tomografia computadorizada (CT) e de tomografia por emissão de positrões (PET). Abordagens diferentes, como técnicas de tradução de imagem para imagem, são usadas, e as redundâncias entre as duas imagens também são investigadas. Para realizar a abordagem de tradução imagem a imagem, recorreu-se à compressão sem perdas das imagens CT e foi desenvolvido um método de tradução de imagem de modalidade cruzada utilizando uma rede generativa adversária para obter uma estimativa da PET correspondente.

Foram estudadas, implementadas e avaliadas duas abordagens diferentes para determinar uma representação compacta de um resíduo PET em conjunto com a CT original. Na primeira abordagem, foi desenvolvido um método baseado no resíduo resultante da diferença entre a PET original e sua estimativa. Na segunda, o resíduo é obtido recorrendo as ferramentas de codificação de predição-*inter* presentes num codificador normalizado. Assim, em alternativa à compressão independente das imagens das duas modalidades, isto é, ambas as imagens do par PET-CT original, no método proposto apenas a CT é codificada de forma independente juntamente com o resíduo de PET. Para além do *pipeline* proposto, um algoritmo de otimização de pós-processamento, que modifica a imagem PET estimada alterando o contraste e a respectiva dimensão, foi implementado com o objetivo de maximizar a eficiência de compressão.

Quatro versões (*subsets*) diferentes de um dataset público de pares de imagens PET-CT foram usadas nos testes e avaliação de desempenho. O primeiro *subset* proposto foi utilizado utilizado na validação da prova de conceito do método. Os resultados obtidos, mostraram ganhos de até 8.9% utilizando o HEVC. Por outro lado, o JPEG 2000 demonstrou não ser o mais adequado, pois não permitiu obter ganho de compressão, tendo atingido apenas uma perda de 9.1% . Para os restantes subsets (mais complexos que o primeiro), os resultados revelam que o esquema de pós-processamento refinado proposto atinge, quando comparado aos métodos convencionais de compressão, até 6.33 % de ganho de compressão usando HEVC e 7.78 % utilizando o VVC.

Palavras-chave: Tomografica computorizada, Tomografia por emissão de positrões, Compressão sem perdas, Rede generetiva adversarial, Tradução de imagem-para-imagem, Compressão de imagens médicas

ABSTRACT

The massive volume of data generated daily by the gathering of medical images with different modalities might be difficult to store in medical facilities and share through communication networks. To alleviate this issue, efficient compression methods must be implemented to reduce the amount of storage and transmission resources required in such applications. However, since the preservation of all image details is highly important in the medical context, the use of lossless image compression algorithms is of utmost importance.

This thesis presents the research results on a lossless compression scheme designed to encode both computerized tomography (CT) and positron emission tomography (PET). Different techniques, such as image-to-image translation, intra prediction, and inter prediction are used. Redundancies between both image modalities are also investigated. To perform the image-to-image translation approach, we resort to lossless compression of the original CT data and apply a cross-modality image translation generative adversarial network to obtain an estimation of the corresponding PET.

Two approaches were implemented and evaluated to determine a PET residue that will be compressed along with the original CT. In the first method, the residue resulting from the differences between the original PET and its estimation is encoded, whereas in the second method, the residue is obtained using encoders inter-prediction coding tools. Thus, in alternative to compressing two independent picture modalities, i.e., both images of the original PET-CT pair solely the CT is independently encoded alongside with the PET residue, in the proposed method. Along with the proposed pipeline, a post-processing optimization algorithm that modifies the estimated PET image by altering the contrast and rescaling the image is implemented to maximize the compression efficiency.

Four different versions (*subsets*) of a publicly available PET-CT pair dataset were tested. The first proposed subset was used to demonstrate that the concept developed in this work is capable of surpassing the traditional compression schemes. The obtained results showed gains of up to 8.9% using the HEVC. On the other side, JPEG2k proved not to be the most suitable as it failed to obtain good results, having reached only -9.1% compression gain. For the remaining (more challenging) subsets, the results reveal that the proposed refined post-processing scheme attains, when compared to conventional compression methods, up 6.33% compression gain using HEVC, and 7.78% using VVC.

Keywords: Computerized tomography, Positron emission tomography, Lossless compression, Generative Adversarial Network, Image-to-image translation, Medical Image Coding

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AAGAN	Attention-Guided Generative Adversarial Network.
AE	Autoencoder.
CG	Compression Gain.
CNN	Convolutional Neural Network.
CT	Computed Tomography.
CTU	Coding Tree Unit.
CU	Coding Unit.
GAN	Generative Adversarial Network.
HEVC	High Efficiency Video Coding.
I2I	Image-to-image translation.
JPEG	Joint Photographic Experts Group.
MAE	Mean Absolute Error.
MRI	Magnetic Resonance Imaging.
PET	Positron Emission Tomography.
PPS	Picture parameter set.
QP	Quantization Parameter.
SPS	Sequence parameter set.

VAE	Variational Autoencoder.
VPS	Video parameter set.
VVC	Versatile Video Coding.

1 INTRODUCTION

1.1 CONTEXT AND MOTIVATION

In recent years, medical imaging has become a crucial resource to assist in different medical tasks, such as diagnosis, surgery, and others. As such, the continuous development and improvement of medical imaging technologies is leading to increasingly demanding requirements, including the acquisition of higher resolution and higher quality images in multiple modalities. Consequently, a vast amount of information is acquired daily, generating terabytes of data for modalities like digital radiography, CT, PET, Magnetic Resonance Imaging (MRI), among others [1]. The growth rate of acquired medical imaging data requires efficient storage systems capable of supporting such amounts of information. For this purpose, a continuous development and implementation of more efficient compression algorithms is paramount in medical acquisition systems. The use of lossy compression may discard fundamental information contained in the original data, thus increasing the analysis difficulty and the probability of critical errors. Therefore, since every detail is important in medical imaging, in most medical compression applications, lossless compression is employed to ensure that no information is lost.

Among the various medical acquisition imaging systems, the hybrid imaging technology has become one of the most interesting to work with. It allows to capture different modalities almost simultaneously, retaining both structural and functional information of the body region under analyisis. The use of hybrid imaging methods has demonstrated to perform more accurate diagnostic when compared to individual acquisition methods. This is the case of CT acquisitions systems, which tend to be mostly hybrid nowadays, acquiring PET images as well [2]. Therefore, for every acquisition even more data is produced, when compared to the traditional approach.

These PET-CT hybrid systems introduce a new challenging paradigm in terms of medical image compression. In traditional acquisition methods, each modality is individually processed and encoded, whereas in the multimodal acquisition systems the information of both modalities is available and can be jointly used to develop efficient multimodal compression algorithms. The main objective is to store the minimum amount of information needed to restore both modalities. To this end, the integration of deep leaning tools along with traditional compression methods is an efficient approach to multimodal compression.

Deep learning has become a prominent research topic, providing solutions to diverse problems including data compression. The increasing attention given to this field of research has fomented the emergence of algorithms such as deep convolutional neural networks (CNN), Autoencoders (AE), and Generative Adversarial Networks (GAN). GANs and AEs have demonstrated interesting results in terms of data compression in recent years [3] [4] [5]. Among all the deep learning research areas, one of the most alluring topic is the Image-to-image translation (I2I) technique. I2I algorithms, rely on the deep neural networks with generative capabilities to learn source domain characteristics onto a target domain. It has been widely used in tasks such as image segmentation [6] , super-resolution [7], image synthesis [8], and domain translation, such as text-to-image translation [9].

1.2 OBJECTIVES

The main objective of this research consisted in the development of an efficient compression method of multimodal biomedical images recurring to a deep learning approach. In the course of this work, the following tasks were carried out to accomplish such objective:

- Study and implementation of a generative adversarial network to perform the CT to PET domain translation.
- Development and evaluation of an efficient pipeline to jointly compress the minimum data needed to recover the original PET-CT pair without loss of information.
- Use of inter prediction tools supported by the standard encoders to improve the proposed pipeline.
- Research and development of a non-linear simplex optimization method applied to predicted images to improve the compression efficiency.
- Performance evaluation of the proposed methods.

1.3 DISSERTATION STRUCTURE

This document is organized as follows: Chapter 2 gives an overview on the multimodal images, with a special focus on the medical imaging techniques used in this work, as well as the state-of-art lossless encoders. Chapter 3 presents a study on deep learning methods, focusing on Image-to-image translation algorithms. Chapter 4 describes the proposed framework, explains the Image-to-image translation GAN structure, and presents the obtained results. Chapter 5 details the modifications to the initially implemented cross modality compression pipeline, explains the optimization technique devised to improve the compression efficiency, and presents the results. Finally, Chapter 6 presents the conclusions of the developed work and several suggestions of improvement for future work.

2

MULTIMODAL IMAGE CODING

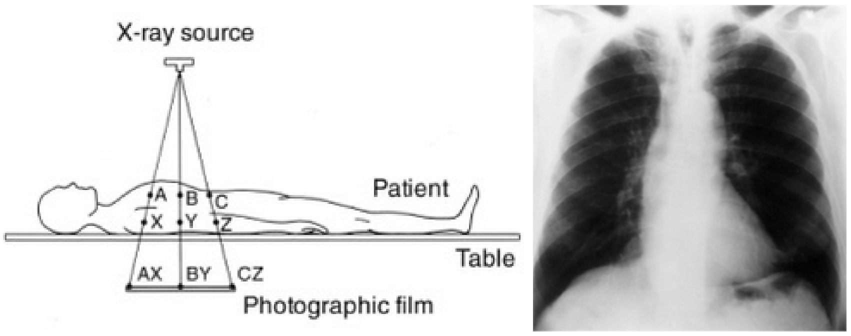
This chapter introduces the characteristics of biomedical images, and associated acquisition systems, considered in the scope of this thesis, as well as the fundamentals of state-of-the-art lossless encoders. Firstly, a brief overview on the biomedical image modalities relevant to this thesis is provided, followed by a detailed description of the image dataset, and the different subsets, that were used in the research. Finally, a description of the lossless image encoders used is presented.

2.1 MULTIMODAL BIOMEDICAL IMAGES

Medical imaging is nowadays recognised as a research and technological area where different techniques are used to obtain visual information from different, internal and external, regions of the human body. This has become a crucial technology in several medical tasks, such as interpretation and diagnosis of diseases. For many years, X-ray radiography was the main, if not the only, non-invasive technique used to provide visual information about inner organs and structures of the body.

Conventional radiography relies on the use of an X-ray beam in projection geometry, so points in the same path of an X-ray (e.g. A, and X) are projected on the same point (e.g. AX) of the photographic film, as depicted in Figure 1(a). The type of image obtained is presented in Figure 1(b), in the case of the chest region. The image contrast arises from differences in radiation absorption: denser regions and/or with higher atomic number (e.g. bones) have higher X-ray absorption rates than soft tissue (e.g. muscles) or air cavities. Accordingly, less absorbing regions lead to areas of the sensor more exposed to the transmitted radiation that consequently will appear darker in the final image (film-based radiography). On the contrary, higher absorption means less film exposure and therefore brighter regions in the image, and these characteristics can be observed in Figure 1(b).

Through continuous research and the development of new technologies, multiple imaging modalities have emerged. These are capable of probing anatomical, physiological, and functional information from the human body, going well beyond the



(a) (b)

Figure 1: (a) Projection radiography acquisition technique (b) Conventional Radiography chest image [10]

conventional X-ray radiography. Indeed, using different physical acquisition processes it is possible to obtain diverse imaging modalities like computed tomography (CT), positron emission tomography (PET), magnetic resonance imaging (MRI), and ultrasound imaging, among others, each one capturing specific characteristics, representing different structural or functional information. While the former (CT) allows to spatially map details of the anatomic structure of the human body such as organs or bones, the latter (PET) is highly dependent on physiological processes thus providing insight on metabolism and biochemical activity. Two of the most common structural imaging techniques are CT and MRI, while PET is a functional imaging procedure.

Nowadays, many biomedical image scanners are hybrid, i.e. different images are simultaneously captured using distinct, complementary, modalities. This is the case, for example, of PET-CT or PET-MRI scanners which simultaneously capture functional and structural details of the organs and tissues within the body, thus providing more detailed information about the same body region. This type of capture has the additional advantage of generating a pair of images whose dimensions and coordinates are geometrically aligned, which does not happen when the acquisition of each modality is performed separately.

In the next sections, the biomedical images used (CT and PET), as well as the datasets used throughout the work are described.

2.1.1 *Computed Tomography*

Computed Tomography (CT), firstly introduced during the 1970s [11], has evolved into a crucial imaging technology in clinical practice. It relies on the use of X-rays but, unlike radiography and its stationary projective geometry, it is based on a moving source-detector system that rotates around the patient body. At each angular position, X-ray information is recorded for the different ray paths crossing the region of interest. Then, with the acquired data it is possible to create a spatial mapping that corresponds to a cross sectional view of the imaged area.

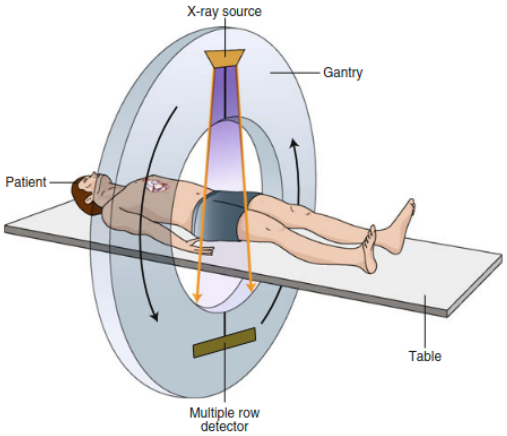


Figure 2: The principle of computed tomography with a moving X-ray system source-detector around the patient body [12].

From the physical point of view, the underlying process that leads to the formation of CT image is still the fact that different regions have different absorption rates. Particularly, in CT images, the value in each pixel is a representation of the so called linear attenuation coefficient [13], which can be mathematically related to the original data by means of the Radon transform [14]. As mentioned before, different tissues (i.e, bone, air, water, etc) have different attenuation coefficients so, to display an image and after the conversion to gray scale the output range is often optimized for a better visualization of particular body tissues. An example of a CT image from the head region can be seen in the Figure 3.

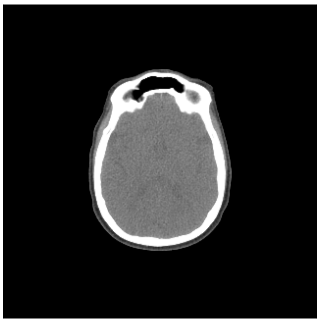


Figure 3: CT image of the head region.

One of the advantages of this modality, when compared with X-ray radiography, is the higher capability of distinguishing tissues according their absorption rate. Additionally, since the CT images are crossectional views, it allows 3D reconstruction using the multiple images of longitudinal scans. These 3D capabilities facilitate the interpretation of the results.

However, and from a health point of view, one should take care, as CT scans cause higher radiation exposure than traditional X-ray radiography. So, if in general both modalities might not be recommended to certain patients (e.g. pregnant women), risk assessment procedures should be more strict with CT.

2.1.2 *Positron Emission Tomography*

Positron Emission Tomography (PET) [15] is a nuclear medicine imaging technique that is used to assess metabolic activity of tissues and body organs. This technique is widely used in a variety of applications including, but not limited to, oncology (cancer detection) field, Parkinson's, Huntington, and Epilepsy diagnose [16].

In PET, a radioactive substance known as radiotracer is administrated to the subject, either orally or via an intravenous injection. This substance is absorbed by the organ or tissues to be examined, emitting its characteristic radiaton, which are positrons in case of PET that annihilate generating X-rays photons. Much like in a CT arrangement, a ring of detectors is used and cross sectional views of the area can be obtained. The whole process naturally depends on the metabolic activity of the region by means of the radiotracer concentration, making PET a functional imaging technique. As a rule, for a range of clinical situations cells with abnormal behaviour tend to absorb higher amounts of radiotracers when compared to healthy cells, and this can be properly assessed with PET An example of a PET image from the head region can be seen in the Figure 4.

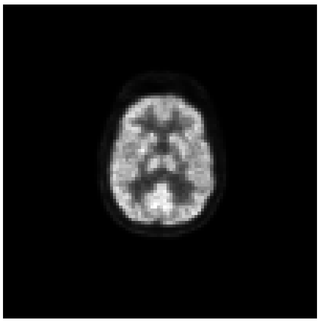


Figure 4: PET image of the head region.

One of the key advantages PET imaging technique is that the correction of issues related to the attenuation caused by overlapping tissues (i.e bones) is easily accomplished. However, the PET scanners presents an expensive operation cost.

2.1.3 *Other image modalities*

Given the prompt availability of a convenient dataset, in the scope of this work, PET and CT images simultaneously obtained with hybrid systems were used. Nevertheless, other interesting modalities, such as MRI or functional MRI (fMRI), could had been considered for the same purpose. In general, magnetic resonance imaging (MRI) [17] is an medical imaging technique based on nuclear magnetic resonance (NMR) [18] in which external magnetic fields are used to probe information about protons and their environment. MRI scanners do not expose the patients to X-ray or other types of ionizing radiation, as CT and PET systems do, and are capable of capturing more detailed information with higher contrast in soft tissue regions, as for example the brain. The MRI modality is commonly used to detect brain anomalies, tumors, and breast cancer. Due to MRI natural high contrast procedure, contrast agents are not strictly necessary to generate high detailed images of the anatomical structures and blood vessels.

2.1.4 *Dataset*

This research work was based on pairs of multimodal PET and CT images from the head region, acquired with hybrid PET-CT systems. The images come from "The Cancer imaging Archive (TCIA)" [19], a publicly available dataset of medical images. The dataset has a total of 2111 PET-CT pairs, where the resolution of PET images is 128 × 128 pixels, while that of CT images is 512 × 512; both PET and CT images include only one colour channel, i.e in grayscale, with eight bit depth. In this work, four different subsets were derived from the source dataset and used in different parts/steps of the work, in order to allow investigation of various training/testing options, some of them inline with previous works in the literature. Each of these subsets was obtained by either changing the resolution and/or PET-CT pair alignment mode. The pairs of images used in both train and test set, are identified with the series instance unique identifier, and the respective Instance number in the Tables 11 and 12, respectively. In order to identify the PET or CT series, the 'Series instance unique identifier' is used. This identifier is shared among the images of the same series. The 'Instance number' identifies an image of a certain series. Both identifiers are available in the PET and CT Dicom file (which is the original format of the images).

The need for creating the four different subsets was the following: the first one was used in initial studies that acted as a proof of concept. Then, the number of images was substantially increased in the remaining subsets ($\sim$ 8 times), allowing higher inter-patient variability. In subsets 2, 3, and 4 different image alignments / downsampled strategies were considered for better system performance.

SUBSET_1 The first subset, comprised 257 PET-CET pairs: 220 used for training, 37 for test. Both PET and corresponding CT were resampled to the resolution of 256×256, as in [20], and then cropped to 136×136. In the cropping process a binary segmentation mask was used to determine the boundaries of the region of interest. Such mask is obtained by first applying two consecutive morphological dilations with different structuring elements to the CT image, followed by a binarisation process. The regions are classified in three groups, namely as background, head, and other artifacts. The classification criterion used to separate these regions is the number of connected pixels in each group, which is consistently higher for the head region and smaller for the artifacts (the background consists of the remaing pixels without connected elements). The resulting binary mask of the head region is further enlarged few pixels (through a morphological dilation operation) to ensure that no boundary information is lost. Figure 5 illustrates the cropping process.

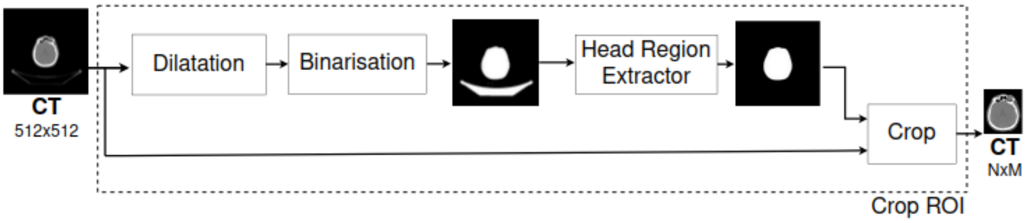
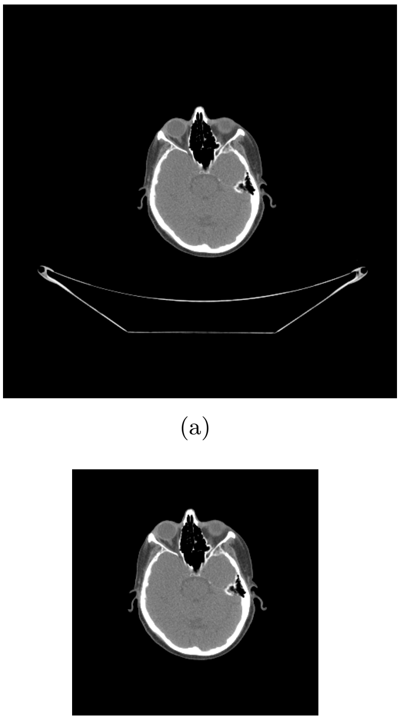


Figure 5: Crop ROI process.

Since the cropped region does not necessarily have the same number of pixels in all images (i.e., its size depends on the size of the scanned head), these are resized to the final resolution (136×136), which was enough to encompass the head region in all images of the dataset, by using zero-padding when necessary. Note that the mask computed from the CT image is also used for its paired PET.

An example of a CT image before and after the crop operation can be seen in the Figure 6.



(a)

(b)

Figure 6: (a) CT image before crop operation (b)CT image after proposed crop operation

SUBSET_2 It is an expanded version of the previous one, comprising 2111 pairs (1391 for training, 721 for test). Besides the higher number of image pairs, both CT and PET images are cropped from their original resolution (512 × 512 and 128 × 128, respectively) rather than resampled beforehand. The process pipeline to obtain the Substet_2 is shown in Figure 7. The cropping process is the same as explained above for Subset_1. Since PET images have lower resolution, the segmentation mask derived from the CT is downsampled to 128 × 128 before being applied to the paired PET images. As previously explained, the cropped PET images do not have the same resolution, therefore zero padding is applied to obtain the final resolution of 100 × 100 pixels, which was enough to encompass the head region in all images of the dataset. The same process is applied to the CT images, but in this case zero-padding is done to obtain the resolution of 320 × 320. For the training procedure, the cropped CT images are downsampled to the same resolution of their cropped PET pairs, i.e., 100 × 100, as to avoid the increase of data in the coding process.

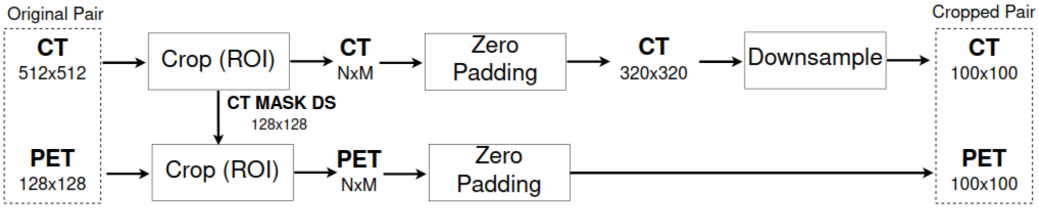


Figure 7: Subset_2 proposed scheme.

Since the head regions of the PET-CT pairs are not aligned amongst themselves, after the cropping process, a re-alignment procedure is carried out. This consists in aligning the centre of mass of the PET head region with that of its corresponding CT. For this purpose the x coordinate of centre of mass is calculated as follows. First all pixels within the same row are added together and weighted by respective row index. The weighted values obtained are also added together and then divided by the sum of all pixel values. The floor operation is used to ensure that the coordinate is an integer. The function to compute the x coordinate of the center of mass can be defined as:

$$Centre\_of\_mass_x = \left\lfloor \frac{\sum_{j=1}^{H}([\sum_{i=1}^{W} I_{ij}] \times j)}{\sum_{j=1}^{H}\sum_{i=1}^{W} I_{ij}} \right\rfloor , \tag{1}$$

where I represents the image, i the column index, j the row index, and $\lfloor\cdot\rfloor$ is the floor operation. To compute the y coordinate, the same equation can be used but columns are used instead of rows.

SUBSET_3 Derives from Subset_2 with an additional alignment step. In all PET-CT image pairs, the centre of mass is enforced to be located at the centre of the image during the zero-padding operation.

SUBSET_4 In Subset_4, the images' centre of mass are located at the central pixel of the image like in Subset_3, however in this case CT images were not downsampled, presenting a final resolution of 320 × 320 pixels after the crop process. As further described in Chapter 5, downsampling to 100 × 100 is implicitly carried out in the proposed GAN.

Table. 1 summarizes the information about the different versions of the dataset.

Table 1: Description of the used PET-CT datasets

Subset	Resolution (pixels)		Nº of image pairs		Specifications
	CT	PET	Training	Test	
Subset_1	136 × 136	136 × 136	220	37	Artifacts removed
Subset_2	100 × 100	100 × 100	1391	720	PET center of mass aligned with the CT
Subset_3	100 × 100	100 × 100	1391	720	CT and PET center of mass adjusted to the image center
Subset_4	320 × 320	100 × 100	1391	720	CT downsampled performed with a GAN

2.2 LOSSLESS IMAGE CODING

In this section, the state-of-art encoders used in this research are described. As this work focuses mainly on biomedical images, lossless compression is understood as a requirement. That way, only encoders with such capability were considered.

2.2.1 *JPEG 2000*

The JPEG 2000 [21] is a still image standard codec, proposed as the JPEG successor with the main goal of creating a coding system capable of efficiently provide compressed images from distinct application fields such as medical, natural, and documents, also to achieve high interoperability. One of the main differences of JPEG 2000, when compared to JPEG, is the use of wavelet transform, instead of the discrete cosine transform (DCT). The overall basic structure of JPEG 2000 encoder and decoder is shown in Figure 8, and Figure 9 respectively.

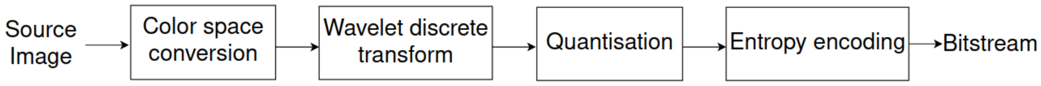


Figure 8: JPEG 2000 encoder basic structure.

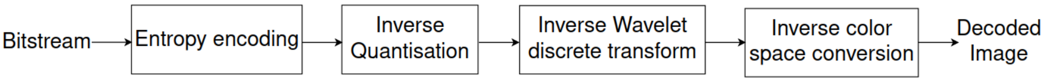

Bitstream
Entropy encoding
Inverse Quantisation
Inverse Wavelet discrete transform
Inverse color space conversion
Decoded Image

Figure 9: JPEG 2000 decoder basic structure.

The JPEG 2000 works as follows. Initially, a DC Level shift is applied to each component in order to ensure zero mean of all pixels before entering the coding process. Then, a color transformation is applied to all components. There are two different component transforms in JPEG 2000, a Reversible Component Transform (RCT) and an Irreversible Component Transform (ICT). For lossless coding, only

RCT can be used, while in lossy applications any of them can be considered. After the color space transformation, the image is split into independent rectangular blocks with no overlap between them, which represent the basic coding unit of JPEG 2000. A Discrete Wavelet Transform (DWT) is then applied to each block, computing the coefficients corresponding to several frequency sub-bands that retain information which describes particular regions of the blocks. The different sub-bands are sampled at different spatial resolutions, which means that JPEG 2000 accomplishes spatial scalability. Finally, the wavelet coefficients are quantised, independently grouped in rectangular coding blocks, and entropy coded.

One of the main functionalities of JPEG 2000, is the possibility for compression of regions of interest with higher fidelity than the remaining regions of an image. Therefore, distinct regions can be decoded with different levels of quality.

2.2.2 *High Efficient Video Coding*

High Efficient Video Coding (HEVC) [22], also known as H.265 is a hybrid video coding standard, jointly developed by ITU-T Video Coding Experts Group and the ISO/IEC Moving Picture Experts Group, as the successor to Advanced Video Coding (AVC - H.264). HEVC was originally proposed with focus on high resolution video and optimizing the processing architecture with parallel computing. Compared to H.264, HEVC improved codding efficiency from 25% to 50%, achieving lower bitrate to the same video quality.

HEVC employs many of the H.264 characteristics, and the same hybrid coding approach that uses both intra and inter prediction. In the intra mode, only I-frames are used, thus pixel predictions are computed along the spatial information within the current frame. The inter prediction exploits the temporal redundancies between multiple frames of the video. Thus, at least one of the frames must be encoded as an I-frame (Reference frame), and the remaining, that can either be a B-Frame or a P-Frame, have their predicted values based on the reference frame, using a motion prediction algorithms. The B-Frame is a bidirectional predicted picture that can use information from two reference frames (e.g., the next and previous frame in display order), while the P-Frame only uses the information from a one single previously enncoded frame.

Previous coding standards used the macro-block as their coding unit, which represent a processing region of 16x16 pixels, whereas the HEVC proposed the coding tree unit (CTU), that can expand to areas up to 64x64 pixels. The CTUs can be fractionated recursively into variable size blocks, known as coding units (CUs), using a tree structure and quadtree syntax [23]. The use of larger CTU sizes has demonstrated to improve compression efficiency, mostly in images with larger resolution.

HEVC also increased the number of modes used for intra prediction, from the 9 used in H.264, to 34 that are supported by all blocks sizes within HEVC coded. It specifies one DC prediction mode, that uses the mean value of pixels within a neighbour block as a prediction to the entire block, and 33 directional prediction modes, as shown in Figure 10. The different directional modes are suited to predict textures with structures such as edge regions, in the specific direction.

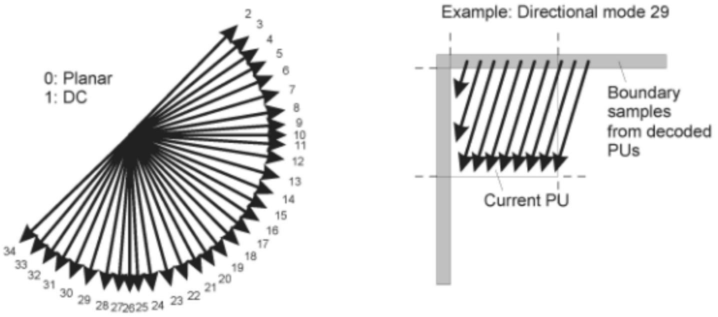


Figure 10: HEVC Modes for intra prediction [22].

The HEVC encoder block diagram is shown in Figure 11. Initially the input frame is divided into CTUs using the syntax described previously. The first frame of every sequence is encoded using only intra prediction. The possible remaining frames can be encoded either with intra or inter prediction. To perform the inter prediction it is applied motion estimation that uses motion vectors to predict the samples of each block based on the selected reference picture. By feeding the motion vectors computed to the motion compensation block, the decoder is capable of computing a inter prediction signal identical to the one computed by the encoder. To the residue of the intra or inter prediction, that results from subtracting the original block with the respective prediction, it is applied a liner spatial transform. The resulting coefficients are then scaled, quantized, entropy encoded, and transmitted in the final bitstream along with the prediction and control information. To ensure that the encoder and decoder generate identical predictions, the decoder processing loop is duplicated in the encoder architecture (gray-shaded boxes in Figure 11). Therefore, to obtain an approximation of the residual signal, an inverse scaling and then an inverse transform is applied to the quantized transform coefficients. The frame prediction is added to the residue and the resulting picture representation can be filtered by one or two loop filters. This picture is stored into a decoded picture buffer and used for future predictions.

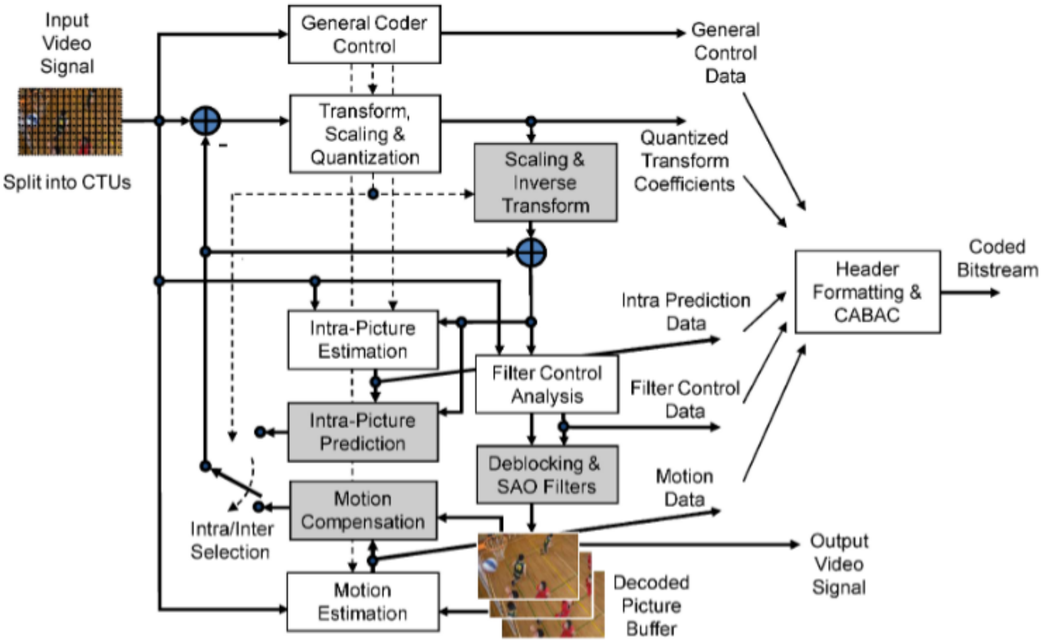


Figure 11: HEVC encoder block diagram (with decoder elements shaded in gray)[22]

The HEVC standard allows either lossy or lossless compression. While operating in lossy compression mode, one of the techniques that control the amount of information loss in order to reduce the compression ratio is the quantisation parameter (QP), which is defined within a range from 0 to 51. The QP controls the level of quantisation applied to the transform coefficients. Larger QP values result in higher quantisation, thus it increases the compression ratio and subsequently reduces the quality of the decoded frames when compared to the original ones. In lossless mode, the transform, quantisation, sample adaptive offset (SAO), and deblocking filters are skipped, thus the intra or inter residue is encoded with the entropy coder. Differently from H.264, in HEVC the lossless mode can be used in both intra and inter prediction mode.

A Range Extension [24] was introduced in the HEVC standard to allow more coding flexibility, specially targeting applications of medical imaging, screen content, among others. This extension tools allow to encode images with higher bit depth, up to a maximum of 16 bits per sample, and also include new types of chroma sampling.

2.2.3 *Versatile Video Coding*

Versatile Video Coding (VVC) [25], also known as H.266, is the successor to HEVC and the most recent video codec standard. It was developed by the Joint Video Experts Team (JVET) in collaboration with the Moving Picture Experts Group (MPEG). VVC is implemented using HEVC baseline architecture with major improvements and with the addition of new coding tools. It was developed to target tasks such as panoramic video compression and streaming, and to improve compression efficiency in higher resolution videos (i.e 4k video). The VVC standard achieves up to 50% bit-rate saving for the same quality relative to HEVC, the previous dominant video coding standard. In Figure 12, the VVC encoder architecture is presented.

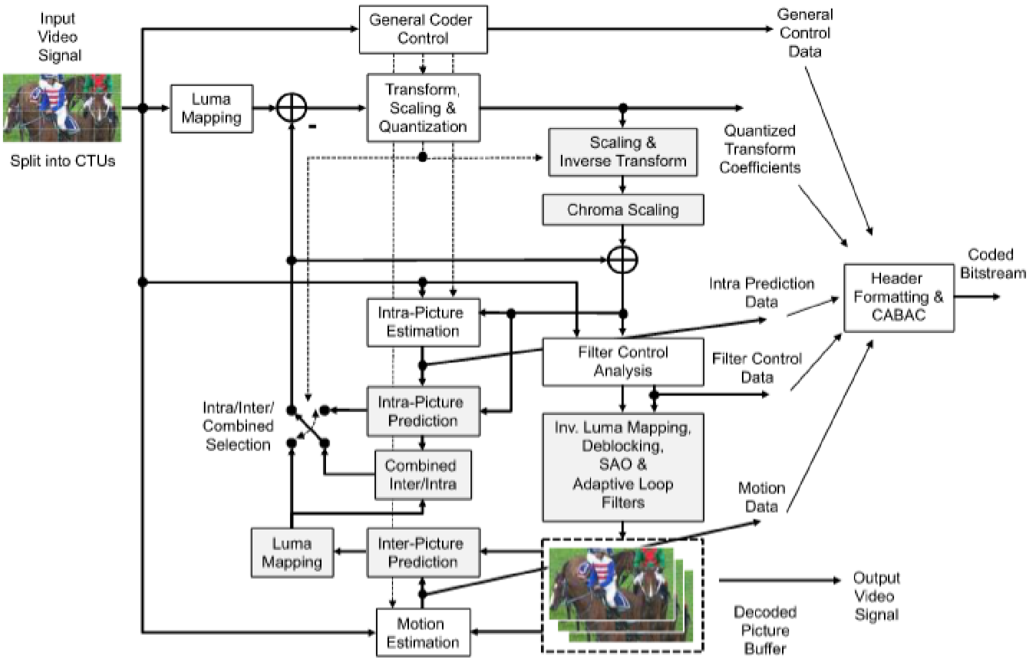


Figure 12: VVC encoder diagram [25].

VVC also uses the CTU as the base coding unit with an increased coding area of 128x128 pixels in comparison with the 64x64 used in HEVC. It has also increased the number of directional intra-predictions modes to 93, improving the compression efficiency in higher resolution image. Figure 13 shows the VVC intra modes.

In order to create more immersive video experiences, the VVC standard proposed the use of independent sub-pictures. These sub-pictures are differentiated in the bitstream generated with the encoder, allowing to independently encode each sub-picture. Thus, the decoder has the capability of decoding only specific sub-pictures of interest. This feature has shown to be efficient in applications such as streaming 360-degree video, since this coding tool allows to only decode the part of the bitstream that the user is watching instead of decoding the full bitstream. Thus, it is not necessary to fully decode the bitstream in order to watch specific viewports of the 360-degree video.

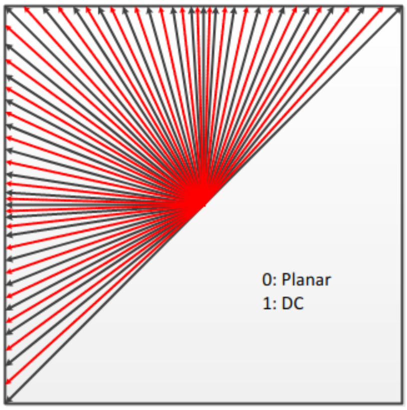


Figure 13: VVC proposed 67 intra prediction modes [26]

As shown in Figure 14, differently from HEVC that uses a single tree structure and a quadtree syntax to split the CTUs into 4 sub-blocks recursively, in VVC the first split is performed using the quadtree approach and then each block can be split either vertically or horizontally into 2 or 3 subblocks. Within the same CTU, the coding-tree structure of luma samples is independent from the chroma samples, allowing the use of different coding blocks sizes according to the sample type. An example of a CTU with multiple coding-tree structures is shown in Figure 14.

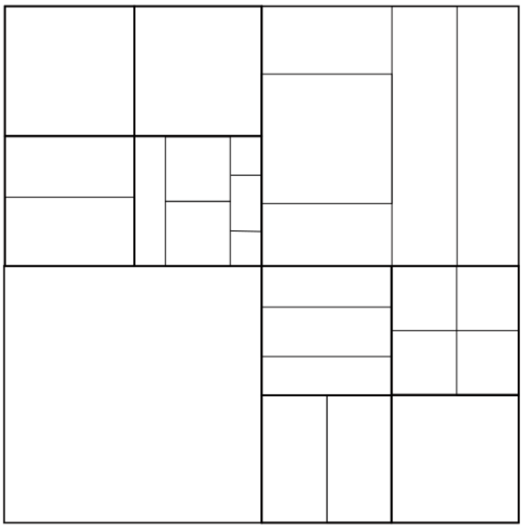


Figure 14: VVC CTU block partition.

One major change in VVC compared to HEVC is the addition of 92 different geometric partitioning modes. Besides the conventional split mode described in Section 2.2.2, where the blocks are split into rectangles and squares, VVC also employs a geometric partition approach [27]. This non-horizontal or non-vertical split of blocks allow a better compression efficiency, since it is capable of a more accurate motion prediction from real life videos. The Figure 15 demonstrates an example of the same CTU split with a rectangular block partition and with a geometric block partition. As shown in Figure 15(b), the geometric block partition allows a better spatial adaptation to the image boundaries and requires less amount of partitions.

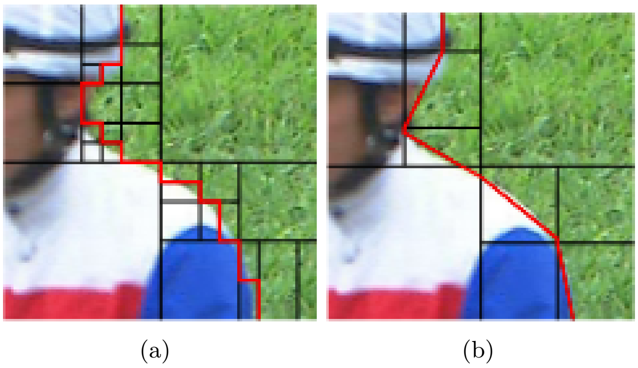


Figure 15: (a) Rectangular block partition (b) Geometric block partition [27].

2.3 SUMMARY

In this chapter, the fundamentals of medical multimodal image coding were described. It started with a detailed description on several biomedical image modalities, with a focus on the CT and PET as the present work focus on these two. Then, the four different subsets based on the same public PET-CT paired dataset, that were employed in the experiments of this thesis, were presented and discussed. The first subset was used to perform the initial studies of the proposed concept. The remaining subsets were created to test the performance with higher number of pairs and different image alignments and downsampling strategies.

The chapter is concluded with a brief review of state-of-art codecs used in this research. JPEG2000 was designed as a still image codec, whereas HEVC and VVC were designed as hybrid video codecs. The hybrid architecture is more suitable to compress image sequences, i.e. video signals, since it is designed with coding tools that explore the redundancies between consecutive frames to improve the compression efficiency. VVC as the successor of the HEVC presents an increase in the architecture complexity.

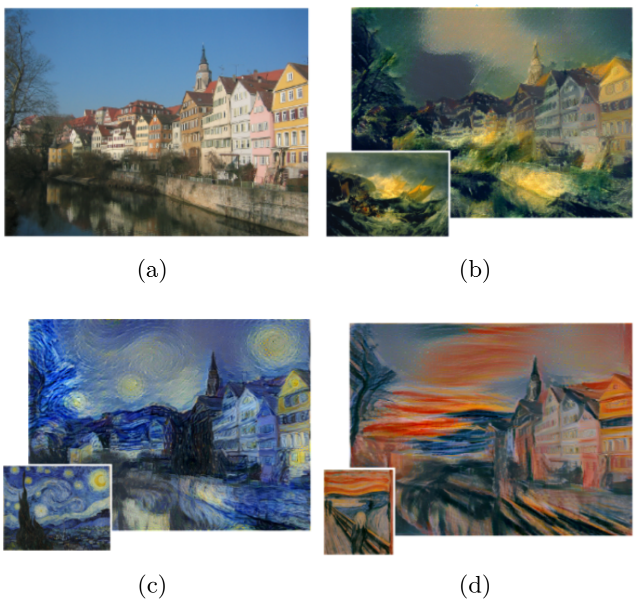
3

DEEP LEARNING FOR IMAGE-TO-IMAGE TRANSLATION

In this chapter, the basic concepts involved in Image-to-image translation based machine learning are presented. It starts with an overview on deep learning architectures related with the generative models. Then, variants of the GAN model are given, and finally several applications using Image-to-image translation architectures are described.

3.1 IMAGE-TO-IMAGE TRANSLATION

The Image-to-image translations (I2I) concept consists in the development of a method that maps an image from a source to target domain, e.g., mapping CT to PET domain. To perform the mapping between different domains, it is required to learn how to distinguish features that are specific from the domain (i.e, textures) from the features that are independent of the domain (i.e, the image content structure). Figure 16 illustrates one example where an image of the Neckarfront Tübingen is translated to different artist painting style [28], such as Van Gogh, J.M.W. Turner, or Edvard Munch. After the mapping process, the structures of the buildings (features invariant to the domain) were preserved, while the style of the image, in this case, the textures of the image (features specific from the domain) changed according to the target artist style. To this end, the use of machine learning tools, such as deep leaning networks, that are capable of efficiently learn the different representations / features within the image domain is paramount.



(a) (b)

(c) (d)

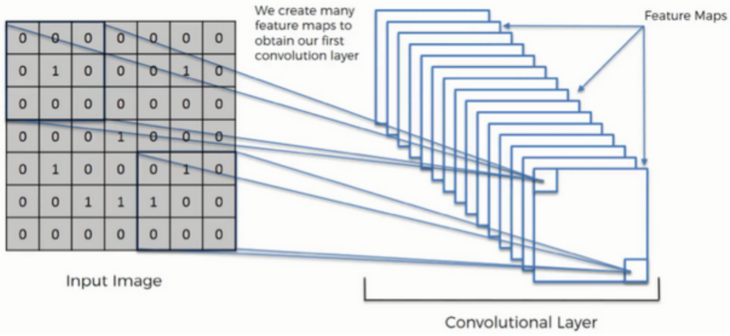
Figure 16: I2I applied to style transfer, where the structures (e.g. buildings) were preserved, and textures changed according the target artist style (a) Original image of the Neckarfront in Tübingen, Germany. (b) The Shipwreck of the Minotaur by J.M.W. Turner, 1805. (c) The Starry Night by Vincent van Gogh, 1889. (d) The Scream by Edvard Munch, 1893. [28]

3.2 DEEP LEARNING ARCHITECTURES

The deep learning research has demonstrated interesting results in the Image-to-image translation tasks, where the networks are capable of translating an input image from a source domain to a target domain. Such operation is achieved due the networks generative capability of learning the meaningful features of both domains (source and target), and perform a domain translation. The generative models, such as variational autoencoders and generative adversarial networks, are part of an emerging research topic, as a reliable approach to Image-to-image translation tasks.

3.2.1 *Convolutional Neural Network*

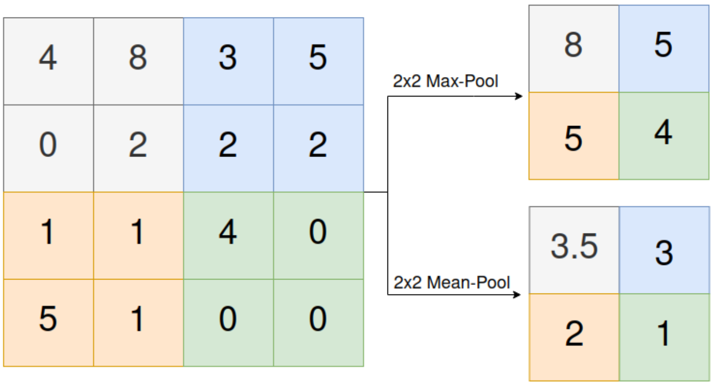
A Convolutional Neural Network (CNN) [29] is a deep learning algorithm composed by multiple layers (convolutional, activation, pooling, and fully connected layers, among others), capable of learning and distinguishing different features within an input image such as object shapes, e.g circular shapes, rectangular shapes. The convolutional layer is composed of different feature maps, each one obtained from the convolution with a specific kernel, as represented in Figure 17. The weights of the kernels are determined and continuously updated during the training procedure using backpropagation, in order to improve the output quality of the algorithm.


We create many feature maps to obtain our first convolution layer
Feature Maps
Input Image
Convolutional Layer

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	1	0	0	0	1	0
0	0	1	1	1	0	0
0	0	0	0	0	0	0

Figure 17: Convolution layer [29].

After the convolutional layer, usually a pooling operation is performed. In this step, each feature map is downsampled reducing the number of features. Models with larger amount of parameters are more complex and have a higher capability of fit and memorize the training data. However, due the existence of overfitting the model is not capable of generalize and produce low quality results when used in data that was not present during the training procedure. Therefore, the reduction of the features dimension is a crucial step to prevent overfitting, since it allows the model to achieve a more generalized solution. Different approaches can be used in the pooling layer. Examples of different layers include max pooling, that extracts the maximum value from a given patch of a feature map and transmits this value to the new downsampled feature map, and average pooling, that instead of extracting the maximum value computes the mean value of the feature patch. The Figure 18 demonstrate an example of a 2 × 2 Max pooling and 2 × 2 Mean pooling operation.



4	8	3	5
0	2	2	2
1	1	4	0
5	1	0	0

8	5
5	4

3.5	3
2	1

Figure 18: Example of 2 × 2 Max pooling and 2 × 2 Mean pooling operation.

3.2.2 *Autoencoder*

An Autoencoder (AE) [30] is an unsupervised algorithm composed by two main functional blocks, as shown in Figure 19. It comprises an encoder that learns to efficiently compress the input data to a latent space, and a decoder that learns to reconstruct the input data from the compressed latent space. In that manner, the network is designed to minimize a reconstruction loss function, that measures the error between the input and reconstructed image.

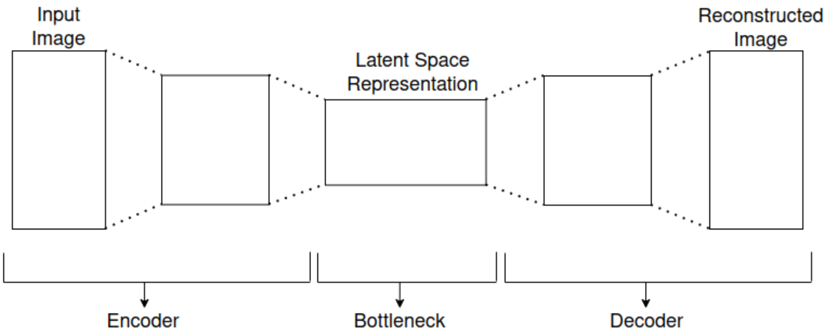


Figure 19: Autoencoder architecture.

Autoencoders are ideal for dimensionality reduction applications due to their capacity of learning the most essential features and from there create a compressed representation of the input data. However, Autoencoders present some limitations, as since they only learn features from a specific image domain, they only perform efficiently on the type of data. For example, an Autoencoder trained with images of cars is not expected to perform well on images of turtles because of the most representative features will be significantly different. Autoencoders are not suitable for lossless image compression as the encoding / decoding process is carried out with information loss.

The Autoencoder model has also been used in many other tasks such as image denoising [31] and anomaly detection [32]. Image denoising might be a crucial step / task in medical imaging, as it removes undesirable noise that degrades image quality with a negative impact on medical diagnosis. Several algorithms of noise removal have been explored in the last decade using different approaches. Gondara proposed a denoising Autoencoder [31] to efficiently remove the noise from X-ray medical images. Images from a mammogram and a dental radiography with added noise were used to train and evaluate the proposed network. Different levels of the noise introduced were also tested. It was demonstrated that higher noise levels negatively impact the performance of the proposed network. Denoising Autoencoders are basically an extension of the traditional Autoencoder, where the output is no longer a reconstruction of the image that is given as input. In this case, the algorithm no longer targets a reconstruction of the input image but rather a denoised version. To achieve so, in the training process, noise is added to the original images and the result fed to the Autoencoder, which then attempts to reconstruct the original image. In order to accomplish it, the reconstruction loss is computed between the reconstructed image and the original one (not the input, noisy, version). Anomaly detection is another important research topic in medical imaging since it is responsible for recognizing abnormal patterns that typically are associated with diseases in the human body. In [33], a chest X-ray anomaly detection algorithm based on AE is proposed. The proposed network was trained with a dataset of 112120 frontal view chest X-ray images (86523 for training, 25595 for test), and utilized a perceptual loss. To evaluate the AE performance, different conventional algorithms to detected anomalies were also tested. It was demonstrated that the trained Autoencoder was capable of outperforming the conventional anomaly detectors tested, and it was also capable detecting even the barely visible abnormalities in the chest X-rays images. Such results can be achieve due the powerful capability of the Autoencoder to learn and capture the most relevant information, that in this case are the chest X-rays anomalies.

One of the disadvantages of the AE is that there is no organization of the latent space. The latent space organization, is how the features of the learned domain are structured along the latent space. The lack of control in the latent space organization does not guarantee that every sample in the latent space has meaningful information. Thus, the sampling of a random point of the latent space could result in an uninterpretable output. For this manner, the AE algorithm is not the most suitable method for generating new content (e.g images).

3.2.3 *Variational Autoencoder*

The algorithms that learn the probability distribution of an input domain and are capable of generating new data (e.g images) from samples of a compressed representation (e.g latent space) are known as generative models.

The Variational Autoencoder (VAE) [34] is a generative deep learning technique for learning latent representations. During the training process, to make sure that the VAE network is generative, the latent space has to be continuous, allowing interpolation and random sampling. This continuity means that two points in the latent space that are close to each other should not give two outputs completely different after being decoded . The presence of discontinuities in the latent space will lead to the generation of unrealistic output data since the decoder is not capable of recognising the features.

VAEs have an architecture similar to the traditional AEs, with an encoder and a decoder. However, instead of learning a compressed representation of the input data, the encoder learns the probability distribution of the input features and provides the mean and standard deviation of the learnable space, as represented in Figure 20. To compute the reconstruction loss, first a latent representation must be sampled from the statistical learned distribution. Afterwards, this representation is decoded generating a reconstructed image.

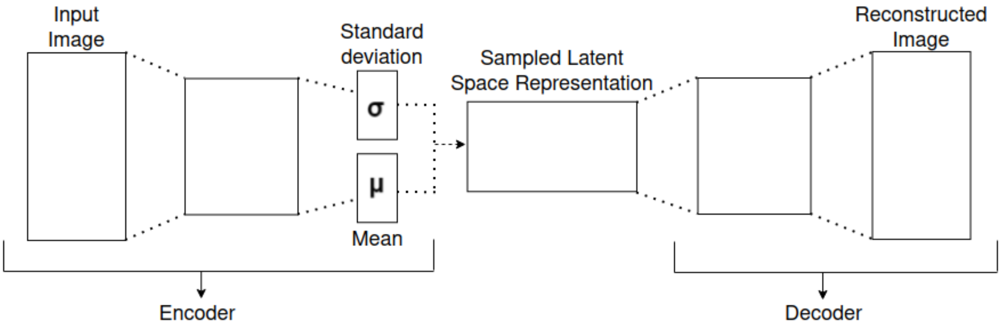


Figure 20: Variational Autoencoder architecture.

One of the disadvantages of using VAE network, is that the $\mathcal{L}_2$ reconstruction loss used during the training often produces blurry results. This loss minimizes the mean per-pixel squared difference, and as the error approximates to zero the computed gradients tend to become smaller. This decrease negatively impacts the network capability of producing sharper details, since it does not penalize as well the smaller deviations between the original and reconstructed image. The combination of dimensionality reduction, its generative properties and the training stability makes the VAE a versatile and interesting network for solving different problems. Choi et al. [35] train a VAE to estimate PET images from the brain region at different ages. The trained network was used to predict future brain metabolic topography by estimating PET images. Choi et al. [35] proposed a method to predict future brain metabolic topography by estimating PET images. For this purpose, a VAE network was trained to estimate PET images from the brain region at different ages.

3.2.4 *Generative Adversarial Network*

A Generative Adversarial Network [36] is a deep learning framework capable of creating data with high level of realism. Traditionally, it is composed by a network that generates data (dubbed generator) and a network that distinguishes if data is real or generated (dubbed discriminator). Figure 21 illustrates the GAN basic architecture.

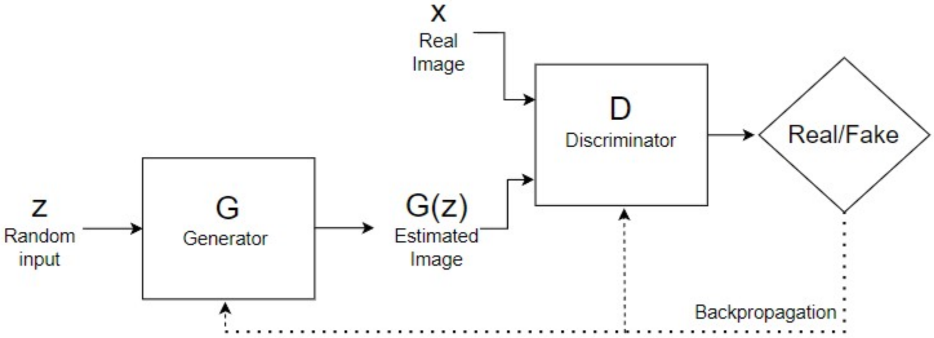


Figure 21: GAN basic architecture.

Both generator (G) and discriminator (D) are trained simultaneously in a minimax game. The generator attempts to deceive the discriminator, by generating data increasingly realistic, thus forcing the discriminator to better distinguish generated from original data. The GAN function can be defined as:

$$\mathcal{L}_{GAN}(D,G) = \mathbb{E}_{x \sim p_{\text{data(x)}}}[logD(x)] + \mathbb{E}_{z \sim p_{\text{z(z)}}}[1 - log(D(G(z)))], \tag{2}$$

where x represents the real data, $G(z)$ the fake samples generated by the generator G (from a random noise vector z), $\mathbb{E}_{x \sim p_{\text{data(x)}}}$, and $\mathbb{E}_{z \sim p_{\text{z(z)}}}$ represents the expected value of the x distribution and z distribution respectively. D is the probabilistic indicator of the discriminator performance, where it is expected to return the value '1' if the input image is real, and '0' if it is estimated. $D(x)$ and $D(G(z))$ indicates the probability that D discriminates the x and $G(z)$ respectively. The GAN is optimized using a minimax optimization function, that can be defined as:

$$\min_{G} \max_{D} \mathcal{L}_{GAN}(D,G). \tag{3}$$

The discriminator D wants to maximize the objective function such that $D(x) = 1$ for the real images, and $D(G(z)) = 0$ so that it identifies the estimated images. On the other hand, the generator G aims to minimize the objective function such as that $D(G(z)) = 1$, so that the discriminator is tricked to identify the estimated data as real.

The GAN model has been widely used in different tasks such as text-to-photo image synthesis [37], super-resolution [38], and Image-to-image translation [39]. Differently from the VAE, because of the minimax optimization game the GANs parameters tend to oscillate, becoming unstable during the training which difficult the training procedure. One of the key advantages is the network versatility, that is the capability of learning detailed features from different domains.

3.3 GENERATIVE ADVERSARIAL NETWORK SCHEMES

The GAN research topic has aroused the scientific community interest, leading to the development of new and improved approaches. In this section, several variations of the GAN model are presented.

3.3.1 *Conditional GAN*

Ian Goodfellow et al. proposed the original GAN [36]. The generator input is composed of a random noise vector z and there are no restrictions of its use. Therefore, this network is sometimes considered an unconditional GAN due the lack of control of the output generated. In order to control the model output, Osindero et al. [40] proposed the conditional GAN. Figure 22 illustrates the conditional GAN architecture. The conditional GAN function can be defined as:

$$\mathcal{L}_{cGAN}(D,G) = \mathbb{E}_{x\sim p_{\mathrm{data(x)}}}[logD(x|y)] + \mathbb{E}_{z\sim p_{\mathrm{z(z)}}}[log(1 - D(G(z|y))], \qquad (4)$$

where the y represents the additional information fed to network, x represents the real data, z the random noise vector, $G(z|y)$ the estimated samples generated with the additional label y by the generator G, $D(x|y)$ the probability that D input is real given the label y, and $D(G(z|y))$ is the discriminator result of the estimated image given the label y. Similar to the original GAN, the $\mathbb{E}_{x\sim p_{\mathrm{data(x)}}}$ and $\mathbb{E}_{z\sim p_{\mathrm{z(z)}}}$ represents the expected value of the x distribution and z distribution respectively.

The conditional GAN also resort to a minimax optimization function, that can be defined as:

$$\min_{G} \max_{D} \mathcal{L}_{cGAN}(D,G) \tag{5}$$

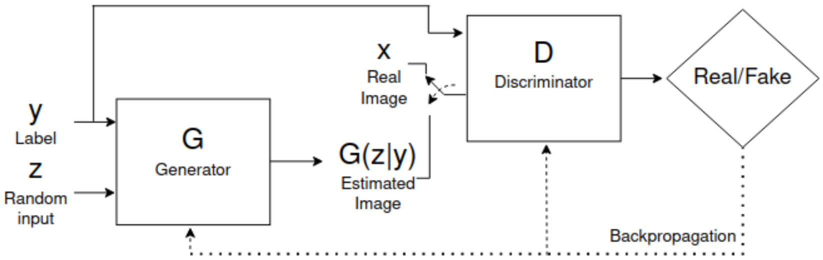


Figure 22: Conditional GAN framework.

Osindero et al. evaluated the proposed framework using the famous MNIST [41] dataset. If an unconditional GAN was trained with MNIST handwritten digits dataset, the network would not be able to control the output (e.g. class of digits) of the estimated samples. On the other hand, with the conditional GAN, it was demonstrated that feeding information such as data labels (e.g class of digits), in both G and D, allows to control the estimated data, thus obtaining the desirable output (e.g handwritten digit). An example of the MNIST digits estimated with the Conditional GAN can be seen in the Figure 23, where each row is conditioned by a different label (y), and in each column there is a different estimated sample ($G(z|y)$).

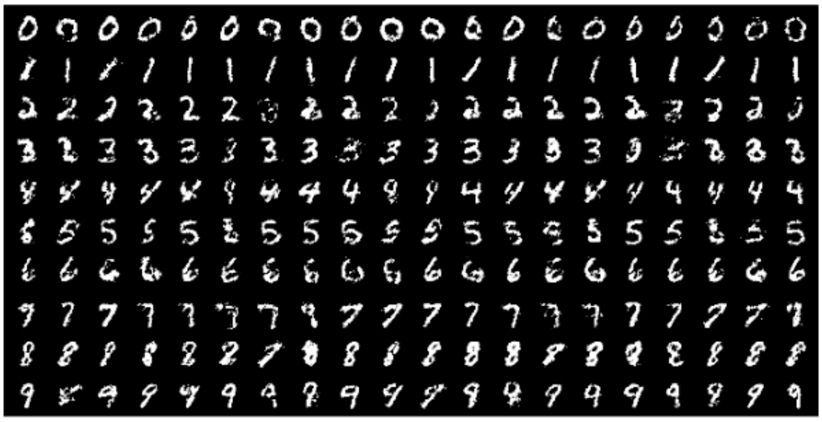


Figure 23: Estimated MNIST digits by the Conditional GAN framework [41].

3.3.2 *Information Maximizing GAN*

Chen et al. proposed an unsupervised approach to learn interpretable and disentangled information, entitled InfoGAN [42]. The objective of this network is to control the output estimated by structuring the latent space for the generator G. For this reasons, the input of the generator is decomposed in the noise vector z and a latent code c that is unknown and it is obtained through the training procedure. Instead of feeding a label (e.g. text) to the generator as in the conditional GAN, the learnable latent code c is used to control the output generated. The InfoGAN framework, proposes to maximise the mutual information between the latent code c and the generated data, $G(z)$, to improve the quality of the learned features. The objective function can be defined as:

$$\mathcal{L}_{iGAN}(D,G) = \mathbb{E}_{x \sim p_{\mathrm{data(x)}}}[logD(x)] + \mathbb{E}_{z \sim p_{\mathrm{z(z)}}}[1 - D(G(z))] - \lambda I(c;G(z,c)), \quad (6)$$

where $I(c;G(z,c))$ measures the amount of information learned in the latent code c about the generator output, and λ is a hyperparameter that controls the importance given to the mutual information exploited. The InfoGAN optimization function can be defined as:

$$\min_{G} \max_{D} \mathcal{L}_{iGAN}(D,G) \quad (7)$$

Since the objective is to maximise the mutual information between c and $G(z,c)$, the $I(c;G(z,c))$ should be high. Therefore, when the minimizing the objective function the $I(c;G(z,c))$ value must be subtracted, so that the an increase on its value represents an improvement on the mutual information learned.

To evaluate the performance of the proposed method, a regular GAN, that serves as anchor and the InfoGAN were both trained with the MNIST dataset. For both networks the generator input was divided in two parts, a noise vector z and a latent code c. The obtained results demonstrate that InfoGAN training procedure is faster when compared with the traditional GAN. An example of the estimated MNIST digits using both networks can be seen in the Figure 24, where in each row is used a different latent code c. It is showed that in traditional GANs, is not guaranteed that the generator will use the latent code, thus the variation on the latent code does not present a clear meaning, as shown in Figure 24(a). In contrast, the results obtained with the InfoGAN network demonstrate that training the mutual information maximization procedure allows to learn disentangled representations, that in this is case is to learn how to distinguish the numbers, where different sampled codes are able to generate different numbers from the MNIST dataset, as it can be seen in the Figure 24(b).

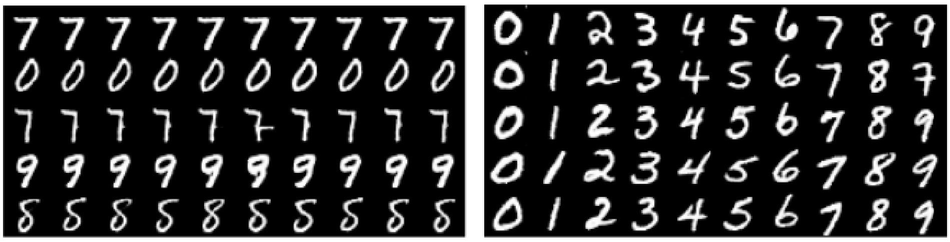

(a) (b)

Figure 24: (a) Estimated MNIST digits by InfoGAN (b) Estimated MNIST digits by regular GAN [42]

3.3.3 *BigGAN*

Although GAN has demonstrated to be a reliable tool for image generation, they are usually restricted to low images resolutions (e.g 64×64 or 128×128) and the success of the training procedure is highly dependent upon the selected hyperparameters. In [43], A. Donahue et al. studied and tested several techniques to efficiently generate high resolution and high fidelity images. This work was carried out with a special focus on real-world images from the ImageNet dataset.

With the resulting framework, named BigGAN, it has been demonstrated that scaling and training models with a higher number of parameters (e.g. more feature maps), an orthogonal regularization [44], and eight times the batch size compared to prior state-of-art GAN training procedure achieved estimated images with higher quality. The orthogonal regularization prevents the vanishing or exploding signals (due the several multiplication operations) on the convolutional layers by enforcing the weight matrices to be orthogonal. This is possible because the norm of a matrix is invariant when multiplied with an orthonormal, thus it preserves its magnitude. Also, instead of jointly updating the discrimination and generator model, the BigGAN uses a slight different approach, where in each training iteration the generator updates after the discriminator has updated twice.

The proposed framework was assessed using images with three different resolutions, 128×128, 256×256, and 512×512, and with an inception score metric that recurs to a pre-trained learning neural network to classify the quality of the image, where higher values indicate better quality. For the 128 × 128 resolution images it used the GAN framework proposed by Zhang et. al. in [45] as baseline. The results obtained demonstrated an improvement from a 52.2 to 166.5 on the inception score, for the 128 × 128 resolution. It also achieved 232.5 and 241.5 for the 256 × 256, and 512 × 512 resolution, respectively. This framework demonstrated to achieve high quality results in the estimation of natural images of multiple categories, as it can be seen in the Figure 25.

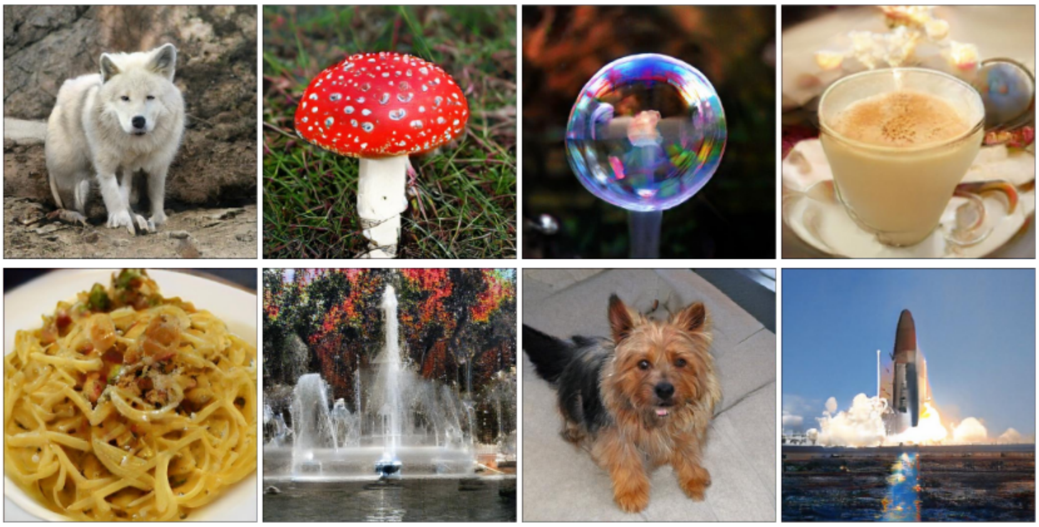


Figure 25: Images estimated by BigGAn Framework at 512 × 512 resolution[43].

3.4 IMAGE TO IMAGE TRANSLATION APPLICATIONS

The GAN framework can be used in a variety of applications, as stated previously. One of the most intriguing of its emerging applications, and that closely related to the main focus on this thesis, is I2I translation. In this section other I2I frameworks are detailed.

3.4.1 *Pix2pix*

Isola et al. proposed Pix2Pix [39], a conditional GAN trained in a supervised manner, to translate an input image to an output image domain. This type of approach requires the use of two datasets with paired images. In the Pix2pix architecture a pair of images is given as input (e.g. x and y), where the x is fed to the generator G, and y the corresponding pair is the target to be estimated, as shown in Figure 26. The generator G learns to translate x to an estimated image $y' = G(x)$. The objective is to estimate an image y' identical to the corresponding ground-truth y.

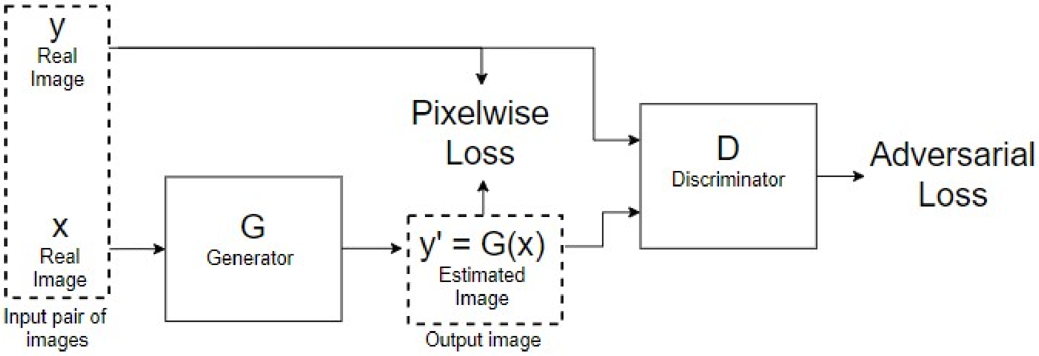


Figure 26: Pix2Pix architecture.

The discriminator D uses a convolutional PatchGAN architecture that penalises structure differences at a scale of $N \times N$ image patches, classifying individually each patch as real or fake. The generator G uses a UNET architecture that exploits skip connections between layers. The Pix2Pix network is trained in an adversarial manner using the conditional GAN loss presented in Equation 4 .*Isola et al.* also introduced a pixelwise loss based on the $\mathcal{L}_1$ of the difference between the estimated image y' and the corresponding ground-truth image y, that can be defined as:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|] \tag{8}$$

The complete objective function can be defined as:

$$\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \times \mathcal{L}_{L1}(G), \tag{9}$$

where λ controls the importance the pixelwise loss has during the training procedure.

Pix2Pix has demonstrated to be an efficient approach to synthetize photos from label maps, image reconstruction, and colorizing from edge maps. Figure 27 demonstrates some examples of the results obtained in the edges-handbags dataset using the Pix2pix model. For this task, the Pix2pix used a paired dataset composed with the edges images of handbags and their corresponding ground truths, and was trained to colorize the edge images. This application can be useful in multimedia applications, where the user draws the edges of a certain object and the Pix2pix network returns a colorized image of the object draw.

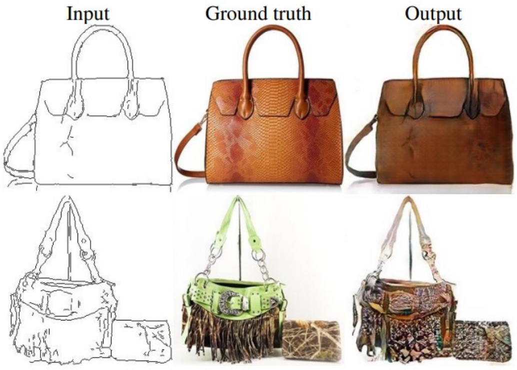


Figure 27: Example results of Pix2pix method on detected edges-handbags dataset [39].

One of the disadvantages of using the Pix2pix model is the need of paired datasets during the training procedure. Depending on the application, paired datasets can be difficult to obtain.

3.4.2 *BicycleGAN*

Zhu et al. proposed BicycleGAN[46], a hybrid framework composed by a conditional Variational Autoencoder GAN (cVAE-GAN) [47] and a Conditional Latent Regressor GAN (cLR-GAN), as shown in Figure 28.

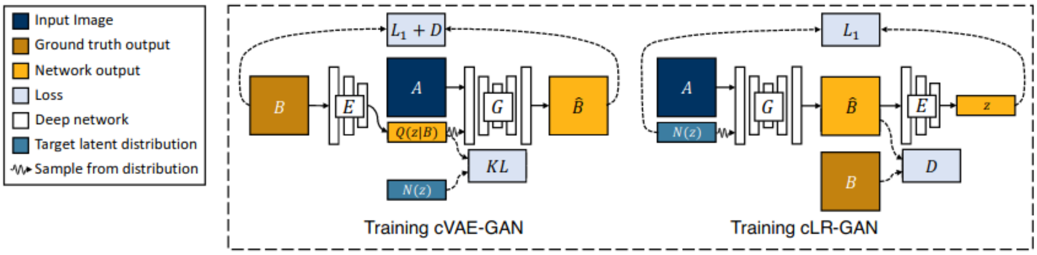


Figure 28: BicycleGAN training procedure [46]

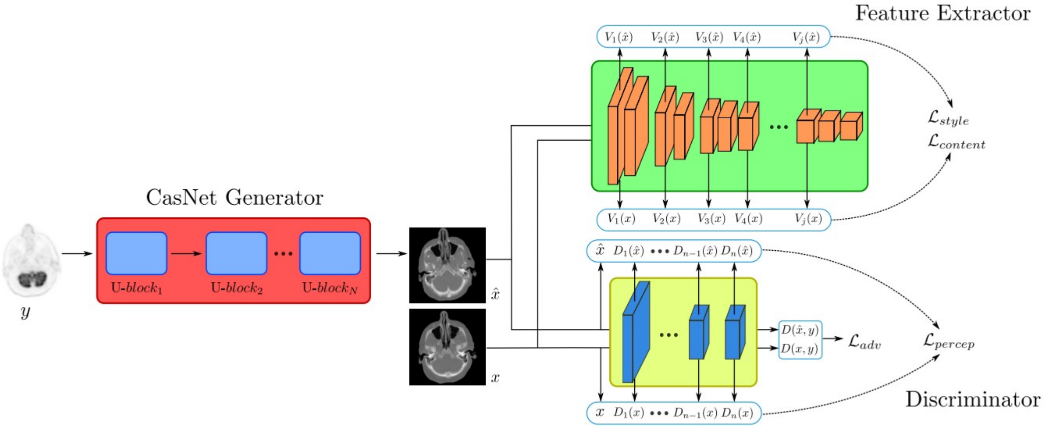
BicycleGAN is a supervised multimodal Image-to-image translation model. It enforces an injective mapping by adding an extra encoder, E, that should learn learn the mapping from the generated data, $\hat{B}$, back to the latent representation z. In this manner, two different latent codes cannot generate the same output. The cVAE-GAN is responsible for encoding a ground truth source B to a latent space, $Q(z|B)$, and then feeding it, along with input source A, to a generator G in order to synthesise $\hat{B}$ that should be identical to the ground truth B.

The cLR-GAN samples a random latent code from the target latent distribution $N(z)$ that, along with the input source A, is given to the generator G to estimate the ground truth B. The generator output is then encoded back to the latent space z with the objective of reconstructing the initial sampled code.

BicycleGAN demonstrated results that were more realistic and with an higher diversity when compared to Pix2Pix. However, Pix2pix is a less complex method and therefore is easier to train. BicycleGAN is trained in a supervised manner, thus it contains the same issue than Pix2Pix, where both require paired dataset that are not always possible to collect.

3.4.3 *MEDGAN*

Armanious et. al, in [48], proposed a supervised GAN, called MedGAN, for medical Image-to-image translation that was applied on PET to CT translation, correction of MRI motion artefacts, and PET image de-noising. The MedGAN also introduced a new generator architecture entitled CasNet. This framework receive as input a pair of images, x and y, where the CasNet generator is used to translate y into the x image domain. The Figure 29 shows the MedGAN framework applied to the PET to CT translation application. Where the PET image (y) is fed to the generator, which translates it to the target domain, thus obtaining a estimated CT $(\hat{x})$.



Feature Extractor
$V_1(\hat{x})$ $V_2(\hat{x})$ $V_3(\hat{x})$ $V_4(\hat{x})$ $V_j(\hat{x})$
$\mathcal{L}_{style}$
$\mathcal{L}_{content}$
CasNet Generator
U-$block_1$
U-$block_2$
U-$block_N$
y
$\hat{x}$
x
$V_1(x)$ $V_2(x)$ $V_3(x)$ $V_4(x)$ $V_j(x)$
$\hat{x}$ $D_1(\hat{x})$ ••• $D_{n-1}(\hat{x})$ $D_n(\hat{x})$
$D(\hat{x}, y)$
$D(x, y)$
$\mathcal{L}_{adv}$
$\mathcal{L}_{percep}$
x $D_1(x)$ ••• $D_{n-1}(x)$ $D_n(x)$
Discriminator

Figure 29: MedGAN Framework [48]

This framework combines the adversarial conditionalGAN loss described in the Section 3.3.1, with three non-adversarial losses (style, content, and perceptual loss). This non-adversarial losses are used to enhance the estimated images global structure, in addition to the fine details and ensure that the target textures are well represented. To compute the style and content loss, it uses a pre-trained feature extractor (V) to obtain rich features and compare those generated by the network ($V_n(\hat{x})$) with those from the target ($V_n(x)$). Combining both style and content losses, forces the network to learn the style transfer (texture and details) of the input to the estimated images. The content loss objective is to minimize the difference between the target and estimated features representations, whereas the style loss is used to minimize the discrepancies in the style (e.g. textures). Instead of using V to extract the features representations, to compute the perceptual loss, the Discriminator D is used. This loss minimizes the mean absolute error (MAE) between the features representations extracted, enhancing the global details of the estimated images.

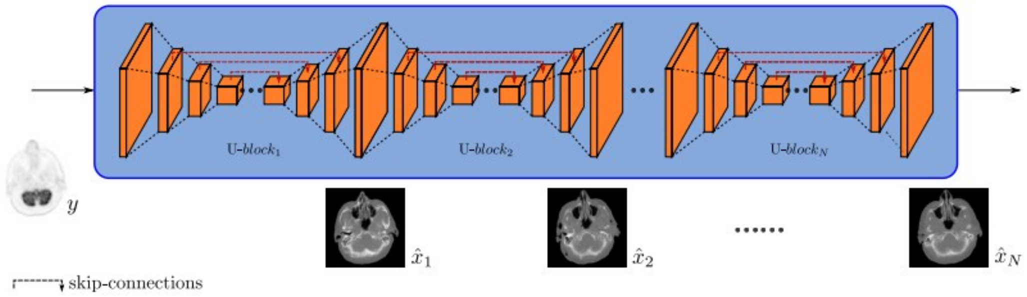


Figure 30: Casnet generator architecture [48]

The CasNet generator consists in the merge of multiple fully convolutional encoder-decoder networks. As shown in Figure 30, the first block performs the PET to CT translation ($y \rightarrow \hat{x}_1$ and the remaining are responsible for progressively enhance the CT quality. The encoder-decoder network, also know as U-block, is based on the U-net [49] architecture. The decoder is similar the encoder architecture with the difference that it uses strided deconvolutions instead of convolutions. The U-block contains skip-connections (identified by the red lines) that are used to pass features from the encoder to the respective decoder. This is achieved by concatenating the output of a convolutional layer from the encoder to the respective layer in the decoder side. The use of such connections prevents the gradient to vanish, by providing an alternative path to the gradients during the backpropagation, aswell as an improvement of the estimated images quality.

To assess the MedGAN performance, different I2I frameworks, such as Pix2pix, were tested for the three applications, and five radiologists performed a perceptual study on the fidelity of the estimated images. An example of the results obtained with MEDGAN for the PET-CT translation, MR motion correction, and PET denoising can be seen in the Figure 31. The results obtained made it clear that the proposed MEDGAN outperformed other I2I frameworks, in every task. The

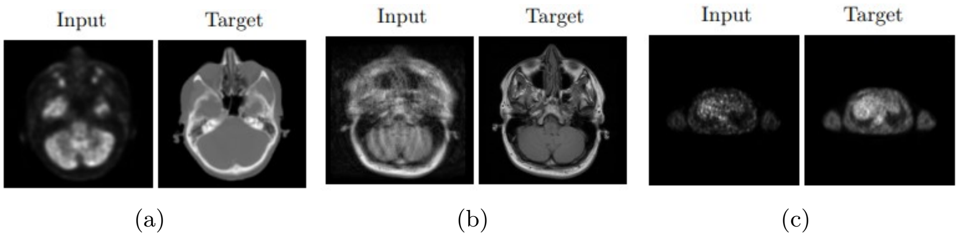


Figure 31: An example of the results obtained with MEDGAN framework for the three different applications. (a) PET-CT translation (b) MR motion correction (c) PET denoising [48]

radiologists rated the ground truth images between 3.7 – 3.8 in a scale of 4, and between 2.8 – 3.2 the estimated images by the MedGAN. The Pix2pix present the worst result, as their estimated images were rated between 1.7 – 2.0. Similar to Pix2pix and BicycleGAN, one of the disadvantages of the MEDGAN is that requires paired datasets, in this case PET-CT paired images which are not easily available.

3.4.4 *CycleGAN*

Zhu et. al proposed CycleGAN [50], a new cycle-consistent adversarial network capable of learning the mapping between two image domains. This network model uses unsupervised training with two different generators (G and F) to translate an input domain X to a target domain Y and vice-versa, as shown in the Figure 32.

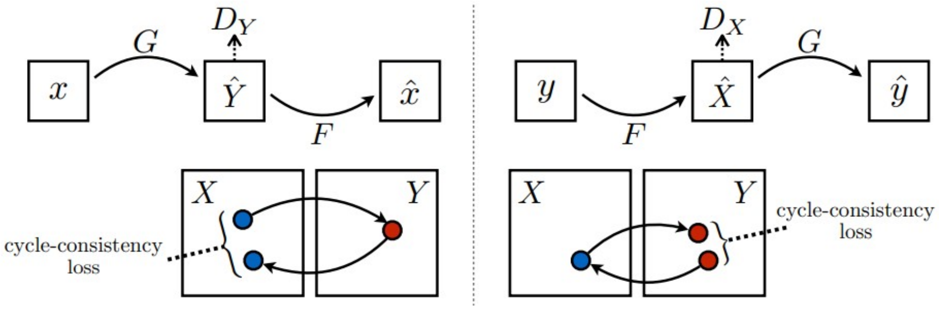


Figure 32: CycleGAN model [50]

The 'cycle-consistent' property defines that an image in the source domain x, can be mapped to the the target domain $\hat{Y}$, and this translated image in the target domain can be mapped back to the original image in the source domain $\hat{x}$. The inclusion of the cycle-consistent loss allows the training process to be unsupervised, therefore the use of unpaired datasets. CycleGAN resorts to an identity loss function to regularise the generator, ensuring that when the input data is part of the target domain, the output generator is similar to the input.

CycleGAN has demonstrated high quality results on tasks that involve color or texture transformation. However for tasks that require geometrical transformation such as cat to dog image-to-image translation, CycleGAN is not effective, as it generates low quality results.

3.4.5 *Attention-Guided Generative Adversarial Network (AGGAN)*

Hao Tang et. al, in [51], proposed the Attention-Guided Generative Adversarial Network (AGGAN), an unsupervised GAN with an embedded attention mechanism to translate images from different domains, e.g. CT to PET.

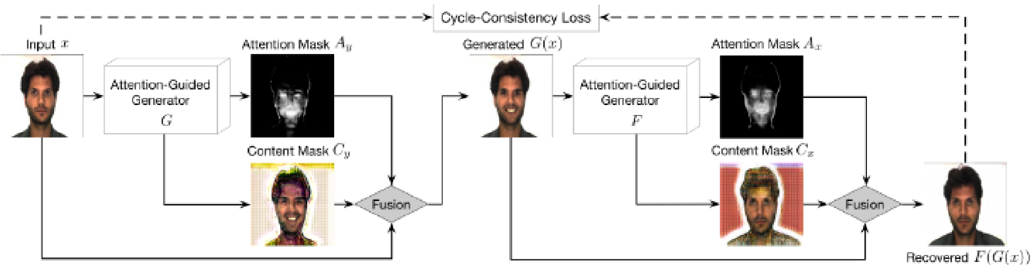


Figure 33: AttentionGAN framework [51]

As shown in Figure 33, the AGGAN presents a built-in attention mechanism, which can detect the most discriminative semantic parts of images in different domains. It consists of two generator networks with an embedded attention method, being the generators adversarially trained with the respective discriminator. The attention-guided generator computes a content mask C_y which is an intermediate representation of the target image, and an attention mask A_y that defines the pixel contribution of C_y. The differences between the mask extraction processes lie in the last layer used, where the kernel size differs. A softmax function is used to determine the attention mask, and a hyperbolic tangent function to the content mask. The combination of the content and attention masks with the input image generates an image from the target domain, $\hat{y} = G(x)$, where $G(x) = x \times A_y \times C_y$. The quality of both masks is improved / learned during the backpropagation process of the generator optimization.

One of the key advantages of this type of approach is that it allows the generator to focus on the most discriminative parts of the image. Thus, the learning process is optimized by focusing on the most important regions of the image. This architecture was found capable of generating more detailed and higher quality images, when compared to Pix2pix and CycleGAN.

3.4.6 *Cycle-MedGAN*

Supervised approaches require medical imaging paired datasets that are not broadly available, therefore *Armanious et. al* further proposed the CycleMedGAN [20], an unsupervised medical image translation network. This architecture is based on CycleGAN [50] with the addition of new non-adversarial cycle losses. The unsupervised Cycle-MedGAN Framework is presented in Figure 34. The generator $G1$ is used to translate an input domain x (e.g. PET) to a target domain y (e.g CT). To guarantee the cycle-consistency, the generator $G2$ is responsible to translate the estimated image ($\hat{y} = G1(x)$), back to the original input domain ($\hat{\hat{x}} = G(\hat{y})$). Similar to MedGAN, Cycle-MedGAN also uses a feature extractor F to compute a style and perceptual loss which penalises feature differences between generated and target features.

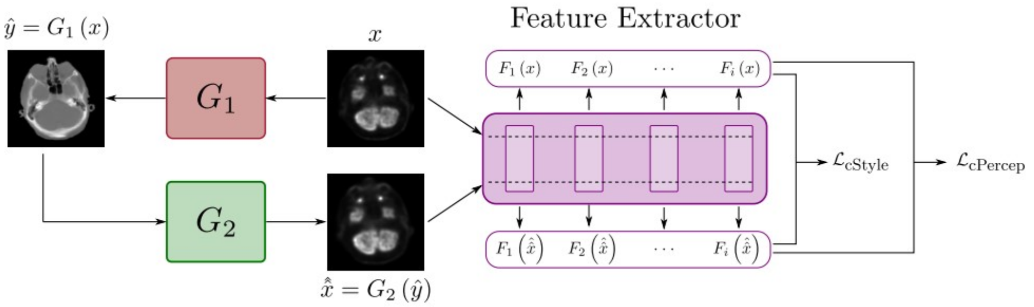

$\hat{y} = G_1(x)$
G_1
x
Feature Extractor
$F_1(x)$ $F_2(x)$ $\cdots$ $F_i(x)$
$\mathcal{L}_{\text{cStyle}}$
$\mathcal{L}_{\text{cPercep}}$
G_2
$F_1(\hat{\hat{x}})$ $F_2(\hat{\hat{x}})$ $\cdots$ $F_i(\hat{\hat{x}})$
$\hat{\hat{x}} = G_2(\hat{y})$

Figure 34: CycleMedGAN Framework [48]

Cycle-MedGAN demonstrated superior quantitaive and qualitative results for PET-CT translation and MRI motion correction when compared to other unsupervised translation methods such as CycleGAN.

3.5 DISCUSSION

The AE was demonstrated to be versatile and a reliable method for compression, denoising and anomaly detection applications. However, due the lack of regularization on the learned latent space, it is not the most suitable network to applications that require the generation of new content. The AE training does not guarantees that every point sampled from the latent space will result in a meaningful result once decoded. To overcome this issue, the VAE introduced regularization in the latent space and during the train procedure it enforces the latent space to be continuous, which means that two neighbour points in the latent space will result in similar results once decoded. One of the disadvantages of the VAE is that the generated images tend to be blurry. The proposed GAN framework tend to produce higher quality and detailed images. Nevertheless, the GANs can become unstable during the training, which makes such type of network harder to be trained when compared to VAE and AE.

For Image-to-image translation, these networks can be trained using either a supervised or an unsupervised approach. The supervised networks such as Pix2pix, BicycleGAN, and MEDGAN required paired datasets that can be very difficult to obtain. BicycleGAN combined a conditional VAE and a conditional Latent Regressor GAN and it demonstrated to estimate higher quality images when compared to the Pix2pix network. The lower complexity level of the Pix2pix framework, allows easier and faster training. The casnet generator that enhances the quality of the estimated image along with the use of a pre-trained feature extractor to compare the features from the original and estimated images allows the MEDGAN to be a versatile network that can be used in different applications. The MEDGAN was capable of outperforming other frameworks such as Pix2pix in terms of image quality estimated. To overcome the need of paired datasets, the cycleGAN framewok was proposed. AGGAN and the Cycle-MedGAN used the cycle-consistency loss proposed in the cycleGAN architecture, to be trained in a unsupervised manner. AGGAN makes use of learnable masks to enforce the generator to focus on the most important regions of the image. Similar to MEDGAN, the Cycle-MedGAN also uses a pre-trained feature extractor, which is allows to compute a loss based on the features extracted. This loss enforces the network to represent more detailed information. Both AGGAN and the Cycle-MedGAN, demonstrated to achieve better results when compared to the cycleGAN.

3.6 SUMMARY

This chapter presented an overview of state of the art methods for I2I based on deep learning. Several deep learning architectures were presented, with special emphasis on frameworks with generative capabilities was first presented, focusing the generative adversarial network. Accordingly, several variants of the traditional GAN approach were detailed. It was pointed out that the GAN algorithms can be modified and tuned in order to control and improve the generator output.

The chapter provides background for the main objective of this thesis, which is the development of an approach to efficiently compress the PET and CT image modalities exploiting methods based on I2I. Accordingly, this chapter is concluded with the a review on I2I frameworks addressing both supervised and unsupervised training procedure. This review showed that deep learning is a relevant topic and it can support multiple applications. Therefore this is also a promising approach with high potential for new contributions in emerging fields of image processing.

4

PROOF OF CONCEPT - CROSS-MODALITY LOSSLESS IMAGE COMPRESSION

In this chapter, a proof of concept was conducted to study and evaluate a lossless compression framework of PET and CT pairs, based on Image-to-image translation methods. First, a detailed description on the proposed baseline of the cross-modality lossless image compression scheme is presented. Then, the network used to translate the CT image to the PET domain is explained, including the generator and discriminator architecture and the losses used during the training process. Finally, three different compression strategies are presented and their compression efficiency analysed.

4.1 PET-CT CROSS-MODALITY IMAGE COMPRESSION

The proposed method resorts to an Image-to-image translation framework which is used to obtain an estimation of the PET image from a CT to reduce the amount of information needed to compress and recover the PET-CT pairs without loss, differently from traditional approaches where both modalities are individually compressed. Therefore, instead of the classical procedure, only the original CT and a residual representation of the respective PET image need to be encoded, as shown in the processing pipeline presented in Figure 35.

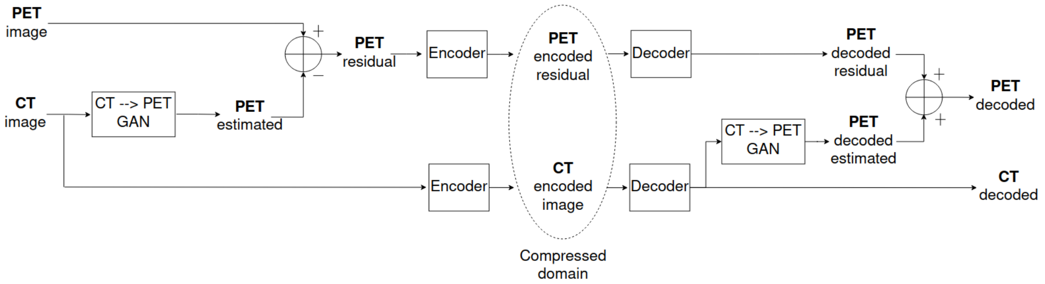


Figure 35: Proposed Cross-modality Lossless Compression scheme.

In the proposed pipeline, the CT image is encoded and stored first. Then, an estimated PET is obtained by translating the original CT image into the PET domain using a GAN. In order to achieve the residual PET representation (PET residual), the difference between the PET estimated and PET image is computed and encoded. To retrieve the original PET-CT pair, it is required to first decode the residual PET and CT image; then, with the GAN network that was used in the beginning, perform a CT to PET domain translation with the decoded CT decoded, thus achieving a PET identical to the one previously used to compute the residual PET representation. Finally, the decoded residual is added with the estimated PET to obtain the final PET image. As this estimated PET is mathematically equal as the previous one, the final PET will also be exactly equal to the original one. It is expected that the PET residual that results from the difference between two PET images requires less information to be compressed when compared with the original PET image.

The functional blocks of the presented workflow are a lossless image codec and a GAN that translates the CT image to a PET image. The implemented GAN architecture is based on the Attention-Guided Generative Adversarial Network (AGGAN) proposed by Hao Tang *et. al* [51], with a reformulation of the generator and discriminator loss functions. Two different lossless codecs, HEVC Intra and JPEG 2000, were used to encode the PET and CT paired images.

4.2 CT TO PET IMAGE TRANSLATION

A supervised GAN is proposed to perform the domain translation from CT to PET, based on the AGGAN framework. This is composed by an attention-guided generator (G) and a discriminator (D), also including two non-adversarial losses, namely perceptual and style. The proposed GAN architecture, produces an attention mask A_y and a content mask C_y that combined generate the estimated PET, as shown in Figure 36. The discriminator is used to distinguish the real (y) and estimated ($\hat{y}$) PET image, and also as a feature extractor. The features maps are extracted from the discriminator layers (D_1,D_2,D_3,D_4,D_5), and used to compute the two non adversarial losses, perceptual, and style loss, which in turn are used to minimize the generator optimization function. The penalisation of the features representation allows the generator to be optimized with the information that the discriminator considers the most discriminative to identify whether the input data is real or estimated. Thus, it improves the capability of the generator to deceive the discriminator.

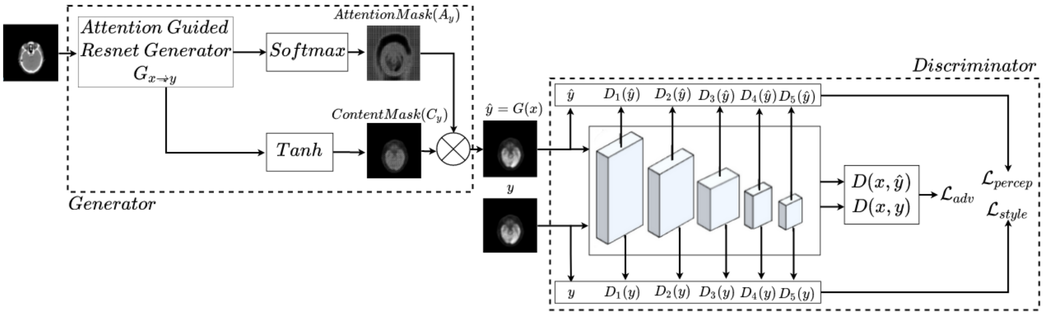


Figure 36: Generative Adversarial Network (GAN) for CT to PET estimation.

Initially, the hypothesis of using the original PET to estimate a CT image and, therefore, compressing the CT residual instead of the proposed PET residual was explored. Unfortunately, the results were not were not good enough, in terms of compression. One of the main limitations of the GAN framework is the difficulty to ensure training stability, as many factors can negatively impact on it performance and results. One of the most common issues is when the Generator is no longer capable of deceiving the Discriminator network. Such difference in the networks performance lead to vanishing gradients problem: as the computed gradients tend to zero, little feedback is provided to the generator, which does not allow the generator to improve. To overcome the this issue, one strategy is to change the cross-entropy adversarial loss used in the traditional GAN, as shown in Equation 2 presented in Chapter 3. Therefore it was decided to adopt the least square loss proposed by Xudong Mao *et al.* [52] for both discriminator and generator, as shown by Equations 10 and 11,

$$\mathcal{L}_{\mathrm{LSGAN}}(D) = \frac{1}{2}\mathbb{E}_{y\sim p_{\mathrm{data(y)}}}[(D(y)-1)^2] + \frac{1}{2}\mathbb{E}_{x\sim p_{\mathrm{data(x)}}}[D(G(x))^2], \tag{10}$$

$$\mathcal{L}_{\mathrm{LSGAN}}(G) = \mathbb{E}_{x\sim p_{\mathrm{data(x)}}}[D(G(x)-1)^2], \tag{11}$$

where x represents to the original CT image, y represents the original PET image, $G(x)$ corresponds to the estimated PET image, $\hat{y}$, obtained with the translation of the original CT image to the PET domain, $\mathbb{E}x \sim p_{\mathrm{data(x)}}$, and $\mathbb{E}y \sim p_{\mathrm{data(y)}}$

corresponds to the expected value of CT and PET data distribution, respectively. The least square GAN objective functions can be defined as:

$$\min_{D}\min_{G}\mathcal{L}_{\text{LSGAN}}(D)+\mathcal{L}_{\text{LSGAN}}(G), \tag{12}$$

Differently from the traditional GAN loss, the least square loss does not employ a minimax game, instead it minimizes the square distance between the discriminator and target result for both networks. It computes the difference between the discriminator probabilistic result and the supposed label (real or estimated). To the real label is assigned the value '1' while '0' is for the estimated. The discriminator function is minimized such that it identifies y as the real input, $D(y) = 1$, and $G(x)$ as the estimated image, $D(G(x)) = 0$. The generator function is also minimized, so that the estimated image is identified as real, $D(G(x)) = 1$.

This loss function penalizes the generated images based on their distance from the assigned value. This network is capable of providing to the generator information of how far generated images are to be classified as real, thus leading to higher gradients for a more effective generator. The lack of this distance information can lead, as pointed out in [53], to the known vanishing gradient problem. Thus, as described in [52], the least square loss improves the training stability.

Adversarial loss functions, naturally, have a fundamental role on the generative process during GAN training. However they do not guarantee that global structures of the estimated image are identical to the target image. For this reason, it is crucial to adopt new non-adversarial loss functions to reinforce the generation of high fidelity features, improving the details and minimising the pixel distance between the estimated images and their corresponding ground truth. Three non-adversarial losses (pixel, perceptual, and style) were implemented to enhance the GAN training quality. To minimize the differences between original (y) and estimated ($\hat{y}$) images, the pixel reconstruction loss (a L1 loss) proposed by *Isola et al.* [39] was adopted. This loss can be expressed as:

$$\mathcal{L}_{\text{Pixel}}(G) = \mathbb{E}_{x\sim p_{\text{data(x)}}}[\|G(x) - y\|_1] \tag{13}$$

where $G(x)$ corresponds to the estimated image $\hat{y}$, generated from a given the original CT image x. This loss measures the sum of the absolute difference between y and $\hat{y}$, which is generated by $G(x)$.

Although the pixel-reconstruction loss improves the capability of the network to generate images structurally more similar to the target, since it is based on the the pixel distance between estimated $\hat{y}$ and target y images, it can still lead to blurry results [54, 55]. The Image-to-image translation process is extremely delicate, in such a way that even if generated and target image are perceptually identical, a spatial shift by a couple of pixels in the generated image will hamper $\hat{y}$ image evaluation when a metric based on the pixel distance is used. Such metrics are not able to properly evaluate the feature representations learned by the GAN network. Thus, the perceptual loss used in [56] was adopted to minimize the discrepancy between high-level perceptual features from the target y and estimated $G(x) = \hat{y}$ images, and consequently to force both feature representations to be similar. Using feature maps extracted from the discriminator D layers, the adapted perceptual loss is given as:

$$\mathcal{L}_{Percep}(G) = \sum_{i=0}^{M} \lambda_{cP,i}(\|D_i(y) - D_i(G(x))\|_1), \tag{14}$$

where D_i represents the intermediate feature map, extracted from the i^{th} layer of the discriminator D, M indicates the number of hidden layers of the discriminator, and $\lambda_{cP,i}$ defines the importance given to the i^{th} layer.

Gatys *et al.* [28] proposed a style reconstruction loss, to assure the generator is capable of learning more detailed patterns. For that manner, the extracted feature maps from the discriminator network (D) are used to compute their correlation over the depth dimension which can be represented by the Gram matrix. The style loss is computed by the squared Frobenius norm of the difference between the Gram matrices obtained from the $\hat{y}$ and y feature representations:

$$\mathcal{L}_{Style}(G) = \sum_{i=0}^{M} \lambda_{cS,i} \frac{1}{4d_i^2} \|Gr_i(y) - Gr_i(G(x))\|_F^2, \tag{15}$$

where $\lambda_{cS,i}$ is a hyperparameter that defines the influence Gram matrix $Gr_i(y)$ extracted from the Discriminator i^{th} layer, M is the number of layers within the

Discriminator, and d_i is the spatial depth of the i^{th} layer. The matrix $Gr_i(y)$ can be determined as follows:

$$Gr_i(y) = \frac{1}{h_i w_i d_i} \sum_{h=1}^{h_i} \sum_{w=1}^{w_i} D_i(y)_{hi,wi} D_i(y)^{\intercal}_{hi,wi}, \tag{16}$$

where $D_i(y)_{h,w,m}$ is the feature map, extracted from the i^{th} layer of the discriminator D, $D_i(y)^{\intercal}_{hi,wi}$ is the transposed of feature map $D_i(y)_{hi,wi}$ and h_i,w_i,d_i are the height, width, and depth of the extracted feature space, respectively. Considering the loss functions discussed in the literature, the composite loss function of the proposed GAN framework, responsible to translate CT to PET images domain, is defined as:

$$\begin{split} L(G,D) = L_{\text{LSGAN}}(D) + \lambda_{\text{LSGAN}} \times L_{\text{LSGAN}}(G) + \lambda_{\text{pixel}} \times \mathcal{L}_{\text{Pixel}}(G) + \\ \mathcal{L}_{\text{Percep}}(G) + \mathcal{L}_{\text{Style}}(G), \end{split} \tag{17}$$

where λ_{LSGAN} and λ_{pixel} are the weights given to the least square and pixel losses, respectively. The optimization function can be defined as:

$$\min_D \min_G L(G,D) \tag{18}$$

The discriminator was implemented using the 70×70 PatchGAN, proposed by Isola *et al.* [39] with five discriminative layers, and for the generator, a ResNet architecture [57] with 9 residual blocks was adopted.

4.3 EXPERIMENTAL ASSESSMENT

To perform CT to PET image translation as a proof of concept, only the first subset described in the section 2.1.4, comprising 220 PET-CT pairs for train and 37 pairs for test, was used. The implemented GAN network was trained with 1500 epochs, with a batch size of 18, and batch normalization. The Adam optimizer [58] was adopted with the momentum terms $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a learning rate of 2×10^{-4}. A linear learning rate policy was used, thus after 250 epochs, the learning rate decreases linearly to 0. The learning rate value on each epoch, $lr(epoch)$, can be calculated by the Equation 19:

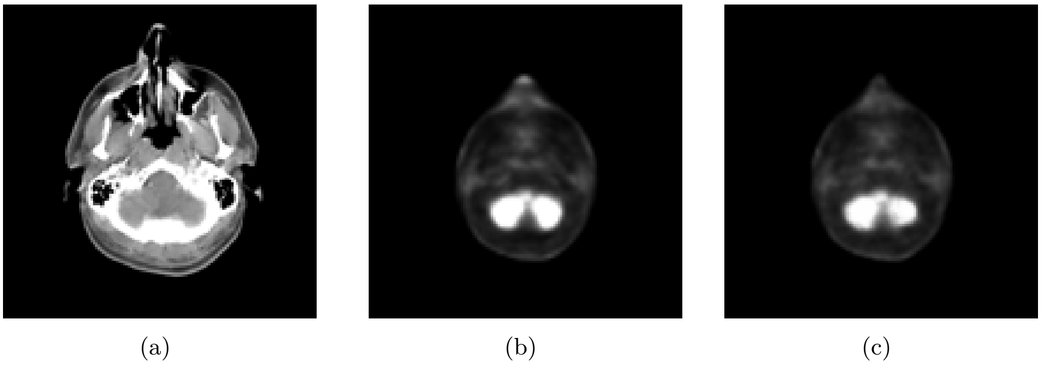
$$lr(epoch) = \begin{cases} lr_{init}, & \text{if } epoch < epoch_{lrinit}. \\ lr_{init} \times \left[1 - \frac{epoch - epoch_{lrinit}}{epoch_{total} - epoch_{lrinit}}\right], & \text{otherwise.} \end{cases} \quad (19)$$

where the *lrinit* is the initial learning rate, *epoch_lrinit* is the number of epochs that the learning rate remains equal, and *epoch_total* is the total number of epochs. Different hyperparameters were tested using a grid search method. The values which obtained the best results are summarized in the Table 2.

Table 2: PET-CT translation GAN selected hyperparameters

	λ_{Pixel}	λ_{LSGAN}	$\lambda_{cP,0}$	$\lambda_{cP,1}$	$\lambda_{cP,2}$	$\lambda_{cP,3}$	$\lambda_{cP,4}$	$\lambda_{cS,0}$	$\lambda_{cS,1}$	$\lambda_{cS,2}$	$\lambda_{cS,3}$	$\lambda_{cS,4}$
Losses weight	10	1	5	5	2.5	1.5	1	5	5	2.5	1.5	1

Figure 37 shows an example of an original PET-CT image pair and the corresponding estimated PET using the proposed GAN network.



(a) (b) (c)

Figure 37: (a) Original CT (b) Original PET (c) Estimated PET

To assess the performance of the proposed pipeline, the results obtained from the following three different coding scenarios were compared.

- Independent encoding of each image from the PET-CT pair: both the original PET and its corresponding CT image were losslessly encoded using HEVC Intra and JPEG 2000.
- Inter-frame coding of the PET-CT pair: HEVC was used to jointly encode both images modalities using lossless inter prediction, using the the CT image as prediction for the corresponding PET.

- Cross-modality coding: the proposed scheme, shown in Figure 35, was used. The original CT and the residual PET were losslessly encoded using HEVC Intra and JPEG 2000.

The efficiency of the proposed strategies was evaluated in terms of the coded data size and compression gain. Concerning the PET image, residual data with less energy will (generally) lead to better compression results. Therefore, since the PET residue is obtained from the difference between the original and its estimate, the similarity between both images is crucial to reduce the energy of the residual signal. To achieve this goal, the GAN is trained to estimate high fidelity images in order to maximise the similarity with the originals. The compression gain is defined as a percentual difference between the compressed data size (in bytes) obtained from the method under evaluation and a predefined reference, i.e.:

$$\text{Compression Gain} = \frac{\text{Size(ref)} - \text{Size(x)}}{\text{Size(ref)}} \times 100. \tag{20}$$

where Size(ref) and Size(x) are the compressed data size of the anchor and the alternative method, respectively.

The individual compression results of CT and PET images are summarized in Table 3. The compression gains presented in this table, use the original images compressed with HEVC Intra as anchor. The table presents average results obtained from the 37 PET-CT pairs used for testing.

Table 3: CT and PET individual lossless image compression results

Files	**Codec**	**Size (bytes)**	**Compression Gain (%)**
CT_Original	HEVC Intra	138610	-
CT_Original	JPEG 2000	170413	-22.9
PET_Original	HEVC Intra	92688	-
PET_Original	JPEG 2000	89468	3.5
PET_residual	HEVC Intra	**84516**	**8.9**
PET_residual	JPEG 2000	101108	-9.1

From these results, several conclusions can be drawn. Firstly, as expected, HEVC is more efficient than JPEG 2000. In this case, Table 3 shows a reduction of 22.9% on the average file size. However, for PET images, JPEG 2000 presents a slightly better efficiency than HEVC Intra (3.5% gain). This is possibly due to the specific type of structures contained in PET images where the intra prediction of HEVC may not perform very well. Nevertheless, in the case of the PET residue, coding with HEVC Intra presents the highest performance (8.9% compression gain). The JPEG 2000 was found to be less effective than HEVC compressing PET residual images (of -9.1%), indicating that is not the most suitable encoding algorithm for this type of residue, most likely due to the rather noisy nature of the data

The overall compression results for the joint PET-CT pair are presented in Table 4 . In this case, the reference to compute compression gains is the average PET-CT file size obtained by independently coding each of them with HEVC Intra in lossless mode. From these results it becomes clear that the proposed schemes leads to higher compression ratios (up to 3.53%). Using HEVC in lossless CT-PET inter prediction mode leads to gains of 2.69%. Finally, JPEG 2000 comes an underperformer which was more or less expected in the light of the results in shown in Table 3.

Table 4: Joint PET-CT lossless image compression results

Files	**Codec**	**Size (bytes)**	**Compression Gain (%)**
CT_Original + PET_original	HEVC Intra	231298	-
CT_Original + PET_residual	HEVC Intra	**223126**	**3.53**
CT_Original + PET_original	HEVC Inter	225080	2.69
CT_Original + PET_original	JPEG 2000	259881	-12.4
CT_Original + PET_residual	JPEG 2000	271521	-17.4

4.4 SUMMARY

In this chapter, a novel scheme for the joint, lossless, compression of PET-CT images based on an Attention-Guided Generative Adversarial Network was introduced. Firstly, the CT to PET translation process was described, focusing on the GAN loss function, which captures the fidelity of an estimated image in multiple dimensions. The proposed approach was used in lossless coding of the original CT and residue of the respective PET image pair. Such residual signal is determined from the difference between the original PET and its estimate generated by the GAN framework.

Different approaches were considered to measure the cross-modality compression pipeline efficiency. The results obtained demonstrate that the proposed compression scheme is able to surpass the state-of-art compression schemes with standard encoders, achieving a compression gain of 8.9% for the PET representation and

3.53% for the PET-CT pair. The best performance was achieved using the HEVC lossless in Intra mode.

5 REFINEMENT ON THE MULTIMODAL IMAGE COMPRESSION

Following the proof of concept to jointly compress PET-CT pairs, which demonstrated better lossless coding efficiency than conventional coding alternatives, this chapter presents the work developed to further improve the results. Despite the satisfactory compression gains, the chosen dataset was inherently limited in size and diversity, so the results obtained in the previous chapter cannot be considered fully unbiased. Therefore, there is room for improvement of proposed coding scheme based on I2I translation through deep learning, which is the aim of this chapter. The improvements targeted both, image processing techniques, and coding strategies.

With the results obtained in the proof of concept, it was verified that the compression efficiency could be compromised by several drawbacks derived from the estimation process. The issues observed consisted on the misalignment, head region size and contrast difference, between the estimated and original PET images.

The chapter, describes several techniques to improve the compression efficiency of the PET-CT pairs, that were devised to address the identified issues mentioned. In particular, it is proposed a new loss term in the GAN training procedure that minimizes the residue entropy; a new approach to efficiently compute the PET residue; and an optimization algorithm that minimizes the residue mean absolute error (MAE) in order to improve the compression efficiency. Moreover, all the experiments were carried out with the enlarged and enhanced versions of the dataset.

5.1 INTER PREDICTION APPROACH FOR PET ENCODING

As previously described, the PET residue that was computed by subtracting the original PET from its estimated version. The subtraction allows to measure the pixel distance between the images, so the resulting amount of information to encode depends on the level of similarity between the images. However, this methodology presents considerable limitations that can affect the compression efficiency and therefore the overall performance of the proposed method.

One of the possible issues is the bit depth increase of the residue representation, that arises from the images subtraction. Furthermore, it may be necessary to shift the dynamic range of the image in order to only use positive values. The need of higher bit depths to represent the residue contributes to reduce the compression efficiency.

Another problem that was identified is that the GAN generator by itself does not guarantee that generated images correctly reproduce the original structure (i.e head region size and location), neither that both images are aligned with each other. Even a slight relative displacement of the image structures, (e.g. one pixel) results in higher frequency components in the residue and consequently less efficient compression.

In order to overcome the issues mentioned above, a new technique to compute a PET residual information using an inter prediction approach is proposed. In this method, instead of computing a residue to be intra coded, a sequence of two frames composed by the estimated and the original PET image is initially generated. In this approach, the PET image is encoded in inter mode as a P-frame, using the estimated PET generated by the GAN as the reference I frame for the original PET. This results in a two-frame IP sequence, where the I frame can be removed from bitstream, because it can be generated by the decoder using the same GAN as the encoder to produce the same PET estimate. The P-frame residue is coded using the motion estimation algorithms of the lossless state-of-art encoders, e.g., either HEVC or VVC. Thus, any possible spatial shift between the estimated and original PET can be compensated by the motion estimation algorithms during the encoding process.

In regard to the PET bitstream obtained from the HEVC or VVC encoder, shown in Figure 38, it is comprised of three parts: The sequence headers, containing different types of parameters, are composed by the Video Parameter Set (VPS), the Sequence Parameter Set (SPS), and the Picture Parameter Set (PPS); the I Frame information, which is composed by one slice and includes the header and the actual coded information; The P Frame information, which has an identical structure to the I Frame, although the slice data is inter-frame type rather than intra-frame.

The VPS conveys information about the layers present in the bitstream, and it is used for handling scalable coding. The SPS packs information, shared by all pictures within a coded video sequence. It carries the video sequence characteristics such as the profile, bit depth, color sampling format, width, height, minimal size of the transform block size, and minimal size of the CTU. The PPS conveys information that is shared between slices within a picture such as number of tile rows, number of tile columns, quantisation parameter, and enabling flags of coding tools. The bitstream structure, in Figure 38, shows the dashed block identifying the data that is extracted to obtain the coded information of the PET frame. As mentioned above, in this scheme the I frame is not necessary at the decoder.

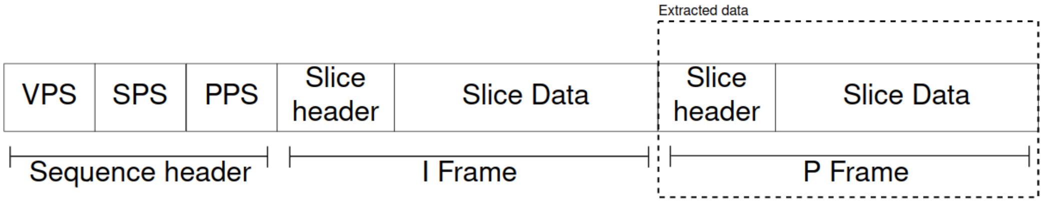


Figure 38: Bitstream Structure of a two-frame sequence and extracted data

The lossless encoder and corresponding decoder comprising the proposed cross-modality codec are represented in Figures 39 and Figure 40, respectively. The original PET-CT pair is encoded as follows. The original CT image (CT_o) is intra-coded using either HEVC or VVC, producing the corresponding compressed data, denoted in Figure 39 as $CT^I_{bitstream}$. The GAN proposed in the previous chapter is used to generate an estimated PET image (PET_e) and then, a two-frame sequence is formed with the original PET image (PET_o) and PET_e. Such sequence, denoted in Figure 39 as PET^{I+P}, is encoded as an I and P frame into bitstream $PET^{I+P}_{bitstream}$, where the coded data of the I Frame prediction is removed, generating $PET^P_{bitstream}$. This only includes sequence header and coded data of the P-frame, which is the residue of the PET_o using the PET_e as reference. The final bitstream is obtained by multiplexing $CT^I_{bitstream}$ with $PET^P_{bitstream}$.

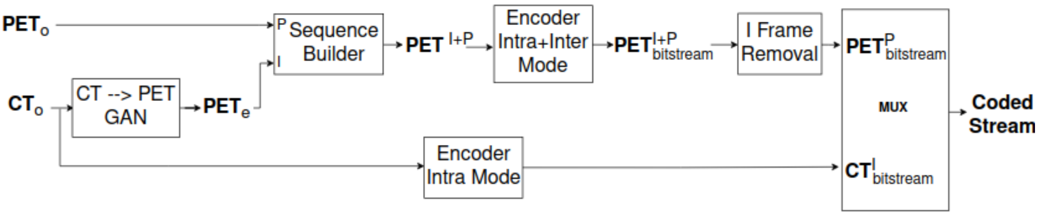


Figure 39: Cross-modality lossless encoding scheme using a inter prediction approach

The first operation performed at the decoder, shown in Figure 40, is demultiplexing the coded stream into $CT^I_{bitstream}$ and $PET^P_{bitstream}$. The CT_o is decoded first and then used to estimate the PET_e using the same GAN as the encoder. Since $PET^P_{bitstream}$ does contain the reference I-frame, it cannot be independently decoded, thus before decoding it is necessary to concatenate the coded I-frame (PET_I) used as reference in the encoder, which is the intra-coded PET_e. This allows to rebuild the $PET^{I+P}_{bitstream}$, which is decodable on its own. The output of such decoder is the two-frame sequence comprising the PET_e and PET_o frames. After demuxing, the original PET image (PET_o) is obtained and paired with the corresponding CT CT_o.

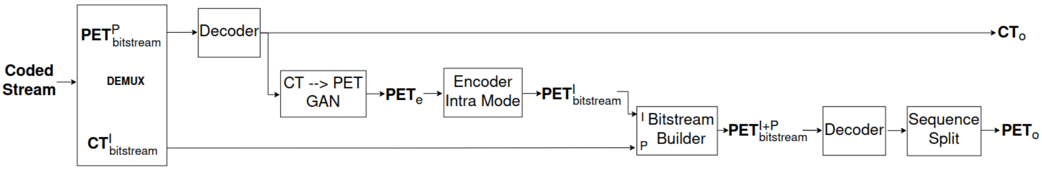


Figure 40: Cross-modality lossless decoding scheme using a inter prediction approach

As already mentioned, the same GAN model, proposed in the previous chapter, was used to perform the CT to PET domain translation. A new entropy loss function was devised, as described in the next subsection.

5.2 ENTROPY LOSS TERM

One of the key steps for efficient image compression using the previously described lossless encoders is entropy encoding. It relies on a lossless data compression algorithm that exploits the statistical redundancies of the symbols to be encode. The use of variable code length is based on the principles of information theory, which state that symbols with higher probability of occurrence can be represented with smaller codes to reduce the amount of data needed to represent the overall information [59].

During the training phase, the GAN network aims to estimate PET images as similar as possible to the original ones. The higher the quality of estimated PET images, the better for coding, due to lower residue entropy. It is known that residues with lower entropy lead to better encoding efficiency. Therefore, the new loss term devised for the generator loss function attempts to minimize the residue entropy, and it is defined as:

$$\mathcal{L}_{Entropy}(G) = -\sum_{i=1}^{M} p_i \log_2 p_i, \tag{21}$$

where p_i is the probability of occurrence of each unique value of the residue $(y - G(x))$, and M is total number of unique values. Then the new loss function of the proposed GAN framework, is defined as:

$$\begin{aligned}\mathcal{L}(D,G) = L_{\text{LSGAN}}(D) + \lambda_{\text{LSGAN}} \times L_{\text{LSGAN}}(G) + \lambda_{\text{pixel}} \times \mathcal{L}_{\text{Pixel}}(G) \\ + \mathcal{L}_{\text{Percep}}(G) + \mathcal{L}_{\text{Style}}(G) + \lambda_{\text{Entropy}} \times \mathcal{L}_{Entropy}(G),\end{aligned} \tag{22}$$

where λ_{Entropy} is the weight given to the entropy loss component. The value used for λ_{Entropy} will be further detailed. The optimization function can be defined as:

$$\min_D \min_G \mathcal{L}(D,G) \tag{23}$$

5.3 OPTIMIZATION OF ESTIMATED PET IMAGES

The optimization procedure consists in minimising the mean absolute error (MAE) between the original and estimated PET image by adjusting the contrast, scaling, and alignment of the estimated PET. Such minimisation contributes to improve the compression gain, regardless the coding approach used for the residue.

In fact, as previously described, the estimation process does not guarantee that both estimated and original PET images are properly aligned with each other, nor that the position and dimension of the head region in both images are the same. This factors negatively impact on the residue coding process by requiring encoding of more motion vectors to compensate such differences, and/or more bits to encode the residue. Another problem that may arise is related with the contrast difference between PET images, as pixel mean values from estimated and original might significantly differ. Given the negative impact of all these factors in what concerns the residue obtained, the post-processing optimization method aims at minimizing the differences, specifically in the head region area, of pixel contrast, size, and alignment to increase the compression efficiency. The Nelder–Mead simplex algorithm [60] was used to perform this task. It is a direct search method that aims to minimize an unconstrained objective function in a multidimensional space. This optimization method, neither requires gradient information nor approximate derivatives of the objective function to search for an optimal solution. In fact, it uses the function values to search for a set of points around a current one, that best minimizes the objective. Therefore, it can be a suitable method to solve nonlinear optimization problems for which the objective function is not be differentiable, or its derivatives are unknown.

The following example shows the problems described above and the results of the optimisation. Figure 41 shows an original PET image, the corresponding estimated PET and the residue computed as the subtraction of the two images. As it can be seen in the Figure 41(a), the estimated head region is larger that the original, leading to a residue that is far from optimal: since the estimated head region is larger than the original one, the residue contains a lot of information which does not favour efficient coding (Figure 41(c)).

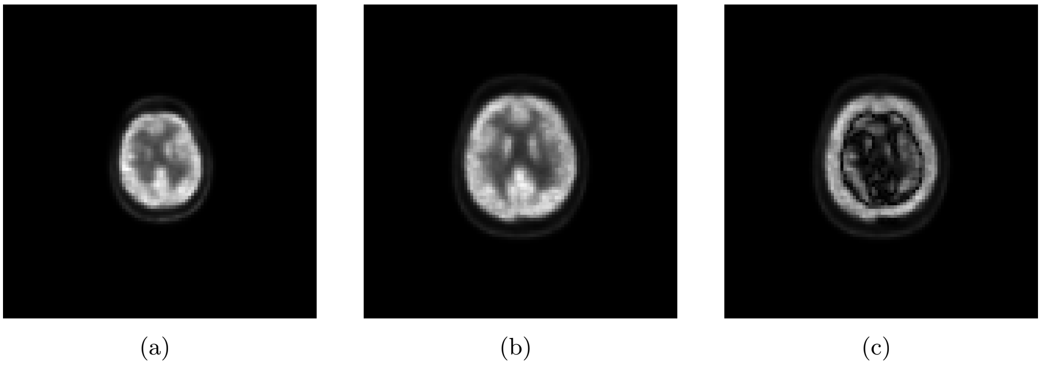

(a)
(b)
(c)

Figure 41: (a) Original PET (b) Estimated PET (c) Residue PET

After the optimization procedure the estimated PET is rescaled to better match the size of the original and the resulting residue becomes much closer to zero in all image regions (see Figures 41(c) and 42(c)). The non overlapping area between the head regions decreases, thus improving the correspondence with the original image. Although not visually perceivable in this example, the contrast and the image alignment was jointly optimized as well, improving the MAE from 11.45 (without optimization) to 2.27.

In the proposed coding frameworks, the optimization algorithm was individually applied to every estimated PET image. Given the lossless requirement, the same adjustments have to be carried out on the images estimated at the decoder side, so, in order to recover the optimized estimated PET in the decoding process, the scaling, contrast, and alignment parameters must be transmitted along with the coding stream as side information.

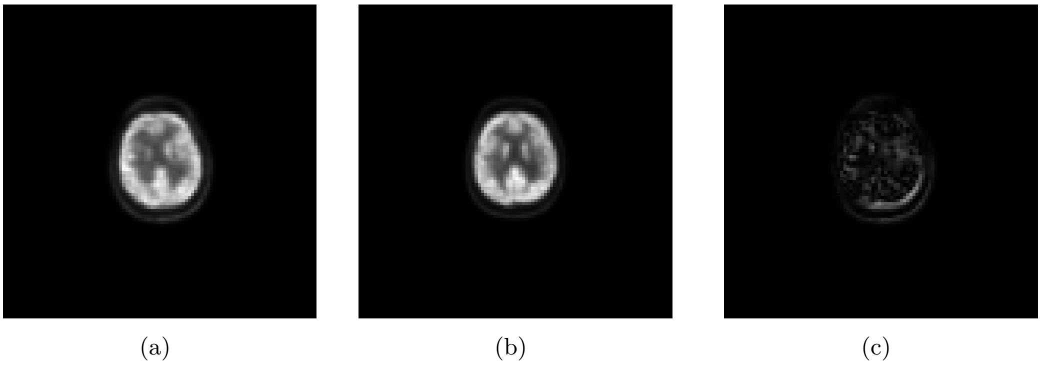


Figure 42: (a) Original PET (b) Estimated PET after optimization (c) Residue PET after optimization

5.4 RESULTS AND DISCUSSION

In order to assess the performance of the two proposed methods to obtain the residue, i.e., difference (Intra residue) and motion prediction (Inter residue), the Subset_2, Subset_3, and Subset_4 were used. The Compression Gain (CG), defined in Equation 20, was used to measure the file size difference of the residue, as a percentage relative to an anchor. For each dataset, the reference size corresponds to the size of PET_o obtained by using HEVC lossless in Intra mode. Due to the limited performance achieved in the experiments of the proof of concept, JPEG 2000 was not used in the performance comparisons of this new scheme, only HEVC and VVC.

The GAN network training process is similar to that described in Section 5.2. However, the new entropy loss term was added to the objective function and different batch sizes were also used. In order to evaluate the impact of the entropy term, six different entropy loss weights were used: $\lambda_{Entropy} = 0$, $2, 4, 6, 8$, and 10. The batch size is one of the most important hyperparameter to tune when training GANs, so six different sizes of 2, 4, 6, 8, 10, and 20 were also considered.

For the Subset_4, a slightly modified network was tested and used, to account for the differences on CT images' size. In fact, the new architecture is basically the one proposed in 4.2, with an additional convolutional layer that is responsible for downsampling CT images so that their size matches that of the PET images.

To evaluate the performance of the proposed refinement methods, four different scenarios were considered for comparison and discussion.

1. Independent intra-coding of the CT-PET pair: both original PET and its corresponding CT image independently encoded using lossless HEVC and VVC.

2. Independent intra-coding of the CT and PET residue: The scheme described in Section 4.1 was used with the PET residue and the original CT independently encoded using lossless HEVC and VVC. The PET residue is computed between the original PET and its $CT \rightarrow PET$ estimate generate by the GAN.

3. Independent intra coding of the CT and inter-coding of the PET as a P-frame: HEVC and VVC were used to encode both the original CT and the PET, which was inter-coded as a P-frame (see Section 5.1).

4. The two previous schemes, but using the estimated PET obtained with the optimization method proposed in the Section 5.3.

The average compression results of the PET_o and CT_o images within each dataset (scenario 1) are summarized in Table 5 and Table 6 respectively, where the results obtained with HEVC were used as an anchor. As expected, VVC outperforms HEVC in all tested datasets, achieving compression gains of 1.02% in Subset_2 and Subset_4 , and 0.96% in Subset_3, for PET_o. For CT_o, a compression gain of 0.13% was achieved. The general better performance of VVC in comparison with HEVC is also confirmed for the special case of CT and PET image pairs. The lower gains of CT in comparison with PET is most likely due the higher level of detail (i.e., high spatial frequencies) originated from the various structures (e.g bones) contained in these type of images. The final compressed size of the CT files is considerably larger than the PET files (up to 8-9 times), not only due to these lower compression gains, but also due to the fact that the original CT images have larger dimensions (width, height) than PET images.

Table 5: PET individual lossless image compression results

Files	**Subset**	**Codec**	**Size (bytes)**	**CG (%)**
PET_o	Subset_2	HEVC Intra	1200399	-
PET_o	Subset_2	VVC Intra	1188154	1.02
PET_o	Subset_3	HEVC Intra	1188536	-
PET_o	Subset_3	VVC Intra	1177110	0.96
PET_o	Subset_4	HEVC Intra	1200399	-
PET_o	Subset_4	VVC Intra	1188154	1.02

Table 6: CT individual lossless image compression results

Files	Subset	Codec	Size (bytes)	CG (%)
CT_o	-	HEVC Intra	10831135	-
CT_o	-	VVC Intra	10816820	0.13

The compression results for the joint PET-CT pair with no optimizations on the PET image are summarized in Table 7. The results consider the size of the independent encoding of the original CT is and the PET using the described strategies. The results for the residual PET compression using the first proposed scheme (scenario 2) are identified as 'Intra residues' in the table, since they are independently encoded. The new proposed scheme (scenario 3), obtained by encoding the PET as a P-frame is identified as 'Inter P-frame'.

From the obtained results, it can be observed that, in this set of experiments, the strategy based on coding the 'Intra residues' does not lead to compression gains. In fact, the best results for this method were obtained by compressing the residue with VVC, achieving -0.81%, -0.92%, and -0.79% for the Subset_2, Subset_3, and Subset_4 respectively. Unlike the results presented in Section 4.3, the compression scheme performed worst than simply compressing the original images with HEVC. This happens so because a much larger number of images was used (about 6 times more in total, as shown in Table 1) and the GAN, which showed signs of overfitting during the training, was no longer able to learn the diversity of the data set and generate high quality images. The training and validation curve of the GAN loss per epoch (blue and orange curve, respectively) obtained with the Subset_2, a $\lambda_{Entropy}$ and a batch size of 4, can be seen in the Figure 43. The analysis of the figure indicates that the training loss decreases at a much higher rate than the validation. Which implies that the network is probably memorising how to estimate the images in the training set to decrease the loss value. Such factor does not allow the network to converge to a more general solution, which negatively affects the capability to represent a larger diversity of images with better quality.

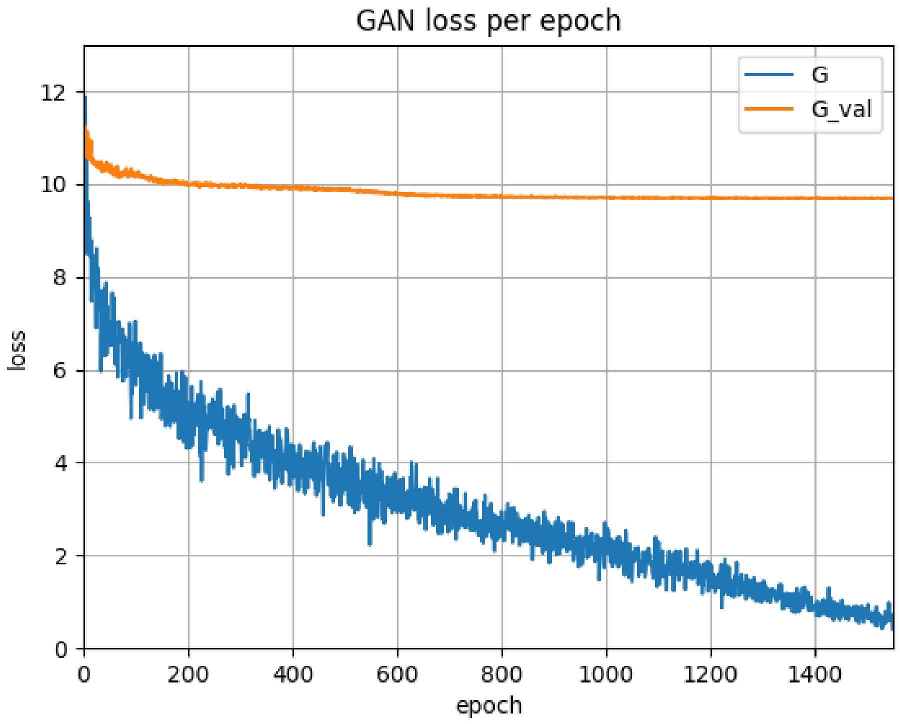


Figure 43: GAN loss per epoch when trained with Subset_2, $\lambda_{entropy} = 4$, and a batch size of 4.

With a compression scheme based on coding the residues, low quality estimates compromise the performance of the algorithm leading to losses rather than gains. The scheme proposed in Section 5.1 that computes the PET residue using the encoder inter prediction techniques, i.e., as a P-frame, outperforms the scheme proposed in Section 4.1, when jointly encoding the PET image (P-frame) and the respective CT using VVC, obtaining a maximum compression gain of 0.87%, 0.88%, and 0.88% for Subset_2, Subset_3, and Subset_4 respectively, as shown in Table 8. Overall, this approach outperforms the anchor strategy using independent intra-coding with HEVC, and also the scheme proposed in Section 4.1.

Table 7: PET-CT lossless image compression results before optimization

Subset	Batch	Entropy	HEVC Intra residue CG(%)	VVC Intra residue CG(%)	HEVC Inter P-frame CG(%)	VVC Inter P-frame CG(%)
Subset_2	4	0	-1.04	-0.81	0.58	0.86
Subset_2	4	2	-0.89	-0.65	0.58	0.87
Subset_2	4	4	-0.89	-0.65	0.58	**0.87**
Subset_2	4	6	-0.90	-0.66	0.58	0.87
Subset_2	4	8	-0.87	-0.63	0.58	0.87
Subset_2	4	10	-0.90	-0.66	0.58	0.87
Subset_2	8	0	-1.12	-0.88	0.57	0.87
Subset_2	8	4	**-0.81**	**-0.57**	**0.59**	0.87
Subset_2	20	4	-1.30	-1.07	0.56	0.86
Subset_3	4	0	-1.32	-1.08	0.57	0.87
Subset_3	4	4	-0.97	-0.72	0.58	**0.88**
Subset_3	4	10	**-0.92**	**-0.66**	**0.59**	0.88
Subset_3	8	0	-1.34	-1.10	0.57	0.87
Subset_3	8	4	-1.02	-0.76	0.58	0.88
Subset_3	20	4	-1.18	-0.93	0.58	0.88
Subset_4	4	0	-1.04	-0.79	0.58	0.88
Subset_4	4	2	**-0.79**	**-0.53**	0.59	0.88
Subset_4	4	4	-0.87	-0.62	0.59	0.88
Subset_4	4	6	-0.89	-0.64	0.59	0.88
Subset_4	4	8	-0.85	-0.60	0.59	0.88
Subset_4	4	10	-0.84	-0.58	**0.59**	**0.88**
Subset_4	8	0	-1.11	-0.87	0.58	0.87
Subset_4	8	4	-0.85	-0.59	0.59	0.88
Subset_4	20	4	-1.76	-1.51	0.53	0.86

Since the CT images have significantly larger resolutions than the PET, the results presented in Table 7 can somehow be masked in regard to the performance of the proposed PET compression scheme. In order to verify the actual impact of the proposed framework it is therefore important to separately analyse the compression results for the PET image. The relative compression performance of both cross-modality compression schemes, using only the size of the PET images as reference and without the optimization method are summarized in Table 8.

Table 8: PET lossless image compression results before optimization

Subset	**Batch**	**Entropy**	**HEVC Intra residue CG(%)**	**VVC Intra residue CG(%)**	**HEVC Inter P-frame CG(%)**	**VVC Inter P-frame CG(%)**
Subset_2	4	0	-10.52	-9.39	5.84	7.53
Subset_2	4	2	-8.99	-7.75	5.86	7.59
Subset_2	4	4	-9.03	-7.79	5.91	**7.64**
Subset_2	4	6	-9.14	-7.92	5.87	7.58
Subset_2	4	8	-8.77	-7.54	5.91	7.61
Subset_2	4	10	-9.07	-7.88	5.90	7.61
Subset_2	8	0	-11.31	-10.14	5.76	7.49
Subset_2	8	4	**-8.21**	**-6.98**	**5.95**	7.61
Subset_2	20	4	-13.17	-11.98	5.62	7.45
Subset_3	4	0	-13.19	-12.04	5.76	7.53
Subset_3	4	4	-9.73	-8.38	5.84	**7.62**
Subset_3	4	10	**-9.19**	**-7.84**	**5.89**	7.61
Subset_3	8	0	-13.44	-12.22	5.73	7.52
Subset_3	8	4	-10.23	-8.84	5.81	7.59
Subset_3	20	4	-11.78	-10.49	5.83	7.59
Subset_4	4	0	-10.40	-9.13	5.85	7.59
Subset_4	4	2	**-7.89**	**-6.55**	5.92	7.66
Subset_4	4	4	-8.70	-7.40	5.92	7.66
Subset_4	4	6	-8.90	-7.56	5.89	7.66
Subset_4	4	8	-8.56	-7.23	5.94	7.67
Subset_4	4	10	-8.39	-7.03	**5.95**	**7.67**
Subset_4	8	0	-11.16	-9.92	5.80	7.59
Subset_4	8	4	-8.51	-7.13	5.95	7.64
Subset_4	20	4	-17.60	-16.29	5.36	7.43

From Table 8, it can be observed that the 'Intra residue' scheme underperforms. In fact, it does not lead to any compression gain, with the best result of -8.21%, -9.19%, and -7.89% for Subset_2, Subset_3, and Subset_4 respectively. The 'Inter P-frame' scheme proposed in Section 5.1 achieves the best results using VVC, with a maximum compression gain of 7.64%, 7.62%, and 7.67% for Subset_2, Subset_3, and Subset_4 respectively, as shown in Table 8. It outperforms the conventional strategies that use either HEVC or VVC to individually encode the original PET (up to 7 times). It is possible to conclude that using the standard motion tools to predict the image differences and consequently the residue is a more efficient approach. The tests performed with the larger batch size on each dataset present the worst overall results. The success of using smaller batch sizes has been observed by Ulyanov et al. [61] to improve the results quality of I2I tasks. From these last results, it is also possible to verify that for the same subset and batch size, training with a $\lambda_{Entropy} = 0$ presents the worst compression results for both intra and inter schemes. Using an entropy weight term equal to zero is the same as using the network proposed in Section 4.2, which did not minimize the residue entropy. Therefore it can also be conclude that adding a term that minimizes the residue entropy allows better compression results within the same batch size used.

The performance of the post-processing method described in the Section 5.3, which uses the optimized estimated PET image to compute the compressed residue, was also evaluated. The compression results obtained for the PET-CT pair and for the individual PET compression are summarized in Table 9 and Table 10, respectively.

Table 9: PET-CT lossless image compression results after optimization

Subset	Batch	Entropy	HEVC Intra residue CG(%)	VVC Intra residue CG(%)	HEVC Inter P-frame CG(%)	VVC Inter P-frame CG(%)
Subset_2	4	2	-0.37	-0.11	0.61	0.88
Subset_2	4	4	-0.33	-0.07	0.62	0.88
Subset_2	4	6	-0.36	-0.09	0.61	0.88
Subset_2	4	8	-0.34	-0.08	0.62	0.88
Subset_2	4	10	-0.36	-0.10	0.62	**0.88**
Subset_2	8	4	**-0.31**	**-0.05**	**0.62**	0.88
Subset_2	20	4	-0.45	-0.19	0.60	0.86
Subset_3	4	4	-0.42	-0.15	0.62	0.89
Subset_3	4	10	**-0.37**	**-0.10**	**0.63**	**0.89**
Subset_3	8	4	-0.57	-0.13	0.61	0.88
Subset_3	20	4	-0.39	-0.13	0.62	0.89
Subset_4	4	2	**-0.29**	**-0.03**	**0.63**	0.89
Subset_4	4	4	-0.33	-0.06	0.63	0.89
Subset_4	4	6	-0.37	-0.11	0.63	0.89
Subset_4	4	8	-0.34	-0.07	0.63	0.90
Subset_4	4	10	-0.32	-0.06	0.63	**0.90**
Subset_4	8	4	-0.32	-0.05	0.62	0.89
Subset_4	20	4	-0.74	-0.47	0.57	0.87

Table 10: PET lossless image compression results after optimization

Subset	Batch	Entropy	HEVC Intra residue CG(%)	VVC Intra residue CG(%)	HEVC Inter P-frame CG(%)	VVC Inter P-frame CG(%)
Subset_2	4	2	-3.76	-2.31	6.22	7.68
Subset_2	4	4	-3.37	-1.94	6.28	7.71
Subset_2	4	6	-3.60	-2.16	6.22	7.68
Subset_2	4	8	-3.47	-2.05	6.23	7.68
Subset_2	4	10	-3.59	-2.20	6.25	**7.70**
Subset_2	8	4	**-3.13**	**-1.69**	**6.32**	**7.70**
Subset_2	20	4	-4.59	-3.16	6.20	7.53
Subset_3	4	4	-4.17	-2.71	6.23	7.72
Subset_3	4	10	**-3.68**	**-2.21**	**6.27**	**7.73**
Subset_3	8	4	-5.73	-2.46	6.09	7.67
Subset_3	20	4	-3.87	-2.45	6.25	7.72
Subset_4	4	2	**-2.94**	**-1.44**	**6.33**	7.76
Subset_4	4	4	-3.28	-1.81	6.29	7.76
Subset_4	4	6	-3.76	-2.29	6.28	7.77
Subset_4	4	8	-3.37	-1.92	6.30	**7.78**
Subset_4	4	10	-3.20	-1.77	6.31	**7.78**
Subset_4	8	4	-3.19	-1.70	6.25	7.72
Subset_4	20	4	-7.43	-5.88	5.72	7.53

The results in Table 9 and Table 10 indicate that the optimization of scaling, contrast, and location of the head region within the PET image, has a positive impact on the quality of the estimated PET image used for prediction, which improves the compression gain results. As expected, the results obtained from the 'Intra residue' compression improve significantly after the optimization process, since the PET_e is transformed with parameters that were optimized to minimize the mean absolute error of the residue. However, even with this optimization, the 'Intra residue' approach is not able to present gains, achieving CG of -0.05%, -0.1%, and -0.03% for the Subset_2, Subset_3, and Subset_4 respectively, for the PET-CT pair size. The compression gain for the 'Intra residue' with the individual PET size was -3.14%, -3.68%, and -2.94% for the Subset_2, Subset_3, and Subset_4 respectively, as shown in Table 9.

The 'Inter P-frame' with VVC using the optimized PET_e, obtained by training the GAN with a batch size of 4 and an entropy weight of 10, achieves the best performance in all the datasets. When the sizes os both the PET and the CT images are considered, experiments with Subset_2 shown a maximum compression gain of 0.88%, while Subset_3 shows a gain of 0.89% , and Subset_4 a CG of 0.90%, as observed in Table 9. When considering only the sizes of the PET images, a maximum compression gain of 7.70% as obtained using Subset_2, 7.73% with Subset_3, and 7.78% with Subset_4, as demonstrated in Table 10.

One of the benefits of using the Subset_4 is that the network learns to downsample the CT image using a convolutional layer instead of downsampling before training. In this manner, the network is capable of adjusting the downsample process of the the PET_o to improve the quality of the estimated PET image. Again, it is clear that smaller batch sizes, rather than larger, perform better in this image translation application.

5.5 SUMMARY

In this chapter several contributions to enhance the cross modality lossless compression of PET-CT paired images were proposed. First, a new loss term to minimize the entropy of the PET residue during the GAN training process was established. Then, a new pipeline that modifies the method to encode a PET residual a standard P-frame is described. In this case, the estimated and original are encoded as a two-frame sequence, benefiting from the standard motion estimation to achieve better inter-frame prediction. One of the key advantages of this approach when compared to the initial one is that the residual information of the P-frame is directly obtained by the encoder, which is optimised by the standard encoding tools and algorithm. A method based on a simplex optimization algorithm was also proposed to minimize the differences between the original and estimated PET image by adjusting the scaling, contrast and alignment of the estimated PET. A new GAN framework was proposed to test the Subset_4, by adding a convolutional layer to the beginning of the network in order to downsample the input CT image to the resolution of the corresponded PET image.

The method that relies on simple subtraction to compute the residue demonstrated to be inefficient, once it did not achieve any compression gains. On the other hand, the obtained results show that the proposed pipeline using the P-frame encoding and estimated PET image optimization achieved the best compression results with a compression gain of 0.9% for the PET-CT pairs and 7.78% for the PET images. This results were obtained using new GAN framework proposed to test the Subset_4.

6 CONCLUSIONS AND FUTURE WORK

6.1 CONCLUSION

The main objective of this work, was to explore and develop efficient lossless multi modal image compression methods with a deep learning approach. Medical image processing has become an alluring research topic over the last years. The development of new sophisticated medical imaging acquisition systems led to an abrupt increase in the amount of data generation, with several Terabytes of medical images from multiple modalities being generated daily worldwide. The proposed coding strategy demonstrated to be able to improve the compression efficiency of PET-CT pairs when compared to other traditional schemes. To accomplish such goal, four different versions of paired PET-CT datasets were used, and different deep learning architectures, based on generative adversarial algorithms, were proposed for I2I applications.

In this work, a GAN with an embedded attention mechanism to translate the CT to PET domain was developed and used to produce estimated PET images. These, in turn, are used to compute a residual representation with lower entropy than the original images. Two different methods were proposed to compute the PET residue. In the first one, the residue is obtained by subtracting the original PET with the estimated PET image; in the second, standard inter prediction coding tools of P-frames was used. The residue is then compressed along with the original CT image. In order to enhance the quality of the compressed PET residue, a method that optimizes the scaling, contrast, and location parameters of the region of interest in the estimated PET was also implemented to minimise the entropy of the residue. The four different versions of the original dataset were used were evaluate the performance of the proposed method .

As a proof on concept, the first version of the dataset (Subset_1) and the residue based on the subtraction of original and estimated PET iamges were used. It was shown that this pipeline was capable of achieving a compression gain of 8.9% for the PET representation and 3.53% for the PET-CT pair, relatively to the conventional compression schemes using standard losses encoders alone. The main limitation of

Subset_1 is the reduced number of images and also the resampling of both PET and CT images. Although the coding approach proposed in this work was capable of surpassing the traditional schemes in terms of compression ratio, the limitations of the subset (both in number and diversity) means that the conclusion could somehow lack generality.

To improve the robustness of the proposed method, and to demonstrate that it can be applied in more realistic scenarios, several modifications were implemented and evaluated. These consisted on a technique of inter prediction to compute a residual representation, a new non-adversarial loss that minimizes the residual entropy loss, and an optimization procedure that scales and modifies the contrast of the estimated image. To assess the compression efficiency three different subsets (Subset_2,Subset_3,Subset_4) with more PET-CT pairs were used. The Subset_2 and Subset_3 used a downsampled version of the CT image to train the network. With Subset_4, a new GAN framework was considered. This framework is adapted to learn how to downsample the CT towards the same resolution of the PET using an additional convolutional layer at the beginning of the network. It was shown that the proposed cross modality compression scheme using the inter residue approach with the optimized estimated PET achieved the best performance in all versions of dataset, when the GAN is trained with a batch size of 4 and an entropy term of 10. A maximum compression gain for the Subset_2, Subset_3, and Subset_4 were 0.88%, 0.89%, and 0.90% respectively for the PET-CT pair, and 7.70%, 7.73%, and 7.78% for the individual PET representation.

Using the encoder to compute the residual representation is more effective than computing the residue simply using the frame difference. The use of lower batches and a higher entropy term to train the GAN network allows to obtain estimated PET images with higher accuracy and therefore it improves the overall compression efficiency. It was also shown, that the optimisation process applied to PET estimates is capable of reducing the entropy of the residual quality and consequently improving the compression efficiency.

Overall, one may conclude that the use of the proposed compression framework, that targets the joint prediction of the PET-CT pairs, resulted in a considerable bit-rate reduction. Such achievement, can be beneficial in the scientific-hospital environment and translate into storage and transmission savings of medical image information.

6.2 FUTURE WORK

In future work, an AE or a GAN framework can be utilized to further enhance the estimated PET quality. To improve the generative capability and reduce the possible overfit, changes to the GAN responsible to translate the CT to PET can be explored. These modifications consist on the decrease of the number of parameters in the generator, so that the model do not become too deep / complex, and implement new architectures such as the CasNet generator described in the Chapter 3. Also, techniques such as data augmentation and weight regularization can further be implemented and tested. When the PET image was used to estimate a CT, the results in terms of compression efficiency of the residual CT were not satisfactory, however with the improvement of the GAN network, this method should be reassessed.

Additionally, unsupervised approaches to remove the dependencies on paired datasets might be analysed. Finally, an interesting application would be to apply and test the developed method with other paired modalities, such as PET-MRI.

6.3 CONTRIBUTIONS

Contributions that resulted from the research work done during this dissertation.

- J. Parracho, L. Thomaz, L. Távora, S. Faria, and P. Assunção, “Cross-modality lossless compression of PET-CT images,”Proceedings of Conf. on Telecommunications - ConfTele, Leiria, Portugal, February 2021
- F. Cunha, S. Faria, J. Parracho, L. Thomaz, and P. Assunção, “Data compression algorithms forbiomedical images,” inImaging Modalities for Biological and Preclinical Research: A Compendium, Volume 2, 2053-2563,pp. III.4.f–1 to III.4.f–11. 2021

BIBLIOGRAPHY

[1] M. Ansari and R. Anand. “Recent Trends in Image Compression and its Application in Telemedicine and Teleconsultation”. In: *National Systems Conference*. Roorkee, India, Dec. 2008, pp. 59–64.

[2] D. Delbeke et al. “Hybrid imaging (SPECT/CT and PET/CT): improving therapeutic decisions”. In: *Seminars in nuclear medicine* 39 (Oct. 2009), pp. 308–40.

[3] A. Habibian et al. “Video Compression With Rate-Distortion Autoencoders”. In: *IEEE/CVF International Conference on Computer Vision*. Seoul, South Korea, Nov. 2019, pp. 7032–7041.

[4] L. Zhu et al. “Generative Adversarial Network-Based Intra Prediction for Video Coding”. In: *IEEE Transactions on Multimedia* 22.1 (Jan. 2020), pp. 45–58.

[5] C. Jia et al. “Light Field Image Compression Using Generative Adversarial Network-Based View Synthesis”. In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9.1 (Mar. 2019), pp. 177–189.

[6] R. Li et al. “Simplified unsupervised image translation for semantic segmentation adaptation”. In: *Pattern Recognition* 105 (Sept. 2020), p. 107343.

[7] Y. Zhang et al. “Multiple Cycle-in-Cycle Generative Adversarial Networks for Unsupervised Image Super-Resolution”. In: *IEEE Transactions on Image Processing* 29 (Sept. 2020), pp. 1101–1112.

[8] T. Park et al. “Semantic Image Synthesis with Spatially-Adaptive Normalization”. In: *CoRR* abs/1903.07291 (Dec. 2019).

[9] S. Gorti and J. Ma. “Text-to-Image-to-Text Translation using Cycle Consistent Adversarial Networks”. In: *CoRR* abs/1808.04538 (Aug. 2018).

[10] L. Lerman, M. Rodriguez-Porcel, and J. Romero. “The development of x-ray imaging to study renal function”. In: *Kidney International* 55.2 (1999), pp. 400–416.

[11] T. Buzug. “Computed Tomography”. In: *Springer Handbook of Medical Technology*. Ed. by R. Kramme, K. Hoffmann, and R. Pozos. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 311–342.

[12] R. Pelberg. “Basic Principles in Computed Tomography (CT)”. In: *Cardiac CT Angiography Manual*. London: Springer London, 2015, pp. 19–58.

[13] Glenn Frederick Knoll. In: *Radiation Detection and Measurement*. New Jersey, USA, Aug. 2010, p. 864.

[14] S. Webb. In: *The Physics of Medical Imaging*. Florida, USA, Jan. 1988, p. 652.

[15] D. Bailey, D. Townsend, and P. Valk. “Positon emission tomography. Basic sciences”. In: *Journal of Neuroradiology - J NEURORADIOL* 33 (Oct. 2006), pp. 265–265.

[16] D. Brooks. “PET: its clinical role in neurology.” In: *Journal of Neurology, Neurosurgery & Psychiatry* 54.1 (1991), pp. 1–5.

[17] L. Lezzoni, O. Grad, and M. Moskowitz. “Magnetic Resonance Imaging:Overview of the Technology and Medical Applications”. In: *International Journal of Technology Assessment in Health Care* 1.3 (1985), pp. 481–498.

[18] J. Hornak. *The Basics of NMR*. Jan. 1997. URL: https://www.cis.rit.edu/htbooks/nmr/inside.htm.

[19] K. Clark et al. “The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository”. In: *Journal of digital imaging* 26 (July 2013), pp. 1045–1057.

[20] K. Armanious et al. “Unsupervised Medical Image Translation Using Cycle-MedGAN”. In: *European Signal Processing Conference*. A Coruna, Spain, Sept. 2019, pp. 1–5.

[21] M. Adams and R. Ward. “Wavelet transforms in the JPEG-2000 standard”. In: *2001 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (IEEE Cat. No.01CH37233)*. Vol. 1. Victoria, BC, Canada, Aug. 2001, 160–163 vol.1.

[22] G. Sullivan et al. “Overview of the High Efficiency Video Coding (HEVC) Standard”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 22.12 (2012), pp. 1649–1668.

[23] H. Samet. “The Quadtree and Related Hierarchical Data Structures”. In: *ACM Comput. Surv.* 16.2 (June 1984), pp. 187–260.

[24] D. Flynn et al. “Overview of the Range Extensions for the HEVC Standard: Tools, Profiles and Performance”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 26.1 (2016), pp. 4–19.

[25] B. Bross et al. “Overview of the Versatile Video Coding (VVC) Standard and its Applications”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2021), pp. 1–1.

[26] E. Alshina et al. *JVET-D1001: Algorithm Description of Joint Exploration Test Model 4*. June 2018.

[27] H. Gao et al. “Geometric Partitioning Mode in Versatile Video Coding: Algorithm Review and Analysis”. In: *IEEE Transactions on Circuits and Systems for Video Technology* PP (Nov. 2020).

[28] L. Gatys, A. Ecker, and M. Bethge. “A Neural Algorithm of Artistic Style”. In: *CoRR* abs/1508.06576 (Aug. 2015).

[29] A. Ajit, K. Acharya, and A. Samanta. “A Review of Convolutional Neural Networks”. In: *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. Vellore, India, Feb. 2020, pp. 1–5.

[30] G. Hinton and R. Salakhutdinov. “Reducing the dimensionality of data with neural network”. In: *Foundations and Trends® in Machine Learning* 313 (Aug. 2006).

[31] L. Gondara. “Medical Image Denoising Using Convolutional Denoising Autoencoders”. In: *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. Barcelona, Spain, Feb. 2016, pp. 241–246.

[32] N. Merrill and C. Olson. “A New Autoencoder Training Paradigm for Unsupervised Hyperspectral Anomaly Detection”. In: *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*. Waikoloa, HI, USA, Feb. 2020, pp. 3967–3970.

[33] N. Shvetsova et al. “Anomaly Detection in Medical Imaging With Deep Perceptual Autoencoders”. In: *IEEE Access* 9 (Aug. 2021), pp. 118571–118583.

[34] D. Kingma and M. Welling. “An Introduction to Variational Autoencoders”. In: *Science (New York, N.Y.)* 12 (Aug. 2019).

[35] H. Choi et al. “Predicting Aging of Brain Metabolic Topography Using Variational Autoencoder”. In: *Frontiers in Aging Neuroscience* 10 (July 2018), p. 212.

[36] I. Goodfellow et al. “Generative Adversarial Networks”. In: *CoRR* abs/1406.2661 (June 2014).

[37] H. Zhang et al. “StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks”. In: *IEEE International Conference on Computer Vision*. Venice, Italy, Oct. 2017, pp. 5908–5916.

[38] M. Bosch, C. M. Gifford, and P. Rodriguez. “Super-Resolution for Overhead Imagery Using DenseNets and Adversarial Learning”. In: *IEEE Winter Conference on Applications of Computer Vision*. Lake Tahoe, NV, USA, Mar. 2018, pp. 1414–1422.

[39] P. Isola et al. “Image-to-Image Translation with Conditional Adversarial Networks”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, United States, July 2017, pp. 5967–5976.

[40] M. Mirza and S. Osindero. “Conditional Generative Adversarial Nets”. In: *CoRR* abs/1411.1784 (Nov. 2014).

[41] L. Deng. “The mnist database of handwritten digit images for machine learning research”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[42] X. Chen et al. “InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets”. In: *CoRR* abs/1606.03657 (June 2016).

[43] A. Donahue and K. Simonyan. “Large Scale GAN Training for High Fidelity Natural Image Synthesis”. In: *CoRR* abs/1809.11096 (Feb. 2019).

[44] A. Brock et al. “Neural Photo Editing with Introspective Adversarial Networks”. In: *CoRR* (Sept. 2016).

[45] H. Zhang et al. “Self-Attention Generative Adversarial Networks”. In: *CoRR* (May 2018).

[46] J. Zhu et al. “Toward Multimodal Image-to-Image Translation”. In: *CoRR* abs/1711.11586 (Oct. 2018).

[47] A. Larsen, S. Sønderby, and O.Winther. “Autoencoding beyond pixels using a learned similarity metric”. In: *CoRR* abs/1512.09300 (Feb. 2016).

[48] K. Armanious et al. “MedGAN: Medical image translation using GANs”. In: *Computerized Medical Imaging and Graphics* 79 (Jan. 2020).

[49] O. Ronneberger, P. Fischer, and T. Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing, Nov. 2015, pp. 234–241.

[50] J.-Y. Zhu et al. “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”. In: *CoRR* abs/1703.10593 (Mar. 2020).

[51] H. Tang et al. “Attention-Guided Generative Adversarial Networks for Unsupervised Image-to-Image Translation”. In: *CoRR* (Mar. 2019).

[52] X. Mao et al. “Least Squares Generative Adversarial Networks”. In: *IEEE International Conference on Computer Vision.* Venice, Italy, Oct. 2017, pp. 2813–2821.

[53] I. Goodfellow et al. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems.* Vol. 27. Montreal, Canada, Dec. 2014, pp. 2672–2680.

[54] D. Pathak et al. “Context Encoders: Feature Learning by Inpainting”. In: *Conference on Computer Vision and Pattern Recognition.* Las Vegas, NV, USA, July 2016, pp. 2536–2544.

[55] R. Zhang, P. Isola, and A. Efros. “Colorful Image Colorization”. In: *CoRR* abs/1603.08511 (Mar. 2016).

[56] J. Johnson, A. Alahi, and L. Fei-Fei. “Perceptual Losses for Real-Time Style Transfer and Super-Resolution”. In: *CoRR* abs/1603.08155 (Mar. 2016).

[57] K. He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Las Vegas, NV, USA, Dec. 2016, pp. 770–778.

[58] D. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. In: *International Conference on Learning Representations.* San Diego, CA, USA, Dec. 2014, pp. 1–13.

[59] E. Shannon. “A mathematical theory of communication”. In: *The Bell System Technical Journal* 27.3 (July 1948), pp. 379–423.

[60] J. Nelder and R. Mead. “A Simplex Method for Function Minimization”. In: *The Computer Journal* 7.4 (Jan. 1965), pp. 308–313.

[61] D. Ulyanov, A. Vedaldi, and V. Lempitsky. “Instance Normalization: The Missing Ingredient for Fast Stylization”. In: *CoRR* abs/1607.08022 (2016).

A

APPENDIX A

Table 11: Identifiers of the PET and CT images used in train set

CT and PET Series Instance UID	Instance number
CT 1.3.6.1.4.1.14519.5.2.1.5099.8010.134132260838950063823639580336 PET 1.3.6.1.4.1.14519.5.2.1.5099.8010.261283266967084478597628473082	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 4; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5099.8010.143537028637640199671884371790 PET 1.3.6.1.4.1.14519.5.2.1.5099.8010.168349987958825242093751863260	17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 53; 54; 55; 56; 57; 58; 59; 60; 61; 62; 63; 64; 65; 66; 67; 68; 69;
	1; 10; 11; 12;

CT 1.3.6.1.4.1.14519.5.2.1.5099.8010.246547274898446333047174112338 PET 1.3.6.1.4.1.14519.5.2.1.5099.8010.207287176598217759947210775301	13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 39; 4; 40; 41; 42; 43; 44; 45; 46; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.111080330119580197908036270989 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.336385428185872585243261838275	10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.199279685651120733509991507733	10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 5; 50; 51; 52;

PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.101091674262507926824569594631	53; 54; 55; 56; 57; 58; 59; 6; 60; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.224476197037863218191030555592 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.335007062870506708630997318969	10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 53; 54; 55; 56; 57; 58; 59;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.230115363317015597961189690066 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.256623848656901547757777237402	10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 53; 54; 55; 56; 57; 58; 59; 60; 61; 62; 63; 64; 65; 9;
	10; 11; 12; 13;

CT
1.3.6.1.4.1.14519.5.2.1.5168.2407.246764360511507529763370532127
PET
1.3.6.1.4.1.14519.5.2.1.5168.2407.167927586521771186022458043143
	14; 15; 16; 17;
	18; 19; 20; 21;
	22; 23; 24; 26;
	36; 37; 38; 39;
	40; 41; 42; 43;
	44; 45; 46; 47;
	48; 49; 50; 51;
	52; 53; 54; 55;
	56; 57; 58; 9;

CT
1.3.6.1.4.1.14519.5.2.1.5168.2407.514293492211796146298241921357
PET
1.3.6.1.4.1.14519.5.2.1.5168.2407.144491913839196102242400890591
	1; 10; 11; 12;
	13; 14; 15; 16;
	2; 27; 28; 29;
	3; 30; 31; 32;
	33; 34; 35; 36;
	37; 38; 39; 4;
	40; 41; 42; 43;
	5; 6; 7; 8;
	9;

CT
1.3.6.1.4.1.14519.5.2.1.5168.2407.178106158980683756379992731272
PET
1.3.6.1.4.1.14519.5.2.1.5168.2407.968450461974998678527699997735
	1; 10; 11; 15;
	2; 21; 22; 23;
	24; 25; 26; 27;
	28; 29; 3; 30;
	31; 32; 33; 34;
	35; 36; 37; 38;
	39; 4; 40; 41;
	42; 43; 44; 45;
	46; 5; 6; 7;
	8; 9;

CT
1.3.6.1.4.1.14519.5.2.1.5168.2407.304892201745070613575318903971
	1; 10; 11; 12;
	13; 14; 15; 16;
	17; 18; 19; 2;
	20; 21; 22; 23;

PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.383318066432803110253519682412	24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 43; 44; 45; 46; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.163889340882526392690974310923 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.296684579355446368740026016736	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 3; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.248797857562521148666306246157 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.159878820906048161823652086012	1; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 5; 6; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.208592381514313913642986136312	10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 37; 38; 39; 40;

PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.679106706031849961074383957791	41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 53; 54; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.278898958559212204650436127371 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.255652014676197197669260945497	10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.321824893604802280707231364791 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.335902376670341157715251162154	1; 10; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 43; 44; 45; 46; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.182228440556455604859009598786 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.286977723110941955029144312258	1; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 43;

	44; 45; 5;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.930675219543891029354497598714 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.125623100970758475890560903622	11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 53; 54; 55;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.154477402534711761978366056287 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.178850337197246272356018881360	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.879512647778996849137351676390	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41;

PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.667473488192231259943928079329	42; 43; 44; 45; 46; 47; 48; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.188640272842614251817546405159 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.399286983392251844694693386787	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.281461477294968392267739949022 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.166240763151202446565669536236	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 43; 44; 45; 46; 47; 48; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.253456014800842387819662988768	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23;

PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.212889426154722101719695450734	24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 43; 44; 45; 46; 47; 48; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.242290316007009170953492762017 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.432307057523084865751445895414	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.665216418533113783055874429906 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.109595065464008295831949400041	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.259796523922086989668664382838	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2;

PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.324699768056399837424657639323	20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 43; 44; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.206472346874222573193881148384 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.226358495676435351658017746759	10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.272657410861156008163677293423 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.155180146157947911501819272776	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 5; 6; 7; 8; 9;
	10; 11; 12; 13; 14; 15; 16; 17;

CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.202171088946510467752674965603 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.190529860885922744883716825366	18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.824058953991801757435325996890 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.453845656490256804438721690175	10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 53; 54; 55; 56; 57; 58; 59; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.110643041644486140274933628561	10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45;

PET
1.3.6.1.4.1.14519.5.2.1.5168.2407.550869537725271111657920230425
	46; 47; 48; 49;
	50; 51; 52; 53;
	54; 55; 56; 57;
	58; 59;

CT
1.3.6.1.4.1.14519.5.2.1.5099.8010.119322609779875323559580878660
PET
1.3.6.1.4.1.14519.5.2.1.5099.8010.249263082661586220285883632385
	1; 10; 11; 12;
	13; 14; 15; 16;
	17; 18; 19; 2;
	20; 21; 22; 23;
	24; 25; 26; 27;
	28; 29; 3; 30;
	31; 32; 33; 34;
	35; 36; 37; 38;
	39; 4; 40; 41;
	42; 43; 44; 45;
	5; 6; 7; 8;
	9;

Table 12: Identifiers of the PET and CT images used in test set

CT and PET Series ID	Instance number

CT
1.3.6.1.4.1.14519.5.2.1.5099.8010.676870978342995025854500830255
	13; 14; 15; 16;
	17; 18; 19; 20;
	21; 22; 23; 24;
	25; 26; 27; 28;
	29; 30; 31; 32;
	33; 34; 35; 36;
	37; 38; 39; 40;
	41; 42; 43; 44;
	45; 46; 47; 48;
	49; 50; 51; 52;
	53; 54; 62; 64;
	65; 66; 67; 68;
	77; 78; 79; 80;

PET
1.3.6.1.4.1.14519.5.2.1.5099.8010.168019048968830267825126787458
	81; 82; 83; 84;
	85; 86; 87; 88;
	89; 90; 91; 92;

CT
1.3.6.1.4.1.14519.5.2.1.5168.2407.362413225455022159449564066196
PET
1.3.6.1.4.1.14519.5.2.1.5168.2407.238146266795916509370199931342
	10; 11; 12; 13;
	14; 15; 16; 17;
	18; 19; 20; 21;
	22; 23; 24; 25;
	26; 27; 28; 29;
	30; 31; 32; 33;
	34; 35; 36; 37;
	38; 39; 40; 41;
	42; 43; 44; 45;
	46; 47; 48; 49;
	50; 51; 52; 53;
	54; 55; 56; 57;
	58; 7; 8; 9;

CT
1.3.6.1.4.1.14519.5.2.1.5168.2407.305612440324917357174708603007
PET
1.3.6.1.4.1.14519.5.2.1.5168.2407.792703612882094378413587520680
	10; 11; 12; 13;
	14; 15; 16; 17;
	18; 19; 20; 21;
	22; 23; 24; 25;
	26; 27; 28; 29;
	30; 31; 32; 33;
	34; 35; 36; 37;
	38; 39; 40; 41;
	42; 43; 44; 45;
	46; 47; 48; 49;
	50; 51; 52; 53;
	54; 55; 56; 57;
	58; 59; 60; 61;
	62; 63; 9;

	1; 10; 11; 12;
	13; 14; 15; 16;

CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.294843762331111078340186327179 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.172732179865477455994355964587	17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 4; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.249365977453199862819433311007 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.192968079346077425924392755581	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 43; 44; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.138006540823343024847150800319 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.703290556591845680979689128316	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 5; 50; 6; 7; 8; 9;
	11; 12; 13; 14;

CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.239284839389957871610926941245 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.175654072425753217435054685652	15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.217678721215537238538786378975 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.296409700753957766902158048455	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 4; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.248971131371449580269976848140 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.296896700333907266472614846742	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 43; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.673495109316769189649734258189	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27;

PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.141816712492899828222731834549	28; 29; 3; 30; 31; 32; 33; 34; 35; 4; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.225974555763622593604657231392 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.362366067560800411671184004428	1; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 43; 44; 45; 46; 5; 6; 7; 8; 9;
CT 1.3.6.1.4.1.14519.5.2.1.5168.2407.330913987034947895745742498802 PET 1.3.6.1.4.1.14519.5.2.1.5168.2407.460079315359441222899811672069	10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 2; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 3; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 4; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 5; 50; 51; 52; 53; 54; 6; 7; 8; 9;
	1; 10; 11; 12; 13; 14; 15; 16;

CT
1.3.6.1.4.1.14519.5.2.1.5168.2407.836332957492252364644062915982
PET
1.3.6.1.4.1.14519.5.2.1.5168.2407.253672990060043871913155470672
	17; 18; 19; 2;
	20; 21; 22; 23;
	24; 25; 26; 27;
	28; 29; 3; 30;
	31; 32; 33; 34;
	35; 36; 37; 38;
	39; 4; 40; 41;
	42; 43; 44; 5;
	6; 7; 8; 9;

CT
1.3.6.1.4.1.14519.5.2.1.5168.2407.222158197040257516790732475812
PET
1.3.6.1.4.1.14519.5.2.1.5168.2407.334394790197753216584134239724
	1; 10; 11; 12;
	13; 14; 15; 16;
	17; 18; 19; 2;
	20; 21; 22; 23;
	24; 25; 26; 27;
	28; 29; 3; 30;
	31; 32; 33; 34;
	35; 36; 37; 38;
	39; 4; 40; 42;
	43; 44; 45; 46;
	47; 48; 49; 5;
	6; 7; 8; 9;

CT
1.3.6.1.4.1.14519.5.2.1.5168.2407.267100731412298810992195121169
PET
1.3.6.1.4.1.14519.5.2.1.5168.2407.324270282124507158140800393032
	10; 11; 12; 13;
	14; 15; 16; 17;
	18; 19; 20; 21;
	22; 23; 24; 25;
	26; 27; 28; 29;
	30; 31; 32; 33;
	34; 35; 36; 37;
	38; 39; 40; 41;
	42; 43; 44; 45;
	46; 47; 48; 49;
	50; 51; 52; 53;

	9;

CT
1.3.6.1.4.1.14519.5.2.1.5099.8010.228419894693506024475980055894
PET
1.3.6.1.4.1.14519.5.2.1.5099.8010.324303256391236074929030378575
	1; 10; 11; 12;
	13; 14; 15; 16;
	17; 18; 19; 2;
	20; 21; 22; 23;
	24; 25; 26; 27;
	28; 29; 3; 30;
	31; 32; 33; 34;
	35; 36; 37; 38;
	39; 4; 40; 41;
	5; 6; 7; 8;
	9;

DECLARAÇÃO

Declaro, sob compromisso de honra, que o trabalho apresentado nesta dissertação, com o título *"Joint Coding of Multimodal Biomedical Images Using Convolutional Neural Networks"*, é original e foi realizado por João Oliveira Parracho (2192600) sob orientação do Professor Pedro Antonio Amado Assunção, Professor Luís Miguel de Oliveira Pegado de Noronha e Távora, e Professor Lucas Arrabal Thomaz.

João Oliveira Parracho