FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Towards Biometrically-Morphed Case-Based Explanations

Maria Manuel Domingos Carvalho

WORKING VERSION

Mestrado em Bioengenharia

Supervisor: Jaime Cardoso

Supervisor: Wilson Silva

Supervisor: Maria João Cardoso

September 17, 2022

Towards Biometrically-Morphed Case-Based Explanations

Maria Manuel Domingos Carvalho

Mestrado em Bioengenharia

September 17, 2022

Abstract

Breast cancer is the second most common cancer among women in Portugal and worldwide. Despite being a disease that mutates and evolves at great speed, it is highly treatable when diagnosed early. In the past years, due to earlier stages of diagnosis and the evolution of breast cancer research, breast cancer conservative treatments (BCCT) have been favoured and almost replaced the traditional, more radical total mastectomies. This treatment is much less invasive, with similar oncological results and much better aesthetic results than total mastectomies. The introduction of conservative treatments and the increase in the survival rate also led to a new focus on the quality of life of the patient, in which the aesthetic outcome of the treatments plays a major role [103].

Unfortunately, a large number of women get dissatisfied with their aesthetic results after a breast cancer surgery or reconstruction procedure. Although this dissatisfaction may originate from objective, technical failures, it often stems from unrealistic expectations prior to the surgery, as it is very hard for patients to have a clear idea of what their outcome will be and images of results of past patients are often not representative due to physical differences between them. Taking this problem into account, the Cinderella project was born, with the aim of building a system based on Artificial Intelligence (AI) that creates customised predictions of the aesthetic results of the breast cancer treatments for each patient, in order to help the acceptatance process of the outcome and improve their self-esteem after surgery. These predictions are generated by taking into account the characteristics of the patient under evaluation and the results of similar, past cases.

This dissertation was developed under this project, and represents an early investigation on methods to create images that represent the physical outcome of a case on the body of a second patient from a current case.

As a starting point, paired transformations were explored, where the aim was to train a deep learning model to introduce or correct common physical consequences of breast cancer procedures that affect the breasts' aesthetic, such as colourization and asymmetry. These experiments showed that, visually, the aesthetic of the images can easily be changed. The next set of experiments consisted of unpaired translations between the domains of 'Excellent' and 'Poor' aesthetic classification, where the goal was to worsen or improve the aesthetic result of images. For this, conditional GANs based on the Pix2Pix models and cycleGANs were explored. Despite the unsatisfactory results in this more challenging setup, these models and methods could be very useful in the future for this project. Finally, a preliminary attempt using a Wasserstein GAN to develop the final model that could be used to create the biometrically-morphed case-based image was introduced.

To conclude, in this dissertation, different methods were explored to alter the aesthetic of the results of a breast cancer treatment, in the hopes of using them in the future to create images that

represent, realistically, probable aesthetic outcomes of BCCTs.

Keywords: Breast Cancer, Aesthetic Evaluation, Generative Adversarial Networks, Image Translation

Resumo

O cancro da mama é o segundo cancro mais comum entre as mulheres em Portugal e no resto do mundo. Apesar de ser uma doença que evolui a grande velocidade, é altamente tratável quando diagnosticada precocemente. Nos últimos anos, devido à sensibilização para rastreios e à evolução da investigação do cancro da mama, os tratamentos conservadores do cancro da mama (BCCT) têm sido favorecidos e quase substituíram na totalidade as tradicionais mastectomias radicais. Este tratamento é muito menos invasivo, com resultados oncológicos semelhantes e resultados estéticos muito melhores do que as mastectomias totais. A introdução de tratamentos conservadores e o aumento da taxa de sobrevivência levou também a um novo foco na qualidade de vida da paciente, em que o resultado estético dos tratamentos é importantíssimo.

Infelizmente, um grande número de mulheres fica insatisfeita com os seus resultados estéticos após uma cirurgia de cancro da mama ou a um procedimento de reconstrução. Embora este descontentamento possa ter origem em falhas objectivas e técnicas, muitas vezes resulta de expectativas irrealistas antes da cirurgia, uma vez que é muito difícil para as pacientes terem uma ideia clara de qual será o seu resultado e que as imagens de resultados de pacientes anteriores são muitas vezes não representativas devido a diferenças físicas entre elas. Tendo em conta este problema, nasceu o projecto Cinderella, com o objectivo de construir um sistema baseado em AI que cria previsões personalizadas dos resultados estéticos dos tratamentos do cancro da mama para cada paciente, de modo a ajudar o processo de aceitação do seu resultado e melhorar a sua autoestima após a cirurgia. Estas previsões são geradas tendo em conta as características da paciente em avaliação e os resultados de casos semelhantes anteriores.

Esta dissertação foi desenvolvida no âmbito deste projeto, e representa uma investigação inicial sobre diferentes métodos para criar imagens que representam o resultado físico de um caso no corpo de um segundo paciente de um caso atual.

Como ponto de partida, foram exploradas transformações emparelhadas, onde o objectivo era treinar um modelo de *deep learning* para introduzir ou corrigir consequências físicas comuns dos procedimentos do cancro da mama que afectam a estética das mamas, tais como a cor e a assimetria. Estas experiências mostraram que, visualmente, a estética das imagens pode ser facilmente alterada. O próximo conjunto de experiências consistiu em traduções não emparelhadas entre os domínios de classificação estética 'Excelente' e 'Fraca', onde o objectivo era piorar ou melhorar o resultado estético das imagens. Para tal, foram explorados GANs condicionais baseados nos modelos Pix2Pix e CycleGANs. Apesar dos resultados pouco satisfatórios, estes modelos e métodos têm um grande potencial no futuro deste projecto. Finalmente, foi introduzida uma tentativa preliminar com uma Wasserstein GAN de desenvolvimento do modelo final que poderia ser utilizado para criar as imagens híbridas.

Para concluir, nesta dissertação foram explorados diferentes métodos para alterar a estética dos resultados de um tratamento do cancro da mama, na esperança de os utilizar no futuro para

criar imagens que representam, realisticamente, os resultados estéticos prováveis de um BCCT.

"Research is formalized curiosity. It is poking and prying with a purpose.
- Zora Neale Hurston"

Contents

List of Figures

List of Tables

Abbreviations

ML	Machine Learning
DL	Deep Learning
PROM	Patient-reported Outcome
ANN	Artificial Neural Networks
ReLU	Rectified Linear Unite
CNN	Convolutional Neural Network
VGG	Visual Geometry Group
ResNet	Residual Neural Network
BCCT	Breast Cancer Conservative Treatment
BRA	Breast Retraction Assessment
BCE	Breast Compliance Evaluation
LBC	Lower Breast Contour
UNR	Upward Nipple Retraction
OBCS	Objective Breast Cosmesis Scale
BAT	Breast Analyzing Tool
BSI	Breast Symmetry Index
SVM	Support Vector Machine
DNN	Deep Neural Network
CBMIR	Content-Based Medical Image Retrieval
PACS	Picture Archiving and Communication System
xAI	Explainable Artificial Intelligence
K-NN	K-Nearest Neighbors
CBR	Content-Based Reasoning
COLE	Contributions Oriented Local Explanations
LIME	Local Interpretable Model Agnostic Explanation
LDA	Latent Dirichlet Allocation
AE	Autoencoder
VAE	Variational Autoencoder
GAN	Generative Adversarial Network
DCGAN	Deep Convolutional Generative Adversarial Network
WGAN	Wasserstein Generative Adversarial Network
WGAN-GP	Wasserstein Generative Adversarial Network with Gradient Penalty
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
MSLE	Mean Squared Logarithmic Error
SSIM	Structural Similarity Distance

Chapter 1

Introduction

1.1 Context

Breast cancer is a disease caused by the abnormal multiplication of tumorous cells in the breast tissue. Both men and women can suffer from breast cancer, but it is most prevalent in the female population (more than 99% of the cases). In 2020, globally, 2.3 million women were diagnosed with breast cancer and 685 000 died from it. By the end of 2020, there were 7.8 million women who were diagnosed with breast cancer in the past five years, making it the most prevalent cancer in the world [7]. The appropriate treatment for breast cancer depends on many factors, such as type of cancer, age of the patient and cancer stage, as more advanced cancers call for more extreme treatments. Standard treatments include chemotherapy, radiation therapy and surgery.

The improvement of breast cancer treatments and increased life expectancy after the diagnosis over the years led to a new focus on the quality of life of the patient after treatments, in which the aesthetic outcome of the treatments plays a huge role [103].

1.2 Motivation

The improvement of breast cancer treatments has led to the use of more conservative treatments, and has fomented a recent interest in the aesthetic outcome of these procedures. Sadly, almost 30% of patients who undergo breast reconstruction are unhappy with the results [103]. This disappointment may be due to objective failures during treatment, but it may also be due to unrealistic expectations: it is very difficult for doctors to explain to the patient, in words or by showing her pictures of previous patients, what aesthetic result she can expect from surgery, because it is difficult for the patient to visualise it. This problem is not constricted to surgeries related to breast cancer, but also other elective cosmetic surgeries, as many patients say they are disappointed with the results of their cosmetic procedure due to the fact that the result didn't match their expectations

[11]. Therefore, there is a need for a system that generates realistic images that represent the aesthetic outcome of surgery, taking into consideration the characteristic of the patient and historical evidence of other patients with a similar history.

Cinderella is an European project that aims to do just this, and provide women with breast cancer with a realistic preview of their surgery options, from conservative to radical and with or without reconstruction, based on past results, which can help patients make a more informed choice about their treatment and manage their expectations regarding its outcome. The impact of this project will be measured through comparisons between patient reports. This master dissertation is developed under this project.

1.3 Goals

The main goals of this dissertation are:

- Investigation of different methods to produce biometrically-morphed case-based explanations
- Update the state of the art with this novel interpretability approach

1.4 Main contributions

This dissertation's main contributions are:

- Development of methods to create synthetic images with introduced characteristics that affect the aesthetic evaluations
- Exploration of different methodologies to perform paired-image translation between images with different aesthetic results
- Exploration of conditional and cycle GANs to perform unpaired-image translation between images with different aesthetic results
- Proposal of a method for the generation of biometrically-morphed images with a Wasserstein GAN

1.5 Dissertation Structure

Besides Introduction, this thesis contains X more Chapters.

Chapter 2 introduces some basic concepts about Machine and Deep Learning, with a particular focus on Convolutional Neural Networks.

Chapter 3 reviews the current methods used for the aesthetic evaluation of breast cancer procedures and automatized methods developed for this task.

Chapter 4 reviews some methodologies used in Content-Based Medical Image Retrieval.

Chapter 5 focuses on four different types of generative models, with special attention to Generative Adversarial Networks.

Chapter 6 presents some results related to paired image translation, with the goal of introducing or correcting defects that alter the breast aesthetic evaluation.

Chapter 7 showcases the results of experiments with unpaired image translation, using conditional GANs and a cycleGAN in an attempt to perform image translation between two domains of aesthetic classification.

Chapter 8 proposes some experiments using a Wasserstein GAN as a method to create biometrically-morphed images.

Chapter 9 concludes this document with some final remarks and future work proposals.

Chapter 2

Background Knowledge

2.1 Machine learning

Arthur Samuel coined the term **Machine Learning** in 1959, referring to 'algorithms that learn from data, and even improve themselves, without being explicitly programmed'. With the high amount of data production and improvement of computational power, machine learning is used everywhere nowadays, from recommendation systems in our phones to large industrial processes.

Machine learning methods include three different approaches: **supervised**, **unsupervised** or **reinforcement** learning [17].

- In **Supervised Learning**, a ML model is trained with labelled data, meaning that each data sample used has a corresponding class. During training, the weights of the model are adjusted in order to approximate the mapping function. After training, the model can be used to predict the class of a new, unlabelled sample. Regressions and classification problems are types of supervised learning.

- In **Unsupervised Learning**, the training data is not labelled or categorized, and the goal is to find patterns and similarities between the samples and, without having a "correct" answer, organize them according to their features. Clustering and learning of association rules are forms of unsupervised learning.

- **Reinforcement Learning** algorithms learn by interacting with their environment and through trial and error. The reinforcement learning algorithm is trained in a dynamic way, with a system of rewards or penalties according to its performance. The final goal of the system is to maximize the total reward.

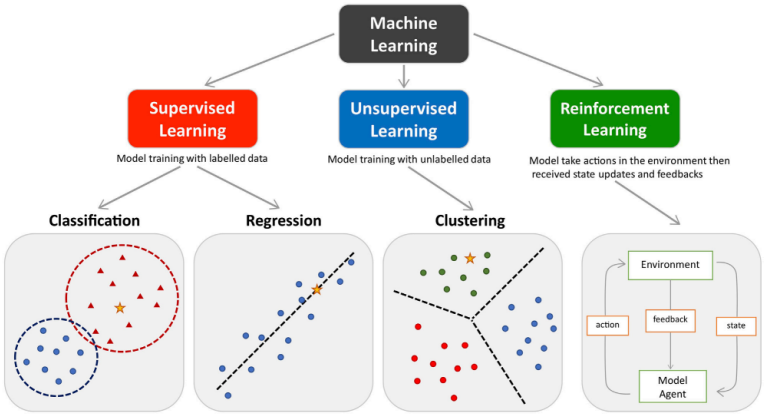


Figure 2.1: Summary of the different learning approaches in ML, from [107]

2.2 Deep Learning

Deep learning is a sub-field of machine learning, based on the use of **Artificial Neural Networks** (ANN), that allow the extraction of complex and abstract features, automatically, without the need for explicit programming. Deep learning has been theorized for many decades, but only rather recently has it made its breakthrough, as it requires large amounts of data and lots of computational power [16, 110]. This automatic feature extraction process simplifies many tasks such as image classification problems, where manual feature extraction can be tiresome. Moreover, ANNs can learn complex patterns from data, even when unstructured. Deep learning has surpassed the state-of-the-art results of many conventional ML algorithms, and has rapidly replaced these traditional methods, especially in complex tasks with large amounts of data.

Artificial neural networks try to mimic the way the brain works. The human brain contains a densely interconnected network of, on average, 86 billion neurons [69], each connected to other neurons. Similarly, neural networks are generally composed of interconnected units that pass information amongst each other, like neurons communicating through synapses. The simplest neural network, the perceptron, shown in Figure 2.2, contains a single unit. Complex neural networks are basically multi-layers perceptrons, where they are stacked and interconnected. A perceptron has an input layer, internal parameters (weights and bias), and an activation function. Weights are used to give more importance to certain inputs for the final decision, and the bias value allows to shift the activation function. The inputs are multiplied by their weights and summed together, along with the bias terms, which is constant. Then, an activation function is applied to the weighted sum. The activation function allows the introduction of non-linearities in the decision boundary, and the approximation to complex functions [14, 28]. Some common activation functions include:

- **Rectified Linear Unit (ReLU)** is an easy-to-implement, linear function, whose output is either zero, when the input is negative or otherwise, equal to the input itself. However, the ReLU function is known to have a problem, called "dying ReLU", when many nodes output an activation value of zero, which can "kill" the entire network, and turn it into a constant

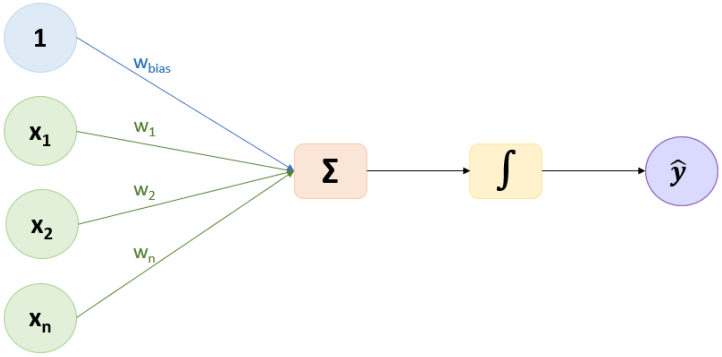


Figure 2.2: Representation of a perceptron

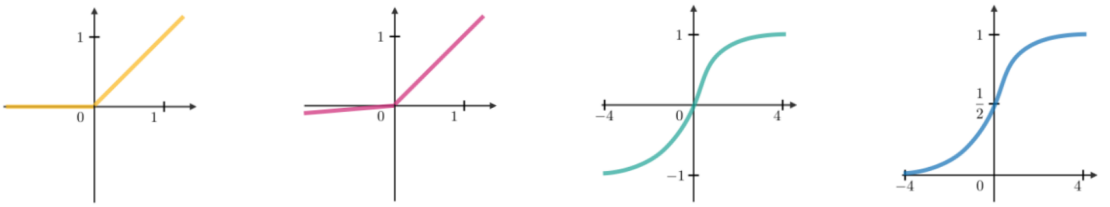
function. **Leaky ReLU** came to solve this by allowing small negative values when the input is also negative.

- **Sigmoid or Logistic Function** outputs values between 0 and 1, which can be interpreted as a probability, making it useful for binary classification tasks, for example. However, this function can also cause a neural network to get stuck during training due to a problem known as vanishing gradients, where the gradient becomes so small and close to zero that prevents the optimisation of the network's weights.

- **Hyperbolic Tangent Activation Function (Tanh)** maps the input into a range from -1 to 1, and centred around 0.

Perceptrons can be stacked together and added in different layers with connections with each other, forming a multi-layer neural network. These layers between the input and the output are known as hidden layers, and the number of hidden layers in a network defines its depth. Deep neural networks often have a high number of hidden layers, each with multiple units and connections.

The network is trained and its performance is improved through a process called backpropagation, where an optimization algorithm, such as gradient descent, guides the process towards a minimum in the loss function [35]. The loss is propagated backwards to the previous units, their



(a) ReLU function graph

(b) Leaky ReLU function graph

(c) Tanh function graph

(d) Sigmoid function graph

Figure 2.3: Graphical representation of some common activation functions. Figures from [98]

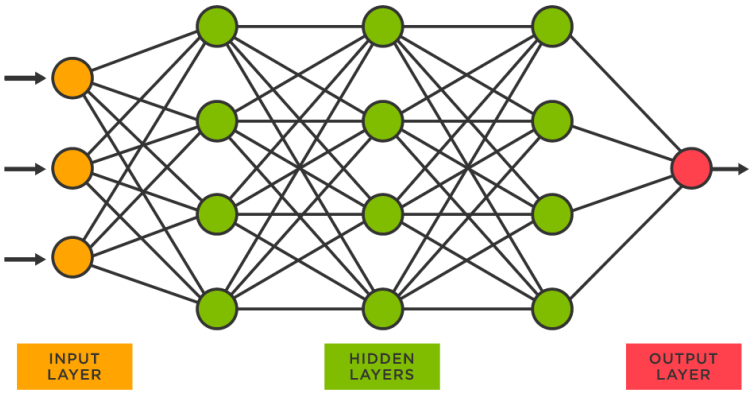


Figure 2.4: Representation of an Artificial Neural Network

gradient is computed and the weights are adjusted. With each iteration, the network becomes more accurate.

Deep neural networks have a high capability to learn and fit the data, which often leads to the problem of overfitting, where the model fits the training data almost perfectly, but is unable to generalize to new data. In other words, the model achieves high accuracy levels during training, but has a high error rate in the validation step, with new data [28]. In order to minimize overfitting, some regularization techniques can be used, such as:

- **L2 and L1 regularization**, the most common strategies. The network's regularization is achieve by adding a term, known as the regularization term, to the loss function.

- **Dropout**, a very useful technique, especially in very large networks. At every iteration, a percentage of total nodes is randomly selected and removed, along with its connections, resulting in a different set of outputs at each training iteration.

- **Early stopping**, a strategy that forces the training to stop immediately when the performance on the validation set starts to decrease.

- Overfitting is often caused when the model is trained with a small dataset. Therefore, **data augmentation** is an easy approach, and new, unlabelled samples can be created through the application of simple operations such as scaling, shifting or flipping to the samples in the dataset. Generative models, discussed in chapter 5, can also be used to create synthetic samples similar to the training set.

There are several types of artificial neural networks, such as Recurrent Neural Networks, Feedforward Neural Networks and Convolutional Neural Networks, which are state-of-the-art approaches in tasks involving images, and will be discussed further in the next section.

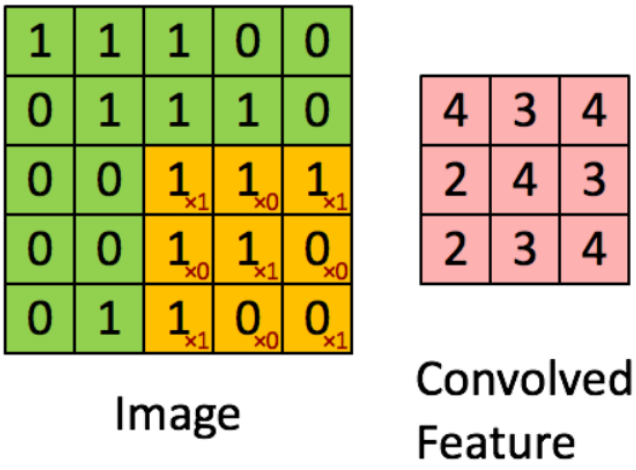


Figure 2.5: Example of the result of a convolutional operation on a 5x5x1 image with a 3x3x1 kernel. Figure from [121]

2.2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a deep learning architecture, usually used for tasks involving images. The CNN design is inspired by the neural organization of the visual cortex in our brain, where neurons only fire action potentials when a stimulus appears in a small portion of their visual field, known as the receptive field. Similarly, a CNN is able to evaluate an image and assign different importance to various elements that constitute the image, and in the end the weights of each region are used for the final classification. CNNs are able to capture the spatial dependency between the pixels in an image through the application of important operations. These relevant operations are performed in the three most important types of layers in a CNN: convolutional, pooling and fully-connected layers [121].

The first layer in a CNN is the **Convolutional layer**, where a set of filters or kernels, with variable sizes but generally much smaller than the image, "slide" through the image and apply a convolutional operation in order to create a feature or activation map.

In the pooling layer, the activation map goes through dimensionality reduction operations, where the spatial size of the features is reduced in order to decrease the computational power required for the training of the CNN. Similarly to what happens in the convolutional layer, a filter sweeps the input in the **Pooling layer**. However, this filter does not contain weights, but instead usually returns either the maximum (max pooling) or the average (average pooling) value from the portion of the image within the receptive field, or, in other words, covered by the filter.

These pooling operations, besides making training more efficient, are also useful for extracting dominant features which are rotational and positional invariant. The first convolutional layer can be followed directly by a pooling layer or by other convolutional layers, and these two can be repeated and mixed many times along the networks. With each layer, the complexity of the representation increases, and as the images go through them, more complex features are detected.

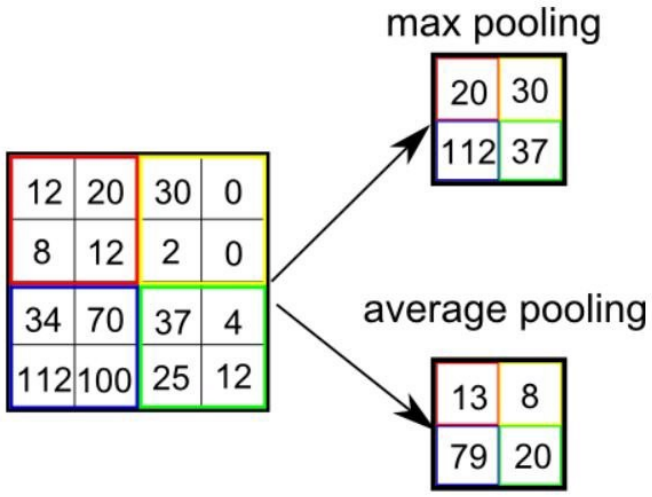


Figure 2.6: Example of the results of max and average pooling with a 2x2 pooling kernel and a stride of 2. Figure from [121]

Nevertheless, the final layer of a CNN is a **fully connected layer**, where each unit is connected to all the units of the previous layer, meaning that each pixel value is accounted for in the final classification.

All in all, we can think of a CNN as a network which is trained to reduce the dimension of images into an easier-to-process size, while maintaining the essential features needed for the classification [15].

A common approach in tasks related to deep computer vision is to use pre-trained CNNs, which were trained on very large and generic datasets, and reuse their weights for a new task, through transfer learning techniques. Some of the most widely used architectures are VGG (architecture in figure 2.8) and ResNet.

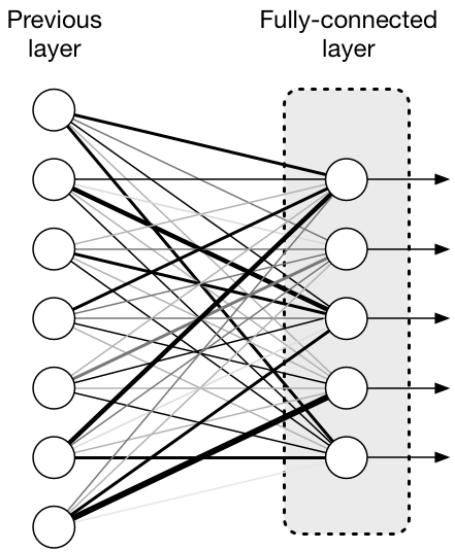


Figure 2.7: Illustration of a fully connected layer

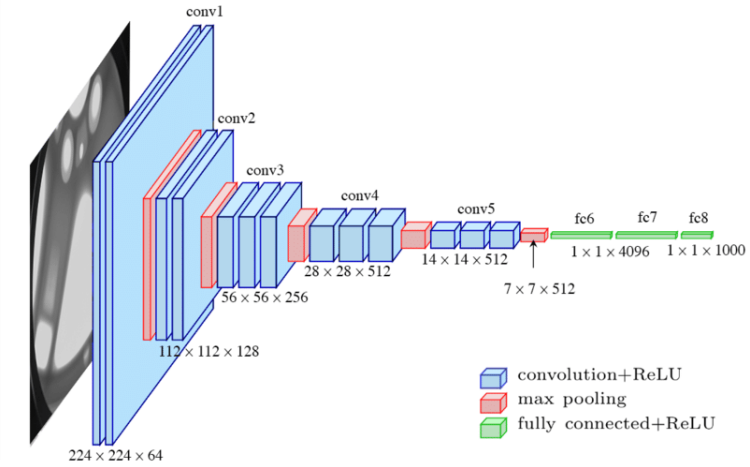


Figure 2.8: VGG Network Architecture. Image from [61]

2.3 Summary

Deep Learning models are based on artificial neural networks, whose architecture is inspired by the human brain, with interconnected units that propagate information to each other. These models allow the extraction of complex features automatically and have achieved state-of-the-art results in many tasks, such as computer vision, where they have almost replaced conventional machine learning models. In computer vision specifically, convolutional neural networks are a very powerful tool that facilitate feature extraction and "learn" the most important characteristics of the image.

Since deep learning models are the state-of-the-art in computer vision, they will also be used throughout this dissertation.

Chapter 3

Literature Review: Breast Aesthetic Evaluation

3.1 Breast Cancer

The human body is made up of trillions of cells. These cells grow and multiply according to the body's needs and die when they are old or defective; these processes are induced by biochemical signals. However, sometimes, this process is disrupted, and abnormal cells grow and multiply instead of being eliminated, leading to the formation of tumours. These tumours can be benign or malignant. Normally, our bodies are able to eliminate cells with damaged DNA before they turn cancerous, but this capacity decreases as we age.

Malignant or cancerous tumours have the ability to spread throughout the body, ignore biochemical signals from the body, promote angiogenesis (formation of new blood vessels to assure a supply of oxygen and nutrients) and "hide" from the immune system to avoid elimination.

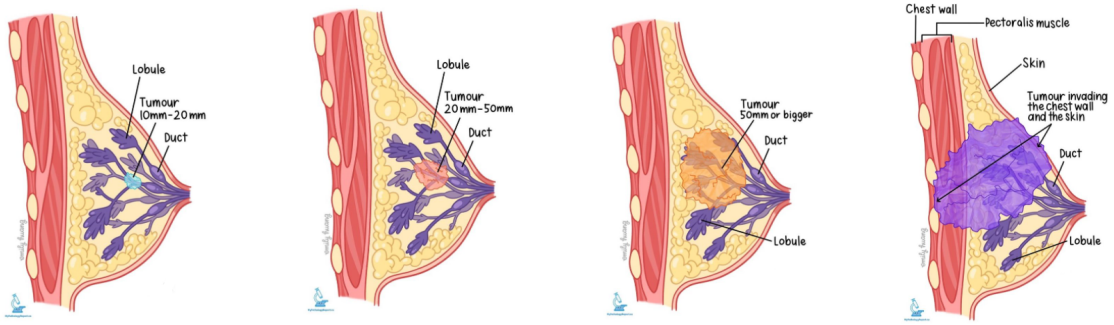
There are more than 100 types of cancer, and they are normally named for the organs or tissues where cancer first develops [23]. The most common types of cancer worldwide are breast, lung, colon and rectum and prostate cancers.

Although experts don't know exactly what triggers the formation of malignant tumours, there are some known cancer risk factors, such as smoking, obesity, alcohol abuse or a family history of cancer [22].

Breast cancer is a disease caused by the abnormal multiplication of tumorous cells in the breast tissue. Both men and women can suffer from breast cancer, but it is most prevalent in the female population (more than 99% of the cases). In Portugal, it is the second most common type of cancer in women: annually, there are 7000 new diagnoses and 1800 deaths [9]. The average risk of a woman developing breast cancer sometime in her life is about 13% [5].

3.1.1 Types and stages of breast cancer

The most common type of breast cancer is **ductal carcinoma** (*in situ* or invasive, according to the extent of proliferation), which originates in the cells of the milk-producing ducts. Other types



(a) Stage 1 Cancer (b) Stage 2 Cancer (c) Stage 3 Cancer (d) Stage 4 Cancer

Figure 3.1: Tumour stage for invasive ductal carcinoma, from [18]

include **lobular carcinoma** (*in situ* or invasive), which begins in in the lobules (glandular tissue) and **inflammatory breast cancer**, a rare and aggressive form of cancer that causes by the blockage of lymphatic vessels [4].

Just like other types of cancer, breast cancer can be diagnosed in different stages, according to factors such as tumour size, extensions of the spread throughout the body and whether it has affected lymph nodes. There are five stages of breast cancer, which can then be sub-grouped according to more specific criteria. The stages range from 0, which describes *in situ* cancer, that is not present in the surrounding breast tissue, to IV, where the cancer is metastatic and has spread to other organs [22]. Figure 3.1 shows four stages of ductal carcinoma, based only on the size of the tumour and the presence of cancer cells in the skin or muscles of the chest wall.

3.1.2 Screening and Diagnosis

Breast cancer is, often times, a silent disease, with no obvious symptoms, especially in earlier stages. It is usually diagnosed in a screening exam or after a woman notices a lump [3]. Because of this, it is of extreme importance for women, especially women over 40 or women who have a known predisposition to develop breast cancer, to perform regular screening exams, such as mammographies and self-examination, and many countries have their own screening programs. The most common methods for diagnosis are: [9]

- **Clinical Breast Examination:** physical exam where a doctor feels the patient's breasts in order to evaluate whether an anomaly exists or to evaluate the nature of a lump, as benign and malignant masses can be distinguished by touch. This method is not recommended by some health boards, for example in the USA and Canada, as when it is used as a complement to mammographies, doesn't add any new information and even increases the rates of false positives [25, 44, 3]. However, it is still recommended in Portugal, for example, as a screening exam for younger women and in developing countries, where this examination might be more accessible [9].

- **Mammography:** medical imaging exam that uses a low-dose x-ray system to visualize the breasts' interior. Mammograms are one of the most common screening exams in many

countries. They present a false negative rate of 1/8 (one in eight cancer is not diagnosed, usually in women with denser breasts, where visibility is compromised), a rather high false positive rate (almost half of the women in the USA are expected to get a false positive over a 10 year period of regular screenings) and a 1-10% rate of overdiagnosis, which results in cancer treatment for women that would not have a clinical complication from small tumours found on the exam [1].

- **Ultrasound:** medical imaging exam that uses high-frequency sound waves in order to assess whether a lump is a cyst or a solid mass, which might be malignant. The ultrasound is a complement to the mammogram's results.
- **MRI:** Magnetic resonance exam, used to complement the mammogram with clearer images, obtained with a magnetic field and computer-generated radio waves.
- **Biopsy:** Removal of a piece of tissue from the lump in order to evaluate if it is cancerous.

  Some of the symptoms associated with breast cancer include swelling, alterations of the breasts' skin, pain, nipple discharge or lumps [5].

3.1.3 Treatment

The appropriate treatment for breast cancer depends on many factors, such as type of cancer, age of the patient and cancer stage, as more advanced cancers call for more extreme treatments. Breast cancer treatments include: surgery, radiation therapy and systemic treatments.

3.1.3.1 Surgery

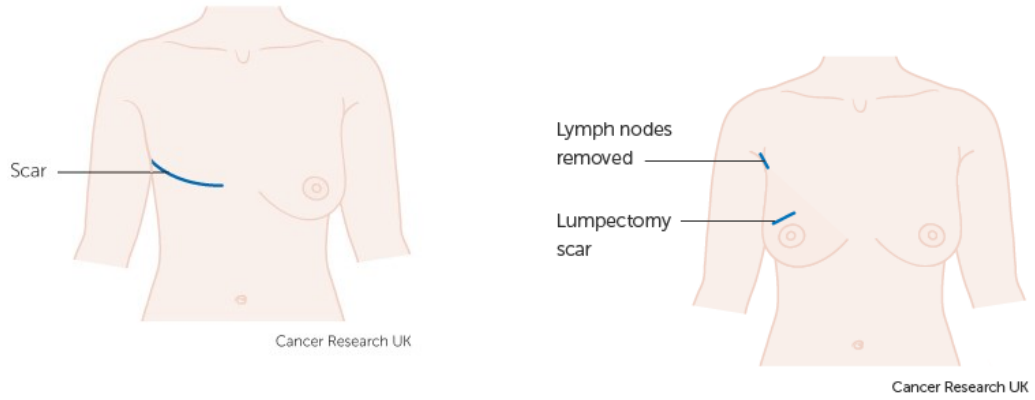
Breast cancer surgery is performed with the goal of removing the tumour. The surgery can be a mastectomy, where the entire breast is removed or a lumpectomy or breast-conserving surgery, where only the cancerous tissue is removed, with a margin of healthy tissue.

There are three types of mastectomies: [2]

- Radical mastectomy: surgical removal of the breast, pectoral muscles and the lymph nodes of the axilla.

  After its introduction in the late 19th century by William Halsted, radical mastectomies were the treatment of choice for breast cancer for almost 80 years, but are rarely performed nowadays [131].
- Modified radical mastectomy: surgical removal of the entire breast and lymph nodes of the axilla.
- Simple or total mastectomy: surgical removal of the entire breast tissue.

Breast-conserving surgery is usually followed by radiation therapy. In the last years research has shown that breast cancer conserving treatments with surgery and radiotherapy, was as effective



(a) Mastectomy (b) Lumpectomy

Figure 3.2: Mastectomy and Lumpectomy surgeries, from [21, 20]

as mastectomies for smaller cancers [131] [2]. However, many women who are eligible for BCCT still choose mastectomies for fear of recurrence or reluctance to do radiation therapy.

Currently, the most common surgical options in Portugal are breast cancer conservative treatments and less invasive mastectomies with immediate reconstruction [103].

3.1.3.2 Radiation Therapy

Radiation therapy is a treatment where high-energy beams are used to kill cancer cells and is normally performed after a surgery, in order to destroy the remaining cancer cells in the breast, chest wall, or axillar area. Radiation therapy is almost always a step in BCCT, as it has been shown that it can reduce the risk of recurrence by about 50% [3].

3.1.3.3 Systemic Therapies

Systemic therapies represent all the treatments where a drug is administered in the bloodstream, affecting every cell of the body, not only the cancerous ones. Systemic therapies include chemotherapy, hormonal and target therapies and immunotherapy. [3].

- **Chemotherapy:**drug treatment that is used to attack fast-growing cells. A combination or a single drug can be used for the treatment [10].
- **Hormone therapy:** treatment with the aim of lowering the estrogen levels in the body, which promote the growth of of the most common subtype of breast cancer.
- **Target Therapy:** treatment with substances that slow or stop the proliferation of cancer cells, by targeting specific molecules that promote their growth and spread.
- **Immunotherapy:** a more recent treatment, that is used to help the patient's immune system recognise and fight the cancer cells.

3.1.4 Breast Cancer Prognosis

When detected in an early stage, breast cancer has a positive prognosis in 95% of the cases [19]. The five-year survival rate is 99% when the cancer is constricted to only one breast, and 65% of patients in the USA are diagnosed at this stage. This survival rate decreases to 86% if the has spread to the lymph nodes and to 29% when cancer has spread to distant parts of the body [6]. In the UK, almost 40% of breast cancer patients are diagnosed with stage 1 cancer, and their five year-survival rate is estimated at 98%, while 37% are diagnosed in stage 2, where this rate is 89.6% [8].

The mortality rate of breast cancer has decreased significantly in the past years in developed countries [8, 19, 125], probably due to prevention screening programs but also due to the evolution in the treatments; in the UK, the survival rate has almost doubled in the last 20 years [8], while in the USA the overall mortality rate has decreased 34% between 1975 and 2010 [125].

3.2 Aesthetic Evaluation

As mentioned in chapter 3.1, breast cancer is the second most commonly diagnosed cancer amongst women. The prognosis for breast cancer has become more optimistic over the years, due to improved treatments and widespread screening programs, which allow the diagnosis in earlier stages.

Since the mortality of breast cancer has decreased significantly in the past years and the disease has become much more manageable, there has been an increased focus on restoring the patient's quality of life. The quality of life of breast cancer survivors is dependent on many factors, and the relationship between the patient and her body image is one of them [55].

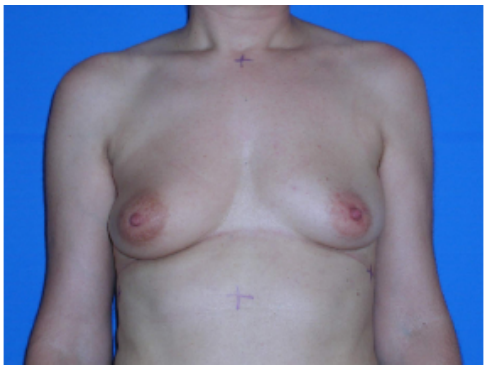
Fortunately, a lot of progress has been made in the past years to improve the cosmetic results of breast cancer treatments. Breast conservative treatments have become a standard practice for smaller cancers, and show similar oncological results to more aggressive techniques, and a much better cosmetic results [75, 131].

The cosmetic result of a breast cancer procedure is an extremely important means of evaluating the success of the treatment, both for the patient and the medical institution where the treatment was performed. Despite the improved cosmetic results with BCCT, the results are still very heterogeneous and depend on a lot of factors [77], both related to the tumour, such as its position and size for example, and the patient, such as age, breast size or weight.

The aesthetic evaluation can be subjective, through photographs or direct observation, by the patient (self-evaluation) and/or a panel of professionals or objective, which is achieved through measurements on photographs, the patient's body (anthropometry) or 3D images of the breasts [85]. However, despite the evolution over the years in techniques for aesthetic evaluation, a gold standard still doesn't exist [78, 79].

However, despite the method of evaluation used, the aesthetic assessment is often done according to the **Harvard Scale** [27, 134], introduced by Harris et al. [80] in 1979. This scale has four grades and is based on the difference between the treated and the untreated breast:

(a) Excellent



(b) Good

(c) Fair

(d) Poor

Figure 3.3: Examples of aesthetic evaluation with the Harvard Scale, from [57]

- Excellent: when both breasts are nearly identical
- Good: when the treated breast has some differences from the untreated one
- Fair: when the treated breast is clearly different from the untreated one, but it is not seriously distorted
- Poor: when the treated breast is seriously distorted

The Harvard scale, although simple and not time-consuming, is not reproducible, with low consensus between observers [134], is affected by the experience level of the observers and their relation with the patient [117, 45] and can only be applied when the surgery is constricted to only one breast, since this scale is based on comparisons between the treated and untreated breast.

In the next section, different methods for aesthetic evaluation will be reviewed.

3.2.1 Subjective evaluation

The subjective evaluation of the aesthetic result by the patient is of utmost importance; after all, the aim is that the patient is happy with her body image. Patient-centered evaluations are usually performed with questionnaires that will measure several aspects of a patient's health status after their treatment. The questionnaires used are usually Patient-Reported Outcome Measures

(PROMs) and Patient-Reported Experience Measures (PREMs). PROMs are used to measure several health-related parameters and assess the outcomes of the treatment, while PREMs evaluate the patients' experience whilst receiving care, by assessing for example the communication between the medical staff and the patient [12]. However, despite its importance, this type of evaluation is not reproducible nor can it be used to compare different treatments or institutions, as it is extremely dependent on many factors related to the patient's history, such as the woman's age, marital status, expectations before surgery or relation with her medical team. The subjective aesthetic evaluation can also be performed by a doctor or a panel of doctors; although they may focus more on clinically relevant factors, it can be equally biased, and vary a lot depending on the doctor or the clinic. Consequently, there is not an appropriate subjective method to use for comparing clinics, doctors or treatments, which is essential in order to assess which institutions have better results [78, 46].

3.2.2 Objective evaluation

However, some objective measures, although not used currently as the standard for cosmetic evaluation, have been developed. They involve measurements and comparisons between fiducial points of the breast, such as the nipples, the breast contour and endpoints [62]. Moreover, these objective methods can be automatized in order for the measurements to be performed through pictures automatically.

The first objective measurement was introduced by Pezner *et al.* [111]; the Breast Retraction Assessment (BRA) is a measure of breast asymmetry and is calculated as shown in figure 3.4, by comparing the position of the nipples. Since then, other methods have been developed and introduced with the aim of evaluating the aesthetic outcome, such as:

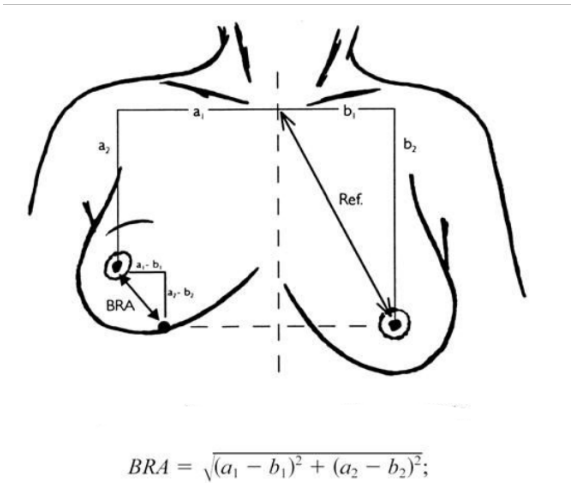
- Breast Compliance Evaluation [127], which uses the difference between the lengths from the inframammary fold to the centre of the nipple of both breasts.

- Lower breast contour and Upward nipple retraction, introduced by Van Limbergen *et al.* [54], which calculates the difference between the inferior breast contour in both breasts and the difference between nipple height, respectively.

- Calculation of the distance of the areolar border from the sternal notch, midline of the sternum and submammary fold, introduced by Stark *et al.* [33].

- Objective breast cosmesis score, which is used to calculate the nipple displacement and the asymmetry in breast dimensions and contour

The introduction and development of objective methods of evaluation have provided a way to increase the reproducibility of the aesthetic assessment. The outcome of these objective measurements correlates highly with the outcome of subjective measures [78].



$$BRA = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2};$$

Figure 3.4: Breast Retraction Assessment [111]

3.2.2.1 Combination methods

Subjective and objective methods can also be used together in order to achieve a more complete evaluation; for example, Noguchi *et al.* [91] introduced a combination method that uses the Moiré effect and nipple deviation as objective measures, and subjective evaluation of breast atrophy, telangiectasia, and scaring and Al-Ghazal *et al.* [120] described a method where the patients final aesthetic evaluation included subjective evaluation through frontal and lateral photographs and objective measures.

Moreover, some methods have been developed for the evaluation with 3D images, which allow the assessment of important parameters that can't be analysed in photographs, such as breast volume, and a broader evaluation of the breast's appearance over time [85]. However, the use of 3D images also translates into higher costs [79].

3.2.3 Softwares for aesthetic evaluation

Although the introduction of objective measures has made the process more reproducible and regulated, the aesthetic evaluation process remained time-consuming for health professional, who have to identify the fiducial points and perform the measures manually. In order to tackle this issue, some automated solutions have been developed: the kOBCS software, the BAT software and BCCT.core.

3.2.3.1 kOBCS Software

kOBCS is a recent software program (introduced in 2020) that calculates the Objective Breast Cosmesis Scale (OBCS) directly with photographs of the patients, which can be non-standardized (without scale calibration, standardized lighting conditions, standardized digital quality, standardized background, or consistent magnification). The OBCS value is the combined result of specific

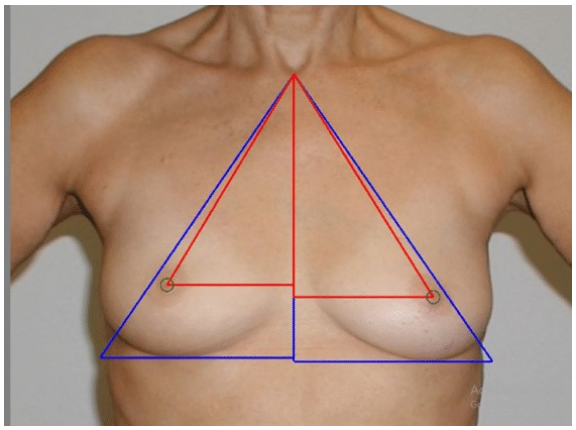


Figure 3.5: Kobcs interface from [124]

measurements that represent geometric asymmetry in the terms of nipple displacement, breast dimensions, and breast contours, as can be seen in fig. 3.5. This software's calculations have proven to be highly correlated with direct measurements, self-evaluation and evaluation by a panel of specialists [124].

3.2.3.2 Breast Analyzing Tool - BAT software

The BAT software was introduced in 2007, and it uses the Breast Symmetry Index (BSI), where the arithmetic mean of the difference between 24 cuts from the nipple to the breasts' contour is calculated. The obtained values can then be converted into a simplified 3-point Harris scale (good, fair, poor) [92]. BSI is proved to be an easy-to-use measure, with a high correlation with subjective evaluation from experts [60].

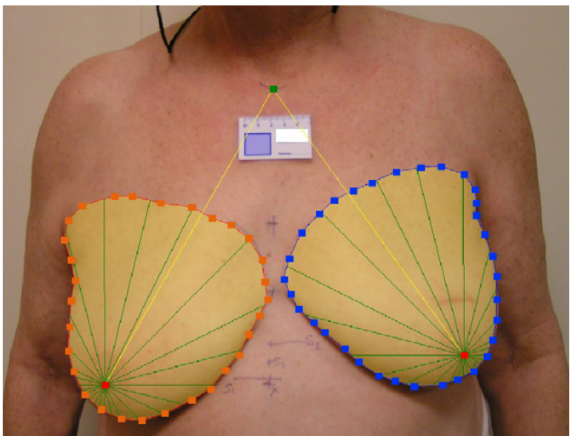


Figure 3.6: BAT interface from [60]

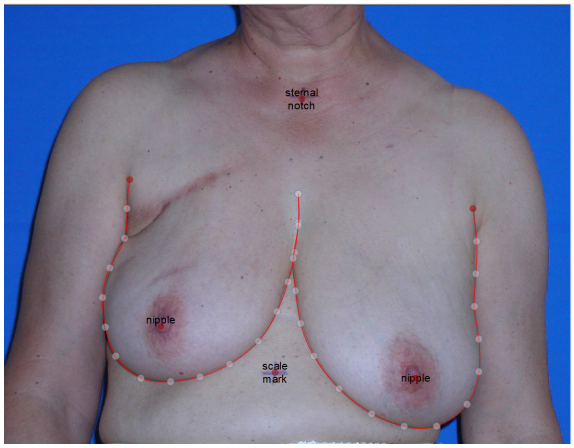


Figure 3.7: BCCT.core interface

3.2.3.3 BCCT.core

BCCT.core was introduced around the same time as BAT. It automatically classifies the aesthetic outcome of a BCCT, by assessing not only the asymmetry measures but also scarring and the colour difference between the two breasts. The user has to manually identify some keypoints and then the breast contour can be automatically adjusted. Then, the software automatically calculates 14 indices related to asymmetry, 8 to colour and another 8 related to the presence of scars. During training, an SVM model was used to find the best combination between all the measures. The final aesthetic evaluation classification is done according to the four classes in the Harvard scale [46].

3.2.4 Keypoint detection

In order to make the process more automatized, methods for the automatic detection of fiducial points (nipples, endpoints and contour), have been developed, with the goal of eliminating their identification in photographs by the user.

3.2.4.1 Breast Contour Detection

The research on automatic detection of the contour of the breasts in photographs was initiated by Cardoso and Cardoso [40]. Before that, breast contour detection, but on digital mammograms, had been investigated in several articles. In this first research, three approaches were mentioned: polynomial modelling, active contours and graph-based computation of the shortest path. However, only this last method showed satisfactory results with an acceptable computing time. This approach requires the manual insertion of the endpoints. Then, an edge detector is applied to the image, which enhances the contour, and the detection is based on finding the shortest path between the two endpoints, following the detected edge.

An extension of this work introduced prior knowledge, in the form of shape priors, with the aim of simplifying the localization of the breast contour [112]. Parametric (parabola, ellipse) and non-parametric (mask, unimodal) priors were tested and it was concluded that the introduction

of prior knowledge was beneficial, and that non-parametric models obtained better results than parametric models.

Another method for the detection of breast contour includes parametric active contour, proposed by Lee *et al* [89]. In this method, a mathematical shape constraint, based on a catenary curve, which previously has been shown to capture the overall shape of the breast contour reliably, is enforced on the image. These methods successfully detect the breast contour, but require the manual annotation of the endpoints.

3.2.4.2 Endpoints Detection

Following the work by Cardoso and Cardoso [40] related to contour detection, a method for the automatic detection of the endpoints [56] was investigated. This framework proposed finding the endpoints of the breast contour by assuming them to be the connection point between the arms' and the trunk's contour. However, due to the patients' position in the photos (arms-down), the contours of these regions are often indistinguishable. Thus, the endpoint is defined as being the highest point of the breast contour.

3.2.4.3 Nipple detection

The detection of the nipples is of utmost importance for asymmetry assessment, since many objective measures, such as BRA, use the location of the nipples. Udpa *et al.* proposed a method where normalized cross-correlation is used with a template bank of variants of Gaussian and Laplace of Gaussian filters; a probability map of likely nipple locations, determined from the database, is used to reduce the number of false positives. Finally, the nipples are considered to be the pixels within the region of interest that show the highest intensity [24]. Cardoso *et al.* [76] proposed a method that consists of detecting the nipple inside the areola area, which needs to be detected as well. Constricting the search for the nipple inside this area facilitates the process, since, although the breast surface is usually described as mainly featureless with the nipple being the most prominent feature, traditional feature detection algorithms may mistake other characteristics on the image as the nipple. In order to limit the search within the areola, candidates for nipples were over-detected. Then, a closed contour was found for each of the candidates. The best pair of closed contour and candidate, which represent the areola and the nipple, is chosen based on an SVM model that is trained with four high-level features: Harris corner quality factor, average magnitude of the directional derivative of the contour, shape of the contour and diameter of the contour.

Many other methods for nipple detection have been developed for mammograms, ultrasounds or obscene image detection, which are not practical for aesthetic evaluation.

3.2.4.4 Keypoint detection with Deep Learning

Deep learning methods allow the automatic extraction of task-specific, abstract and complex features from data, as it was explained in chapter 2. Deep learning has been rapidly replacing traditional machine learning algorithms for more challenging tasks, as is the detection of the breast's fiducial points.

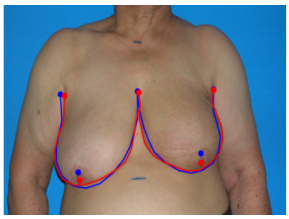
Moreover, the methods for detection of the keypoints reviewed in section 3.2.4 are separate, and information might get lost; this issue is may be solved with DL, since it follows an integrated learning approach that uses context information [62].

In [59], Silva *et al.* reported two methods, a deep and a hybrid one, that use deep neural networks for the task of detecting the fiducial points. One of the problems with DL is the fact that it requires a large amount of labelled data, which is not always possible in tasks related to health or biomedical engineering. In order to mitigate the effect of overfitting, the authors explored in this work the idea of learning an intermediate representation of the keypoints and an iterative process of refinement. The first module is what we call Heatmap Regression and Refinement. The intermediate representation was generated based on heatmaps, which are first obtained with a Gaussian kernel and then go through a regression algorithm with the U-Net model [119], which was initially proposed for biomedical image segmentation. The original images are multiplied by these heatmaps and then are fed to a regression module composed of three blocks: VGG16, four convolutional layers and three dense layers, for the final keypoint detection.

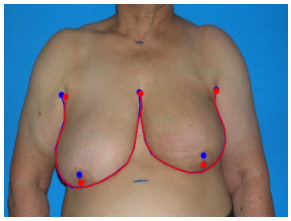
The DNN model obtained better results in all the detection tasks than the traditional methods, except for the contour detection in one of the tested datasets. In order to tackle this problem, a hybrid model was proposed, where the endpoints are detected with the DNN model and then the shortest path with shape priors algorithm [112] mentioned in subsection 3.2.4.1. This approach showed better results than the DNN model for the contour detection task. However, this hybrid model is not very efficient time-wise, and it requires more processing time than a deep learning model, which takes a while to train, but is fast to apply. In order to overcome this issue, Gonçalves *et al.* attempted and succeeded in the development of a deep learning method with better performance for the contour detection than the hybrid model. The method consists of the use of two DL models: a segmentation model and a keypoint detection model. Segmentation masks were generated with a U-Net++ model, which had been previously trained and fine-tuned. Then, the contours of the masks are extracted, since it is assumed that they contain points of the region of interest. The final contour points are determined based on a minimization of the distance between the mask contour and the predicted contour keypoints by the algorithm described in [59]. This method showed better results than the hybrid model, and is much more time efficient [57].

3.2.5 Aesthetic Evaluation with Deep Learning

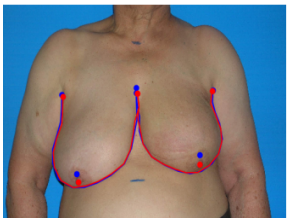
Given the evolution of AI algorithms towards deep learning, the next logical step in the breast aesthetic evaluation field is to improve the evaluation itself, resorting to deep learning methods.



(a) Deep model from Silva *et al.* [59]



(b) Hybrid model from Silva *et al.* [59]



(c) Deep model from Gonçalves *et al.* [57]

Figure 3.8: Keypoint detection examples with the algorithms from Silva *et al.* and Gonçalves *et al.*. Prediction is in blue and ground-truth is in red. Figures from [57]

Silva *et al.* [58] proposed the first deep learning method for aesthetic evaluation. Binary classification was considered (Excellent and Good vs Fair and Poor) due to the small size of the dataset. A deep neural network with a simple design (262,908 learnable parameters) was used in order to avoid intense overfitting, and with a "conv-conv-pooling" scheme. Intermediate supervision regarding the detection of the fiducial points, and the conversion of these keypoints into measures of asymmetry, such as LBC, BCE, UNR and BRA, was also introduced to decrease the effect of overfitting. Then, another CNN was trained for the simultaneous optimization of keypoint detection and aesthetic classification performance. For comparison of results, four SVM models, such as the one used in [46], were used, with different variations of inputs and kernels. The proposed method showed better accuracy and balanced accuracy results than the other SVM models.

Moreover, this work also presents a retrieval system, with a model that has the ability to retrieve similar cases, with the same or adjacent class (in the typical Harvard 4-class setting) and similar types of asymmetry to the test image being considered. This research will be incorporated into the work that will be developed during this thesis, with the goal of retrieving similar cases of past patients.

3.3 Summary

Breast cancer is the second most common cancer worldwide, affecting almost entirely women. In the past years, with the improvement of breast cancer treatments and the increased survival rate of this disease, there has been a new focus on the quality of life of the patient, in which the aesthetic outcome of the surgeries can play a major role. However, this focus on the aesthetic evaluation is hindered by the fact that a gold standard for the evaluation still doesn't exist. It is usually done by an assessment of the doctor and the patient, which although it is important, is not reproducible. To overcome this issue, many objective measurements that assess important aspects related to the physical outcome of breast cancer treatments, particularly the symmetry between the treated and the untreated breast, have been established. In an attempt to automatise this process, some softwares, which perform these measurements automatically and classify the aesthetic outcome of the procedures according to the Harvard scale, have also been developed.

However, these softwares still require the manual input of some keypoints of the breast. In an attempt to make the process even more independent, a lot of research on automatic detection of these keypoints, the endpoints, the contour, and the nipple has been done, with great results being achieved with deep learning.

Finally, a deep learning model which performs binary aesthetic evaluation has been proposed by Silva *et al.* [58] and this model will be used in some experiments of this dissertation.

Chapter 4

Literature Review: Content Based Medical Image retrieval

4.1 Content-Based Medical Image Retrieval

Content-based medical image retrieval (CBMIR) has been a huge, and rapidly growing area of research in medical image analysis.

Medical image interpretation consists in observing the medical image, interpreting its characteristics, using those findings to make a diagnosis and finally recommending the most suitable course of treatment [26]. There is a great potential for assistance in the interpretation and decision making process, not only due to the increased number of medical imaging exams being performed, but also because of possible lack of training in new physicians, fatigue, and in order to reduce inter-observer variations. Studying past cases can help doctors in the decision-making process, but, in the present era of big data, with the widespread use of digital images archives and communications systems (PACS) in hospitals, the size of medical data repositories is constantly increasing, making it hard for physicians to find the cases of interest.

For this reason, image retrieval systems can cause an extreme impact on a physician's workflow; in addition to enabling similarity-based indexing, they can provide extra information for the diagnosis or support the decision of a Computer Aided Diagnosis system, by retrieving similar past cases.

The retrieval of the medical images may be done according to text-based and/or content-based methods. With image retrieval based on text, the user needs to input specific descriptions that need to correspond to the annotation in the database, which is not very efficient, since it requires clinical expertise and time [71].

With content-based image retrieval methods, the images are retrieved based on their similarity with the query image, in terms of colour, texture, shape, and so on, which helps physicians save

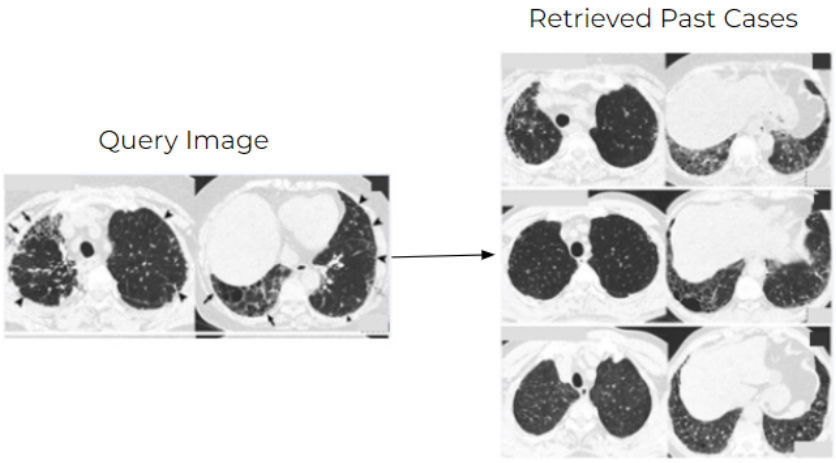


Figure 4.1: Example of a CBMIR system based on deep learning methods to assist in the diagnosis of interstitial lung disease. After the input of the query image, the three most similar images are retrieved. Image from [82]

time and do their search more efficiently. However, some of these features, such as colour, are not very useful in medical tasks, where many of the images are in grayscale.

CBMIR usually consists of two tasks: **feature extraction** and **indexing and image selection**.

4.1.1 Feature Extraction with Traditional Methods

In order to compare two images, it is necessary to extract their features first. Feature representation's aim is to obtain a low-dimensional but informative description of an image [123]. Traditional feature extraction methods include handcrafted and conventional computer vision detection methods, such as corner, line or circle detection. Depending on the case being studied, the most relevant type of features (colour, shape, blob analysis) are selected and analysed in the images. However, a combination of two methods can be used, and for example, the search for specific keywords can complement content-based querying, in order to make the search faster and more specific. Additionally, some methods use domain-specific knowledge to make this process more efficient, by using, for example, prior knowledge of physicians, labels or relational graphs to identify in the image the region of interest for the diagnosis.

4.1.2 Feature extraction with Deep Learning

Nowadays, however, deep learning methods are rapidly replacing traditional hand-crafted feature selection methods, especially in more challenging tasks [97]. Since deep neural networks are trained and not programmed, they require less expert analysis and provide more flexibility, as they can be re-trained with custom datasets, while conventional computer vision algorithms tend to be more domain-specific. For these reasons, deep learning solutions have been explored in order to minimize the main issue with feature representation, the well-known "semantic gap"

between low-level image pixels captured by machines and high-level semantic concepts perceived by humans. Since neural networks aim at mimicking the way the brain works, they seem to be a logical approach to this problem.

In some earlier studies, CNNs showed their potential as a useful tool for feature extraction. Since then, many approaches have been suggested for this task, such as deep autoencoders, Siamese networks, which exploit the distance between features of pairs of images, recurrent neural networks, generative adversarial networks and the use of attention networks, where saliency maps are used in order to decrease the effect of the background. However, most of the developed methods include the features learnt by the autoencoders and convolutional neural networks. With autoencoders, the latent space is used for feature representation, while with CNN, the images are represented by the activation space of the final layer, which corresponds to the penultimate layer, previous to the fully connected layer and previous to the classification [52].

Regarding supervision, supervised, unsupervised and semi-supervised techniques have been studied. Supervised learning usually generates better results, but sometimes it is hard to have a large labelled medical dataset. For this reason, unsupervised and semi-supervised learning, where labelled and non-labelled images are combined, can be used.

Moreover, feature representation from traditional and DL methods can be combined or fused together.

4.1.3 Image Selection

After obtaining the features of the input image, the next step is to identify the most similar past case image in the repository. In order to achieve this, it is necessary to measure the similarity between images, which is done by calculating the distance between feature spaces; the smaller the distance between feature spaces, the greater the similarity. Different distance metrics can be used, such as for example:

4.1.3.1 Per-pixel measures:

- Euclidean Distance: Simple and commonly used reference metric. N represents the dimension of vectors Q and T.

$$d_2(\mathbf{Q},\mathbf{T}) = (\sum_{i=0}^{N-1}(Q_i - T_i)^2)^{\frac{1}{2}} \tag{4.1}$$

- Manhattan Distance: The distance between two points is the sum of the absolute differences in their Cartesian coordinates.

$$d_1(\mathbf{Q},\mathbf{T}) = \sum_{i=0}^{N-1}|Q_i - T_i| \tag{4.2}$$

- Cosine distance: Difference in direction between vectors, irrespective of their norms. The distance is given by the angle between the two vectors.

$$d_{\cos}(\mathbf{Q},\mathbf{T}) = 1 - \cos\theta = 1 - \frac{\mathbf{Q}'\mathbf{T}}{|\mathbf{Q}||\mathbf{T}|} \tag{4.3}$$

- χ^2 distance: Statistical test that calculates whether two images are dependent on each other. Pearson's chi test is often used in radiology to compare categorical images (for example, comparing an image with images with or without disease).

$$d_{\chi^2}(\mathbf{Q},\mathbf{T}) = \frac{1}{2} \times \sum_{i=0}^{N-1} \frac{(Q_i - T_i)^2}{Q_i + T_i} \tag{4.4}$$

- Histogram intersection: The similarity is assessed by comparing color histograms.

$$d_{hi}(\mathbf{Q},\mathbf{T}) = 1 - \frac{\sum_{i=0}^{N-1} \min(Q_i, T_i)}{|\mathbf{Q}|} \tag{4.5}$$

- Quadratic Distance: Unlike the other metrics, where similarity is only taken into account for between each dimension, the quadratic distance considers similarity across dimensions. A is a matrix of similarity coefficients.

$$d_{qad}(\mathbf{Q},\mathbf{T}) = [(\mathbf{Q}-\mathbf{T})^T \mathbf{A^{-1}} (\mathbf{Q}-\mathbf{T})]^{\frac{1}{2}} \tag{4.6}$$

- Mahalanobis Distance: Special case of the Quadratic distance metric in which the similarity matrix is given by the covariance matrix of the feature vectors.

$$d_{mah} = [(\mathbf{X_Q} - \mathbf{X_T})\Sigma^{-1}(\mathbf{X_Q} - \mathbf{X_T})]^{\frac{1}{2}} \tag{4.7}$$

4.1.3.2 Perceptual measures:

Classic per-pixel measures are insufficient for assessing complex images, as they assume pixel-wise independence. For example, blurring an image causes large perceptual changes, but not big alterations in the per-pixel distance measures. In order to evaluate "perceptual distance", some methods, which aim at comparing images in a way that matches human judgement [136].

- HDR-VDP: visual metric that compares a pair of images and predicts the probability of an average observer to see the difference between the images and their quality

- Structural Similarity Index: measures the similarity between two images, taking into consideration structure, luminance and contrast

4.2 Case-Based Explanations

One of the main problems with some ML methods, especially deep learning, is their 'black-box' behaviour, which makes the model's nature and reasoning extremely hard to comprehend [108]. Despite the proven accuracy of deep learning algorithms, DL-based solutions are still not used on a greater scale in high-stake fields, such as medicine or the justice system, due to their lack of interpretability, as in these situations, it is mandatory to know whether the model's decisions are based on relevant factors or on confounding information. This characteristic hinders the use of ML algorithms in these fields, since it makes it hard for professionals to trust the decision of these algorithms in such high-stakes situations. This lack of acceptance has led to a rise in research on the topic of explainable and interpretable artificial intelligence in recent years. Although often used as interchangeable terms, many authors attribute slightly different meanings to interpretability and explainability; with an explainable model, the features and algorithms used for the model prediction can be explained in normal, 'human' terms [50]. A model is considered interpretable when a human can consistently predict its outcome, since he discerns the mechanisms and algorithms used in the model. But, all in all, the goal of xAI is to design models with both of these characteristics. And this is a goal of extreme importance, since even the General Data Protection Right (GDPR) from the European Union includes a right to explanation in its clauses.

Interpretability methods can be intrinsic, when models are inherently interpretable, or post-hoc, when explanations are generated after model training, usually by a second, independent model [94].

Inherently interpretable deep learning models, such as explainable deep neural networks (xDNN), are typically prototype (data instance representative of all the data in a class, selected manually to cover the centres of the data distribution)-based. The prototype is used to classify the new observation and is given as an example to explain the prediction [95] [109].

Post-hoc explanations consist of approximations, saliency maps or derivates, and require the development of a new model for these explanations, separate from the classification model used to classify the samples. Because of this, it is argued that these explanations might not be very reliable as they might not reflect the real reasoning behind a model's decision.

In the medical field, content-based image retrieval has the potential to not only be directly observed and analysed by doctors during the decision making process, but also be used as a provider for case-based explanations for ML models' outputs. The retrieval of similar past cases by the model can provide intuitive explanations and increase doctors' trust in the algorithms [67].

Case-based explanations consist of the retrieval of samples from the training data as an explanation, which can be **factual**, **counterfactual**, **typical** or **semi-factual**. Factual examples represent the data belonging to the same class as the new observation that is most similar to it, typical represent the prototype that is more similar, also in the same class, counterfactual examples are the most similar cases that do not belong to the same class as the current case and semi-factual the case closest to the decision boundary of the same class as the current case. These different types of examples can be used as explanations, according to the task [95].

Case-based explanations can be achieved with intrinsic or post hoc methods. The intrinsic approaches involve the design of models with case-based or prototypical reasoning while the post hoc methods search the data and retrieve the training sample with the most explanatory value.

4.2.1 *Post hoc* methods

A *Post hoc* interpretability method refers to methods where the explanation is retrieved or created after the algorithm's classification. *Post hoc* methods are usually associated with model agnostic methods, where the explanations are generated by a second model, independent from the classification one. This kind of methods has some advantages in comparison with intrinsic approaches, such as their flexibility and the fact that they can be used in already trained models, without the need to change their architecture and possibly compromise their performance [94]. However, many authors claim that *post hoc* explanations don't reflect the true reasoning behind a decision since the explanations for a certain classification can change based on the model used and given that many different explanations can be created to justify how the network classifies a data instance without any of them representing the correct reason for why the object was classified that way [101]. Moreover, *post hoc* models are susceptible to adversarial attacks [53].

4.2.2 Traditional Machine Learning

Post hoc explanations in machine learning can involve, as explained in [41], using the trained model, in this case a decision tree, as a distance metric for k-nearest neighbor to retrieve similar cases. In a decision tree, these similar cases refer to data instances with zero "decision tree" distance to the query image, or, in other words, data samples that are classified in the same leaf node. This method of using the trained model as a distance metric can also be used with ANN or other frameworks.

The Explanation Oriented Retrieval (EOR) [39] method is another framework used to generate case-based *post hoc* explanations that tries to retrieve samples based on their utility, not on their similarity to the query image. First, the new observations are classified with a K-NN algorithm. Then, the nearest neighbours are evaluated according to an utility measure based on the task and reordered. Finally the most informative case is retrieved.

4.2.3 Deep Learning

As mentioned in this chapter, an easy and intuitive way to explain the outcome of a DL model is to retrieve the most similar data instance with the same classification as the new observation. ANN-CBR Twin-Systems [84] are a hybrid system that combines content-based reasoning and neural networks, in which the ANN is mapped into a 'white-box' CBR model, with both using the same dataset, as pictured in the example of figure 4.2. The ANN is analyzed to discover the feature weights that contributed to the final decision, and then these are used by the 'twin' model to retrieve a nearest-neighbour case that justifies this decision. The quality of the extracted weights is

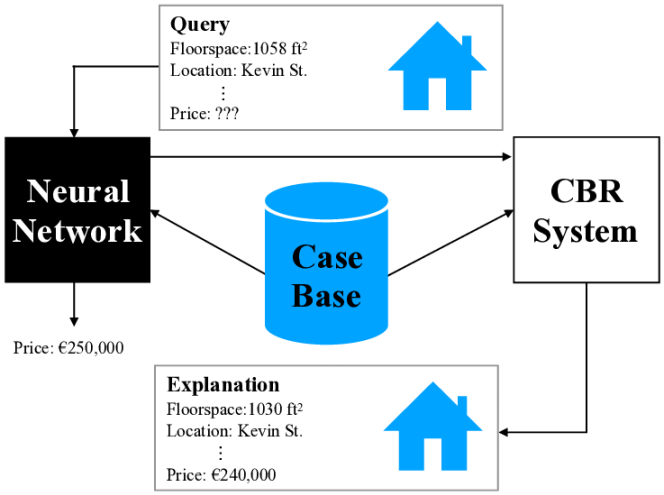


Figure 4.2: Example of an ANN-CBR twin system for the prediction of house prices, from [84]

analysed through a number of methods, including the novel COLE (Contributions Oriented Local Explanations) technique.

An adapted version of the famous interpretability method LIME (Local Interpretable Model-Agnostic Explanation) with content based reasoning was developed by Recio-Garcia *et al.* [114]. LIME is a *post hoc* local interpretability method, meaning that it explains each outcome individually instead of explaining the global model, that modifies the query input and evaluates the impact of this alteration on the final classification. Given a new query image, the CBR-LIME approach searches for similar images in the database using the SSIM, retrieves K of these images and their configurations are used to generate an explanation with LIME.

The Interpretability-guided Content-Based Image Retrieval (IG-CBIR) was originally developed for medical image retrieval by Silva *et al.* [123]. The *post hoc* explanations are retrieved from similar examples from the dataset. However, since the task is related to the medical field, the similarity between images is calculated by using the image regions that are medically significant to the classification, which are obtained with saliency maps.

4.2.4 Intrinsic Methods

Intrinsic interpretability methods are, by definition, model-specific, as interpretability is achieved through the model's design and the techniques that are used are specific to the architecture in consideration. Since intrinsic methods require alterations to the model's architecture, they might compromise the model's performance [51].

4.2.4.1 Traditional Machine Learning

One of the most well-known intrinsically interpretable ML models is the K-nearest neighbours, where observations are classified according to their distance to labelled training samples. The observation is then classified based on the majority of the labels of its k-nearest training samples.

K-NN can be used to classify images by computing the distances between feature vectors, which can be obtained with DL or ML methods. Decision trees also have an interpretable architecture, as long as they are relatively short [94], which means that they lose interpretability as the task gets more complex, which makes them less suitable for image classification tasks.

Other methods include Bayesian models such as Latent Dirichlet Allocation [36] or the Bayesian case model [31]. Both methods can be used to cluster data points. LDA clusters data in an unsupervised way, while the Bayesian case model is prototype-based and provides typical examples.

4.2.4.2 Deep Learning

A deep version of the K-NN algorithm has been developed by Papernot *and al.* [105] and, such as the traditional K-NN, it is a highly interpretable model, that also shows high accuracy.

Prototype-based methods are often used to create inherently interpretable neural network. Prototypes are data instances that cover the centre of the data distribution and are representative of all the data from a certain class. Prototype-based networks can be achieved in a number of ways, some of which will now be reviewed.

Li *and al.* [101] developed an architecture that uses an autoencoder, and the distances between the samples and the prototypes are calculated in the latent space. In the autoencoder, the encoder allows the creation of the low-dimensional feature space, while the decoder allows the visualization of the learned prototypes. Calculating distances in the latent space admits the use of more suitable distance metrics other than the more traditional ones, such as Euclidean distance.

In [109], the authors developed an Explainable Deep Neural Network (xDNN), based on prototypes. These prototypes are derived automatically by selecting training samples that represent local peaks of the empirical data distribution, called typicality (figure 4.3). The samples are then classified based on the most similar prototype's class. The authors of this study later introduce the Deep Machine Reasoning model [29], an improved xDNN that includes the use of a Decision Tree to determine the final classification and balances the classes by synthesising data with interpolation between samples around the prototypes determined from the available training data. DMR shows higher accuracy values than the original xDDN.

However, the prototype-based logic can also be applied to parts of the image, and not globally. ProtoPNet [38] is an attention-based interpretable model, that classifies images by identifying segments of it and comparing them to prototypical part of some class, as it can be seen in figure 4.4. In the end, the prediction is based on a weighted combination of the similar prototypes between the different parts of the image. The segmentation performed in this network by finding identifiable characteristics and the combination of evidence from the different prototypes matches rather similarly the reasoning behind the way a human would perform a challenging image classification task.

The Hierarchical Prototype based-model [65] is a network that decomposes complex problem into a series of smaller, simpler ones, by aggregating prototypes in a pyramidal hierarchy form, where the specificity and detail of prototypes increases as the levels decrease. This architecture is a promising interpretability solution, especially with high-dimensional and complex problems.

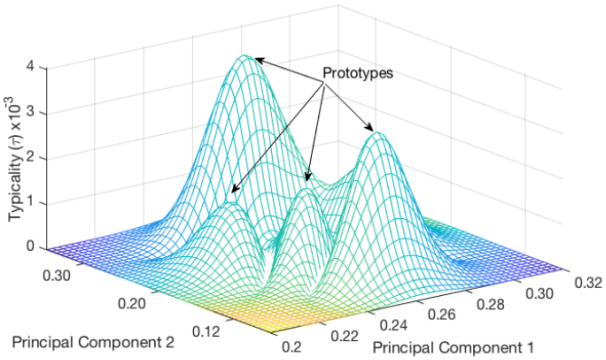


Figure 4.3: Prototype selection, from [109]

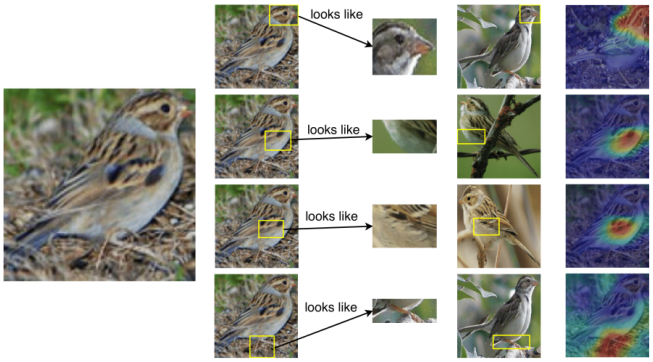


Figure 4.4: Classification of an image of a clay colored sparrow based on learnt prototypical parts

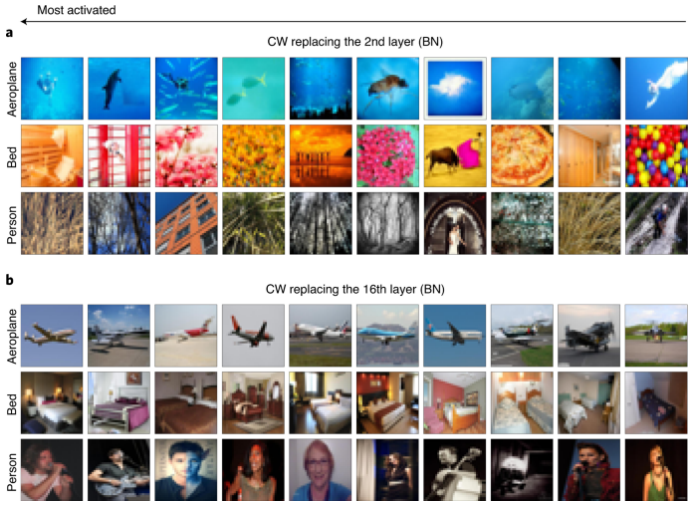


Figure 4.5: Top-10 Image activated on axes representing different concepts, with CW replacing the second (a group) and sixteenth (bgroup) layer

However, sometimes prototypes are not enough to describe data, especially for real-life, complex problems. In order to increase the explainability of case-based networks, Kim *et al.* [32] added criticisms, which are data instances that are not well represented by the prototypes. The final classification is assigned according to the closest prototypes, but the retrieval of criticisms, along with the prototypes, to explain the decision helps humans understand it.

The Concept Whitening method [135], developed by Chen *et al.*, introduces a mechanism that can be used in a given layer of a Neural Network to disentangle its latent space and align its axes with the known concepts of interest. This approach makes it easier to figure out the relations between the input features and the final classification, and understand how the networks gradually learns concepts over the layers (figure 4.5).

Intrinsically interpretable models can also be based on monotonicity. A monotonic model is a model that has a set of features (monotonic features) whose increase always leads the model to an increase in the probability of the outcome being the positive class. Silva *et al.* [132] developed a classification model that is inherently interpretable due to monotonicity restriction, but that generates explanations after the decision, by searching the closest samples of the same and opposite class of the new observation in the latent space to serve as an example and counterexample. In this algorithm, the monotonic and non-monotonic features are separated and serve as an input to two different networks. Then, both of these networks are concatenated into a new monotonic network, which creates a latent space where all the features are monotonic and where the search for examples is conducted.

4.3 Summary

When a doubt emerges during the diagnosis process, doctors typically search in databases for previous similar cases that could help them in their decision-making. However, given the increasing

size of digital medical repositories, this search can be extremely time-consuming and represent a massive burden to the physician's workflow. In the specific cases when the search is for images, content-based medical image retrieval (CBMIR) systems can be used to facilitate the process, and a lot of research on this topic has been conducted. CBMIR usually consists in two tasks: feature extraction and image selection. Nowadays, deep learning methods are favoured for the feature extraction. The latent representation of the image is then compared to the images in the repository, and the most similar one(s) are identified and retrieved.

However, content-based retrieval is not only useful in the medical field, but it can also be used to make machine learning models more interpretable. Interpretability in AI is a rising field of research, since the use of ML models in real-life situations, especially in high-stake fields, is hindered by their black-box behaviour and consequential lack of interpretability. Content-Based retrieval can be used to retrieve samples from the training data as an explanation to the model's output. These explanations can be obtained with intrinsic or post hoc interpretability methods.

In the context of this dissertation, content-based image retrieval will be used to identify the most similar case in the dataset, which will be used to create the morphed image.

Chapter 5

Literature Review: Deep Generative Models

5.1 Deep Generative Models

Deep generative models are a class of deep learning methods that model the probability distribution of the data and generate new samples from the learned distribution. Generative modelling has many direct applications including image synthesis for the purpose of data augmentation for example, resolution improvement, and attribute manipulation, among others [37]. All in all, the ultimate goal of deep generative models is to synthesize new samples with high-quality, which can be achieved through a number of approaches, each of which with its own advantages and disadvantages. In this section, four common approaches to deep generative models will be reviewed: Autoencoders (AE), Normalizing Flows, Autoregressive Models and Generative Adversarial Networks (GAN).

5.2 Autoencoders

Autoencoders are a type of feed-forward neural networks consisting of an encoder and a decoder (figure 5.1), that learn data encodings in an unsupervised manner. The encoder is a set of convolutional blocks that has the job of compressing the input image(s) into a low-dimensional latent space. Then, this reduced feature representation is fed into the decoder, consisting of a series of upsampling modules, that tries to reconstruct the original input from it [34]. The autoencoder is trained to learn the best encoding-decoding scheme using an iterative optimisation process, which aims at minimizing the difference between the output of the decoder and the original input [118].

After training, autoencoders can be used to perform large-scale non-linear dimensionality reductions and to denoise images, if the model is trained with noisy images as input and trained to reduce the difference between the decompressed latent space from the decoder and a noise-free

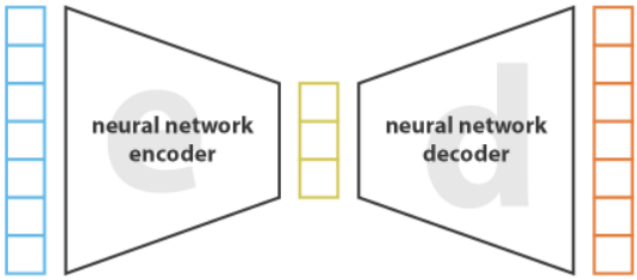


Figure 5.1: Illustration of the architecture of an autoencoder, from [118]

image [48]. However, simple autoencoders do not have the ability to produce new content based on the input data, only to generate new samples that are approximated to it.

Typically, in the latent space created by the encoder, the distribution of the real data is sparse, which means that it is difficult to sample a realistic latent vector, with information to reconstruct an informative and realistic data sample. In a regular, "well-organized" latent space, a random vector could be randomly used to decode into new content, that was similar to the input but not a reconstruction. The *variational autoencoder* tackles this limitation, by providing a probabilistic way to describe an instance in the latent space, which regularizes the latent distribution, and enables the generative process. So the difference between a VAE and a regular autoencoder is that, in the VAE, the input is encoded as a distribution over the latent space, with n means and n variances, not as a single point [115]. From this distribution, a point is sampled and decoded. Then, the loss is computed and backpropagated through the network.

The loss function in a VAE is composed of a "reconstruction term", where the similarities of the input and output are evaluated, just like in the autoencoder, and a “regularisation term” on the latent layer, which uses the Kullback-Leibler divergence to measure the distance between the distribution returned by the encoder and the original data distribution [118].

Since VAEs are probabilistic models, they often generate blurry images. In order to tackle this issue, the VQ-VAE-2 [113] model was proposed. This model combines the architecture of a Vector Quantized Variational Autoencoder (VQ-VAE), originally proposed by Oord *et al.* [130] and an autoregressive model as prior. The encoder and decoder architectures are simple as in [130], but

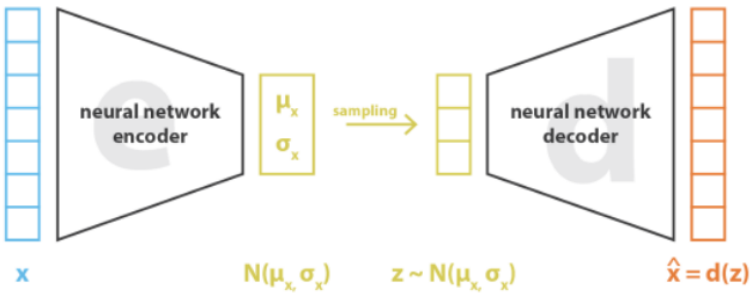


Figure 5.2: Illustration of the architecture of a variational autoencoder, from [118]

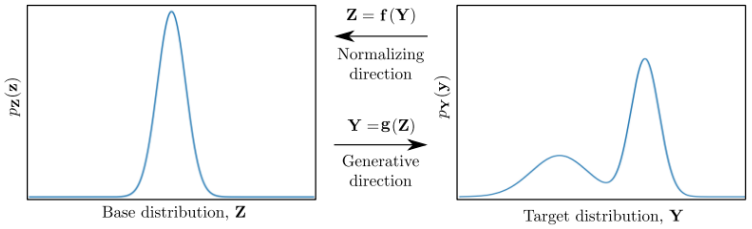


Figure 5.3: Exemplification of a normalizing flow between a base and a target distribution. There exists an invertible function g, such that p(Y) = gp(Z)

hierarchical multi-scale latent maps for increased resolution were used.

The Nouveau VAE (NVAE) [128] is a model proposed by Vahdat and Kautz that achieved state-of-the-art results amongst non-autoregressive likelihood-based models. The NVAE is a deep hierarchical VAE that uses depth-wise separable convolutions in the decoder and batch normalization in the encoder. During training, the model is stabilized by spectral regularization. Moreover, NVAE requires less memory during training, making it more time efficient.

5.3 Normalizing Flows

Normalizing flows are generative models that transform a simple probability distribution such as a Gaussian into the complex data distribution of a dataset. In order to achieve this, the model applies a series of simple, invertible and differentiable transformations to the real distribution of the data. This method allows the creation of tractable distributions with efficient and exact sampling [37].

Different kinds of flows can be used to model this type of network; Element wise, linear, planar and radial flows are simple but limited examples, as they show many drawbacks. Residual flows are reversible networks that use residual connections and can be viewed as a discretization of a first-order ordinary differential equation. This architecture allows an unbiased estimate of log-likelihood and its training is memory-efficient. Infinitesimal or continuous flows are methods that attempt to model the complete continuous dynamical system instead of considering the discrete approximation like in the residual flows architecture [88]. Coupling and autoregressive flows support non-linear transformations and are the most widely used architectures. Both of these methods require coupling functions to apply these transformations with high expression power. In the coupling flows architecture, given a disjoint division of the data x in two subsets (xA, xB) and an invertible function h(·, θ), where θ, which is known as a conditioner function, can be any arbitrary function whose only input is xB, the normalizing flow can be defined by the equation 5.1.

$$\begin{cases} y^A = h(x^A, \theta(x^B)) \\ y^B = x^B \end{cases} . \tag{5.1}$$

The advantage of this method is that Θ can be a very complex, and is usually modelled as a

Figure 5.4: Synthetic images generated with the Glow model [86]

neural network, since its inverse doesn't need to be computed in order to obtain the inverse, normalizing function of the flow. Some examples of models that use coupling flows are realNVP [49], GLOW [86] and Flow++ [70]. RealNVP uses real-valued non-volume preserving transformations for density estimation. The conditioner function consists of a scale and a shift transformation. Glow is a simplification of realNVP, where the reverse permutation operations with invertible 1x1 convolutions. In Flow++, the conditioner in the coupling layers is a CNN.

In an autoregressive flow, the normalizing flow is framed as an autoregressive model, where each output only depends on data observed in the past. The flow is conditioned on the previous entries of the input, meaning that the input of the conditioner function is the past observations (equation 5.2).

$$y_t = h(x_t, \theta_t(x_{1:t-1})) \tag{5.2}$$

The disadvantage of this architecture is that, since the flow is sequential, the inverse function is also a sequential operation and thus cannot be parallelised, which makes it hard to implement efficiently.

Classic autoregressive flow models are the Inverse Autoregressive Flow (IAF) [87] and the Masked Autoregressive Flows (MAF) [104]. In IAF, the flow is conditioned by the previous outputs y, not by the past inputs. The MAF model is based on the idea that it is possible to computer all the entries of the flow using a single network with appropriate masks.

5.4 Autoregressive models

Autoregressive models are used to generate sequential data, since their output is conditioned by the previous inputs. Autoregressive models are extremely useful to generate speech, text and video, but also images [37].

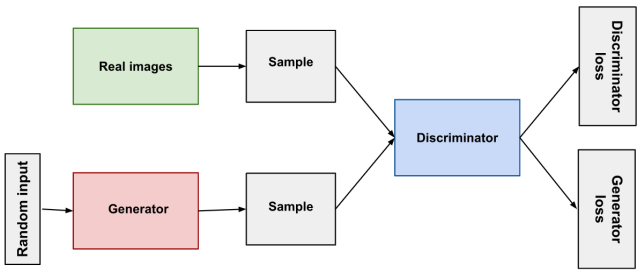


Figure 5.5: Illustration of the architecture of a Generative Adversarial Network, introduced by Goodfellow *et al.* Image from [13]

Thinking of an image as a sequence of pixels, an autoregressive model can be used to generate images where the intensity of a certain pixel is determined based on the values of the past pixels, and the distribution of the image is the combination of the probability of all its pixels. Autoregressive models turn the distribution modelling problem into a sequence problem. These models allow extremely powerful density estimation but, due to the fact that this process is sequential, it can be very slow in high-dimensional data.

PixelCNN [102] is an autoregressive model used for image generation that outputs one pixel at a time in an image along its spatial dimension. In the convolutional layers of this network, the model is able to learn the distribution of all the pixels in the image. Thus, all the pixels around the central pixel are considered for the computation of its output intensity. However, masks are then applied in order to block the use of the information of the following pixels, making the final output only dependent on the previously predicted pixels. PixelCNN matched state-of-the-art performances of PixelRNN [129] with a great reduction in computational cost. PixelCNN+ [122] is an improvement over the PixelCNN, which uses a discretized logistic mixture likelihood on the pixels, among other modifications, to improve the performance of the model.

5.5 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a class of generative models that can be used to generate incredibly realistic synthetic images. This type of network represented a breakthrough in the deep generative models' field and, as stated by Meta's Chief AI Scientist and the "godfather of deep learning" Yann Lecun represents "the most interesting idea in the last 10 years in ML" [93].

The GAN model architecture, illustrated in figure 5.5 was first introduced by Goodfellow *et al.* [64] and involves two sub-models based on neural networks: a generator G, that is responsible for estimating the data distribution and generating new samples from the problem domain, and a discriminator D, which is a classifier model that estimates the probability of a sample coming from the training data or the generator. The two sub-models are trained in an adversarial manner, corresponding to a minimax two-player game, where G is trained to maximize the error rate of D, by "tricking" the discriminator into classifying the generated synthetic images as original from the training set.

Figure 5.6: Example of images produced by StyleGAN [83]

So, essentially, GANs learn how to capture the distribution of the data implicitly through indirect training by the discriminator, which forces the generator to create images as similar to the training set as possible [96]. With GANs, the goal is to get a model that approximates the input distribution, while VAEs attempt to model this distribution from a latent space. This is one of the reasons why GANs are generally able to generate more realistic samples in comparison to VAEs.

As seen in figure 5.5, the generator's output corresponds to the direct input of the discriminator. The sub-models are trained independently: during the training phase of the discriminator, the generator is kept constant and vice-versa, otherwise the generator would be trying to hit a moving target and might never converge [13]. Through backpropagation, the discriminator's classification is used to update the weights of the generator.

GANs can be used for image synthesis, more specifically for generating human faces or increasing the resolution of an image for example, image-to-image translation, image editing and for the generation of cartoon-like figures [133]. Since the original GAN, many architectures have been developed. One of the most important architectures is the DCGAN, which stands for Deep Convolution GAN, which introduced some changes to the original architecture, such as batch normalization, transposed convolutional layers in the generator and the use of the Adam optimizer. Most of the new GANs' architectures are based on the DCGAN [63].

Some important GAN variations include **Progressive GANs**, where the computational cost is reduced, since the generator's first layers produce very low-resolution images, and their quality and details are only incorporated in the last layers. This approach is used for example in the Lightweight GAN [90], which showed results similar to the state-of-the-art with a small dataset and reduced computational cost. StyleGAN [83] is also a model that uses the progressive approach and is able to synthesize very large and high-quality images, as seen in figure 5.6.

In **Conditional GANs** the input data is labelled or includes extra information, which conditions the generation. For example, a classic GAN trained on the MNIST dataset, which is a large database of handwritten digits, would generate random digits, while the conditional GAN would allow only the generation of a specific digit. The Pix2Pix [74] GAN is an example of a conditional GAN used for image-to-image translation tasks, such as transforming maps into satellite photographs or colouring black and white photographs. In this model, the discriminator is provided with the source image and the target image, and must assess whether the generated image is a plausible transformation of the source image into the target's domain.

Figure 5.7: Example of a cyclic image translation from horse to zebra by a CycleGAN. The first image is the original, while the middle one represents its transformation into the zebra domain. The last image depicts the translation of the middle image back into the horse domain, illustrating the idea of cycle consistency. Image from [138].

CycleGANs are also used for unpaired image translation tasks, and are trained to transform images from one domain into images that are more likely to belong to a second domain (figure 5.7. The training set has to include the two different sets of images, but no labels or correspondence between images is required [13, 138]. They can almost be seen as a build-up on the Pix2Pix model, with the difference that, in CycleGANs, two generators and two discriminators are trained simultaneously: one generator takes images from domain A and generates images belonging to domain B, while the second generator does the opposite. A key concept behind the power of CycleGANs is cycle consistency, which regulates the translations. Cycle consistency enforces that a new image generated by the first generator, which for example translates images from domain A to B, can be used as input to the second generator, which is being trained to translate images from domain B to A, and the output of this second generator should match the original image. Cycle consistency in CycleGANs is implemented by an additional consistency loss that compares the output of an image that goes through both generators with the original image.

5.5.1 Common problems with GANs

Despite their incredible high-quality results, the fact that GANs must combine the training of two models leads to various problems and often makes training unstable. Some of the common problems associated with GANs include mode collapse, vanishing gradients and failure to converge [133, 13].

Mode collapse happens when the GAN fails to produce diverse outputs and instead generates the same image, or very similar images, for different inputs. This may happen if the generator outputs a very plausible instance, that easily fools the discriminator. Since the generator is "competing" against the discriminator, it will start to only output that sample, and the system will over-optimize on that one output.

Vanishing Gradients is a problem that occurs when the gradients that are being backpropagated are very small, close to zero, which prevents the update and optimisation of the weights of the network. In GANs, this usually happens when the discriminator is very accurate at classifying the synthetic samples from the generator as fake. This leads to a near null gradient of the loss of the gradient, and the training of the generator fails as it doesn't have enough information to progress.

As the generator improves with training, the discriminator performance gets worse, and its feedback gets less meaningful over time. For this reason, it is important to stop the training before the point where the discriminator is just giving random feedback and the generator's performance starts to decrease. Because of this, **convergence is a fleeting and unstable state**. This convergence state is achieved when a Nash Equilibrium is found between the discriminator and the generator [63]. The Nash equilibrium is a state in game theory where the elements or players in the game have nothing to win by changing their strategy, meaning that a change in their actions does not change the outcome of the game. In GANs, Nash equilibrium is achieved when both the discriminator and the generator do not change their behaviour despite the actions of the other network.

A lot of research in GANs is focused on mitigating these common problems. The Wasserstein GAN (WGAN) was proposed by Arjovsky *et al.* [30] to fix mode collapse, prevent vanishing gradients and stabilize the network. In WGAN, the discriminator is known as a critic, and instead of classifying the images as real or fake, it classifies their level of "realness", where images that are considered real (belong to the training set) receive a higher score. Moreover, the critic and generator are trained with the Wasserstein distance as a loss. The Wasserstein distance, also known as Earth Mover's distance, measures the distance between two probability distributions. It is called Earth Mover's distance because it can be interpreted as the amount of energy it takes to transform a probability distribution into another [43]. The Wasserstein distance is continuous and differentiable in its domain, meaning that the critic can be trained until optimality [30].

Moreover, the critic model weights are clipped in order to be constrained to a limited range and enforce a Lipschitz constraint. Despite the improvements in training stability brought by the WGAN, the model still failed to converge sometimes, which is probably due to the weight clipping. In [66], weight clipping is replaced by a gradient penalty, which showed better results.

Another field of research of interest in GANs is the means of evaluation of the generated images. Despite the popularity of GANs, the evaluation and comparison of the generated images is still a challenging task. Earlier studies included only subjective assessments, but objective evaluation methods have also been developed, such as the Inception score (IS) and Fréchet Inception Distance (FID), which are the most widely used objective metrics [133].

5.6 Summary

In this chapter, a review on deep generative models, a class of deep learning methods that are able to generate new samples from a learned probability distribution of a dataset is presented. In this review, some of the most common deep generative models are introduced, with a focus on Generative Adversarial Networks, which have achieve state-of-the-art results.

A Generative Adversarial Network, also commonly known as GAN, is a model composed by two sub-models, a generator and a discriminator, in which the generator is responsible, as the name suggests, for generating new samples based on the data distribution, while the discriminator is a classifier that labels the samples as real (original from the dataset) or as fake (synthetic

images generated by the model). The two models are trained an adversarial manner, and the generator is trained to maximise the error of the discriminator, into a point where it is longer able to differentiate real or synthetic images.

Inside the class of GANs, many different methods with different architectures, goals and characteristics exist, such as the Progressive, conditional, cycle or Wassertein GANs. The Pix2Pix model, a type of conditional GAN, and the cycleGAN are two models used for image-to-image translation, a task where images from one domain or class suffer a transformation and, while maintaining their basic structure, incorporate the style of images from a second domain. Given that one of the goals of this project is to morph the characteristics of two patients, one who hasn't started her treatments and one after surgery, these two methods were explored as an approach to this goal. Moreover, some brief experiments with the Wasserstein GAN, a more stable version of the original GAN, were also conducted.

Chapter 6

Experiments with paired-image Translation

As mentioned along chapter 1, this dissertation aims to explore deep generative methods for the ultimate goal of generating biometrically morphed-images that represent the probable aesthetic outcomes of breast cancer treatment. In other words, the aim is to perform a first exploratory analysis of methods that, given an image of a patient, before cancer treatment, and an image from a past patient after treatment, can generate a morphed image of the current patient with the aesthetic result of the patient in the second patient. In these experiments, given the fact that all the images in the dataset are post surgery, alterations will be introduced to the images in order to either worsen or improve the aesthetic evaluation, and in this way create a new version of dataset that is meant to represent either the pre or post operation images. The generated image must be realistic, contain the physical characteristics of the current patient and accurately portray the possible physical alterations of certain medical procedures.

In order to accomplish this goal, deep generative adversarial models, namely the CycleGAN, conditional GAN (Pix2Pix model) and a Wasserstein GAN were explored.

However, before exploring GANs, some experiments with simpler generative models were performed, consisting of transformations between paired images. The goal of these experiments was to assess whether it was possible for a generative model to either introduce or correct two typical physical alterations as a consequence of breast cancer treatment: breast asymmetry, probably the most important and obvious measure for the aesthetic evaluation, and changes in the colour of the breast due to radiotherapy. In order to explore these topics, geometric and colour-related alterations were introduced to the images. In this chapter, the methodology, discussion and main conclusions of these experiments will be presented, as well as the dataset used in this project.

6.1 Dataset

The dataset used throughout this dissertation includes 143 photographs of women who had undergone breast cancer conservative treatment. These images were selected from two datasets, PORTO

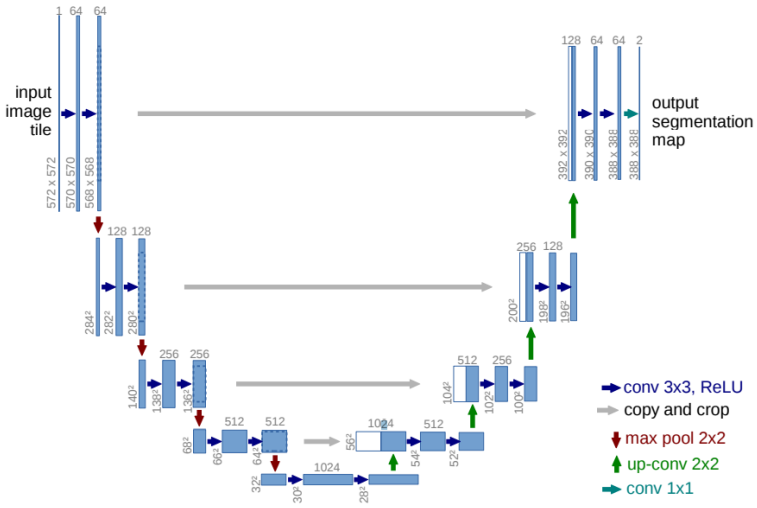


Figure 6.1: U-Net architecture, from [119]

and TSIO. From the PORTO dataset, 113 images were used, which correspond with the 113 images that were given a consensual aesthetic evaluation in [78], and the remaining 30, also evaluated by an experienced breast surgeon, are from TSIO. Both datasets show women in a frontal position, with their arms down and a clean background, which can be dark or bright blue. The images have a 256x384 dimension and have three-channel RGB nature. Moreover, the keypoints obtained in the methods described in [57] and [59], which include 37 coordinates of the breast contour, the position of the nipples and of the sternal notch, exemplified in figure 6.3 were also used.

6.2 U-Net Model

For this task, the U-Net architecture [119] was used. The U-Net was originally developed for biomedical image segmentation, becoming the state of the art for this kind of task. This network has an encoder-decoder architecture, where the encoder or contracting path is a stack of convolutional and max pooling layers and the decoder or expanding path is composed of transposed convolutions as well as regular convolutions. At each step of the expanding section, skip connections are used, and the result of the transpose convolution operations are concatenated with the features from the encoder at the same level and go through two consecutive convolution layers. The two paths are almost symmetrical, and the levelled skip connections make this architecture resemble an 'U' shape, as seen in figure 6.1, hence the name U-Net.

Figure 6.2: Example of images from the dataset used in this work

Figure 6.3: Example of breast keypoints considered in this work

6.3 Image Denoising

Before training the model for the task of correcting image distortions, a simpler task of image denoising was completed, in order to understand the potential of the U-net and which loss functions could be better to reach our goal.

For this, Gaussian noise, centred on zero and with a standard deviation equal to 1 and multiplied by a noise factor of 0.3, was added to the original images, as can be seen in figure 6.4.

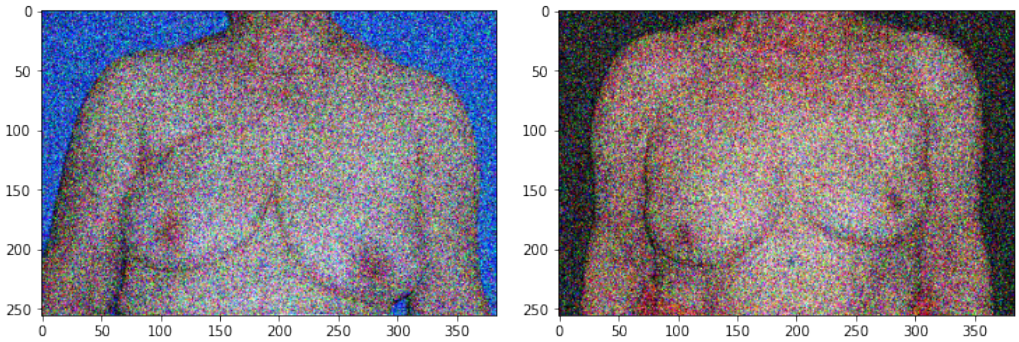


Figure 6.4: Examples of images with Gaussian Noise

Then, a U-Net (architecture in figure 6.5) was trained to approximate the noisy images to the original ones, in hopes of obtaining a denoising model. Of the 143 images, 20% were used for testing, 16% for validation and 64% for training; MSE and SSIM were tested as loss functions, and the models were trained until there was no training loss improvement after 15 epochs. Some final results can be seen in figures 6.6 and 6.7, respectively.

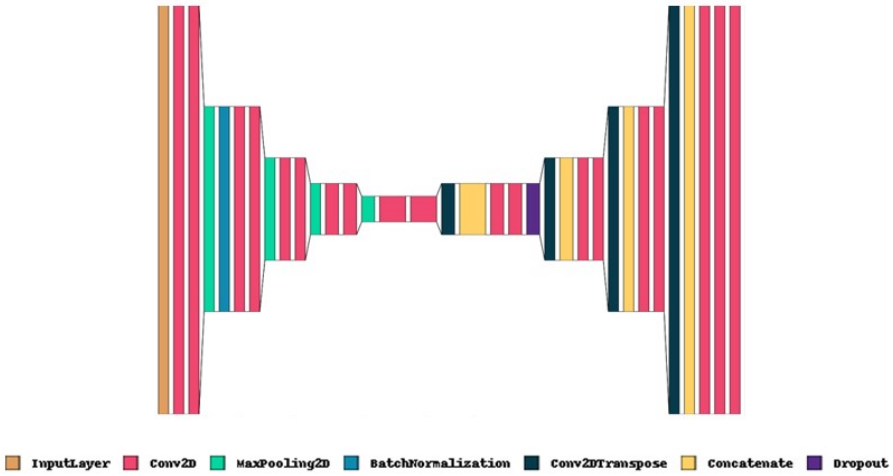


Figure 6.5: U-net architecture used in this project

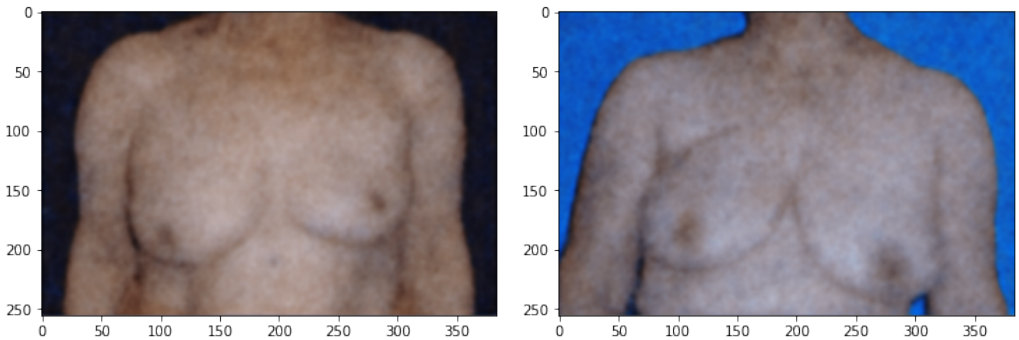


Figure 6.6: Image Denoising with MSLE loss

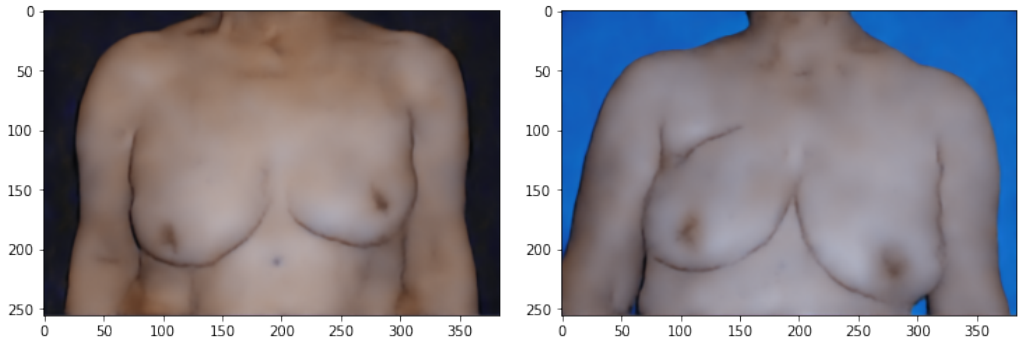


Figure 6.7: Image Denoising with SSIM loss

As it can be seen, the U-Net architecture is useful for the task of image denoising, and shows very satisfactory results in this task, especially with the SSIM loss. In order to point out the power of the U-Net and the importance of skip connections, a simple autoencoder, with the architecture represented in figure 6.8, was trained for the task of image denoising with the SSIM loss, with equal conditions to the U-Net. Two results can be seen in figure 6.9, and it is safe to say that the U-Net provides much better results.

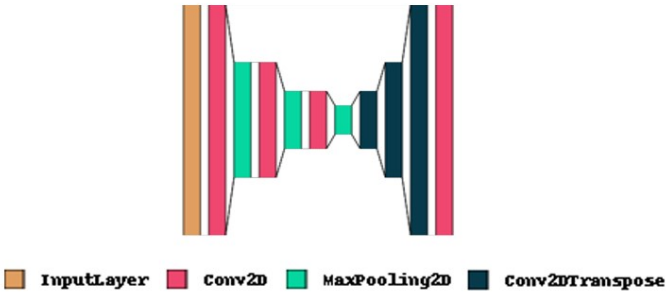


Figure 6.8: Autoencoder architecture

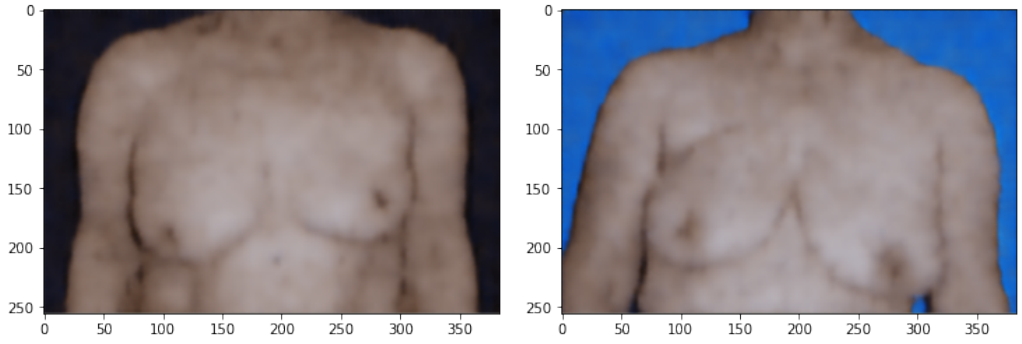


Figure 6.9: Image Denoising with an Autoencoder and SSIM loss

However, an important aspect of aesthetic evaluation is the observation of the breast contours, which got a bit blurred during the denoising process. In an attempt to try to better preserve the contours a modified loss, which took into account the similarity between the target and generated

images, but also the similarity between their edges. At first, two different losses were tested: using SSIM in the images domain and Mean Squared Logarithmic Error (MSLE) (equation 6.1) in the image's gradient and using MSLE in both figures (equation 6.2). The image error accounts for 90% of the loss, while the other 10% correspond to the error between their edges. The loss functions used are the ones in the formulas 6.1 and 6.2. The image gradient was obtained using a Sobel filter.

$$\mathscr{L} = 0.9 \times SSIM(image_{target}, image_{generated}) + 0.1 \times MSLE(edges_{target}, edges_{generated}) \quad (6.1)$$

$$\mathscr{L} = 0.9 \times MSLE(image_{target}, image_{generated}) + 0.1 \times MSLE(edges_{target}, edges_{generated}) \quad (6.2)$$

In figures 6.10 and 6.11 it is possible to see two examples of the results of this operation. It can be concluded that, for the SSIM loss, the addition of the penalty between the images' gradients was not useful. However, when we compare the images in figures 6.6 and 6.11, it is possible to see that the addition of the second loss component helped the preservation of the breast contours, with the cost of losing some realism. All in all, this exercise was important to understand the potential of the U-Net and to explore possible loss functions to use in the following tasks.

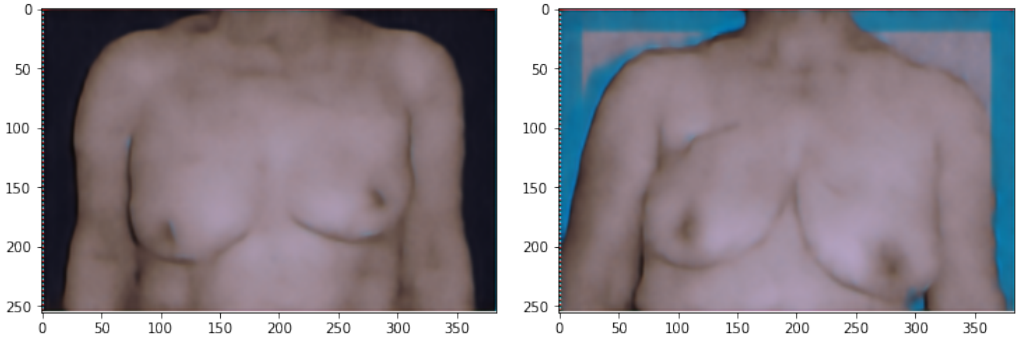


Figure 6.10: Image Denoising with SSIM loss in the image domain and MSLE in the gradient domain

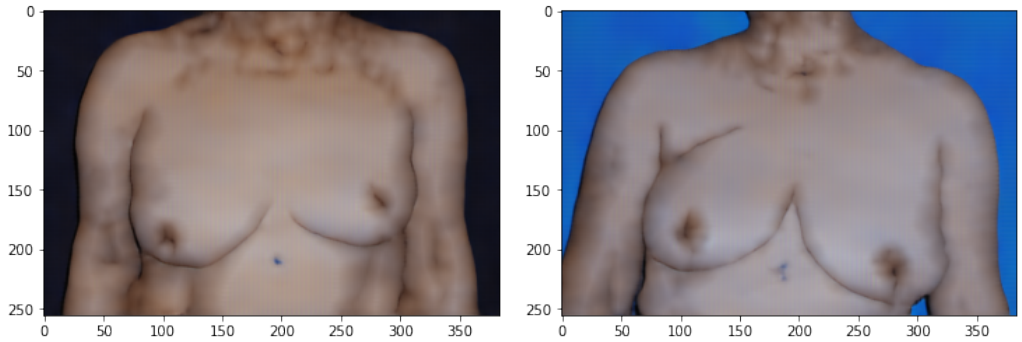


Figure 6.11: Image Denoising with MSLE loss in the image and gradient domain

6.4 Geometric Image distortion

For the experiments that will be further explained in the next sections, new images had to be created. First, distortions, corresponding to the lowering of the breast that had a lower nipple, were introduced to the new images. To create these distortions, a closed contour of the lower breast was first obtained by forming a closed shape with the keypoints that make up the contour.

The goal of this task was to distort the breast area vertically, and to do so, each pixel inside of the close contour was shifted vertically into a new position. This shift took into account the distance between the pixel being moved and the position of the breast endpoints: the bigger the distance to the endpoints, the bigger the shift. The pixel intensities of the new image are determined according to the following formula, where K is a constant equal to 0.2 by default. This constant is only decreased in images where the lower breast is too close to the margin, in order to assure that the distorted breast's contour stays within the image limits.

$point_i = (x_i,\ y_i - k \times (y_i - y_{endpoint}))$
if $distance(point_i, contour) == 0)$ **then**
    $newImage(x_i, y_i) = image(x_i, y_i)$
else
    $newImage(x_i, y_i) = image(point_i)$

Similarly, a new dataset of symmetrical images was created. In this task, instead of lowering the lowest breast to increase asymmetry, the highest breast is lowered to reach the height of the other, in order to create images with a symmetric contour. However, these transformations were not applied to images with an 'Excellent' aesthetic evaluation. The distortion process is similar to the one explained above, with the main difference being that the K value is not predetermined, but calculated for each case taking into account the contour dislocation needed to equalize both breasts' contour heights. A schematic representation of the distortions introduced to create this new dataset can be seen in figure 6.12.

Some results of these distortion tasks can be observed in figure 6.13.

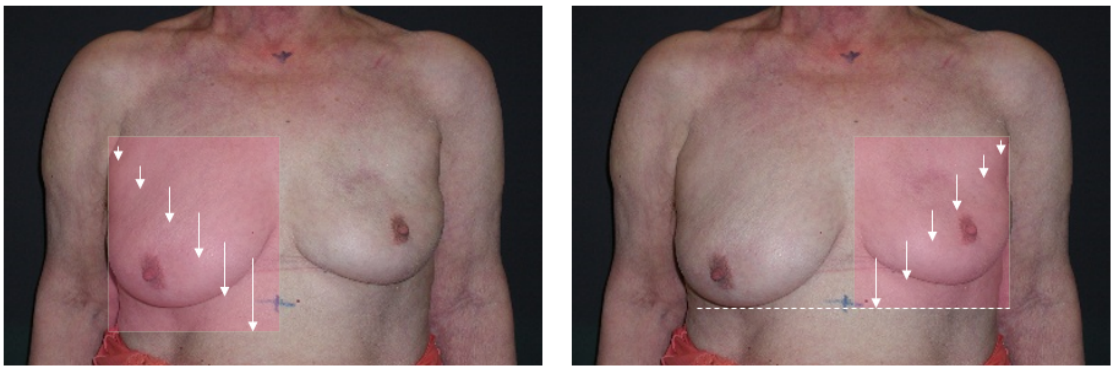


Figure 6.12: Schematic representation of the alterations applied to the dataset to create the new images with increased asymmetry (left) and with symmetry between both breasts

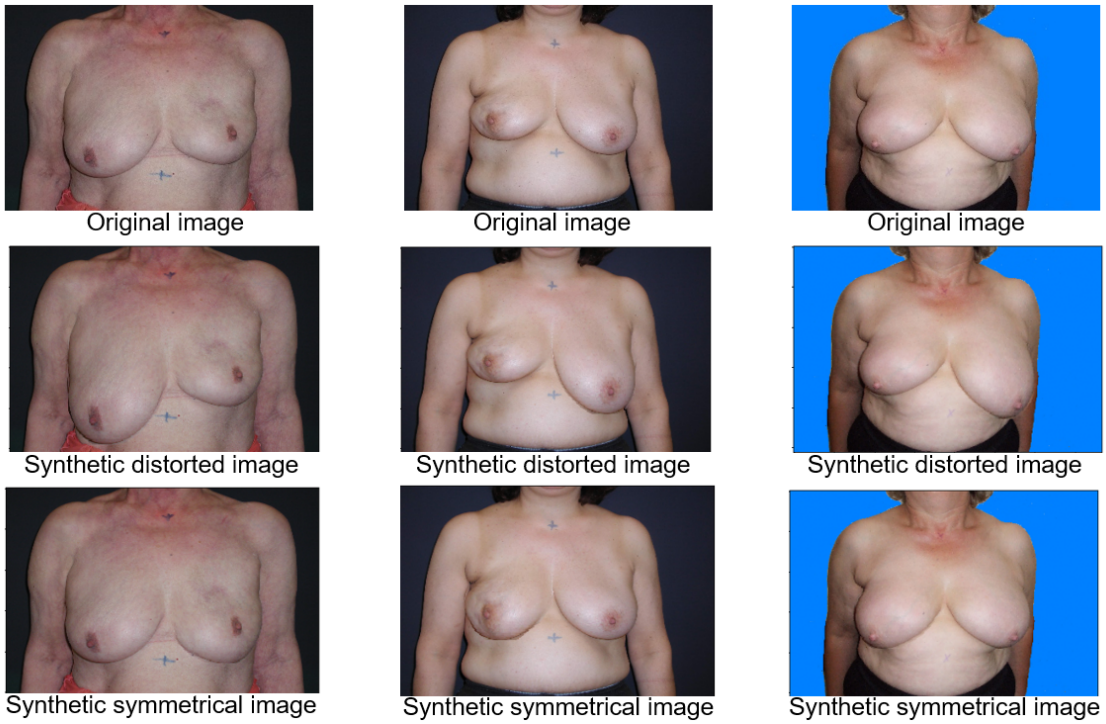


Figure 6.13: Example of new synthetic asymmetrical and symmetrical images created

6.4.1 Distortion correction

After exploring the potential of the U-Net, the first task was to train an U-net model to learn how to correct the synthetic distortions, by approximating each distorted image with its original.

The division of the dataset was equal to the one in the previous section. In order to increase the size of the training set, flipping operations and noise addition were used. The final augmented training set has 183 images.

Just like in the previous task, we first tested the losses in 6.1 and 6.2. One of the final results of this experiment can be seen in the left images in figures 6.15 and 6.16, while the original image

is pictured in figure 6.14. All the corrected images were fed to the denoising model with the SSIM loss mentioned in the previous section in order to remove some artifacts introduced by the correction process. It was concluded that overall, the second loss with just MSLE produced better results.

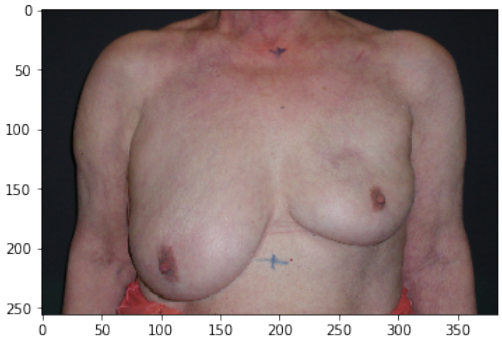


Figure 6.14: Synthetic distorted test image used as input

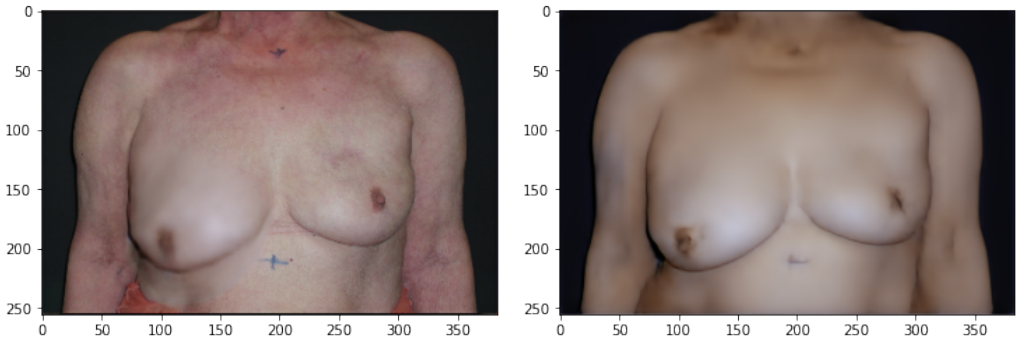


Figure 6.15: Corrected image, with the loss function in 6.1 (left) and after the denoising model (right)

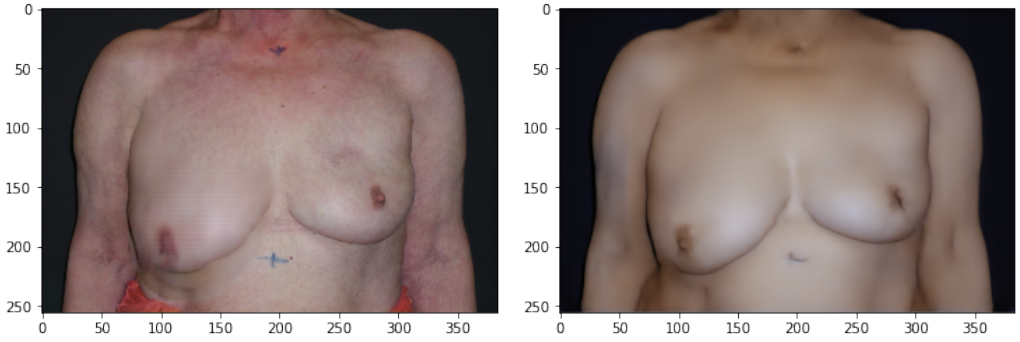


Figure 6.16: Corrected image, with the loss function in 6.2 (left) and after the denoising model (right)

After that, the effect of the weights of each loss component were tested, and were altered to 70/30 and 80/20.

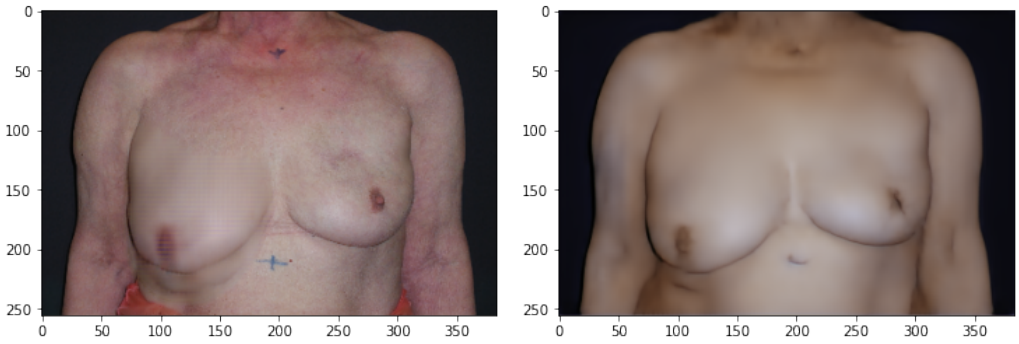


Figure 6.17: Corrected image, with the loss function 6.2, with a 70/30 weights (left) and after the denoising model (right)

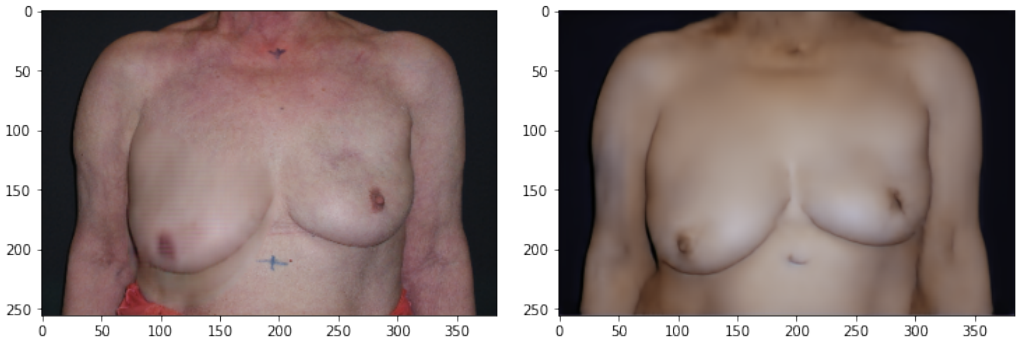


Figure 6.18: Corrected image, with the loss function in 6.2, with a 80/20 weights (left) and after the denoising model (right)

Finally, a new loss component, which used the binary classification model for aesthetic evaluation developed in [58] was introduced. The new loss, seen in equation 6.3, aims at compelling the model to generate images classified as having a good aesthetic evaluation, which happens if both breast are symmetrical.

$$\mathscr{L} = 0.9 \times MSE(image_{target}, image_{generated}) + 0.1 \times MSLE(edges_{target}, edges_{generated}) + 0.001 \times BCE(0, aestheticclassification_{generated}) \quad (6.3)$$

The image in figure 6.19 was generated using the loss function in 6.3. However, as it can be seen, this new component was not very helpful for the image reconstruction process.

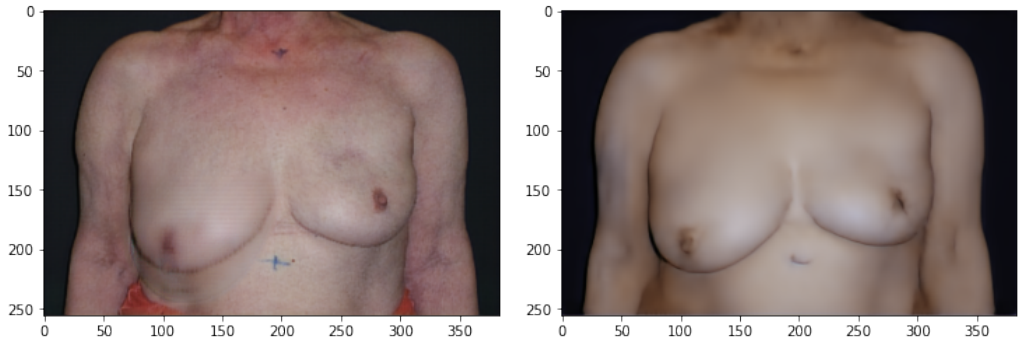


Figure 6.19: Corrected image, with the loss function in 6.3 (left) and after the denoising model (right)

A last set of experiments was performed with a change in the target images; instead of trying to approximate the distorted images to the original ones, the targets are symmetrical synthetic images (except in the cases of aesthetically pleasing results, as explained in section 6.4). Two new models were trained, with the losses in 6.2 with a 80/20 weights and in 6.3.

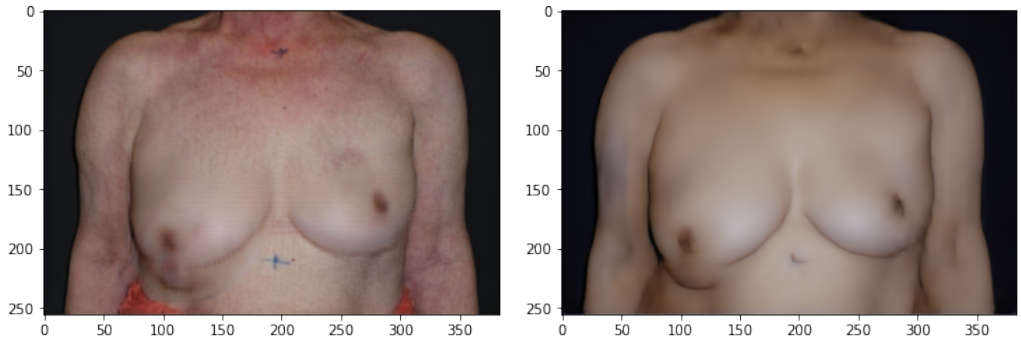


Figure 6.20: Corrected image, with the loss function in 6.3 and with symmetrical images as target (left) and after the denoising model (right)

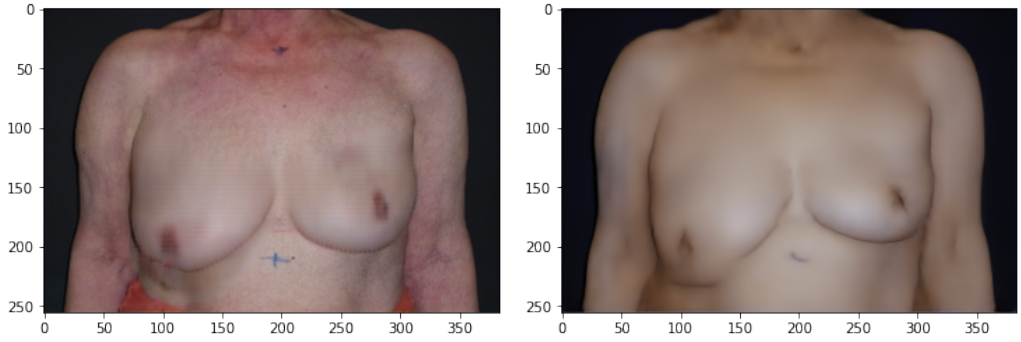


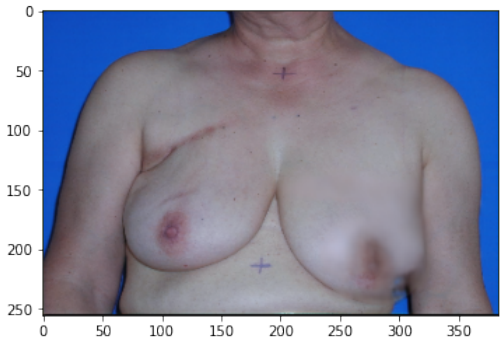
Figure 6.21: Corrected image, with the loss function in 6.2 and with symmetrical images as target (left) and after the denoising model (right)

Table 6.1: Root Mean Squared error between the generated corrected images by the models and the target images, with the samples in the train and test sets

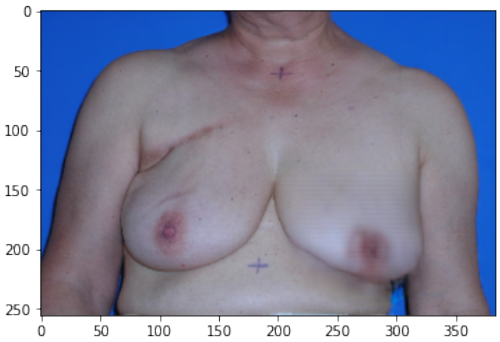
Loss function used during training		**RMSE**
0.9 x SSIM (image domain) + 0.1 x MSLE (edge domain)	Train	0,01219
	Test	0,02064
0.9 x MSLE (image domain) + 0.1 x MSLE (edge domain)	Train	0,01097
	Test	0,02124
0.7 x MSLE (image domain) + 0.3 x MSLE (edge domain)	Train	0,07171
	Test	0,07249
0.8 x MSLE (image domain) + 0.2 x MSLE (edge domain)	Train	0,01124
	Test	**0,02057**
0.9 x MSLE (image domain) + 0.1 x MSLE (edge domain) + 0.001 x BCE (0, aesthetic classification)	Train	0,02253
	Test	0,03113
0.9 x MSLE (image domain) + 0.1 x MSLE (edge domain) + 0.001 x BCE (0, aesthetic classification) (trained with symmetrical images as targets)	Train	0,02537
	Test	0,02917
0.8 x MSLE (image domain) + 0.2 x MSLE (edge domain) (trained with symmetrical images as targets)	Train	0,01748
	Test	0,02796

In order to have a quantitative analysis of the results, the root mean squared error between the generated and the target images was computed, both for the training and test sets. The results can be analysed in table 6.1.

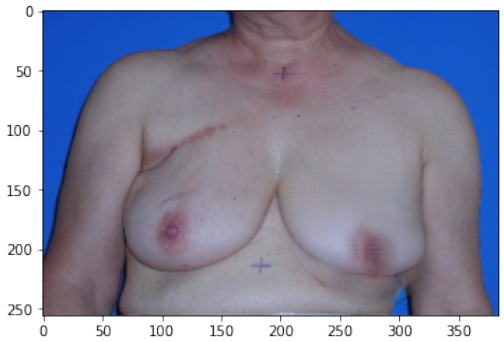
Finally, all the models were tested on real test images with a poor aesthetic evaluation, as it is important to test whether these correction models are efficient at correcting real images and not just synthetic ones. A summary of these results can be seen in figure 6.22.



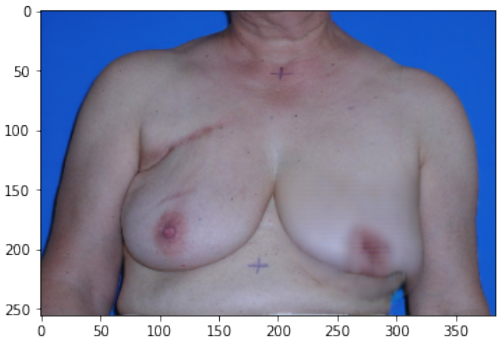
(a) Corrected image with the loss function in 6.1



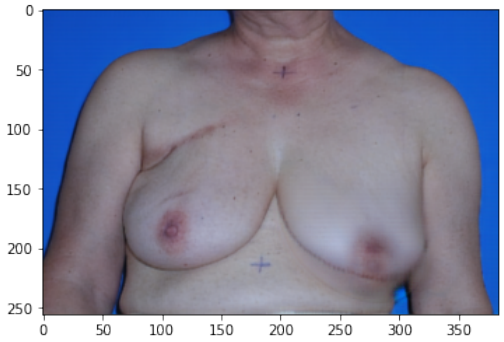
(b) Corrected image with the loss function in 6.2



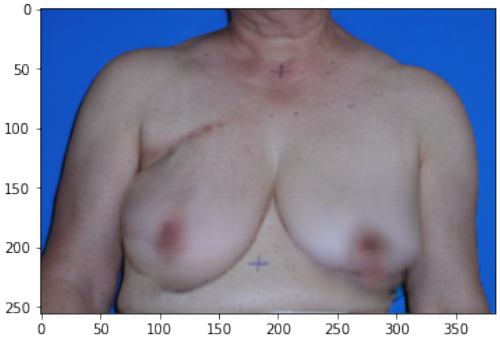
(c) Corrected image with the loss function in 6.2 with 70/30 weights



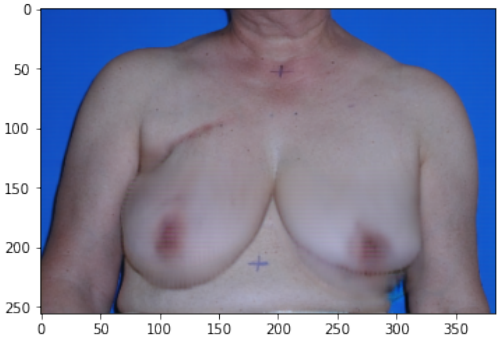
(d) Corrected image with the loss function in 6.2 with 80/20 weights



(e) Corrected image with the loss function in 6.3



(f) Corrected image with the loss function in 6.3, trained with symmetrical images as target



(g) Corrected image with the loss function in 6.2, trained with symmetrical images as target

Figure 6.22: Summary of the distortion correction models, tested on a real image with a 'Poor' aesthetic classification

6.4.2 Distortion introduction

After verifying that it is possible to train a model to correct image distortions and generate images with a better aesthetic result, the goal is to test whether it is possible to do the reverse and train a model that, given an image, returns a more distorted version of it. This task is very important for the final goal of this project, given that one of the main objectives is to develop a method that distorts an image in order for it to represent the expected physical outcome of a breast cancer treatment, which often creates asymmetry in the breasts.

The methods developed are similar to the ones explained previously. The models are trained to approximate the input, which corresponds to the dataset of only symmetrical images, to the target images, which correspond to the distorted dataset. The losses in 6.2 with 80/20% weights division and in 6.3 were used to try to distort the original images into the distorted ones. As done in the previous section, the denoising model was used on the final generated images in order to create a smoother representation of it and eliminate eventual artefacts.

Some results can be seen in figures 6.24 and 6.25, where the input image is represented in figure 6.23.

Figure 6.23: Test image with 'Excellent' aesthetic classification used as input

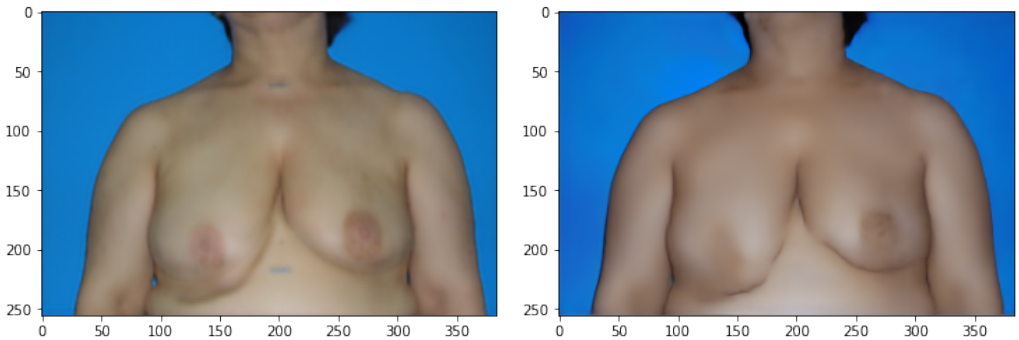


Figure 6.24: Distorted image, with the loss function in 6.3 and after the denoising model (right)

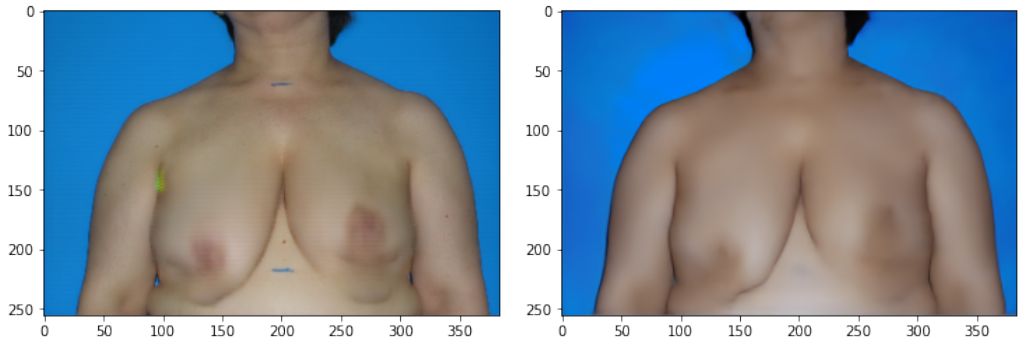


Figure 6.25: Distorted image, with the loss function in 6.2 and after the denoising model (right)

However, as it is possible to see in figure 6.25, since the input is an image with symmetrical or almost symmetrical breasts, the output is a distorted image where both sides are altered on the same level, resulting in a semi-symmetrical images. This error persisted across all symmetrical images, which is normal since the model is not able to identify the pattern of lowering only the breast with the lowest nipple, since they're both at similar heights.

In order to try to solve this issue, the problem was simplified and instead of attempting to distort the lowest breast, which in symmetrical cases is extremely hard to identify, the plan was changed to training a model that introduced a distortion always on the left breast of the patient. With this strategy, it became easier to alter the aesthetic evaluation of the breasts, although control was lost over which breast suffer the alteration. This however, may be fixed in the future by supplying the model with extra information about which breast undergoes the surgery, which will be the one that will be altered.

The results of this experiment can be seen in figures 6.26 and 6.27.

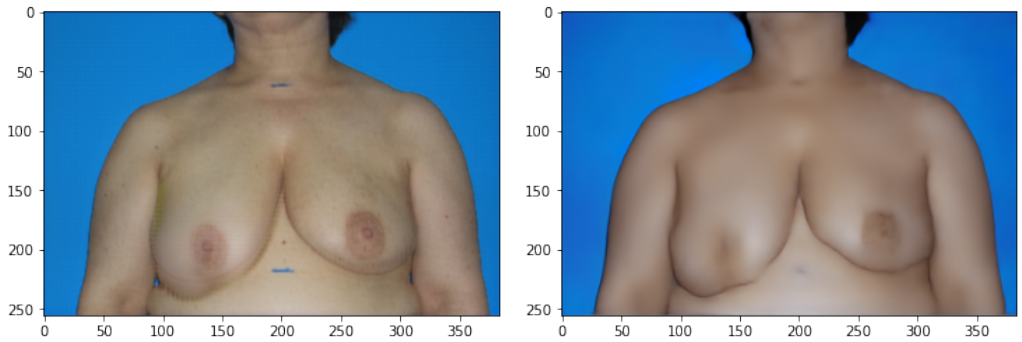


Figure 6.26: Image with distorted left breast, with the loss function in 6.3 and after the denoising model (right)

Table 6.2: Root Mean Squared error between the generated distorted images by the models and the target images, with the samples in the train and test sets

Loss function used during training		**RMSE**
0.9 x MSLE (image domain) + 0.1 x MSLE (edge domain) + 0.001 x BCE (0, aesthetic classification)	Train	0,02235
	Test	**0,02710**
0.8 x MSLE (image domain) + 0.2 x MSLE (edge domain)	Train	0,02073
	Test	0,02726
0.9 x MSLE (image domain) + 0.1 x MSLE (edge domain) + 0.001 x BCE (0, aesthetic classification) , (distortion always on the same side)	Train	0,01848
	Test	**0,02096**
0.8 x MSLE (image domain) + 0.2 x MSLE (edge domain) (distortion always on the same side)	Train	0,02093
	Test	0,02327

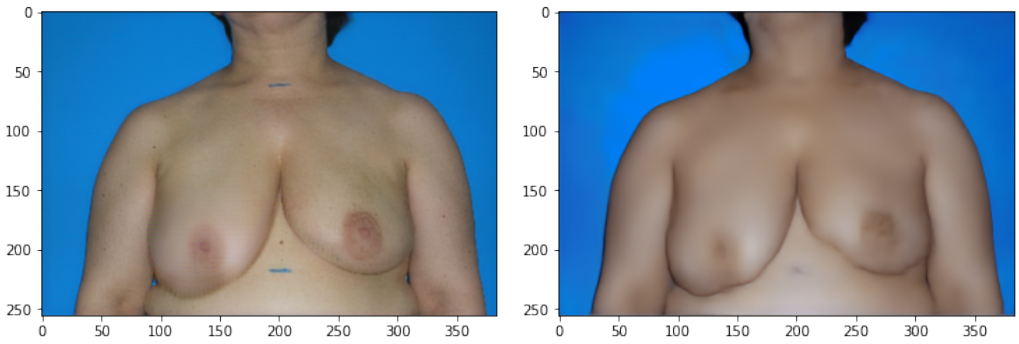


Figure 6.27: Image with distorted left breast, with the loss function in 6.2 and after the denoising model (right)

As presented before, some quantitative results can be seen in table 6.2.

6.5 Colour alterations

In this chapter, the experiments concerning the introduction and correction of colour-related anomalies in the breast will be discussed. After radiotherapy, the treated breast's colour often changes, gaining a darker tone. Similarly to the previous experiments with shape distortions, a new dataset of altered images that tried to mimic this consequence was created. In order achieve this, each pixel inside the contour of the right breast was considered and its intensity was changed gradually: as the distance to the height of the endpoints increased, the intensity of the pixels decreased. The changes to the image were applied according to the following pseudo code:

$ColourFactor_i = 1 - (\frac{y_i - y_{endpoint}}{y_{maximum} - y_{endpoint}} * K)$

if $distance(point_i, contour) == 0)$ **then**

$newImage(x_i, y_i) = image(x_i, y_i)$

else

$newImage(x_i, y_i) = image(x_i, y_i) \times ColourFactor$

In this formula, K is a random number between 0.25 and 0.35, constant for each image. The transformation is applied to the three colour channels. Some final images can be seen in figure 6.28.

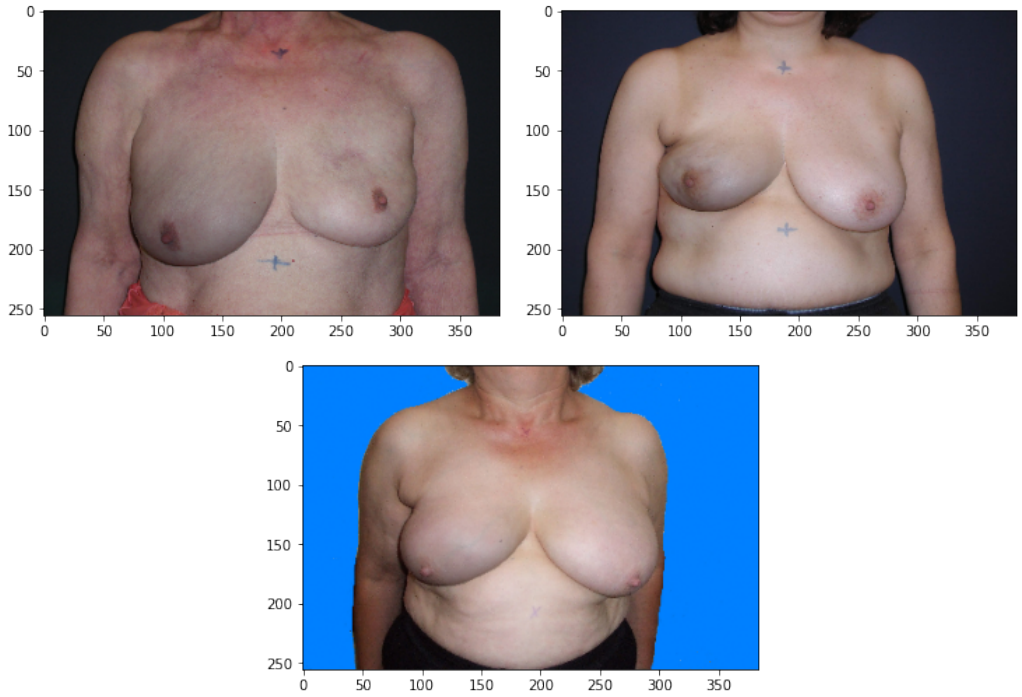


Figure 6.28: Example of new images with breast colourization

6.5.1 Colour correction

Just as it was done for the task related to geometric distortions, a U-Net model was trained to remove the dark colouration introduced to the images. Two loss functions were tested: Mean Squared Error and Mean Absolute Error. Since the contour of the images was not affected, the loss was only considered in the image domain, and the image edges were not regarded. Some results on a test sample can be seen in figures 6.30 and 6.31, while the input image is compared in figure 6.29. As it is possible to see, both approaches showed pretty satisfactory results.

The RMSE values between the target and generated images can be seen in table 6.3.

6.5.2 Colour introduction

The next step was to perform the opposite task, and train a model to introduce colour-related defects into the image. For this, the input and target images were switched, and the model was trained to approximate the original images to the colour distorted ones. The MSE and MAE losses

Table 6.3: Root Mean Squared error between the generated colour-corrected images by the models and the target images, with the samples in the train and test sets

Loss function used during training		**RMSE**
MSE	Train	0,01716
	Test	**0,01749**
MAE	Train	0,01945
	Test	0,01958

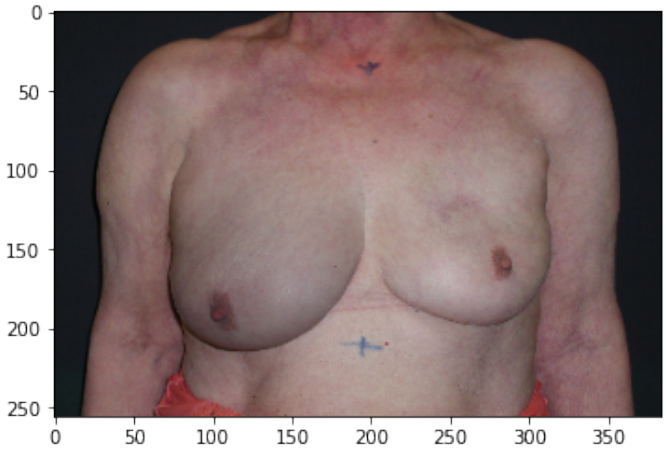


Figure 6.29: Synthetic test image used as input

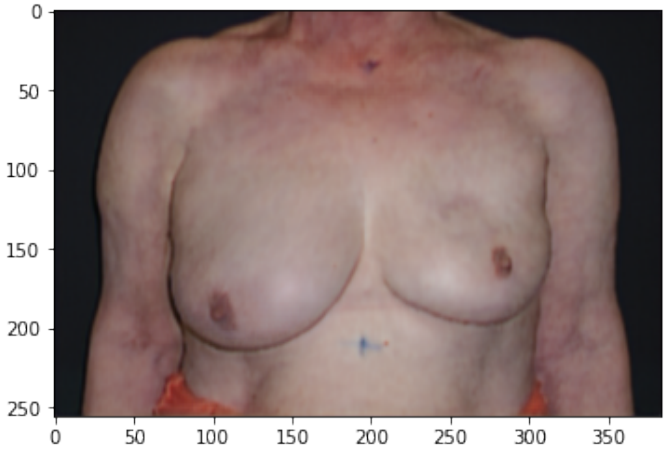


Figure 6.30: Colour corrected image with MSE loss

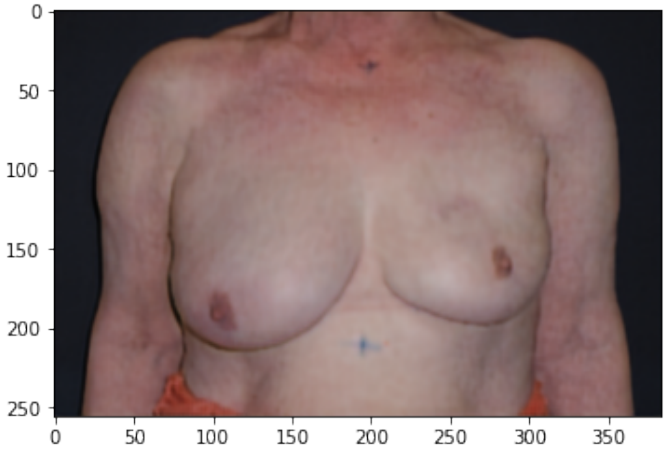


Figure 6.31: Colour corrected image with MAE loss

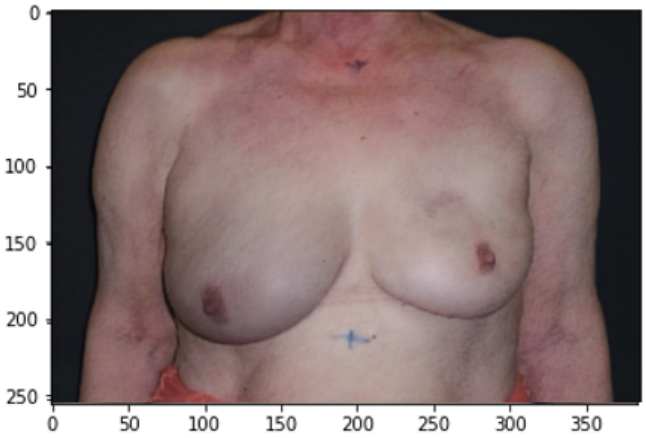


Figure 6.32: Generated image with colourization, with MSE loss

were once again used, and some final results are presented in figures 6.32 and 6.31. As before, the generated images by both models are very satisfactory, as they contain the gradual colour change that we aimed to introduce. However, as it is possible to see, the generated images by the U-Net trained with the MAE loss are blurry and the RMSE between the target and the generated images by this model is almost the double as the one calculated for the model trained with MSE. Because of this, it is safe to conclude that the MSE is a better loss function for the training of models for this kind of colour-related tasks.

Table 6.4 presents the RMSE values between the target and generated images.

6.6 Colour and geometrical alterations

As a closing note, two final models were trained to correct and introduce simultaneously the two types of distortion mentioned in this chapter. Once again, the U-Net model was used, with the loss in equation 6.2, which showed the most positive results throughout the conducted experiments. The results are represented on the test image in figures 6.34 and 6.35.

Quantitative results of these final experiments are summarized in table 6.5 and 6.6.

Table 6.4: Root Mean Squared error between the generated colourized images by the models and the target images, with the samples in the train and test sets

Loss function used during training		**RMSE**
MSE	Train	0,01583
	Test	**0,01666**
MAE	Train	0,03121
	Test	0,03036

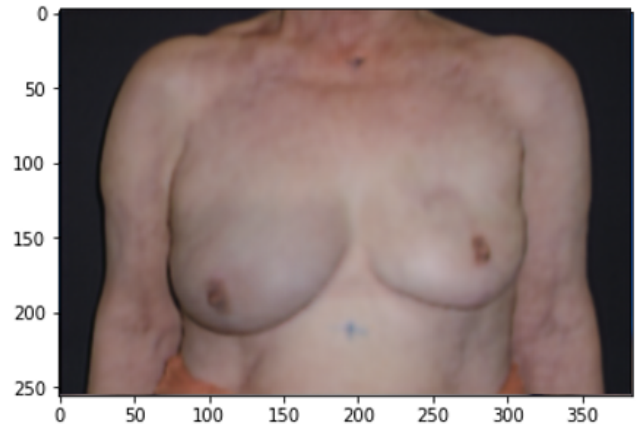


Figure 6.33: Generated image with colourization, with MAE loss

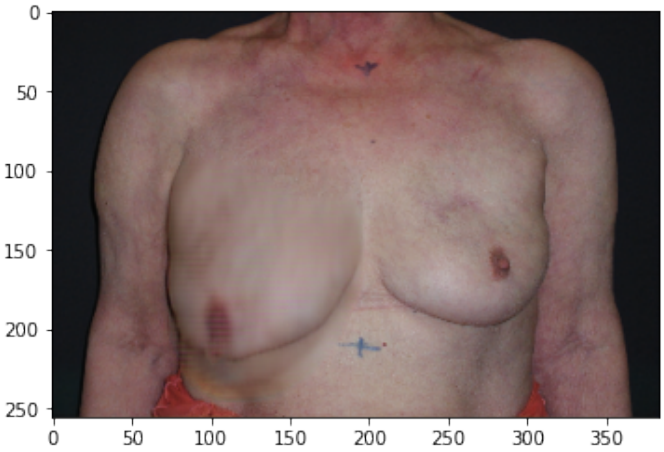


Figure 6.34: Corrected Image - distortion and colourization removal

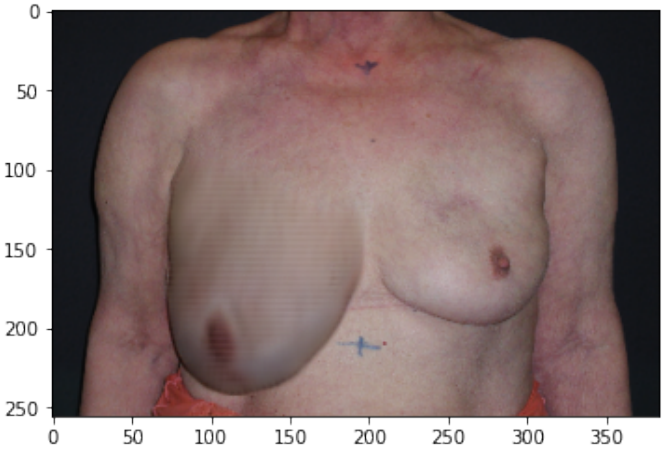


Figure 6.35: Distorted Image - distortion and colourization introduction

Table 6.5: Root Mean Squared error between the generated colour and geometric distortion corrected images by the models and the target images, with the samples in the train and test sets

Loss function used during training		**RMSE**
0.8 x MSLE (image domain) + 0.2 x MSLE (edge domain)	Train	0,02990
	Test	0,03227

Table 6.6: Root Mean Squared error between the generated colourized and distorted images by the models and the target images, with the samples in the train and test sets

Loss function used during training		**RMSE**
0.8 x MSLE (image domain) + 0.2 x MSLE (edge domain)	Train	0,02061
	Test	0,02392

6.7 Discussion

With these experiments, we showed that it is possible to train a rather simple model to alter breast symmetry by either increasing the asymmetry or reducing it, and consequentially altering the aesthetic evaluation of the breasts. Moreover, it is also possible to alter the breast colour and use a U-Net to clean or introduce darker tones in the breast's skin, a typical aesthetic consequence after some treatments.

Regarding the geometric distortions, the model which produced the best results, both visually and objectively, was the one trained with a MSLE loss on the image and edge domain when the goal was to correct the distortion, and the one trained with an extra term of binary cross-entropy when the goal was to introduce the distortion. When the goal was to introduce or correct the colour of the image, the model trained with the MSE as the loss function produce better results than the one with MAE.

As mentioned before, since the final goal is to alter an image so it mimics the aesthetic result of another one, these first experiments represent an initial step towards this goal, since, although only through simple lifts, distortion operations and pixel intensity changes, we were able to change the aesthetic evaluation of the breasts by approximating it to another image.

Chapter 7

Experiments with unpaired-image Translation

In this chapter, we propose the use of the conditional GAN, specifically models based on the Pix2Pix model, and the cycleGAN to perform image-to-image translation between two domains: images with an excellent and a poor aesthetic evaluation. With the conditional and cycle GANs, the goal is to demonstrate that the model can "learn" the characteristics that influence the aesthetic result evaluation and is able to generate distorted images that represent poor aesthetic results from images that represent excellent results and vice versa. Contrary to the experiments reported in the previous chapter, the image translation is not paired, or in other words, we are not trying to approximate an image of one domain to a specific image of the other domain, but instead the goal is to find a general transformation function that approximates the first domain (images with a good aesthetic evaluation) to the target domain (images with a poor aesthetic evaluation).

In the context of the Cinderella Project, these excellent results, due to their symmetry, are meant to represent images of patients before the BCCT. The image translation to the domain of a poor result can lead to creating representative previews of a possible outcome of the treatment, which is one of the major goals of the project.

7.1 Conditional GAN

Five models based on the original Pix2Pix model were tested: one where the generator is the U-net used in the previous chapter, one where the generator is a modified U-Net with a ResNet backbone, two where the U-Net has a pretrained encoder and finally one where the generator is a fully pretrained U-Net for the task of image to image translation. Each of the experiments will be further elaborated on the following sections. However, some constants were kept across all the experiments:

- The discriminator's architecture is the same for all experiments. The discriminator is a convolutional "PatchGAN" classifier, which, instead of classifying the entire image as real or fake, breaks the image into 70 X 70 patches and classifies each patch. This strategy is

known to lead to the generation of less blurry images. The discriminator's architecture is represented in figure 7.1. As it can be seen, the discriminator takes two inputs, the 'real' images from one domain and the generated or real image from the other domain, which are concatenated before going through the network.

- The loss function used to train the discriminator is the binary cross entropy, while the generator is trained to minimize the mean absolute error.

- The generator and discriminator were trained with Adam optimizer with a learning rate of 0.0002.

- Each model was trained twice for 300 epochs and a batch size of 1. The second training round started with the best weights of the previous one, which were selected during the validation process, based on the similarity of the aesthetic classification of the generated and of the validation images.

7.1.1 Dataset

The dataset used in these experiments was the same as the one mentioned in the previous chapter but, in order to create the two datasets of two different domains used in conditional GANs, the dataset was split, first into images with excellent and poor evaluations. However, only a small number of images contain these extreme evaluations - in the whole dataset of 143 images, 10 are labelled as 'excellent' and 27 as 'poor'. In order to balance the two domains, 32 images of the 'good' category were added to the 'excellent' group and 15 images classified as 'fair' were considered in the 'poor' domain; these images were selected as the ones showing better and worse scores with the deep learning aesthetic classification model. Then, 25% of the synthetic images mentioned in chapter 6 were added to the set; 35 synthetic symmetrical images were considered for the 'excellent' domain and 35 synthetic distorted images were added to the 'poor' domain. This set was divided into train, validation and test in the same proportions mentioned in the previous chapter. Finally, some extra data augmentation was performed in the training sets through horizontal image flipping (a third of the dataset was flipped and added to the training set) and through jittering (half of the dataset, including the flipped images), where the samples are resized to 110% on both dimensions, and then are randomly cropped back into its original size. The final training set contained 202 images, 101 of each domain. In order to try and facilitate the task, the background of the images with the light blue background was changed to black.

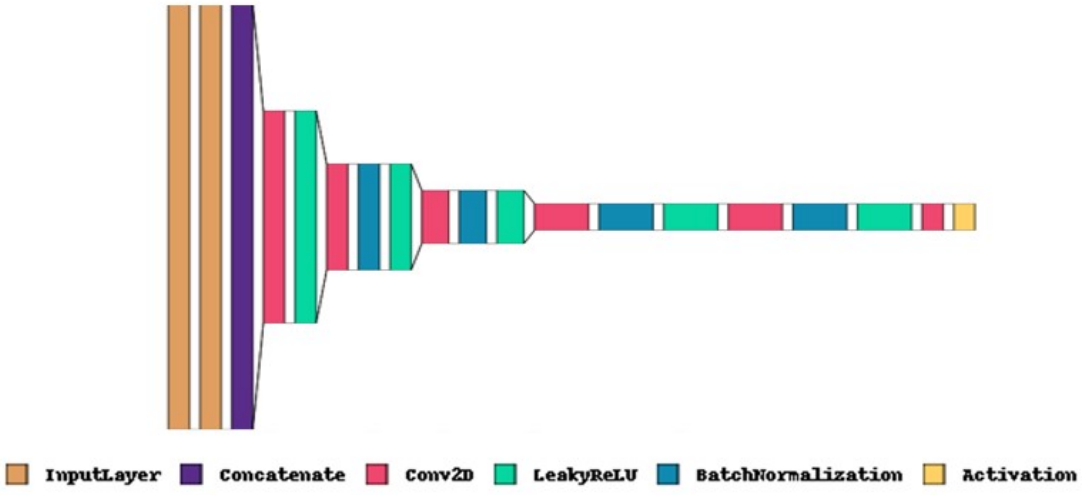


Figure 7.1: Discriminator Architecture

7.1.2 Results

In this section, a summary of the results obtained, with and without the use of transfer learning techniques, will be presented.

7.1.2.1 Without transfer learning

Firstly, the results obtained by two models based on the U-Net will be presented. These models were trained from scratch for this particular task.

7.1.2.2 Original U-Net as generator

The first attempt at training conditional GANs uses the proposed methodology in [74], where a U-Net architecture is used as the generator of the GAN. In figures 7.2 and 7.3 it is possible to see some results of images generated by these models, tested on the test set containing images considered 'excellent'.

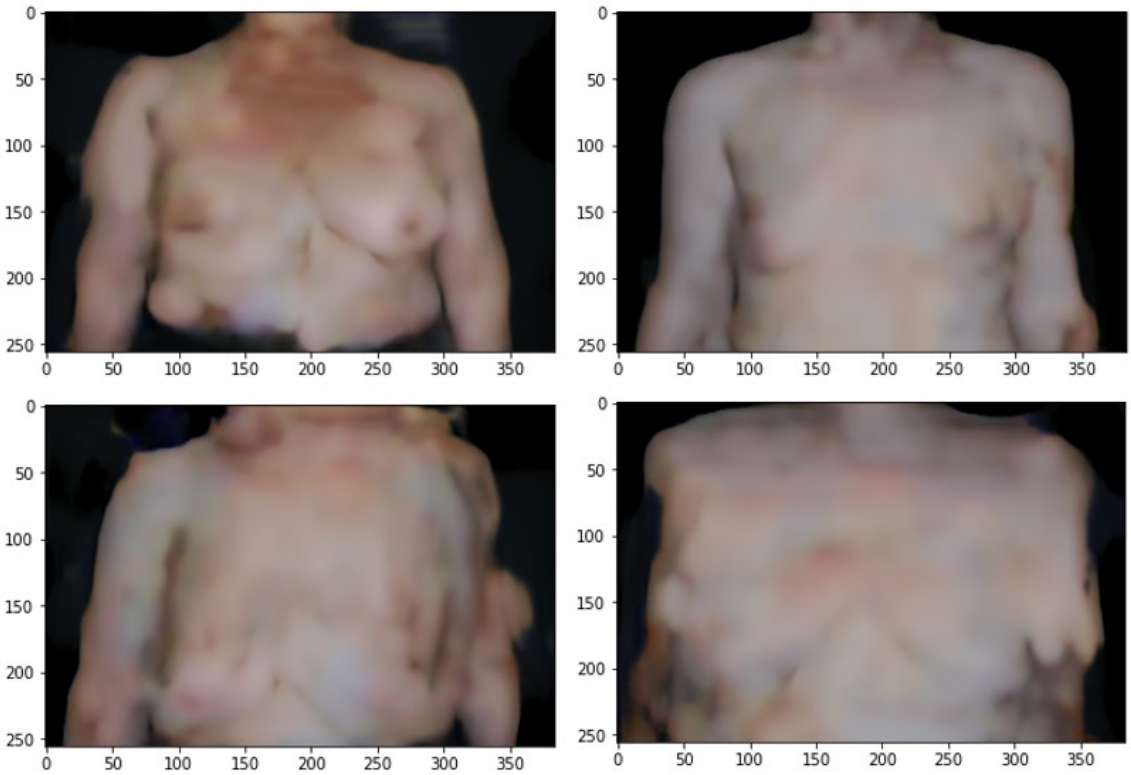


Figure 7.2: Generated images with a U-Net trained from scratch, after the first training round (254 epochs)

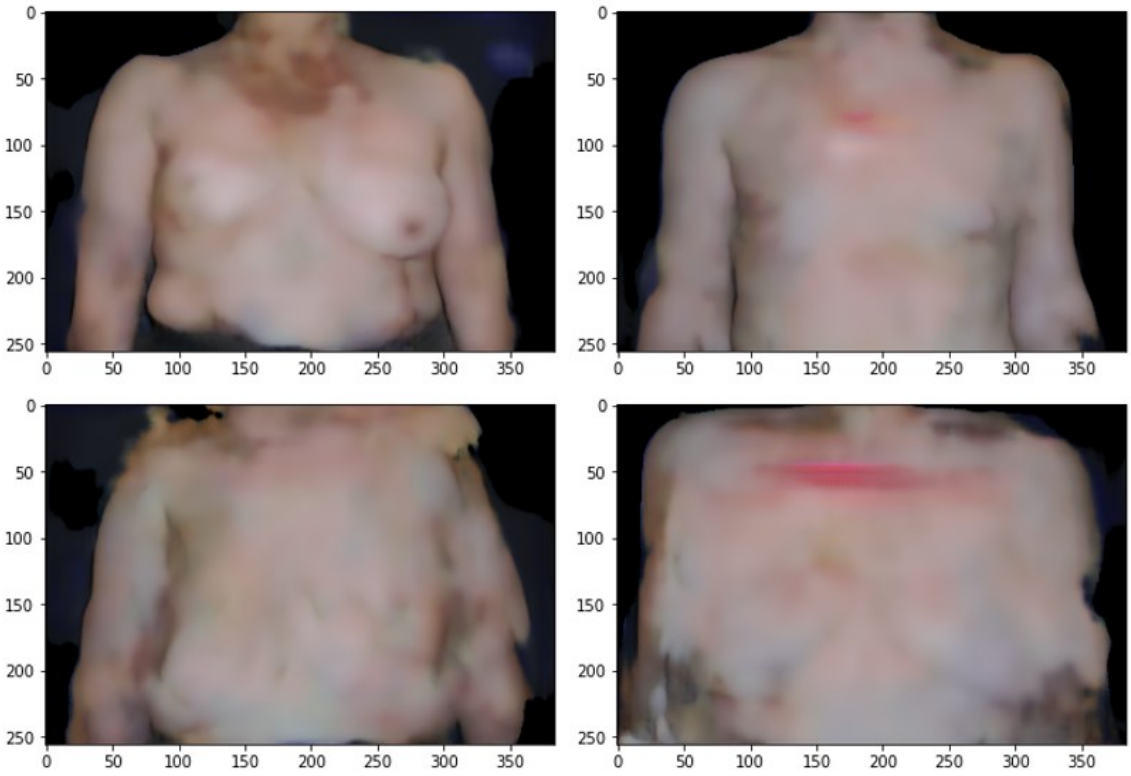


Figure 7.3: Generated images with a U-Net trained from scratch, after the second training round (550 epochs)

As it can be seen, most of the generated images by this model are blurry and some of the most realistic ones (such as the top left one in figure 7.3) are almost copies of the training images of the 'Poor' domain.

7.1.2.3 ResNet backbone

The second model explored was a U-Net with a ResNet50 backbone, which means that the encoder or contracting path of the U-Net has the architecture of the well-known ResNet50 model. The final architecture is represented in figure 7.4 and some of the generated images can be seen in figures 7.5 and 7.6.

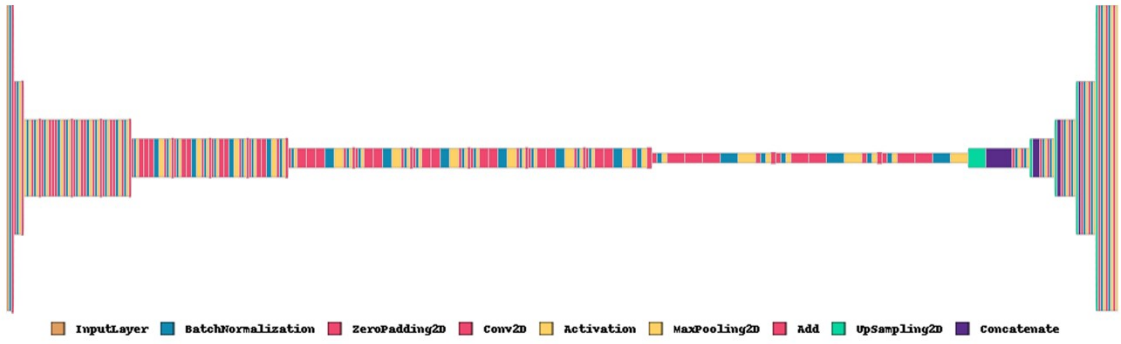


Figure 7.4: U-Net with ResNet50 backbone Architecture

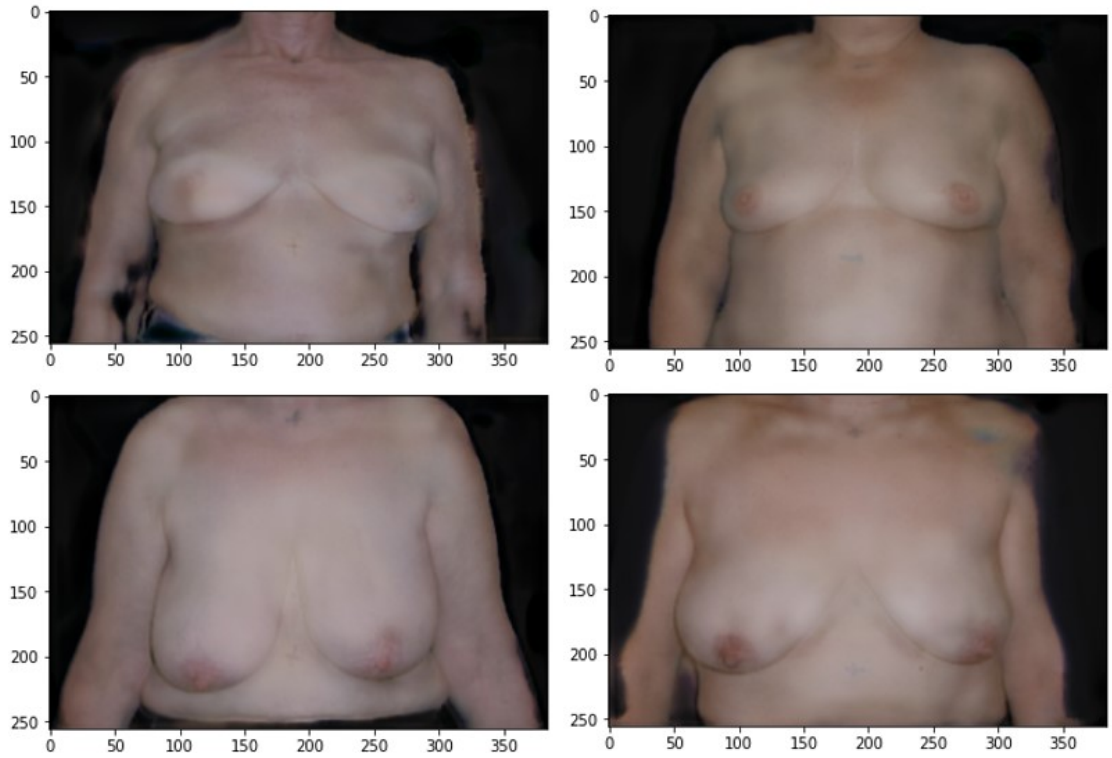


Figure 7.5: Generated images with a U-Net with a ResNet50 backbone, trained from scratch, after the first training round (259 epochs)

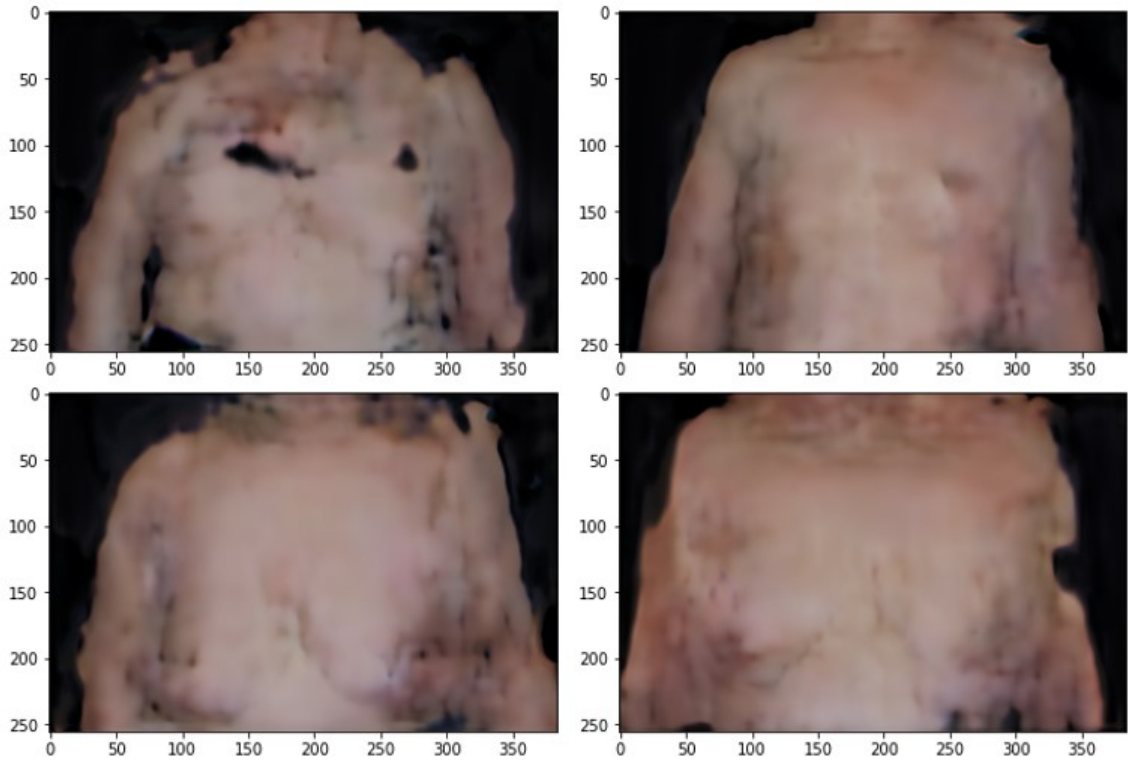


Figure 7.6: Generated images with a U-Net with a ResNet50 backbone, trained from scratch, after the second training round (551 epochs)

With this model, the final generated images are also blurry and unintelligible. However, it is interesting to see the evolution between the generated images after the first and second rounds of training.

7.1.2.4 With transfer learning

Since the size of the dataset is so small in comparison with other problems using GANs, where models are trained with hundreds or thousands of images, it was decided to use transfer learning for (partial or complete) weight initialization. Using transfer learning allows the models to converge faster, and using pretrained generators in GANs can lead to a more efficient training [100, 137, 68].

Two of the approaches include using a U-Net with a pre-trained encoder (or backbone) on the ImageNet dataset. This approach is often used to improve the accuracy of the U-Net, especially, with smaller datasets [73, 116]. For the final approach, the weights of a fully trained U-net on an image translation task are used for initializing the network.

7.1.2.5 U-Net with MobileNet backbone

The first experiment with transfer learning in the Pix2Pix model consisted of using a Mobile Net version 2 (MobileNetV2), trained on ImageNet. The MobileNetV2 model is easy to train since it has fewer parameters than other well-known pre-trained models. The model's architecture can be seen in figure 7.7, and was based on the implementation in [126].

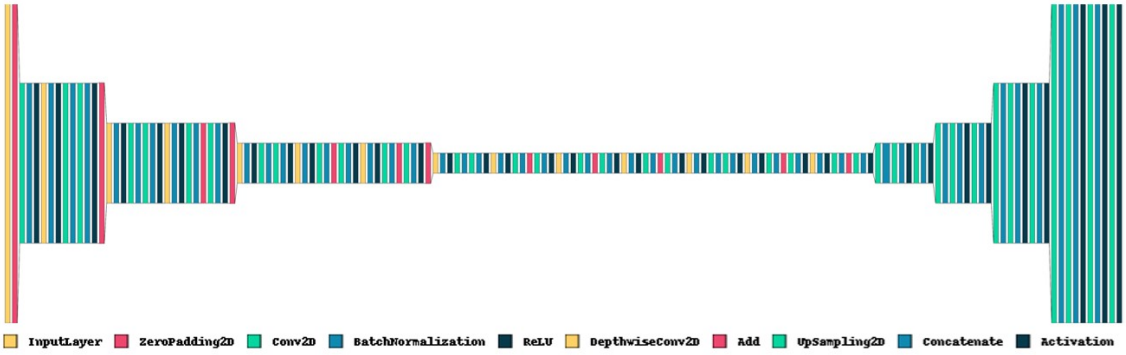


Figure 7.7: U-Net with MobileNetV2 backbone Architecture

In figures 7.8 and 7.9, it is possible to see some results of the generated images by this network.

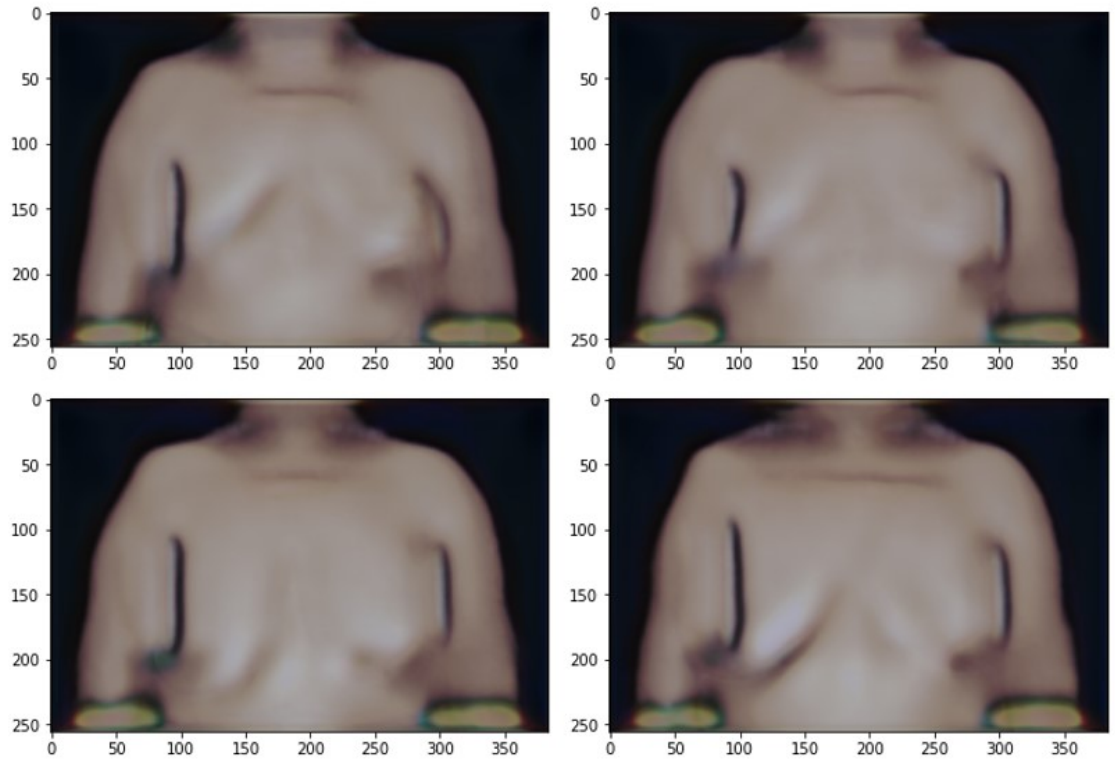


Figure 7.8: Generated images with a U-Net with a pre-trained MobileNetV2 backbone after the first training round (297 epochs)

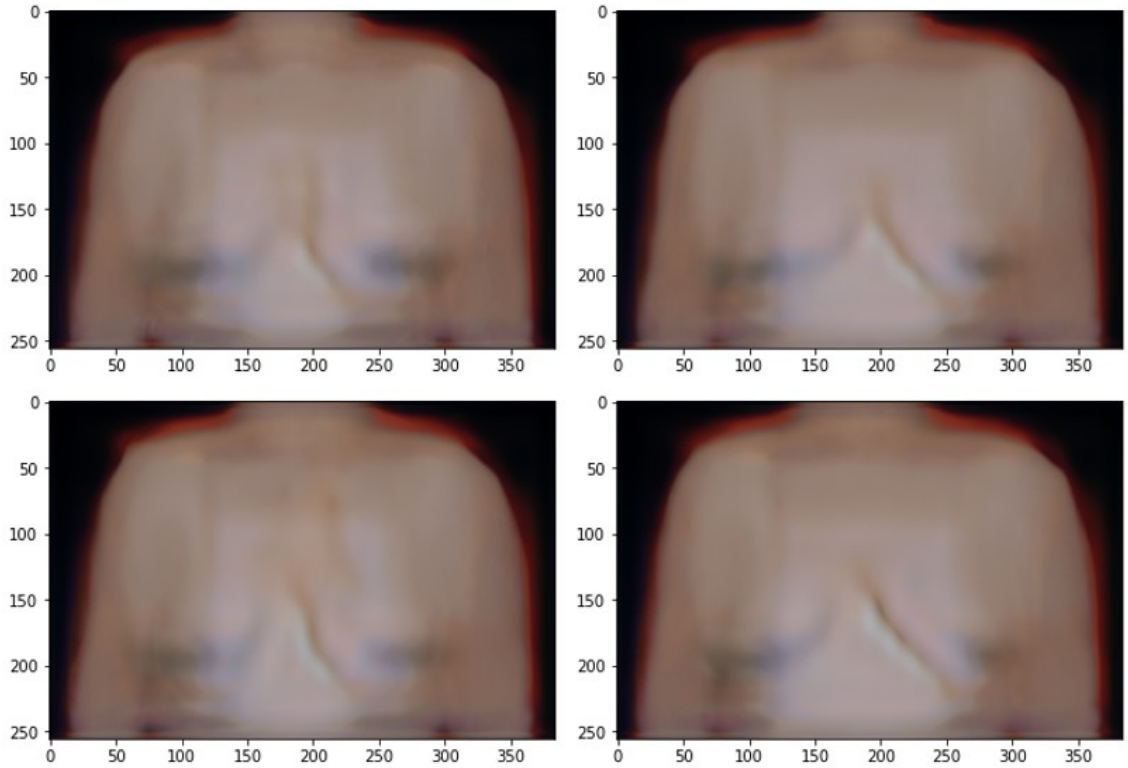


Figure 7.9: Generated images with a U-Net with a pre-trained MobileNetV2 backbone after the second training round (597 epochs)

The U-Net with the pre-trained MobileNet encoder also fails to create acceptable examples of images from the 'Excellent' domain with the style of the 'Poor' domain. Moreover, it also fails at generating different images from different inputs.

7.1.2.6 U-net with Resnet50 backbone

After observing the results of the MobileNet, we decided to bet on a larger, more complex network to use in the encoding portion of the U-Net: a ResNet-50. The model was created with the Segmentation Models library [72]. Some of the generated images are in figures 7.10 and 7.11

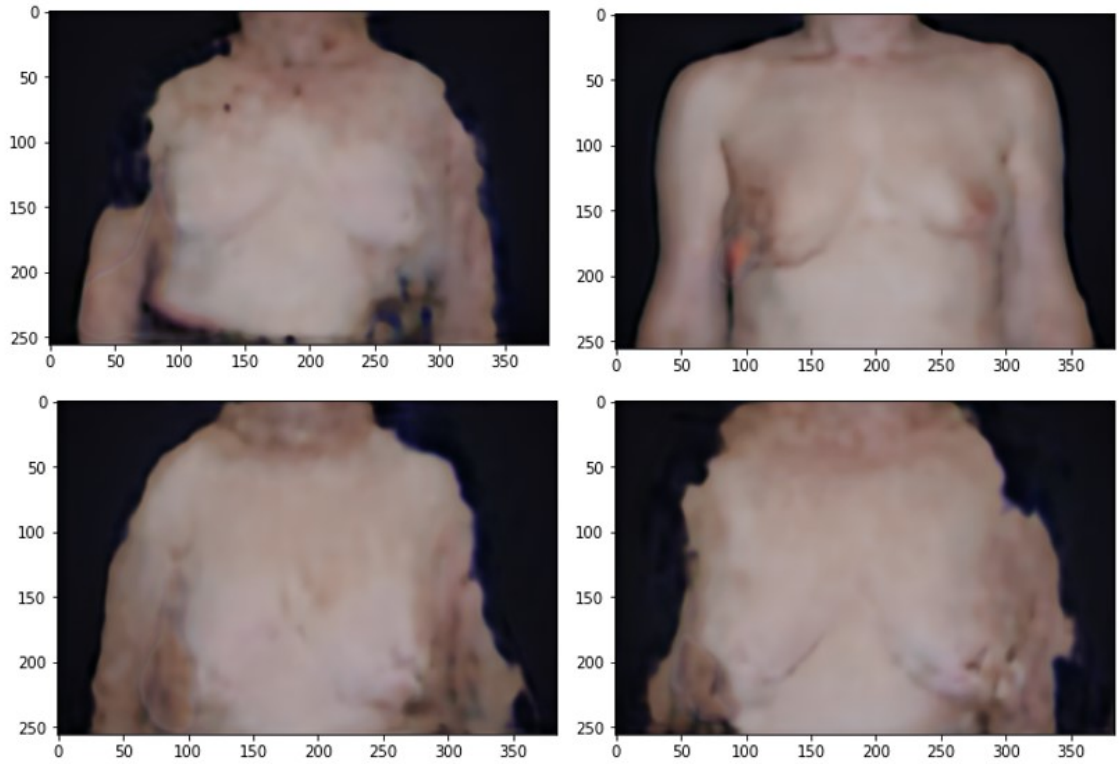


Figure 7.10: Generated images with a U-Net with a pre-trained ResNet50 backbone after the first training round (264 epochs)

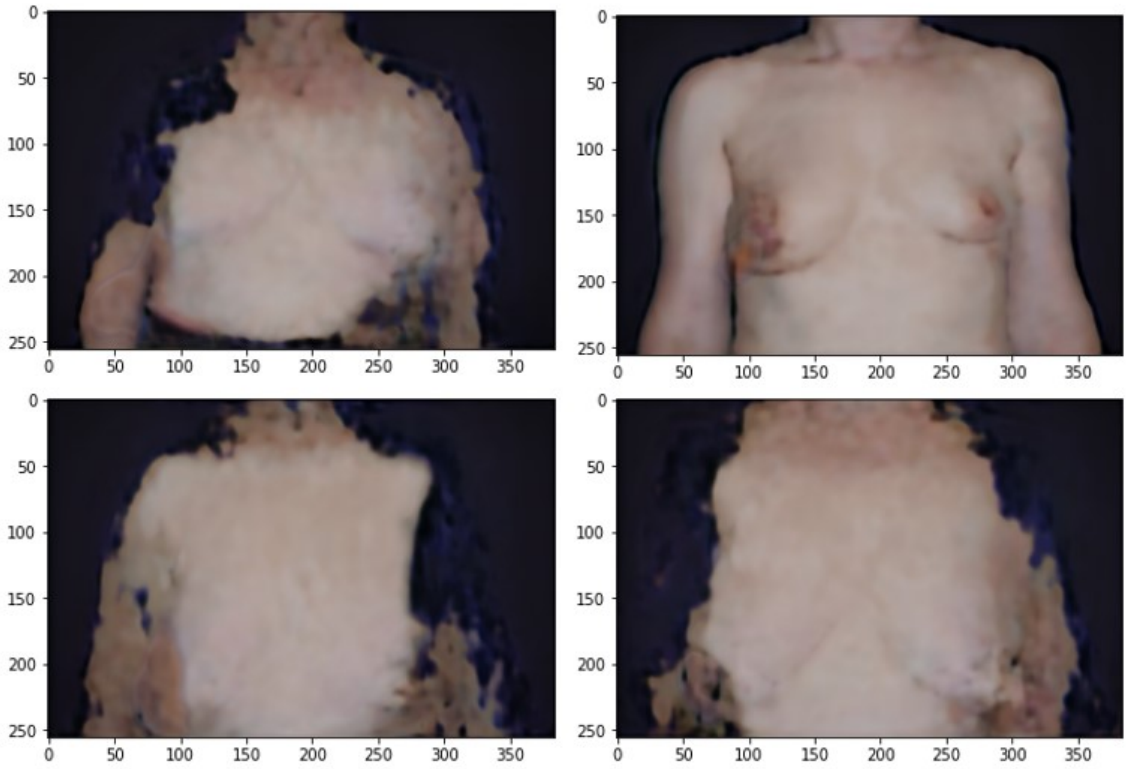


Figure 7.11: Generated images with a U-Net with a pre-trained ResNet50 backbone after the second training round (510 epochs)

The resulting images by this model show blurriness and have large portions of the bodies removed.

7.1.2.7 Pre-trained Pix2Pix generator

Finally, the last experiment with transfer learning consisted of using a fully trained U-Net on an image translation task with the goal of colouring images using a Pix2Pix [99]. The U-Net architecture is similar to the one used in the previous experiments. Although the dataset and goal of this Pix2Pix project are very specific and different to the one being exploited in this dissertation, it was believed that the fact that it was trained in an adversarial manner could benefit the task.

The generator's input size was set to 256x256, so it was necessary to cut the images in the dataset. In order to obtain centred images, the keypoint coordinates were used, and the image was centred using the contour points. Due to the size of the generated images, the validation step with the aesthetic evaluation model could not be completed, and the weights were saved after the 100th, 200th and 300th epochs and analysed later.

Some of the generated images can be seen in figures 7.12 and 7.13.

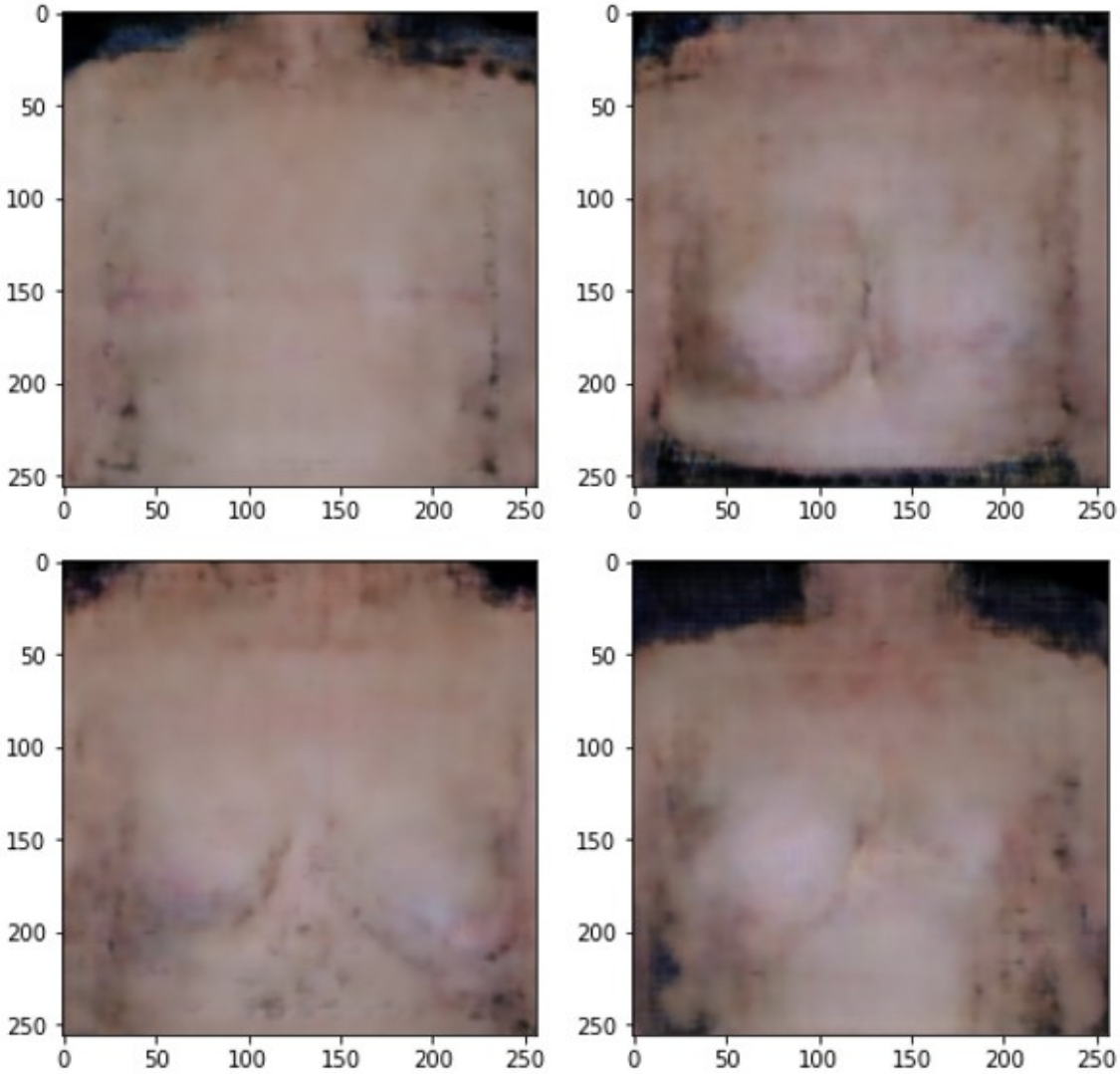


Figure 7.12: Generated images with a pre-trained U-Net after the first training round (300 epochs)

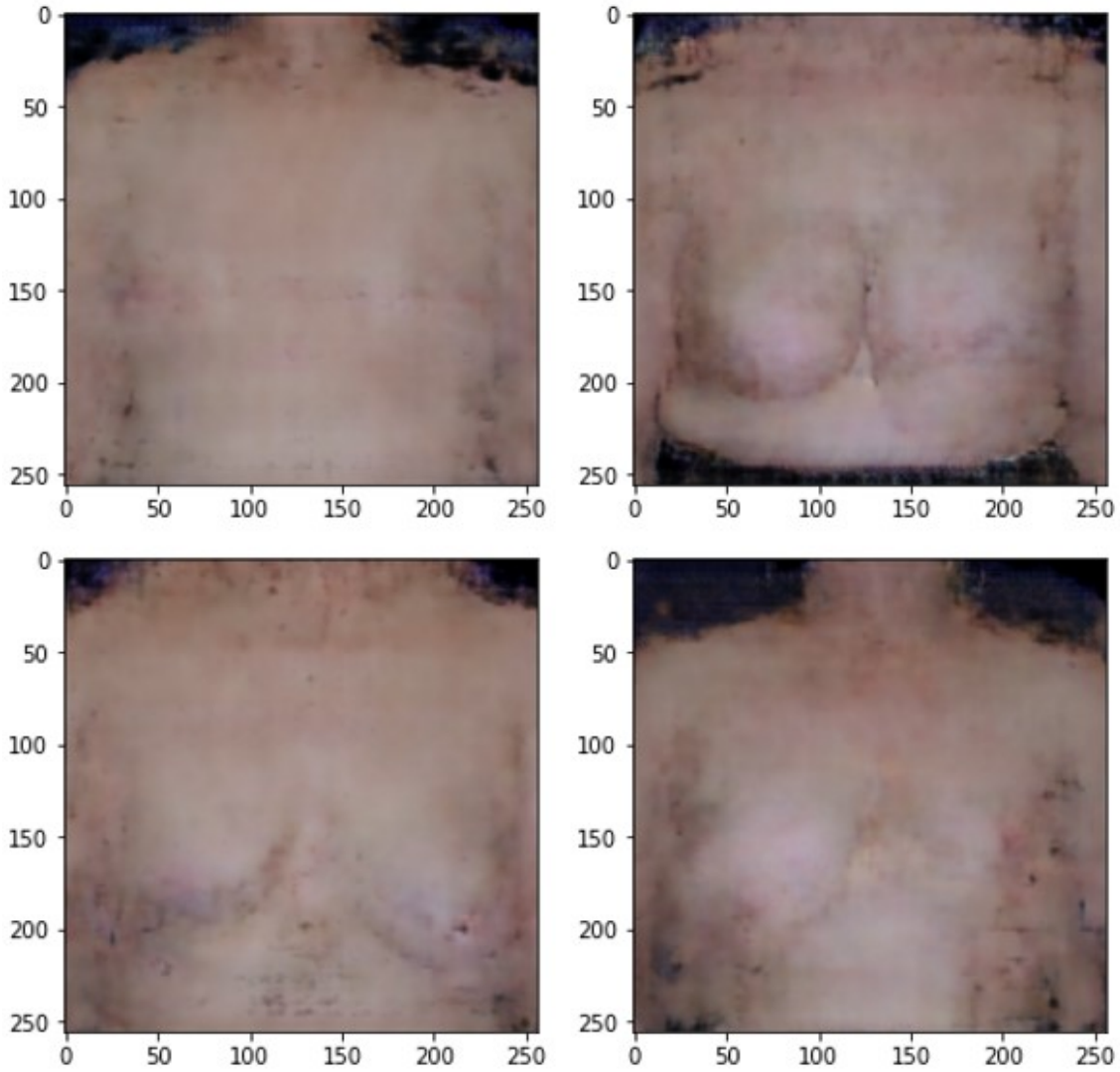


Figure 7.13: Generated images with a pre-trained U-Net after the second training round (600 epochs)

As it can be seen, the generated images by the GAN with the fully trained generator aren't realistic and most of the breast's keypoints are hard to identify.

In order to obtain a more objective analysis of the results, and similarly to what was done in chaper 6, the RMSE value between the generated images by the models after the second training round and the images of each domain was computed, both in the training and test sets. The results can be seen in table 7.1.

Table 7.1: Root Mean Squared error between the generated images by the different Pix2Pix models and the images from each domain, with the samples in the train and test sets

Model	Train		Test	
	'Excellent' domain	**'Poor' domain**	**'Excellent' domain**	**'Poor' domain**
U-Net	0,152497	0,179542	0,161683	0,129109
U-Net with Resnet backbone	0,131363	0,160663	0,128919	0,136529
U-Net with pre-trained MobileNet backbone	0,145645	0,143346	0,147071	0,147152
U-Net with pre-trained Resnet backbone	0,180891	0,195025	0,162087	0,065159
Pre-trained Pix2Pix generator	0,117242	0,119773	0,11455	0,121097

As expected, the test results generated by the U-Net and the U-Net with the pre-trained Resnet backbone show a greater similarity to the 'Poor' domain than the 'Excellent' one, which makes sense given that some of generated images were almost copies of samples of the 'Poor' domain. However, this didn't happen during training, showcasing signs of overfitting. In the U-Net with the Resnet backbone trained from scratch and the model with the fully pre-trained generator, the opposite happens, and the generated images are slightly more similar to the images in the 'Excellent' domain. The overall lower error in the images generated by this second model is most likely due to the smaller size of the generated images, since the background was almost eliminated. Finally, with the U-Net with the MobileNet encoder, the generated images are almost equally similar to the images in both domains.

7.2 Cycle GAN

In the present section, an extension of the work introduced in the previous one will be presented. The use of cycleGANs is proposed as an alternative for the task introduced in 7.1 of performing image-to-image translation between the two domains of binary aesthetic valuation of a BCCT treatment.

The complete cycleGAN method is composed of two generators and two discriminators, which form two composite GAN models, each composed by the two generators and one of the discriminators. The main advantage of the cycleGAN over the Pix2Pix is that the first is more stable due to the cycle consistency loss, which can help improve the results obtained with the conditional GAN. However, due to its architectural complexity, it takes a lot more time to train.

The dataset used in these experiments is the same as the previous chapter.

The architecture used is similar to the one in the original cycleGAN paper [138], with 3 consecutive blocks of sequential convolutional, instance normalization and ReLU activation layers, 9 residual blocks from the ResNet model and again 3 blocks, similar two the first 3 but with a final tanh activation layer. The instance normalization layer is a standardizing method, similar to Batch Normalization, but instead of scaling the values across a batch of samples, the scaling is done individually for each feature map. This architecture, represented in figure 7.14, has shown impressive results for neural style transfer and super-resolution in [81]. The discriminator, such as in the Pix2Pix, is a 70x70 Patch GAN, with the difference that in the case of the cycleGAN, only one batch of samples is fed to the model, either the real samples or the fake/generated ones.

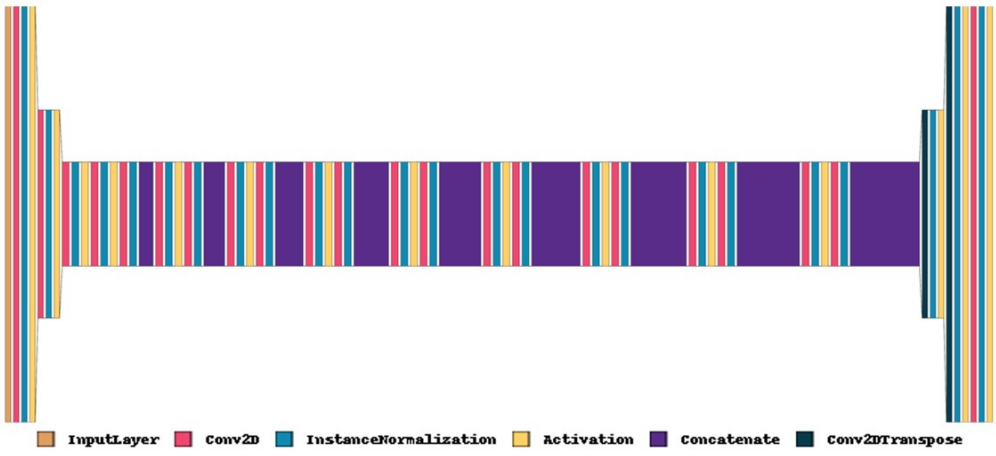


Figure 7.14: Architecture of the CycleGAN's generator

The discriminator is trained to minimize the MSE, while the generator is trained indirectly to minimize four losses:

- **Loss 1:** corresponds to the adversarial loss, where the discriminator is used to train the generator, and consists of the MSE between the discriminator's classification and the target classification, which is set to 'trick' the discriminator.
- **Loss 2:** the identity loss, computed as the MAE between the original image and the reconstructed image, generated by the generator of its domain:

$$Image\ Domain\ A \rightarrow Generator\ B\ to\ A \rightarrow Reconstructed\ Image\ Domain\ A$$

$$Identity\ loss\ =\ MAE(Image\ Domain\ A, Reconstructed\ Image\ Domain\ A)$$

- **Loss 3:** The forward cycle loss, a consistency loss consisting on the MAE between the original image from the input domain and its reconstruction after translating it into the target domain.

$$Image\ Domain\ B \rightarrow Generator\ B\ to\ A \rightarrow Generated\ Image\ Domain\ A \rightarrow Generator\ A\ to\ B \rightarrow Reconstructed\ Image\ Domain\ B$$

$$Forward\ cycle\ loss\ =\ MAE(Image\ Domain\ A,\ Reconstructed\ Image\ Domain\ A)$$

- **Loss 4:** The backward cycle loss, which is similar to the forward loss but with an image from the target domain as the first input.

$$\begin{aligned} &\textit{Image Domain A} \rightarrow \textit{Generator A to B} \rightarrow \\ &\qquad \textit{Generated Image Domain B} \rightarrow \textit{Generator B to A} \rightarrow \\ &\qquad\qquad \textit{Reconstructed Image Domain A} \end{aligned}$$

$$\textit{Backwards cycle loss} = \textit{MAE}(\textit{Image Domain B}, \textit{Reconstructed Image Domain B})$$

These four losses have different weights in the final generator's losses, with loss 1 to 4 being attributed to a weight of 1, 5, 10 and 10, respectively. Both the discriminator and the GAN model were trained with the Adam optimizer, with a learning rate of 0.0002. The models were trained for 250 epochs with a batch size of 1.

Once again, the cycleGAN trains simultaneously two GAN models, and each is trained with the same losses, only changing the domain that is considered 'original' and the 'target', and the examples used above to exemplify the losses would be related to a GAN that translates images from domain B to domain A. In the present case, the GAN models will have excellent images as input and poor images as a target and vice-versa. From now on, the GAN with the task of translating images from the excellent domain to the poor will be addressed as GANE2P while the GAN with the opposite task will be called GANP2E.

7.2.1 Results

After training, it is possible to retrieve multiple results from the cycle GAN: the reconstruction of the inputs, reconstruction of the targets and the of course the translation between domains. The reconstructions are related to the three losses mentioned above, as each GAN outputs three images: the identity generated image, the forward reconstruction image and the backward reconstruction image, which are used during the training to compute the identity, forward cycle and backwards cycle losses.The forward reconstruction of the GANE2P corresponds to the backward reconstruction of the GANP2E and vice versa. The generated images representing the translations between domains are obtained with the generators that compose the GAN.

In figures 7.16 and 7.15, it is possible to see the identity-generated images of each of the GANs. Image 7.15 depicts examples of the identity reconstruction of the GANE2P, while image 7.16 represents the example of the GANP2E.

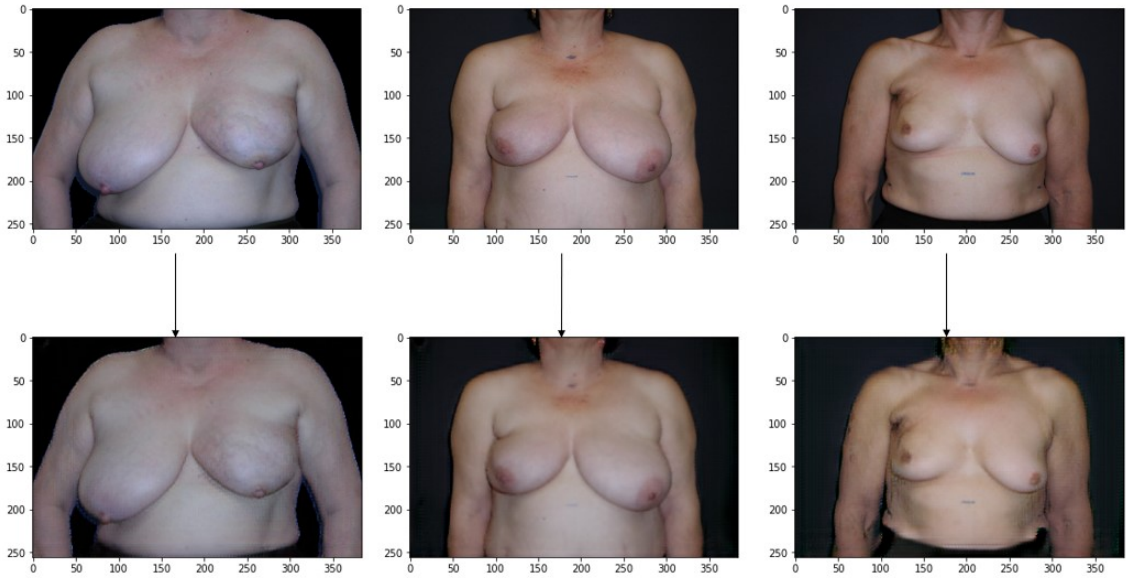


Figure 7.15: Identity reconstruction by GANE2P

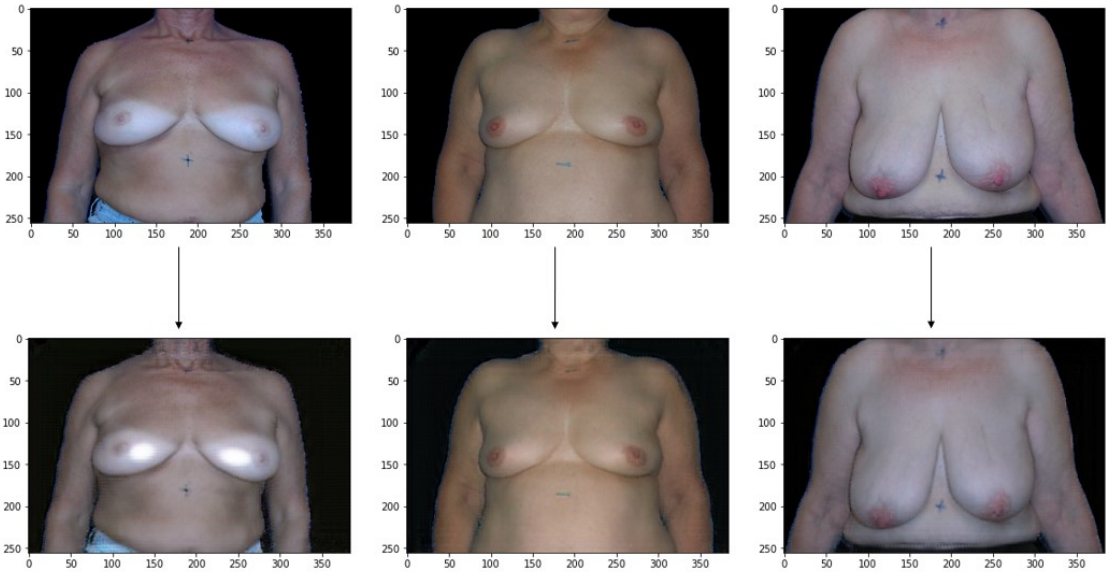


Figure 7.16: Identity reconstruction by GANP2E

In figures 7.17 and 7.18, it is possible to see the forward and backward reconstructions of the targets and inputs.

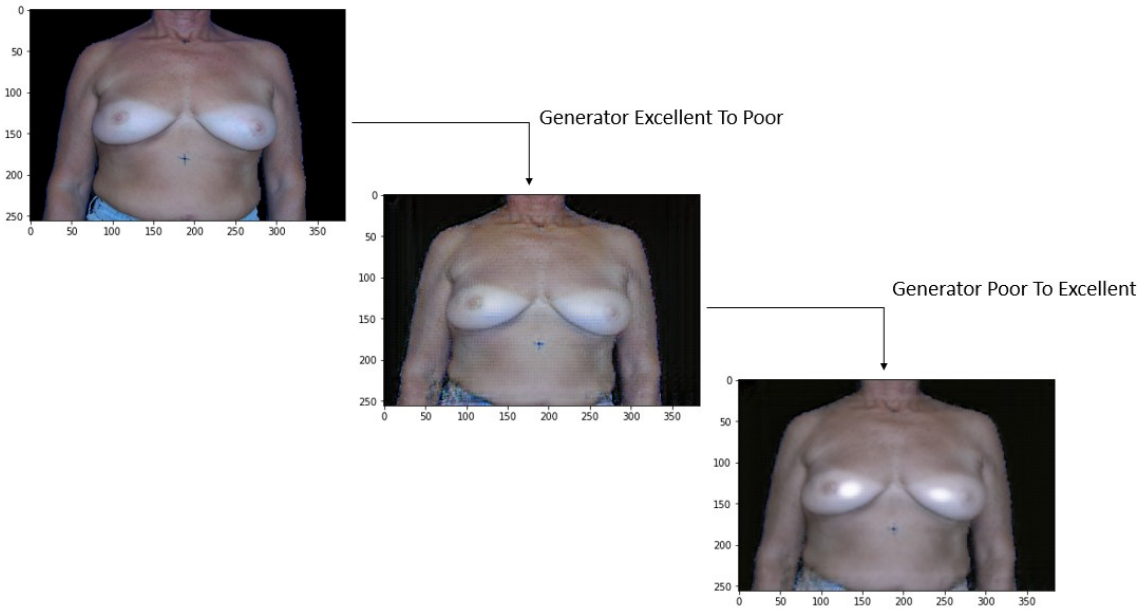


Figure 7.17: Forward reconstruction of the GANE2P and backwards reconstruction of the GANP2E

Figure 7.18: Forward reconstruction of the GANP2E and backwards reconstruction of the GANE2P

Finally, some examples representing the translations between domains can be seen in figures 7.19 and 7.20.

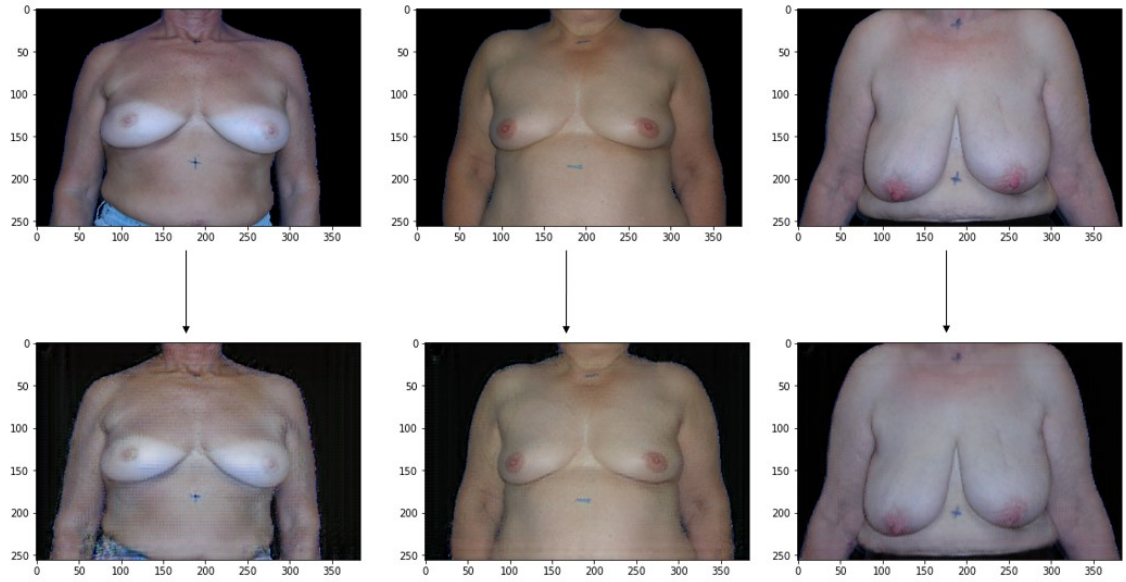


Figure 7.19: Translation from the excellent domain (input images in the top row) to the poor domain (generated images in the bottom row)

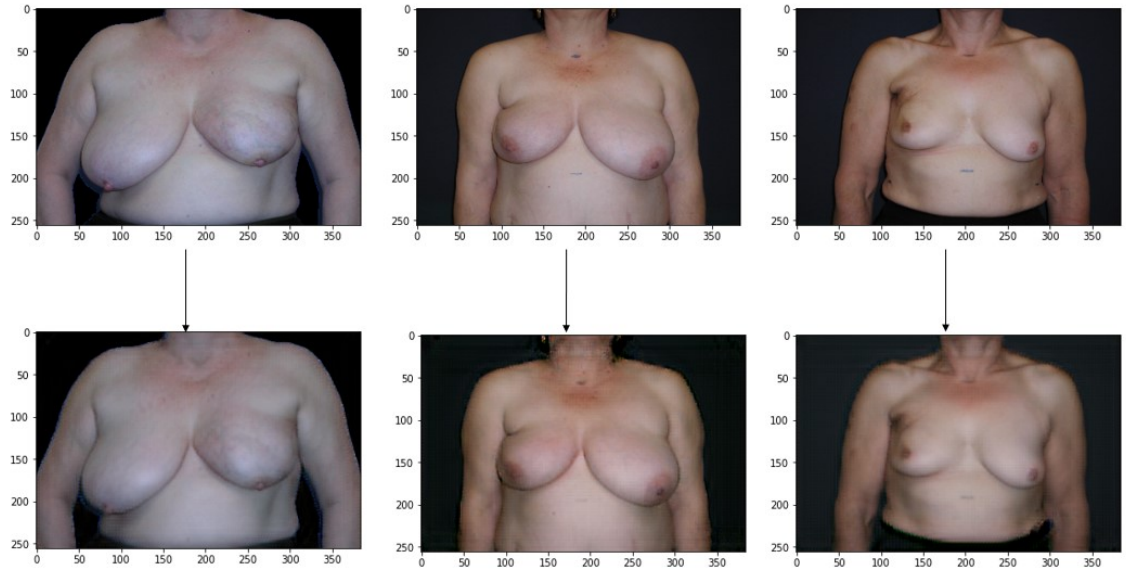


Figure 7.20: Translation from the poor domain (input images in the top row) to the excellent domain (generated images in the bottom row)

As it is clear to see, the results of the cycleGAN are not satisfactory, as the model was not able to learn the features of each domain and distort the input images in order to alter their aesthetic aspect.

7.3 Discussion

Overall, it is safe to conclude that the generated images by the two kinds of GANs tested are not acceptable.

Regarding the Pix2Pix-based models, the generated images are clearly altered, but do not possess realism nor clear signs that breast aesthetic-related transformations are being applied. While it is hard to completely assess what is happening with the training, it is clear to see that there is a general problem of overfitting, since that generally the model seems to be learning to copy the examples of poor cases used during training, instead of applying transformations on the excellent images used as input. This phenomenon is especially noticeable in the generated images by models without a pre-trained backbone or generator or in more complex models with more parameters, such as the one where the U-Net's backbone is a ResNet50. This makes sense, since it is known that models with a pre-trained weights would converge slower, and that models with lots of trainable parameters often lead to overfitting. This assumption is also corroborated by the quantitative analysis showcased in table 7.1. However, for example in the results obtained with the GAN where the generator is the U-Net with a Resnet50 backbone without pre-trained weights, it is possible to see that, after the first round of training, the model produced images that are essentially equal to the input images, while in the second round, the images show similar poses but with distortions. Although these distortions are noisy and the generated images are not intelligible, these changes could show promise that, with further training, they would represent obvious distortions related to changes in the aesthetic results, such as asymmetry. In the case of the smaller model, the U-Net with the MobileNet backbone, the problem seems to be different, with the model producing very similar images despite the different input images. It can be concluded that the model suffers from model collapse and got stuck on images similar to the ones in figure 7.9, which, according to the RMSE results, represent an equilibrium between the images of both domains. Finally, the model with the fully trained generator created blurry images, where even the breast contours are difficult to see, and showing probably the worst results.

In the future, these experiments could be repeated, given the great results obtained by the Pix2Pix in other tasks. It is suggested that, first and foremost, the dataset size is increased, not only through image augmentation techniques, but with new, real images. Then, the best strategy can be chosen based on the size of the dataset: models with pre-trained portions can lead to better results, but also can lead to over or underfitting if the model used has too much or not enough complexity, which is what seemed to have happened with the tested examples. Still, given the specificity and novelty of the task, it can be useful to train the model from scratch if enough computational resources are available. It can also be interesting in the future to use the breasts' keypoints to guide the process, and give extra information on the most relevant areas to transform.

All in all, the Pix2Pix can prove to be a good approach for one-way image translations, and in the context of the Cinderella Project could even be used to perform paired image translations, where the source image is the image of the current patient and the target is the retrieved image whose aesthetic outcome we wished to imitate.

Concerning the cycleGAN, despite its proven efficiency in image-to-image translation tasks, the results obtained in this case were not satisfactory, as the translated images remained almost equal to the input images, as it is possible to see in the examples in figures 7.19 and 7.20. These disappointing results may be caused by many factors, but it is important to point out once again the reduced size of the dataset and the fact that, due to the high cost of computational resources required, the model was trained for a rather small number of epochs.

Despite the unsatisfactory results obtained in this task, it is also important to mention that the experiments conducted are a novelty in the research field of BCCT aesthetic evaluation, and these results are a helpful starting point for this important subject.

All in all, given the accuracy shown in other image translation experiments, these two types of network can still be a good approach to creating biometrically morphed images that represent the features related to the aftermath of a BCCT, such as an asymmetry, in the body of a second patient and further work on this subject could be of great potential for the Cinderella Project.

Chapter 8

Experiments with Wasserstein GAN

In this final chapter, some preliminary results of a more complex approach, which aims at targeting most of the requirements for the Cinderella project, will be presented. These requirements include the search for the most similar image in the dataset to the query image and finally an attempt at the approximation of this image to the aesthetic results of the most similar one, while maintaining the physical characteristics of the original. The network used for these experiments is based on the Wasserstein GAN, which as mentioned in chapter 5, is more stable and easier to train. The methods for all of these steps will be further explained in the following sections. First the general methods, which were similar in all the experiments will be discussed, and then each individual experiment will be further elaborated on.

8.1 Search for the most similar image

As mentioned before, one of the final goals in the Cinderella project is to create a method that generates images representing the aesthetic outcome of a breast cancer procedure, by morphing two images, one of the current patient, before the treatment, and one post-treatment of a patient in the system's database. This second retrieved patient will be chosen based on their similarities with the current patient, taking into account a number of characteristics, namely their physical resemblance before the surgery, since, as explained in chapter 3, the aesthetic outcome is dependent on a number of physical factors, such as weight and breast size.

For the retrieval of these most similar 'past' patients, we use the method described in [58], and calculate the Euclidean distance between the final activation layer of the aesthetic classification model developed by Silva *et al.*.

8.2 WGAN model

The Wasserstein GAN with gradient penalty used was developed based on the implementations of [66] and [42].

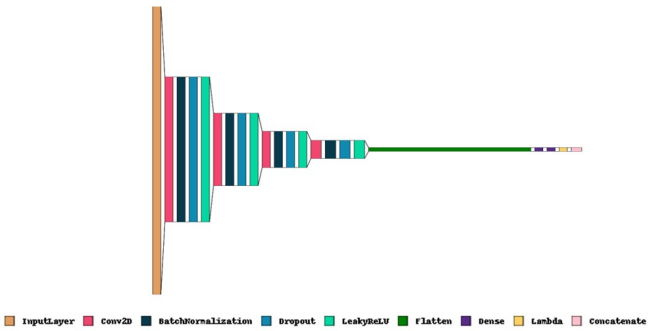


Figure 8.1: WGAN encoder architecture

Such as in [42], the generator is a variational autoencoder, and not a 'traditional' generator. As can be seen in figure 8.1, the encoder consists of a set of blocks of strided convolutional layers and batch normalization and dropout layers for regularization, followed by a LeakyReLU activation layer. Then, two fully connected layers are used to calculate the mean and standard deviation. Finally, a vector is sampled from the latent distribution and is fed to the decoder model (architecture in figure 8.2), which consists of blocks of transposed convolutional layers and Batch Normalization with a Leaky ReLU activation. The final activation layer has a Tanh function, and the generated images contain values in the range of -1 to 1. The discriminator or critic is a CNN classifier with 4 blocks of strided convolution layers with LeakyReLU activation. The final classification layer is fully connected with a linear activation. The critic's architecture is represented in figure 8.3.

8.3 Losses

The critic and the generator are trained separately, and a different loss function is implemented for each model. The critic is trained to minimize the Wasserstein Loss and to penalise the gradients

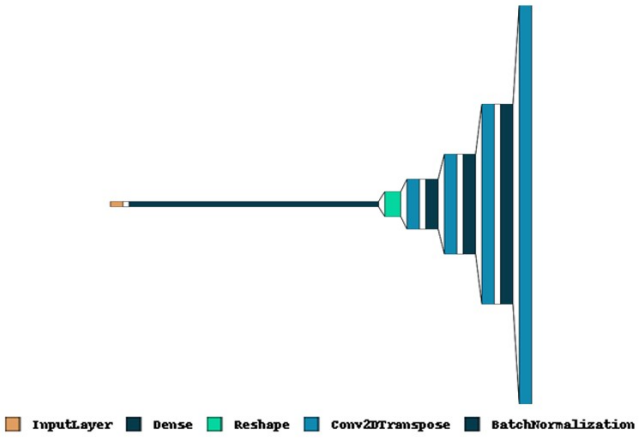


Figure 8.2: WGAN decoder architecture

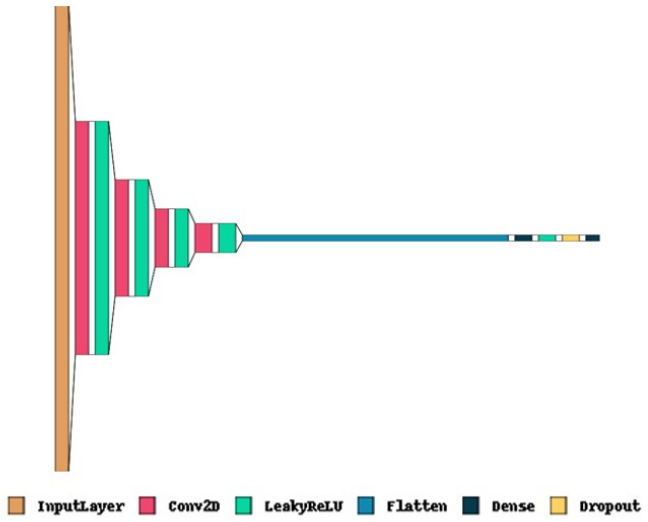


Figure 8.3: WGAN discriminator architecture

in order to enforce a Lipschitz constraint, which stabilizes training.

The generator is trained indirectly through the complete GAN model. The GAN's loss function contains four components: Kullback-Leibler divergence loss, to measure the distance between the encoder's distribution and the original distribution, SSIM loss between the generated image and the input image, categorical cross-entropy between the aesthetic classification of the new image and the retrieved most similar one, and finally the Wasserstein loss.

Each loss component has the same weight in the final loss function in both models, except the weight of the gradient penalty which is set to 10, as recommended in [66].

8.4 Training

During the training process of the WGANG-GP, as mentioned previously, the discriminator and the generator are trained separately, and the discriminator's error, along with the other factors, will be used to update the generator. During the training loop, we start by training the discriminator; the discriminator is trained on 3 different batches: first, only real samples, then only generated and finally on samples that are obtained by a weighted average between real and generated images. The discriminator's loss in this last case corresponds to the gradient penalty. This cycle is repeated 5 times, as in the original Wasserstein GAN paper, which suggests that is beneficial to train the discriminator more times than the generator. Then, the GAN model is trained twice. During this second loop, the discriminator is frozen and although it is used to update the GAN model, its weights are kept constant.

The model is trained for 125 epochs with a batch size of 12.

8.5 Validation process

It is important to constantly control he quality and the relevance and quality of the images generated by the GAN model, since it is a sensible model with a fleeting stability point. In order to collect the most useful results from the GAN, a validation step was introduced after each training loop. During this validation process, the images from the validation step go through the variational autoencoder. Then, the aesthetic classification of the generated images is obtained. The final loss corresponds to the sum of the SSIM value between the generated validation images and the original ones and the binary cross entropy value between the classification of the generated images and the most similar images retrieved from the validation set.

8.6 Results

In the following section, some results of the two experiments conducted with the Wasserstein GAN will be presented.

8.6.1 Experiment 1

At the first attempt at using the WGAN-GP to generate biometrically-morphed examples, the first step was to rescale the images to a range between -1 and 1. This step is fundamental, since the decoder will also generate images in this range, and we wish to train the discriminator into not being able to differentiate between the generated and the real images. Then, the dataset is divide into training (64%), validation (16%) and test (20%). From each of these sets, the similarity between each sample was calculated, and the most similar images were identified in order to be retrieved during the training. An example of a generated image after 123 epochs can be seen in figure 8.4.

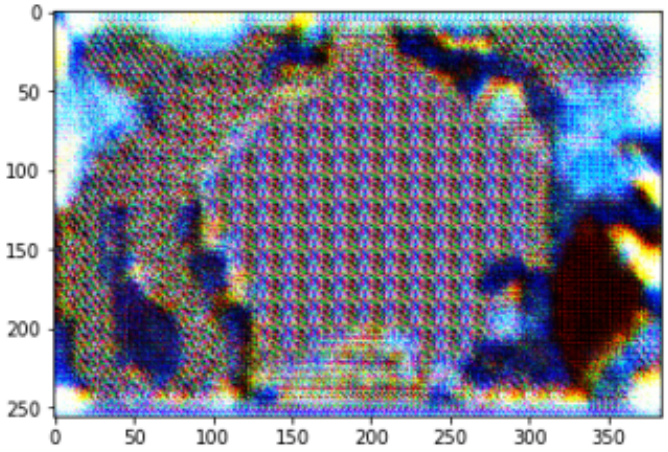


Figure 8.4: Generated image by the WGAN

8.6.2 Experiment 2

In this experiment, the set of 143 images is also divided into training, validation and test and rescaled to the range between -1 and 1. Just like the previous experiment, the most similar image on the dataset is retrieved from each separate set. The main difference in the methodology used in this experiment is that the two images, the original and the retrieved, are combined by interpolation, similarly to what happens in [47], where the goal is to morph faces. Each of the images is fed independently to the encoder, which generates a latent representation of each of the images. Then, a final vector of features is created by interpolating both of the representations, where each representation is equally accounted for. Finally, the decoder translates the final vector into an image, which should contain characteristics of both of the images. The rest of the training process remains constant. An example of a generated image after 122 epochs can be seen in figure 8.5.

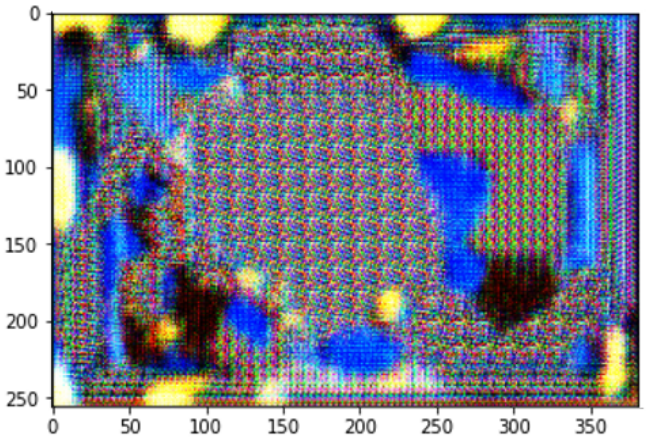


Figure 8.5: Generated image by the WGAN where both images are interpolated to create the new image

8.7 Discussion

The experiments with the Wasserstein GAN represent preliminary attempts at checking all the requirements for the Cinderella project and creating a more controlled model that took into account each of these requirements. Due to the high computational power required to train this model, the models were trained for a rather small number of epochs and the main point of this section was to expose the methodology that could be used in the future to create the morphed images.

Chapter 9

Conclusions and Future Work

9.1 Overview

Breast cancer is one of the most common cancers worldwide, affecting mostly women. For years, the most common treatment for breast cancer was a radical mastectomy, where all of the breast tissue was removed. Over the years, however, treatments and screening exams have evolved, and radical mastectomies are rarely performed, being substituted by more conservative treatments, such as the Breast Cancer Conservative Treatment, where there is an attempt at removing only the tumour and saving the breast tissue. With this evolution came also a new preoccupation with the aesthetic aspect of breast cancer surgery, which is often extremely important for the patient's quality of life. However, this aspect is still often overlooked when a patient is choosing the best course of treatment, which often leads to frustration over the aesthetic result and low self-esteem. Cinderella is a new project that aims to create automatized and personalized realistic predictions of the aesthetic results of a breast cancer treatment in order to show patients probable outcomes, so as to facilitate their choice and acceptance. These predictions are created based on the results of past patients.

This dissertation is framed under this project and serves as a preliminary study where different methods were explored to create biometrically morphed examples of possible aesthetic outcomes of breast cancer procedures.

For this, a literature review on the subjects of aesthetic evaluation of breast cancer treatments, content-based medical image retrieval and deep generative models was performed, which made it possible to acknowledge both the most traditional approaches and the most innovative ones.

Firstly, in chapter 6, transformations between paired images, where factors related to colour and asymmetry, that affect the aesthetic evaluation, were performed. It was concluded that, using a U-Net model and rather simple loss functions, it was possible to introduce alterations to the images that affected the breast aesthetic.

Next, in chapter 7, we attempted at performing unpaired image translation, where the goal was to approximate images of patients with excellent aesthetic results to images of patients with poor aesthetic results. For this task, models based on the Pix2Pix and on the cycleGAN were used.

This task proved to be much more challenging than the previous one, and showed less satisfactory results. With the cycleGAN, the images did not suffer a transformation that altered their aesthetic classification, while with the Pix2Pix, the resulting images suffered alterations, but do not possess realism and are often impossible to decipher and thus do not illustrate an unsatisfactory physical outcome.

Finally, in the last chapter, another complex and challenging experiment with a Wasserstein GAN was performed. This experiment represents a preliminary version of the model that accomplishes most of the needs of the Cinderella Project: a model that morphs two images, and generates a new one, similar to the one that represents the "current" patient but with the aesthetic characteristics of the other image, representing the retrieved one, from past cases. So far, the generated images are not representative and do not possess useful information.

9.2 Future Work

Despite the proven power of GANs, one of their most well-known limitations, common to most deep learning models, is the need for huge amounts of data. This represents the biggest obstacle in this dissertation, since the dataset used throughout the experiments contains only 143 images. For this reason, the future work based on this dissertation should include as a priority the acquisition of more labelled images of patients post BCCT. Then, these experiments can be repeated with the increased dataset and it will be expected that the results will be much more satisfactory.

Besides the tested models, other methods with great potential can be tested, such as Contrastive Unpaired Translation (CUT) [106], a successor method to cycleGAN that is lighter and faster to train and Style transfer, using for example the state of the art model StyleGAN [83]. The StyleGAN allows style mixing, where two latent spaces from two images are used to generate a morphed image containing characteristics from both images, which could be useful for this task.

Another possible approach that could be used to facilitate the task is to use the available breast keypoints or other extra information related to aesthetic evaluation, such as the indices used in the BCCT.core software to score the aesthetic outcome to guide the transformations.

9.3 Final remarks

In conclusion, several experiments related to image translation, either paired or unpaired were conducted, where the goal was to alter the characteristics that affect the aesthetic result after breast cancer treatments. These experiments represent novel approaches towards the problem target by the Cinderella project. Despite the unsatisfactory results obtained in the unpaired translation experiments, the work developed in this dissertation serves as a preliminary and first study for the project and contributes towards the understanding of different methods that can be used to reach the goal of creating biometrically morphed images and hopefully help women feel more informed and confident during their fight against breast cancer.

References

[1] American cancer society - limitation of mammograms. Available at =https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/limitations-of-mammograms.html.

[2] American cancer society. breast cancer facts figures 2017-2018. Available at =https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2017-2018.pdf.

[3] American cancer society. breast cancer facts figures 2019-2020. Available at =https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470.

[4] Breast cancer. Available at =https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470.

[5] Breast cancer facts and statistics. Available at =https://www.breastcancer.org/facts-statistics.

[6] Breast cancer: Statistics. Available at =https://www.cancer.net/cancer-types/breast-cancer/statistics.

[7] Cancer. World Health Organization. Available at =https://www.who.int/news-room/fact-sheets/detail/cancer.

[8] Cancer research uk - breast cancer survival statistics. Available at =https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/survivalheading-Three.

[9] Cancro da mama. Liga Portuguesa Contra o Cancro. Available at =https://www.ligacontracancro.pt/cancro-da-mama/.

[10] Chemotherapy. Available at =https://www.mayoclinic.org/tests-procedures/chemotherapy/about/pac-20385033.

[11] Do you regret having cosmetic surgery? Available at =hhttps://www.medicalaccidentgroup.co.uk/news/do-you-regret-having-cosmetic-surgery/.

[12] Everything you need to know about proms and prems.

[13] Google machine learning education - generative adversarial networks.

[14] Ibm - introduction to deep learning. Introduction to Deep Learning, =https://developer.ibm.com/articles/an-introduction-to-deep-learning/, journal=IBM developer, author=Madan, Piyush and Madhavan, Samaya.

[15] Ibm - what are convolutional neural networks?

[16] Introduction to deep learning. Geeks for Geeks, =https://www.geeksforgeeks.org/introduction-deep-learning/, year=2019, month=Apr.

[17] An introduction to machine learning. Geeks for Geeks. Available at =https://www.geeksforgeeks.org/introduction-machine-learning.

[18] Invasive ductal carcinoma (breast). Available at =https://www.mypathologyreport.ca/breast-invasive-ductal-carcinoma/.

[19] Médis - 6 mil novos casos de cancro da mama por ano. Available at =https://www.medis.pt/mais-medis/cancro/6-mil-novos-casos-de-cancro-da-mama-por-ano/.

[20] Surgery to remove breast cancer (breast conserving surgery). Available at =https://www.cancerresearchuk.org/about-cancer/breast-cancer/treatment/surgery/remove-just-area-cancer.

[21] Surgery to remove your breast (mastectomy). Available at =https://www.cancerresearchuk.org/about-cancer/breast-cancer/treatment/surgery/remove-whole-breast.

[22] Understanding cancer risk. Available at =https://www.cancer.net/cancer-types/breast-cancer.

[23] What is cancer? Available at =https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[24] Min Soon Kim Gregory P. Reece] Nitin Udpa, Mehul P. Sampat and Mia K. Markey. Objective assessment of the aesthetic outcomes of breast cancer treatment: toward automatic localization of fiducial points on digital photographs. *SPIE 6514, Medical Imaging 2007*.

[25] Nordic Cochrane Centre 2012. Screening for breast cancer with mammography.

[26] Napel S Beaulieu CF Greenspan H Acar B Akgül CB, Rubin DL. Content-based image retrieval in radiology: current status and future directions. *J Digit Imaging 2011*.

[27] Anthony E. Dragun Allison M. Hunter. Rethinking the harvard scale of breast cosmesis: An analysis of patients treated with breast-conserving therapy on a phase ii clinical trial. *2013 Breast Cancer Symposium*.

[28] Alexander Amini. Introduction to deep learning, mit 6.s191.

[29] P. Angelov and E. Soares. Towards deep machine reasoning: a prototype-based deep neural network with decision tree inference, 2020.

[30] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[31] C. Rudin B. Kim and J. Shah. he bayesian case model: A generative approach for casebased reasoning and prototype classification. *NIPS'14, page 1952–1960, Cambridge, MA, USA, 2014. MIT Press*.

[32] R. Khanna B. Kim and O. O. Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Neural Information Processing Systems*.

[33] N Olivari B Stark. Breast asymmetry: an objective analysis of postoperative results. *European Journal of Plastic Surgery, 1991*.

[34] Hmrishav Bandyopadhyay. Introduction to autoencoders [types, training, applications].

[35] Nilesh Barla. Deep learning 101: Introduction [+pros, cons amp; uses].

[36] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*

[37] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[38] D. Tao A. Barnett C. Rudin C. Chen, O. Li and J. K. Su. This looks like that: Deep learning for interpretable image recognition. *Neural Information Processing Systems*.

[39] D. Doyle C. Nugent and P. Cunningham. Gaining insight through case-based explanation. *J. Intell. Inf. Syst.,06 2009*.

[40] Jaime Cardoso and Maria Cardoso. Breast contour detection for the aesthetic evaluation of breast cancer conservative treatment. *Advances in Soft Computing*.

[41] R et al Caruana. Case-based explanation of non-case-based learning methods. *AMIA Symposium (1999)*.

[42] Jiawei Chen, Janusz Konrad, and Prakash Ishwar. Vgan-based image representation learning for privacy-preserving facial expression recognition.

[43] Kowshik chilamkurthy. Wasserstein distance, contraction mapping, and modern rl theory., Oct 2020.

[44] João Miguel Doutor Covas. *O papel do Exame Clínico da Mama como método de rastreio de Cancro da Mama*. PhD thesis, Faculdade de Medicina Lisboa.

[45] J. A. Christie T. Kron S. A. Ferguson C. S. Hamilton D. R. H. Christie, M. Y. O'Brien and J. W. Denham. A comparison of methods of cosmetic assessment in breast conservation treatment. *The Breast*.

[46] J. A. Christie T. Kron S. A. Ferguson C. S. Hamilton D. R. H. Christie, M. Y. O'Brien and J. W. Denham. A comparison of methods of cosmetic assessment in breast conservation treatment. *The Breast*.

[47] Naser Damer, Fadi Boutros, Alexa Moseguí Saladié, Florian Kirchbuchner, and Arjan Kuijper. Realistic dreams: Cascaded enhancement of gan-generated images with an example in face morphing attacks. 2019.

[48] Arden Dertat. Applied deep learning - part 3: Autoencoders, 2017.

[49] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp, 2016.

[50] S.; Besold T.R. journal = arXiv 2017 Doran, D.; Schulz. Mwhat does explainable ai really mean? a new conceptualization of perspectives.

[51] Dr. Wei Wei, Prof. James Landay. ML Interpretability and Intrinsic Models - Lecture Notes CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning, Stanford University.

[52] Shiv Ram Dubey. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32:2687–2704, 2022.

[53] Emily Jia Sameer Singh Dylan Slack, Sophie Hilgard and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*.

[54] E. van der Schueren E. Van Limbergen and K. Van Tongelen. Cosmetic evaluation of breast conserving treatment for mammary cancer. *Cosmetic evaluation of breast conserving treatment for mammary cancer*.

[55] Zheng W et al. Epplein M, Zheng Y. Quality of life after breast cancer diagnosis and survival. *Journal of Clinical Oncology*.

[56] Cardoso et al. Automatic breast contour detection in digital photographs. *In Proceedings of the First International Conference on Health Informatics*.

[57] Tiago Gonçalves et al. A novel approach to keypoint detection for the aesthetic evaluation of breast cancer surgery outcomes. *Health and Technology 10.4*.

[58] Wilson Silva et al. Deep aesthetic assessment and retrieval of breast cancer treatment outcomes. *IbPria*.

[59] Wilson Silva et al. Deep keypoint detection for the aesthetic evaluation of breast cancer surgery outcomes. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*.

[60] H. Trischler L. Wutzel O. Riedl U. Kühbelböck B. Wintersteiner M.J. Cardoso P. Dubsky M. Gnant R. Jakesz F. Fitzal, W. Krois and T. Wild. The use of a breast symmetry index for objective evaluation of breast cosmesis. *The Breast*.

[61] Max Ferguson, Ronay ak, Yung-Tsun Lee, and Kincho Law. Automatic localization of casting defects with convolutional neural networks. pages 1726–1735, 12 2017.

[62] Tiago Gonçalves. *Deep Aesthetic Assessment of Breast Cancer Surgery Outcomes*. PhD thesis, FEUP, 2019.

[63] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks, 2017.

[64] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[65] X. Gu and W. Ding. A hierarchical prototype-based approach for classification. *Information Sciences, 2019*.

[66] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017.

[67] J. S. Cardoso H. Montenegro, W. Silva. Towards privacy-preserving explanations in medical image analysis.

[68] Hyungrok Ham, Tae Joon Jun, and Daeyoung Kim. Unbalanced gans: Pre-training the generator of generative adversarial network using variational autoencoder, 2020.

[69] Suzana Herculano-Houzel. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences*.

[70] Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design, 2019.

[71] Choi D Hwang KH, Lee H. Medical image retrieval: past and present. *Healthc Inform Res 2012*.

[72] Pavel Iakubovskii. Segmentation models. https://github.com/qubvel/segmentation_models, 2019.

[73] Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation, 2018.

[74] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2016.

[75] A. Hong D. Stoney M. Law J. James, E. Ip. "the perfect breast": Measuring cosmetic outcomes after breast-conserving therapy. *International Journal of Surgery Open*.

[76] Hélder P. Oliveira Jaime S. Cardoso, Inês Domingues. Closed shortest path in the original coordinates with an application to breast cancer. *International Journal of Pattern Recognition and Artificial Intelligence*.

[77] Maria J. Cardoso Jaime S. Cardoso. Cosmetic outcomes and complications reported by patients having undergone breast-conserving treatment. *Astro*.

[78] Maria J. Cardoso Jaime S. Cardoso. Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. *Artificial Intelligence in Medicine*.

[79] Maria J. Cardoso Jaime S. Cardoso, Wilson Silva. Evolution, current challenges, and future possibilities in the objective assessment of aesthetic outcome of breast cancer locoregional treatment. *The breast*.

[80] Goran Svensson Jay R. Harris, Martin B. Levene and Samuel Hellman. Analysis of cosmetic results following primary radiation therapy for stages i and ii carcinoma of the breast. *International Journal of Radiation Oncology*Biology*Physics, February 1979*.

[81] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016.

[82] Joon Beom Seo Sang Min Lee Jihye Yun Min-Ju Kim Jewon Jeong Youngsoo Lee Kiok Jin Rohee Park Jihoon Kim Howook Jeon Namkug Kim Jaeyoun Yi Donghoon Yu Jooae Choe, Hye Jeon Hwang and Byeongsoo Kim. Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest ct. *Radiology 2022 302:1, 187-197*.

[83] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2018.

[84] Eoin Keane, Mark Kenny. How case based reasoning explained neural networks: An xai survey of post-hoc explanation-by-example in ann-cbr twins.

[85] Reece GP Miller MJ Beahm EK Markey MK Kim MS, Sbalchiero JC. Assessment of breast aesthetics. *Plast Reconstr Surg. 2008.*

[86] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions, 2018.

[87] Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow, 2016.

[88] Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, nov 2021.

[89] Reece GP Markey MK Lee J, Muralidhar GS. A shape constrained parametric active contour model for breast contour detection. *Annu Int Conf IEEE Eng Med Biol Soc*.

[90] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized {gan} training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*, 2021.

[91] Y. Mizukami A. Nonomura N. Ohta N. Koyasaki T. Taniya M. Noguchi, Y. Saito and I. Miyazaki. Breast deformity, its correction, and assessment of breast conserving surgery. *Breast Cancer Research and Treatment*.

[92] Thomas Wild Wilfried Krois Florian Fitzal. Maria João Cardoso, Jaime S. Cardoso. Comparing two objective methods for the aesthetic evaluation of breast cancer conservative treatment. *Breast Cancer Research and Treatment*.

[93] Cade Metz. Google's dueling neural networks spar to get smarter, no humans required, 2017.

[94] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[95] Helena Montenegro. *privacy-preserving framework for case-based interpretability in machine learning*. PhD thesis, FEUP.

[96] Helena Montenegro, Wilson Silva, and Jaime S. Cardoso. Privacy-preserving generative adversarial network for case-based explainability in medical image analysis. *IEEE Access*, 2021.

[97] C Muramatsu. Overview on subjective similarity of images for content-based medical image retrieval. *Radiol Phys Technol 11 (2018).*

[98] Ladislas Nalborczyk. A gentle introduction to deep learning in r using keras - lecture notes, aix marseille university, cnrs, lpc, lnc.

[99] Michael Nation. pix2pix-edges-with-color. https://github.com/michaelnation26/pix2pix-edges-with-color, 2019.

[100] Frank Nielsen and Fré déric Barbaresco, editors. *Geometric Science of Information*. Springer International Publishing, 2021.

[101] C. Chen O. Li, H. Liu and C. Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. *AAAI, 2018*.

[102] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders, 2016.

[103] Daniela Ovadia. Ai will help cinderella to see herself in the mirror. *Cancer World*.

[104] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation, 2017.

[105] McDaniel P. Papernot, N. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*.

[106] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation, 2020.

[107] Junjie Peng, Elizabeth Jury, Pierre Dönnes, and Coziana Ciurtin. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: Applications and challenges. *Frontiers in Pharmacology*, 12, 09 2021.

[108] Florian Perteneder. Understanding black-box ml models with explainable ai. Available at =https://engineering.dynatrace.com/blog/understanding-black-box-ml-models-with-explainable-ai/.

[109] Eduardo Soares Plamen Angelov. Towards explainable deep neural networks (xdnn). *Neural Networks Volume 130, 2020*.

[110] Adnan Qayyum, Syed Muhammad Anwar, Muhammad Awais, and Muhammad Majid. Medical image retrieval using deep convolutional neural network. *Neurocomputing*, 266:8–20, nov 2017.

[111] L R Hill N Vora K R Desai-J O Archambeau J A Lipsett R D Pezner, M P Patterson. "breast retraction assessment: an objective evaluation of cosmetic results of patients treated conservatively for breast cancer. *International Journal of Radiation Oncology*Biology*Physics, March 1985*.

[112] J. F. Pinto da Costa R. Sousa, J. S. Cardoso and M. J. Cardoso. Breast contour detection with shape priors. *2008 15th IEEE International Conference on Image Processing*.

[113] Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. *CoRR*, 2019.

[114] Díaz-Agudo B. Pino-Castilla V. Recio-García, J.A. Cbr-lime: A case-based reasoning approach to provide specific local interpretable model-agnostic explanations. *Watson, I., Weber, R. (eds) Case-Based Reasoning Research and Development. ICCBR 2020. Lecture Notes in Computer Science(), vol 12311. Springer, Cham*.

[115] Lyle Regenwetter, Amin Heyrani Nobari, and Faez Ahmed. Deep generative models in engineering design: A review. *arXiv*, 2021.

[116] Sakib Reza, Ohida Binte Amin, and M.M.A. Hashem. Transresunet: Improving u-net architecture for robust lungs segmentation in chest x-rays. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 1592–1595, 2020.

[117] Nayana L. Vora Richard D. Pezner, James A. Lipsett and Kanta R. Desai. Limited usefulness of observer-based cosmesis scales employed to evaluate patients treated conservatively for breast cancer. *International Journal of Radiation Oncology*Biology*Physics*.

[118] Joseph Rocca. Understanding variational autoencoders (vaes), 2021.

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

[120] J. Stewart A. Morgan S. Al-Ghazal, R. Blamey. The cosmetic outcome in early breast cancer treated with breast conservation. *European J Surg Oncol Ejso*.

[121] Sumit Saha. A comprehensive guide to convolutional neural networks - the eli5 way, Dec 2018.

[122] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications, 2017.

[123] Poellinger A. Cardoso J.S. Reyes M. Silva, W. Interpretability-guided content-based medical image retrieval. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science(), vol 12261*.

[124] Lancellotta V. Kovács G. et al. Soror, T. kobcs©: a novel software calculator program of the objective breast cosmesis scale (obcs). *Breast Cancer 27*.

[125] Anthony B.Miller Steven A.Narod, JavaidIqbal. Why have breast cancer mortality rates declined? *Journal Cancer Policy*.

[126] Nikhil Tomar. U-net with pretrained mobilenetv2 as encoder. https://github.com/nikhilroxtomar/Unet-with-Pretrained-Encoder/blob/master/U-Net_with_Pretrained_MobileNetV2_as_Encoder.ipynb.

[127] L. I. Tsouskas and I. S. Fentiman. Breast compliance: A new method for evaluation of cosmetic outcome after conservative treatment of early breast cancer. *Breast Cancer Research and Treatment*.

[128] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *arXiv*, 2020.

[129] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *CoRR*, abs/1601.06759, 2016.

[130] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *CoRR*, 2017.

[131] Mariani L Greco M Saccozzi R Luini A Aguilar M-Marubini E. Veronesi U, Cascinelli N. Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. *The New England Journal of Medicine*.

[132] M. J. Cardoso W. Silva, K. Fernandes and J. S. Cardoso. Towards complementary explanations using deep neural networks. *Understanding and Interpreting Machine Learning in Medical Image Computing Applications, pages 133–140, Cham, 2018. Springer International Publishing*.

[133] Lei Wang, Wei Chen, Wenjia Yang, Fangming Bi, and Fei Richard Yu. A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access*, 8, 2020.

[134] Jang NY et al. Yu T, Eom KY. Objective measurement of cosmetic outcomes of breast conserving therapy using bcct.core. *Cancer Res Treat*.

[135] Y. Bei Z. Chen and C. Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence, Dec 2020*.

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

[137] Runtong Zhang, Yuchen Wu, and Keiji Yanai. Pre-trained and shared encoder in cycle-consistent adversarial networks to improve image quality. In Shivakumara Palaiahnakote, Gabriella Sanniti di Baja, Liang Wang, and Wei Qi Yan, editors, *Pattern Recognition*. Springer International Publishing, 2020.

[138] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017.