POLITECNICO DI TORINO

Master's Degree in Biomedical Engneering

Master's Degree Thesis

Improving Optical Coherence Tomography Angiography through GAN-based Techniques

Supervisors

Prof. Kristen M. MEIBURGER

Dr. Giulia ROTUNNO

Prof. Massimo SALVI

Candidate

Vilma DOGA

Academic year 2023/2024

Summary

Optical Coherence Tomography Angiography (OCTA) is a powerful imaging technique that provides non-invasive visualization of vascular networks, primarily in the eye but also in the skin. By capturing multiple optical coherence tomography (OCT) B-scans at the same location over time, OCTA identifies motion contrast arising from flowing blood cells.

In this thesis, we propose a novel approach to accelerate OCTA data acquisition by leveraging Generative Adversarial Networks (GANs). The primary objective is to generate high-quality OCTA enface images from lower-quality inputs, which are typically derived from only two OCT volumes, compared to the higher-quality images used as ground truth that require four OCT volumes.

The dataset utilized in this thesis includes images obtained from skin samples in both healthy individuals and patients diagnosed with Chronic Venous Insufficiency (CVI). The chosen format for the study is the Median Intensity Projection (MIP) of 5 slices, obtained in the enface plane. To ensure robustness and generalizability, the dataset is properly partitioned into training (70%), validation (20%), and test (10%) sets. This work presents two different approaches to image quality improvement of OCTA using GAN: super-resolution GAN (SRGAN) and pixel2pixel GAN. The SRGAN proposed is more specifically an Enhanced Super-Resolution GAN wherein the input and output images are of the same resolution allowing enhancement of image quality. This methodology integrates a Residual-in-Residual DenseNet as a generator with a Patch GAN serving as the discriminator. In addition, the introduction of the VGG19 model further enhances perceptual quality by aligning generated images with high-level features extracted from the reference image.

The Pixel2Pixel GAN (Pix2Pix) architecture employing a U-Net as its generator represents a robust framework for image translation from low to high-quality versions. The U-Net structure facilitates precise reconstruction by capturing fine details from the low-resolution input and transforming them into high-resolution outputs with enhanced fidelity. Complementing this, a PatchGAN discriminator assesses local image patches to ensure the generated images exhibit realistic textures and structures akin to authentic high-quality counterparts. In addition to visual assessment, objective metrics such as PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index), and MSSSIM (Multi-Scale Structural Similarity Index) were utilized to quantitatively evaluate image quality. Across all evaluations, MSSSIM exceeded 90%, indicative of high fidelity in image reconstruction. PSNR values hovered just below 30 dB, further affirming the overall quality of the reconstructed images.

In conclusion, our proposed methods utilizing SRGAN and Pixel2Pixel GAN manage to speed up acquisition times as it manages to achieve good image quality even with the acquisition of a lower number of OCT volumes.

Table of Contents

List of Figures

Part I

Introduction

Chapter 1

Optical Coherence Tomography

Optical Coherence Tomography (OCT) is a cutting-edge imaging modality that has significantly impacted biomedical imaging, particularly in the field of ophthalmology due to its non-invasive nature. OCT is an optical analogous of ultrasound B-mode imaging which uses light instead of sound waves to image a three-dimensional, subsurface, micro-structural image of soft tissue. The technique exploits two main properties of light to produce detailed images:

Coherence For two waves to be considered coherent, both need to have same frequency and waveform and their wavelengths need to have a fixed phase relationship. The propagation distance over which the coherence significantly decays is called coherence length.

Interference It occurs when several waves meet while traveling. The effect of this interaction can be constructive, if the waves are in phase, or destructive, if the waves are out of phase.

An important optical configuration which is used in obtaining information about the sample reflection to study its microstructure with coherence method is interferometry, more specifically low coherence interferometry.

1.1 Low Coherence Interferometry

Low coherence interferometry (LCI) represents a highly advanced optical technique leveraging the Michelson interferometer configuration to perform high-resolution, depth-resolved measurements and imaging. The core principle of LCI revolves around the use of a low coherence light source, typically a broadband source such as a superluminescent diode or a femtosecond laser, which possesses a short coherence length. This short coherence length is crucial because it allows the system to discriminate between different reflecting surfaces along the optical path, effectively reducing the impact of multiple reflections and enhancing measurement accuracy. The key difference is the light source that is used. Traditionally, a Michelson interferometer uses a continuous wave laser source. A continuous wave laser has a very large coherence length. The coherence length is defined as:

$$l_c = \frac{2ln2}{\pi}\frac{\lambda_0^2}{\Delta\lambda} \tag{1.1}$$

Where λ_0 is the light's central wavelength and $\Delta\lambda$ its spectral width.

Considering for example a monochromatic source, it will have a small bandwidth ($\Delta\lambda \to 0$) and so a very large $\mathrm{l_c}$. The interference signal, in this case, will present a lot of maxima caused by the constructive interference of the different sample beams at different positions. The intensity of interference is dependent on the optical path length difference. Therefore, the longer the coherence length, the bigger the sampling volume into the depth of the sample. This can be detrimental to the resolution of the system. Depth resolution, particularly in axial scanning, is often a key issue for many biological and industrial samples. By using a low coherence light source, and thus reducing the coherence length, the interference function becomes a δ function and the coherence length becomes the sampling volume in the sample. This allows the coherence length and subsequently the resolution to be tailored to the needs of the sample. This is a very important factor for OCT where depth resolution in imaging biological tissue is critical.

1.1.1 Michelson interferometer

In an LCI system, light from the broadband source is typically directed into a Michelson interferometer, where it is split into two paths by a beamsplitter. One path, known as the reference arm, directs light to a reference mirror, while the other path, known as the sample arm, directs light to interact with the sample under investigation. The reflected or backscattered light from both arms is then recombined at the beamsplitter, creating an interference pattern only if the path length difference between the two arms is within the coherence length of the light source. By scanning the reference arm or the sample, or both, the system can obtain an interference signal as a function of depth, enabling the precise localization and characterization of internal structures within the sample.

The average intensity of the signal collected from the detector at the interferometer output is:

$$\Big\langle I_E(t;\Delta t)\Big\rangle = \Big\langle I_S(t)\Big\rangle + \Big\langle I_R(t)\Big\rangle + G_{SR}(\Delta t) \tag{1.2}$$

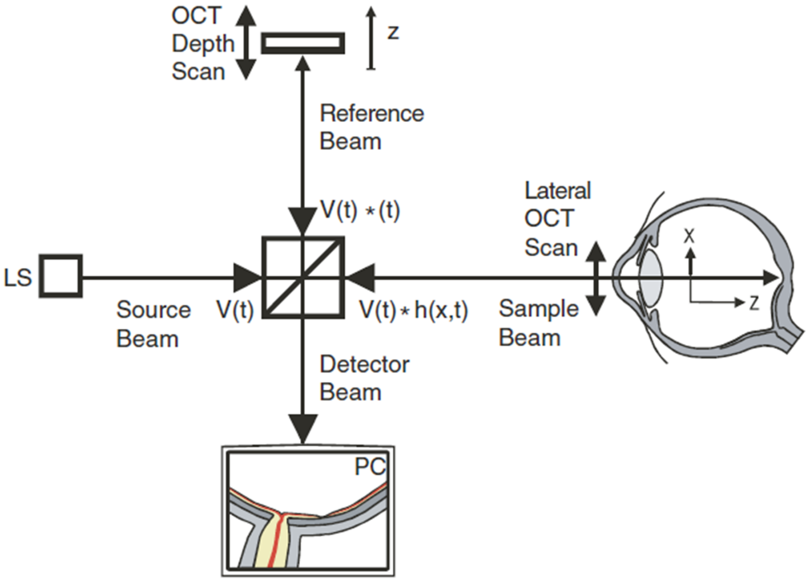


Figure 1.1: Standard OCT scheme based on a low time-coherence Michelson interferometer [1].

$I_R(t)$ is the intensity of the beam arriving at the detector after hitting the reference mirror, $I_S(t)$ is the intensity of the beam arriving at the detector if there would be just the scattering object and not the reference mirror and $G_{SR}(\Delta t)$ is the interference term or cross-correlation term. The Michelson interferometer has been widely employed in many forms throughout the evolution of OCT. The unique ability to vary the difference in optical path of the sample and reference arm has made it an invaluable tool for the development and study of new OCT techniques.

1.2 OCT Techniques

There are two types of OCT techniques: Time-Domain OCT (TD-OCT), Fourier-Domain OCT (FD-OCT). The Fourier-Domain OCT can then be classified in Sprectal-domain OCT (SD-OCT) and Swept-Source OCT (SS-OCT). They mainly differ in the sample scanning protocol adopted.

1.2.1 Time Domain OCT (TD-OCT)

The time-domain optical coherence tomography (TD-OCT) technique was the first OCT technique and was widely used clinically, but has been largely supplanted by various forms of Fourier-domain OCT. In TD-OCT, the reference arm mirror is moved to change the path length of the reference arm and the interference spectrum is measured as a function of the reference mirror position. Thus, the TD-OCT requires a broadband light source, a moving reference mirror and a single photo detector. The need to mechanically change the positions of the laser and of the reference mirror to get the whole volume image makes the acquisition extremely time-consuming and prone to additional noise.

1.2.2 Fourier-Domain OCT (FD-OCT)

To overcome the Time Domain's system constraints, new techniques were introduced such as the Fourier Domain OCT. FD-OCT removes the need for a moving reference arm, enabling a much faster imaging speed, up to 100 times faster than TD-OCT, and with a much higher signal-to-noise ratio (SNR). This is achieved by substituting the time axis with a frequency axis, by using a broad bandwidth light source. This allows the detection of all reflections from every point in the tissue simultaneously.

Spectral Domain OCT

In spectral domain OCT (SD-OCT), scanning is performed by spectrally modulating and measuring the interference term of the reference and sample beams. SD-OCT separates the spectrum with the use of a grating and detects the light with a CCD camera. So instead of the photo detector, a spectrometer is used to separate the spectrally dispersed light into distinct wavelengths and their respective amplitudes and phases carrying the information on sample structure are detected [2]. SD-OCT is apparently faster and more sensitive than TD-OCT [3], giving it the advantage of producing in-vivo, real-time cross-sectional images of biological tissues with higher quality. These advantages come at expenses of a higher complexity, cost of production and high roll-off, i.e. the gradual loss of sensitivity as a function of depth.

Swept-Source OCT

SS-OCT uses a tunable narrowband source to scan through different wavelengths, only allowing a single wavelength to reach the sample at any one time. This method relies solely on a photodetector rather than analyzing an entire spectrum of light. By focusing on the intensity of light detected at specific points, it achieves high detection sensitivity, making it more efficient and effective in identifying subtle changes or signals. A key advantage of SS-OCT is that data acquisition speed is limited only by the detection electronics. This is the kind of equipment located at the Center for Medical Physics and Biomedical Engineering at the Medical University of Vienna (MUW) and used to acquire all the data used for this thesis.

1.3 OCT Volume Acquisition

A single OCT scan produces the reflectivity profile of the sample in depth. This is also called A-Scan a unidimensional axial scan in the z direction. To create a bidimensional scan, also called B-scan, a series of A-scans are collected adjacently in the x direction. The x direction, representing the laser direction, is also referred as the fast-scanning direction. Consequently, the y direction is also called the slow scanning direction.

A B-scan lays in the xz plane. A series of B-scans acquired along the y direction produce a C-scan so the entire 3D volume or also called tomogram.

From the volume, it is possible to visualize the bidimensional images laying in the xy plane. This visualization from the top is called en-face and is the main representation that will be showed in the coming chapters.

Moreover, mostly MIPs will be shown. MIP stands for Median Intensity Projection that is a method of generating a two-dimensional image from a three-dimensional data set, by choosing the median pixel value for intensity along the projection ray. It is more effective in diminishing the effects of noise and can give a better depiction of the internal structures.

Formally, for an XY-plane projection (enface), the median intensity M(x,y) at pixel (x,y) is calculated as:

$$M(x,y) = median\{V(x,y,z)|z \in range\ of\ z\} \tag{1.3}$$

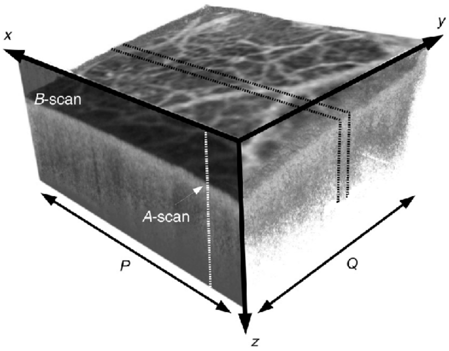


Figure 1.2: OCT image volume and associated coordinate system. Each B-scan consists of P A-scans, acquired sequentially along the x -direction. An OCT volume consists of Q B-scan images acquired along the y -direction. Adapted from [4].

Chapter 2

Optical Coherence Tomography Angiography

Optical coherence tomography angiography (OCTA) is a quickly developing imaging technique that provides depth-resolved information on the vasculature of a tissue bed. The development and validation of the technology have come about from the availability of Fourier-domain OCT systems. It is a functional extension of optical coherence tomography (OCT) with the capability to extract information about flow from tissue using endogenous light scattering. The idea of differentiating tissue structure from blood flow has been described previously, but with the development of OCTA we are able to visualize the vasculature at a higher resolution and contrast, and without the need for contrast agents as required in previous angiography techniques.

2.1 OCTA volume reconstruction

OCTA enhances OCT by focusing on temporal variations in the OCT signal. Under the assumption that erythrocytes are the only moving scatterers in biological tissue, by capturing multiple OCT B-scans at the same location over time, OCTA identifies motion contrast arising from flowing blood cells

While conventional OCT separately detects and compares the strength of light waves returning from the tissue to a reference reflection, OCT Angiography records the changes which occur to signal when light is back-scattered from the vasculature that OCT aims to detect. The location of these changes in OCT and its magnitude represents the structure and depth of a blood vessel and its flow. OCT Angiography generates contrast by detecting the movement of particles (i.e. red blood cells) in response to a gradient in theoretical force and functioning as a high-pass filter to differentiate vascular flow from static tissue. The fast axial movement of particles causes phase variation in the complex coherence signal, whereas slow flow or static tissue will result in attenuation of higher frequency signals.

This motion contrast is detected using either intensity-based or phase-based methods.

2.1.1 Intensity-Based OCTA

These methods, such as speckle variance and decorrelation, rely on changes in the intensity of the Fourier transform of the OCT signal between consecutive B-scans. The motion contrast volume A(x, y, z) is obtained by averaging over pairwise differences of subsequent logarithmically-scaled intensity tomograms log(A(x,z)) from the same set at given position y [5]:

$$A(x,y,z) = \frac{1}{N-1} \sum_{i=1}^{N-1} |A(x,z)_{i+1} - A(x,z)_i| \tag{2.1}$$

where A is the amplitude of the OCT signal, N the total number of consecutive B-scans compared in pairs, i is the i-th B-scan and (x, z) the tomogram coordinates along the fast scan direction and the axial direction, respectively. This kind of approach is of easy implementation and robust because of the lower sensitivity to phase noise and trigger jitter.

2.1.2 Phase-Based OCTA

These methods analyse the phase stability of the OCT signal. Changes in the phase of the reflected light indicate motion and they can be calculated as:

$$\Delta\phi(x,z)_j = arg\Big\{exp[-i\phi(x,z)_{j+1}]exp[-i\phi(x,z)_j]\Big\} \tag{2.2}$$

OCTA is obtained as the average of the phase differences at each location [5]:

$$A(x,y,z) = \frac{1}{N-1} \sum_{j=1}^{N-1} |\Delta\phi(x,z)_j| \tag{2.3}$$

The independency from the intensity usually results in better contrast for the images but not in absolute. In facts the higher sensitivity to phase noise entails an attention to sample displacement. Therefore bulk motion correction is required: for each A-scan, the circularly averaged phase difference $\Delta\phi(\mathrm{x})_\mathrm{j}$ is subtracted from the respective A-scan located at x.

2.2 Advantages and limitations

One of the main advantages of OCTA over conventional angiography imaging techniques is the non-invasive nature of the imaging. In computed tomography angiography and conventional angiography, it is necessary to inject a contrast material into the bloodstream and produce a two-dimensional projection image of three-dimensional vasculature. This carries with it certain risks including anaphylaxis in those who have a contrast material allergy, kidney damage in those with pre-existing kidney disease, and risks associated with the invasive nature of venipuncture for the injection. This makes OCTA ideal for monitoring disease progression and response to treatment over time, as it allows for consistent and repeatable imaging without cumulative risk.

In addition to this, a more common drawback of conventional angiography is the inability to effectively visualize the microvasculature. Instead, OCTA provides high-resolution, down to a few μm, and a penetration depth of about 1-1.5 mm allowing for detailed visualization of blood flow and vascular structures.

There are though also some limitations associated. In dermatology application, where the skin thickness can reach 5 mm, the penetration depth could not be enough to fully view the deeper layers.

Also, as any imaging technique, OCTA is affected by different types of artifacts such as projection and motion artifacts. The first occurs when shadows or signals from superficial blood vessels are projected onto deeper layers, creating false representations of blood flow or vessel structures [6]. This can lead to misinterpretations of the images. With in-vivo imaging there's always the possibility of motion artifacts which are distortions in the imaging data caused by patient or tissue movement during the scanning process. More often they present as vertical white lines in the en-face image, meaning a total loss of information for the b-scans affected.

Moreover, despite the scanning process being quite fast, typically taking only a few seconds, the overall session duration is prolonged due to the time needed to save data on the computer, which scales with the number of repetitions of each B-scan. However, conducting a high number of repetitions is crucial for achieving high-quality OCTA.

This thesis aims to address these challenges by aiming to shorten scan times and reduce the amount of data collected. This approach seeks to achieve several benefits: faster examination times for patients, reduced memory storage requirements, maintaining high OCTA quality, and lowering the likelihood of both motion artifacts and artifacts caused by laser power instability (resulting in more pronounced vertical lines) .

2.3 Clinical Application

OCTA was primarily applied in the domain of ophthalmology; however, it has started to gain various other fields of use recently. This technique makes it possible to assess blood flow different structures, such as skin. It could provide functional angiographic information that is helpful in distinguishing between diseases that clinically present as similar in skin. This promising capability puts OCTA as an effective and important tool in dermatology allowing early diagnoses and improved patients' prognosis.

2.3.1 The skin

The skin, the body's largest organ, is composed of three primary layers: the epidermis, the dermis, and the hypodermis all of which have unique function and composition.

The outermost layer, the **epidermis**, is a self-renewing epithelial tissue organized into four sub-layers:

- Basal layer
- Spinous layer
- Granular cell layer
- Stratum corneum

The epidermis ranges from 60 to 600 μm [7][8] in thickness depending on the factors such as age, skin colour and location of the body part.

Below the epidermis is the **dermis** layer that has connective tissues with an abundant number of nerves, blood vessels, and lymphatics. The dermis as well has sub divided into two layers that are the papillary and the reticular layer the two however are very closely linked. This layer also contains dense capillary network which is crucial for blood circulation that are subpapillary plexus, reticular dermal plexus, and deep dermal plexus. The dermis is usually comparatively thin varying from 1-4 mm in thickness [7][8].

The inner most layer is the **hypodermis** or subcutaneous tissue which include mainly of adipose tissue with the principal functions of insulation against heat and cold, cushioning and energy reserve. Originally, it ranges from 5 mm to 20 mm, depending on the part of the body where it is fixing.

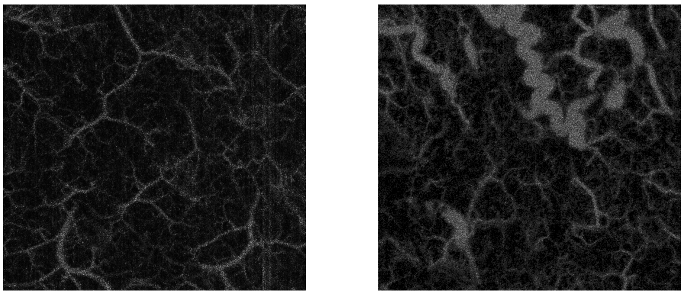
OCT, with its depth penetration, mainly focuses on the epidermis and dermis compartments. In particular, it is possible to capture detailed images of the skin that illustrate the two distinct plexuses within the dermis: the subpapillary and the deep dermal plexus. This process involves creating two separate Median Intensity Projections (MIPs). These MIPs are essential for highlighting and characterizing the differences in the blood vessels under various conditions. By analyzing these projections, we can better understand the structural and functional variations in the dermal vasculature, which can be crucial for diagnosing and studying different skin conditions and vascular disorders such as Chronic Venous Insufficiency (CVI).

2.3.2 Chronic Venous Insufficiency

Octa is an emerging approach for diagnosing Chronic Venous Insufficiency (CVI). It offers a painless alternative to traditional methods like venography that are more invasive and can be uncomfortable for patients.

CVI is a common medical condition that involves the veins of the lower limbs and results from inadequate functioning of the veins' valves. This condition worsens the venous pressure and venous reflux that can show in various clinical symptoms. CVI is defined in to seven stages, starting with the stages that are not showing any symptoms to the stages where there are active ulcers. Undergoing various changes, intermediate stage causes symptomatology including: capillary bed dilation, elongation of capillary tubes, coiling and thickening of the basement membranes [9]. Such changes in the venous system cause major effects on the daily living of the patients, which makes it imperative to diagnose and treat the disease with precision.

OCTA's high resolution can potentially detect early signs of venous insufficiency before clinical symptoms become apparent, allowing for earlier intervention.



(a) Healthy subject

(b) CVI patient

Figure 2.1: Comparison of en-face OCTA images between a healthy subject (left) and a patient with CVI (right), highlighting vascular density differences and pathology.

Chapter 3

Generative Adversarial Networks

Generative Adversarial Networks (GANs) represent a significant breakthrough in the field of artificial intelligence, particularly within the realms of deep learning and unsupervised learning. Introduced by Ian Goodfellow and his colleagues in 2014 [10], GANs have revolutionized the way models generate data, enabling the creation of highly realistic images, audio, and text. This chapter delves into the theoretical foundations, architecture, training dynamics, and various applications of GANs in OCTA imaging.

3.1 The concept of adversarial training

Generative Adversarial Networks (GANs) are a class of machine learning frameworks designed for generative modelling. It aims to extend the traditional task of generative models into learning, not only the conditional distribution between data and labels, but the full and more complicated probability distribution of the data p(x). It departs from the conventional solutions to learn a distribution proposed by generative models, such as maximizing the likelihood, often too simple to approximate the real distribution or too sophisticated and impossible to optimize exactly.

The setup introduced by GANs consists of two neural networks, the generator and the discriminator, that contest with each other in a game-theoretic scenario:

1. Generator(G) - The generator takes random noise z as input and generates data G(z) that resemble the training data. The goal of the generator is to produce outputs that are indistinguishable from real data to the discriminator.

2. Discriminator(D) - The discriminator takes both real data and generated data from the generator as input and outputs a probability value representing the likelihood that the input data is real. The discriminator is trained to maximize the accuracy of distinguishing real data from generated data.

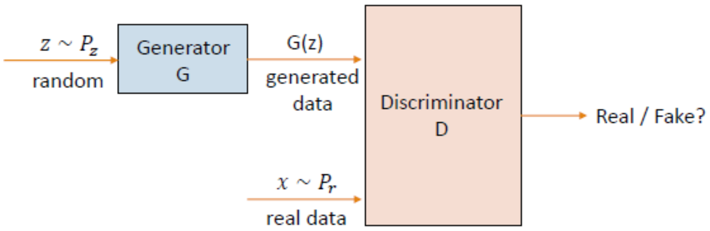


Figure 3.1: Schematic illustration of the basic functioning of a GAN, showing the interplay between a generator (creating synthetic data) and a discriminator (evaluating authenticity)"

The GAN learns the distribution by being able to sample from it. As shown in fig.3.1, it starts by sampling from a simple prior z, usually vector with components drawn Independent and identically distributed random variables from a standard Gaussian. It is then put through the generator neural network G to generate an output G(z) similar to the real data. Real and generated data is both given as input to the discriminator neural network D to decide if the current input is real or fake. In this way the GAN is turning the unsupervised problem of estimating the data distribution into a supervised one of classifying the data.

3.2 MinMax Game and Objective Function

The generator and the discriminator, as stated, have competing objective: the generator tries to maximize the probability that the discriminator makes a mistake, while the discriminator tries to minimize this same probability. This implies that GANs are trained using a two-player minmax game, i.e. a min max optimization problem. Mathematically, the training process can be formulated as follows [10]:

$$\min_G \max_D E_{x\tilde{P}_r}[log(D(x))] + E_{z\tilde{P}_z}[log(1 - D(G(x)))] \tag{3.1}$$

The way to solve this problem is alternating between optimizing the generator for a fixed discriminator and optimizing the discriminator for a fixed generator:

1. Fix the generator(G) and update the discriminator (D) - The discriminator is updated to maximize the probability of correctly classifying real and generated

data. This can be achieved by optimizing:

$$\max_{D} E_{x\tilde{P_r}}[log(D(x))] + E_{z\tilde{P_z}}[log(1 - D(G(x)))] \tag{3.2}$$

For a real input x, the first term is maximized when it converges to 1 and the second term is maximized to 0. This expression is basically a binary cross-entropy.

2. Fix the discriminator(D) and update the generator (G) - The generator is updated to minimize the log probability of the discriminator correctly classifying the generated data as fake. This is done by optimizing:

$$\min_{G} E_{z\tilde{P_z}}[log(1 - D(G(x)))] \tag{3.3}$$

The expression is maximized when D(G(z)) is close to 1, so to force the discriminator to switch the classification label for the fake sample.

This process can be quite unstable since there's no guarantee of convergence between the two operations. To try stabilizing the training problem, the generator's objective can be modified to:

$$\min_{G} E_{z\tilde{P_z}}[-log(D(G(x)))] \tag{3.4}$$

This alternative formulation, known as the non-saturating loss [11], often provides more stable gradients for training.

3.2.1 Jensen-Shannon Divergence

The minimax GAN loss is also called « adversarial loss » because of the competing objective of the two nets. GANs are implicitly minimizing a divergence between the generated distribution and the real data distribution by only using expectations over samples (instead of defining a likelihood). Given the optimal discriminator, GANs are optimizing this loss [10]:

$$\min_{G} 2JSD(P_r||P_g) - 2log2 \tag{3.5}$$

Being JSD the Jensen Shannon divergence

$$JSD(P_r||P_g) = \frac{1}{2}KL((P_r||\frac{(P_r + P_g}{2}) + \frac{1}{2}KL((P_g||\frac{(P_r + P_g}{2}) \tag{3.6}$$

Where KL denotes the Kullback-Leibler divergence

$$KL(P_r||P_g) = \int P_r(x)log\frac{P_r(x)}{P_g(x)}dx \tag{3.7}$$

JSD has two really important properties that makes it suitable for comparing distribution in GANs symmetry and boundedness. JSD is symmetric, meaning that JSD(P||Q)= JSD(Q||P) and it always results in a finite value between 0 and 1, making it more practical for training GANs. The use of Jensen-Shannon divergence has several implications for GAN training:

-Training Dynamics Minimizing JSD encourages the generator to produce data that is as indistinguishable as possible from the real data, leading to high-quality generated samples.

-Stability JSD helps in providing a more stable training process compared to other divergence measures.

3.2.2 Challenges in Training

Training a GAN poses a number of issues that must be taken into account:

- Mode Collapse- where the generator produces limited varieties of outputs. Since JSD can be small when the distributions do not overlap, the generator might still fail to capture the full diversity of the real data distribution.

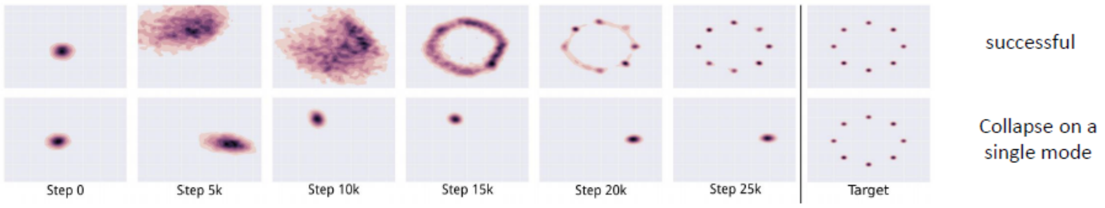


Figure 3.2: Visual representation of mode collapse, where the generator fails to capture diverse output modes, resulting in repetitive or limited variation in generated samples [12]

- Non-convergence: the optimization process involving JSD can be unstable. Small changes in the generator can lead to large changes in the discriminator, causing oscillations and divergence in training. Some mitigation strategies might involve the use of optimization techniques such as Adam, that will be later presented, and proper initialization of the network weights.

- Vanishing Gradients, when gradients become too small for effective learning causing a slow or a stalled training progress.

3.3 Training algorithm

A basic training algorithm is structured as shown in table 3.1 :

```
GAN Training Algorithm
Initialize G, $\theta_{\mathrm{G}}$ (Generator Parameters)
Initialize D, $\theta_{\mathrm{D}}$ (Generator Parameters)
N = number of epochs
m = batch size
$\alpha$ = learning rate
**for** i=1 **to** N **do**
| Sample Z $p(\tilde{Z})$
| Sample X $\tilde{P_R}$
| $\theta_{\mathrm{D}} = \theta_{\mathrm{D}}$ - $\alpha \cdot \nabla\ \theta_{\mathrm{D}} \cdot \frac{1}{m} \sum_{j=1}^{m}$log(D(X)) + log(1 - D(G(Z)))
| $\theta_G = \theta_G$ - $\alpha \cdot \nabla\ \theta_G \cdot \frac{1}{m} \sum_{j=1}^{m}$ log(1 - D(G(Z)))
**end**
```

Table 3.1: Basic Algorithm for training a GAN network

Both G and D are initialized with the parameters θ. These parameters include weights and biases that define the strength and direction of the connection between neurons in the networks. The weights are often initialized from a normal distribution since it helps to prevent gradients from vanishing too quickly during training. Biases meanwhile can be initialized to a small constant value or sometimes they are left uninitialized (set to zero).

N, m and α are the hyperparameters of the training:

- N is the number of epochs representing how many times the entire training dataset will be passed through the network. It determines how long the model will train. More epochs allow the model to learn more but can also lead to overfitting if set too high.

- M is the batch size representing the number of training samples used in one forward and backward pass. It affects the stability and speed of the training process. Smaller batch sizes provide more frequent updates but can be noisy, while larger batch sizes provide smoother updates but require more computational resources.

- α is the learning rate which determines the step size for updating the model parameters during training. It determines how quickly or slowly the model parameters are updated and it is associated to the optimizer chosen.

The core of the GAN training process is a loop that runs for the specified number of epochs N. Each repetition consists several steps:

1. Forward pass which computes the output using the current weights (p(z) and $\mathrm{p_R}$)
2. Loss calculation which measures how far the current output is from the desired one using a loss function
3. Backpropagation which computes the gradients of the loss (∇ θ)
4. Parameter update which adjusts the weights and biases using an optimizer to minimize the loss

The loop continues for N epochs, repeatedly updating the parameters of the generator and discriminator to improve their performance.

3.3.1 Adam Optimizer

Optimizers are essential constituents in the training process of GANs since they dictate how the generator and discriminator's parameters are updated to minimize their respective loss functions. One of the most commonly used optimizers in GAN training is Adam (Adaptive Moment Estimation), which adapts the learning rate for each parameter based on its gradient history, allowing efficient learning across the entire network.

In standard gradient descent, the learning rate (α) is fixed globally for all parameters. The update equation is:

$$\theta_{t+1} = \theta_t - \alpha \cdot g_t \tag{3.8}$$

where:

- θ is the models parameters;
- $\mathrm{g_t}$ is the gradient of the cost function with respect to the parameters;
- α is the learning rate
- t is the iteration step.

Adam adapts the learning rate for each parameter individually. It combines advantages from other optimization methods like AdaGrad and RMSProp. The key idea is to compute adaptive learning rates based on first and second moments of gradients. The update equation for Adam is:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{v_t} + \epsilon} \cdot m_t \tag{3.9}$$

where:

- $\mathrm{m_t}$ is the first moment (mean) of gradients;
- $\mathrm{v_t}$ is the second moment (uncentered variance) of gradients;
- ϵ is a small constant to prevent division by zero.

Adam adapts learning rates dynamically, leading to faster convergence compared to fixed-rate gradient descent. It is a robust method that performs well across different problems and architectures. It is computationally efficient and requires minimal memory.

3.4 GANs variants

Many variants of GANs have been developed to address specific challenges in image generation, translation, and enhancement. These variants take the basic framework of GAN and adapt them to specialized applications like image super-resolution and image-to-image translation. The methodology of two of the most well-known GAN variants, super-resolution GAN and pixel2pixel GAN, will be taken into consideration.

3.4.1 Super-Resolution GAN

Super Resolution GANs (SRGANs) are designed to generate high-resolution images from low-resolution inputs [13]. The key challenge in super resolution is to generate visually plausible details that are missing in low-resolution images. SRGANs typically utilize deep neural networks with skip connections or residual blocks to efficiently learn the mapping from low-resolution to high-resolution images.

This kind of networks use perceptual loss functions that combine the typical adversarial loss with content loss (based on features extracted with pre-trained networks like VGG).

Beyond the upscaling of OCTA images [14], SRGANs are often used to enhance the visual quality of images. This makes this kind of architecture suitable for the aim of this thesis.

3.4.2 Pixel2Pixel GAN

The Pixel2Pixel GANs variant, Pix2Pix, is designed specifically for image-to-image translation [15]. While traditional GANs will generate an image given random noise, a Pix2Pix will generate an image given an input image. It basically means that it has learned the mapping from input to output images in a supervised manner. Many such variants are particularly very handy when there is a direct mapping of input to output images, like image colorization, style transfer, and semantic segmentation.

Pix2Pix uses a conditional GAN framework which incorporates conditional information into both the generator and discriminator networks such as class labels, attributes, or even other images.

It employs a combination of adversarial loss and task-specific loss functions to ensure that generated images are not only realistic but also accurate in terms of the desired output.

The image-to-image task can be adapted to fit the purposes of image enhancement.

Part II

Materials and Methods

Chapter 4

Dataset

This chapter deals with the dataset used for training and evaluating the generative adversarial networks.

4.1 Data Acquisition

4.1.1 MUW OCT system features

The OCT system employed in this study is a custom-built, fiber-based setup incorporating an akinetic swept-source laser from Insight Photonic Solutions, USA. This advanced system is designed for high-resolution imaging and precise data acquisition, making it ideal for the requirements of this research. Figure 4.1 provides a schematic illustration of the OCT system setup. Key components include:

Laser source: akinetic swept-source laser from Insight Photonic Solutions, USA. The system operates with a central wavelength of 1310 nm and a bandwidth of 29 nm. The power emitted by the laser is 70 mW, of which due to dispersion along the path less than 20mW actually reach the target. The duty cycle is 100% ensuring continuous sweeping across the bandwidth without interruption. For this reason, alternate sweeping cycles are recorded, with the next cycle being skipped to allow time for laser alignment.

Fiber Coupler: Splits the source beam into the reference arm (25% of power) and the sample arm (75% of power).

Sample Arm: Equipped with a rotatable imaging probe and a lens system to adapt the bidimensional plane to the skin surface.

Recombination and Detection block: Beams from both arms are recombined using a 50/50 fiber coupler. A dual-balance detector then records the cross-correlation term, essential for generating high-resolution images.

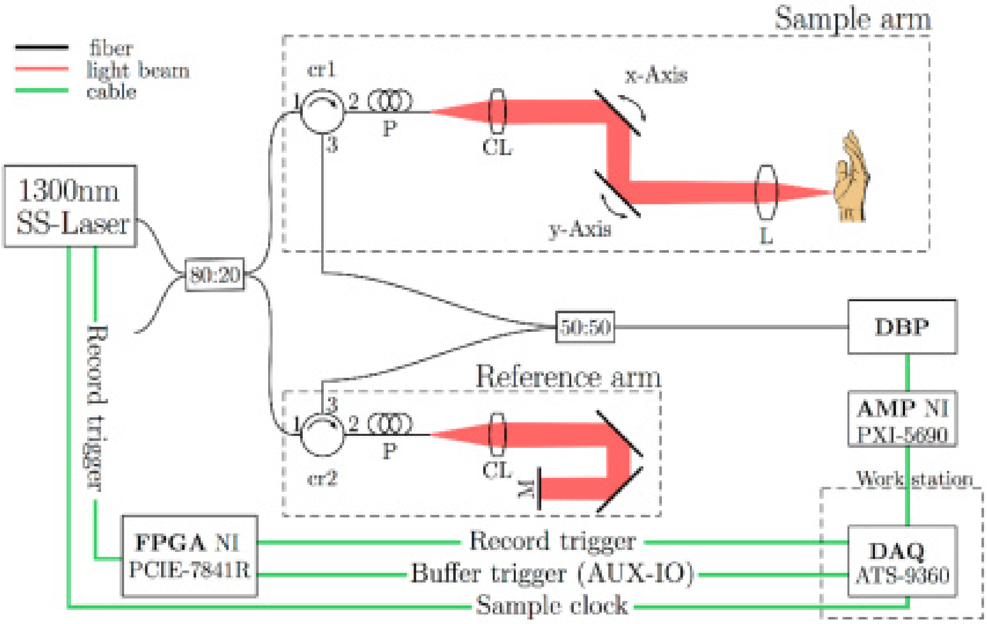


Figure 4.1: Schematic rappresentation of the OCT system located at the MUW [adapted from [5]]

4.1.2 Standard Acquisition

The volumes acquired from this system are composed by 2048 points for each A-line, of which only around 1500 valid, 512 A-lines for each B-scan and 512 Bscans for each C-scan. In terms of image resolution this is translated in axial and lateral resolution respectively of 9,6μm and 19,5μm. The FOV obtained is 10 mm x 10 mm.

The protocol of scanning chosen for OCTA is the BM-scan, i.e. acquiring the same B-line N-times before moving to the next lateral position (in opposition of MB-scan protocol where each A-scan is acquired multiple times [16]). N has a direct impact on the quality of the images: having a big number of repetitions generates high contrast images but also increases the time of the acquisition session and the probability of motion artifacts. For these reasons, the comprise has been found for N=4.

The system achieves a sweep rate of 222.22 kHz. In ideal conditions, the frame rate is determined by dividing the sweep rate by the number of A-lines acquired across the 4 volumes so $(222.2 \cdot 10^3)/(512 \cdot 512 \cdot 4) = 0.212$Hz. This results in an acquisition time of 4 volumes of 4.243 seconds. Adding the trasport and saving time related, the total time is around a minute.

4.1.3 OCTA reconstruction

At the end of the acquisition session, the system provides the raw spectral data from the 4 OCTs collected. The data are organized in 32 files containing 64 B-scans each so 4 repeated B-scan for 16 consecutive positions. The algorithm [17] that delivers the OCTA volume is implemented in two MATLAB codes structured as follows.

The first step is the **OCT reconstruction**.

Along the raw data collected, the laser calibration provides a file that contains the positions of the valid sweep points of the laser. This is used to eliminate all the invalid data points from each A-line.

Then a background subtraction is performed. This is done by subtracting the mean A-line for each lateral location from each A-line in that same lateral position. This operation is useful because it helps to lessen the effect of noise and consequently improve the signal-to-noise ratio.

Lastly a fast Fourier transform is performed with removal of the symmetric copy. The final output consists of OCT volumes of complex numbers.

The second step is the actual **OCTA reconstruction**. As previously stated, either intensity based or phase-based OCTA can be performed. For the purposes of this thesis only intensity-based algorithm is presented.

As only the intensity of the signal is used, from each complex datapoint the magnitude is extracted. A region of interest (ROI) containing the most information is empirically set and the surrounding points are deleted.

The OCT volumes are now separated to extract the N tomograms at different timing. Two thresholds are chosen from the histograms to eliminate the noisy pixels with intensity greater than one or lower than the other.

Afterwards, a fundamental operation is set the zero level of each volume at the skin surface points. To do so a median filter is applied and a threshold is set for each A-scan being the 85% of its quantile. The first pixel for each A-scan to go beyond this value is identified as a surface point. All these points are stored and interpolated through a linear polynomial curve. The process is repeated for each B-scan to recreate the whole tomogram surface.

In the end equation 2.1 is applied and the OCTA is available to be displayed.

4.2 Data Preparation for GAN

4.2.1 Input and Target

The aim of this thesis is to shorten the time of the acquisition session. Even if the sweep itself is really fast, the time-consuming part is the process of saving.

Furthermore, OCT is extremely sensitive to the smallest patient movements, which can lead to artifacts. The appearance of white vertical lines may also result from laser power instability. Therefore, reducing exposure times also reduces the occurrence of both motion artifacts [18] and laser-induced artifacts. However diminishing the number of starting volumes results in a loss of image quality. In this scenario, GANs could offer a possible solution to enhance this loss.

To do so, the input data for the GAN consists of OCTA images generated from 2 OCT volumes (N=2). For convenience, we will refer to this dataset as low quality (LQ). In this way the time of each acquisition should be halved.

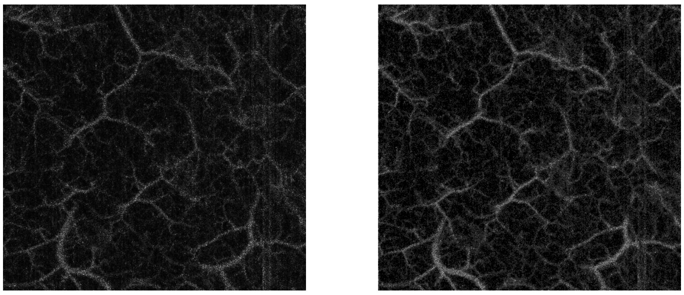
The target data for the GAN consists of OCTA images generated following the standard protocol (N=4). For convenience, we will refer to this dataset as high quality (HQ).

The LQ OCTA is reconstructed using the first and third OCT volumes acquired for the HQ version and follows the same algorithm described before (figure 4.2).

The data provided at the end of the reconstruction algorithm, in both cases, is composed of OCTA volumes with dimension 100x484x512 (x,y,z). The networks are trained using 2D images. The plane chosen for this application is the en-face (y,z) given that it is the most straightforward and easy way to read an angiography.

From the 100 en-face images, the first 10 single slices are usually not reaching the relevant skin layers so lacking any useful information. The last 10 single slices, on the other hand, are affected by excessive noise to be relevant. For this reason, these two groups of single slices are discarded leaving 80 single slices.

Moreover, the images fed to the networks are not single slices from the OCTA volume. To increase the clearness of the angiography, it was chosen to use MIPs of 5 slices. In this way for each volume, instead of 80 images, there are 16 available ones. The final images are therefore saved in png format as grayscale images with integer values between 0 and 255.



(a) LQ OCTA **(b)** HQ OCTA

Figure 4.2: Comparison of en-face OCTA images from two OCT volumes (LQ) and four OCT volumes (HQ), demonstrating enhanced visualization.

4.2.2 Dataset Division

The data employed in this work belong to class of patients: healthy and chronic venous insufficiency patients at different stages. Volunteers range in age from 22 to 90 years and are equally split into male and female.

There's an over-all of 52 subjects, 16 healthy and 36 CVI, from which 224 volumes were extracted for a total of 3584 images (16 MIP for each volume).

The subjects are first divided in training and test set. In particular 5 patients, 3CVI and 2 healthy, randomly chosen were included in the test set. In more detail, the test set comprises 16 volumes from the CVI patients and 7 volumes from the healthy ones (23 volumes or 368 images, around 10% of the total). This set is kept separate from the training set to provide an unbiased assessment of the model's capabilities.

The training set is further divided in training and validation set. In this case the division was not made by patients, but it was chosen to randomly divide the volumes without taking into account the source subject.

The training set is made of the 70% of the total data set: 161 volumes in 2576 images. This set includes a wide range of images to ensure that the network learns to generalize in different conditions and scenarios. Diversity of training set is crucial to prevent over-sizing and ensure model robustness.

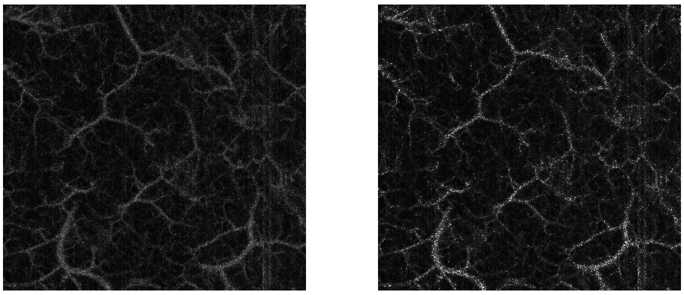
The validation set, consisting of 20% of the dataset (40 volumes in 640 images), is used to fine-tune the model and prevent overfitting. This set is used to validate the model's performance during training and to adjust hyperparameters. It serves as a checkpoint to ensure that the model generalizes well to unseen data.

4.2.3 Pre-processing

To be fed to the networks, the dataset requires a series of actions to be carried out:

- To standardize the images, the first ad the 99th percentile are calculated. Saturation is applied: the pixel with intensity lower than the first are set to 0 and the ones with intensity higher than the 99th percentile are set to 255 (figure 4.3).
- The images are converted to RGB format with 3 channels and converted to tensor.
- To facilitate all the operations inside the layers of the network, multiples of 2 are required for the dimensions of the images. Resizing is thus applied to obtain images with resolution 256x256 or 512x512.

- Lastly batch normalization is applied. It is helpful to stabilize the training. The value of each pixel is now a float number between 0 and 1.



(a) Raw input **(b)** Saturated input

Figure 4.3: Illustration of the pre-processing operation applied to the input

Chapter 5

GAN Training

This chapter discusses the two neural network architectures employed for picture enhancement: a Super-Resolution Generative Adversarial Network (SRGAN) and a Pixel-to-Pixel (Pix2Pix) Network. It details their respective architectures, hyperparameters tuning and loss functions used in training.

5.1 Enhanced Super resolution (ESRGAN)

The model here presented is an enhanced super resolution GAN (ESRGAN) [19] [20], a further development of the original SRGAN architecture, in which several important enhancements were introduced to improve performances in visual quality and accuracy.

5.1.1 Network Architecture

Feature Extractor

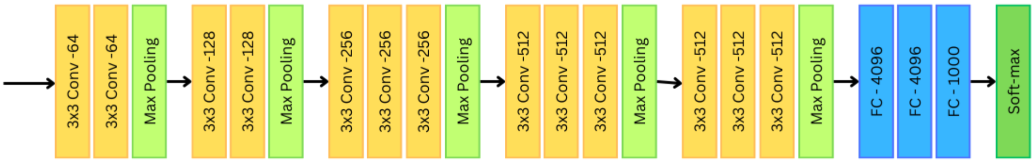


Figure 5.1: VGG19 feature extractor architecture adapted from [21]

The feature extractor plays a crucial role to ensure that the generated images are perceptually similar to the ground truth, focusing in high-level features that are more aligned with the human visual perception than mere pixel-wise similarity. The extractor employed is a pre-trained VGG19 model. VGG19 has 19 layers: 16 of these are convolutional, and 3 are fully connected. In the convolutional layers, small filters 3×3 are used with a stride of 1 pixel, and the max-pooling layers are used to reduce spatial dimensions.

Generator

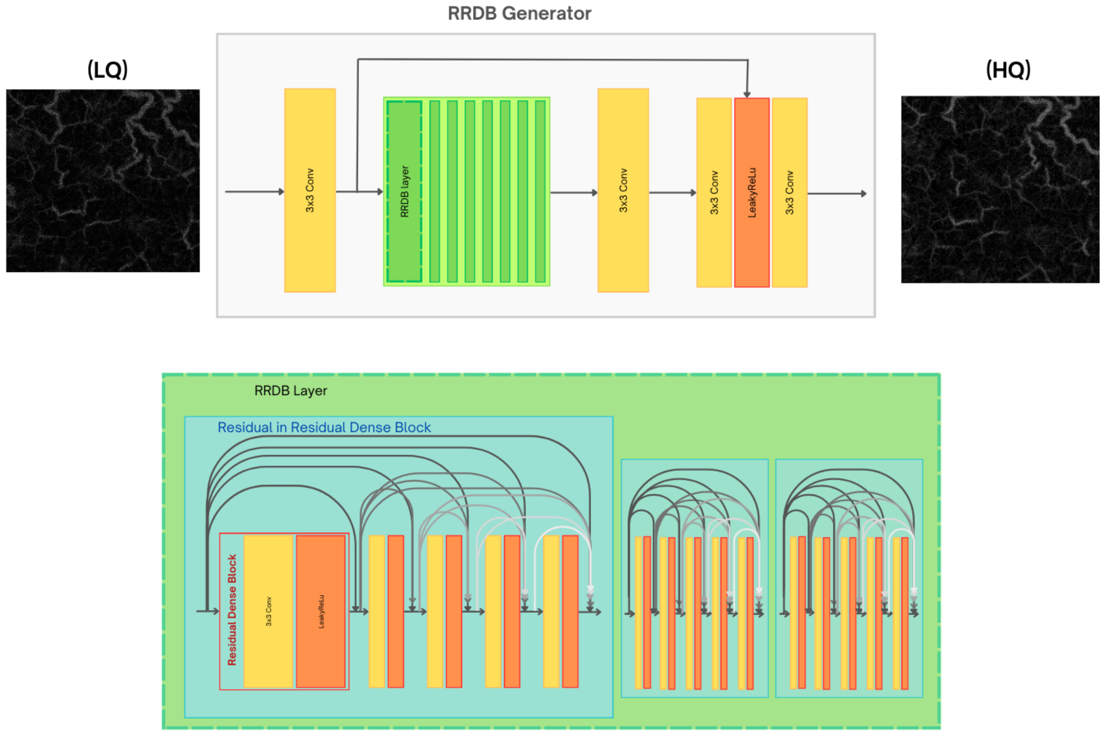


Figure 5.2: RRDB Generator architecture

The Generator is the core component of ESRGAN. It incorporates Residual-in-Residual Dense Blocks (RRDB), which combine dense connections and residual learning to enhance the network's capacity and stability.

DenseResidualBlock is responsible for feature extraction through a series of convolutions. It consists of 5 convolutional layers progressively increasing the number of filters, each followed by a LeakyReLU activation function. The inputs to these layers are concatenated, allowing for the reuse of features and enhancing the representational capacity of the block.

ResidualInResidualDenseBlock integrates 3 DenseResidualBlocks, adding another level of residual learning. The dense connections within each block facilitate efficient gradient flow.

Finally, the GeneratorRRDB is composed as follows:

- Initial Convolution Layer: The generator begins with a convolution layer that increases the number of feature maps, setting the stage for subsequent processing;
- RRDB Layers: Following the initial layer, the network includes a series of ResidualInResidualDenseBlocks. These blocks allow the generator to learn fine details and textures necessary for high-quality image reconstruction;
- Intermediate Convolution Layer: After the RRDB layers, another convolution layer is applied to refine the feature maps before the final reconstruction;
- o Output Block: The generator concludes with a final block that generates the high-resolution image. This block includes a convolution layer followed by an activation function, producing the final output image with enhanced resolution.

All the convolutional layers employ 3×3 filters, with stride and padding settled at 1. The numbers of filters and the number of RRDB are defined at the beginning of each training.

The network structure notably lacks upsampling layers, ensuring that the output maintains the same size as the input to enhance image quality without altering the resolution.

Discriminator

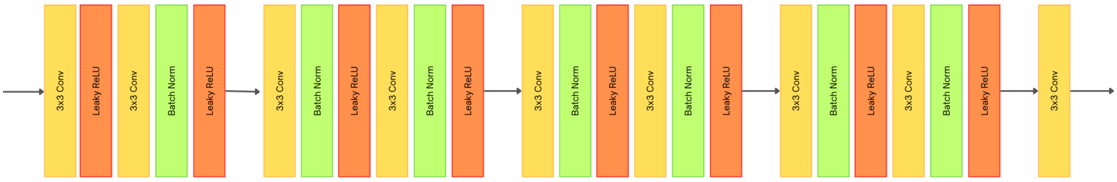


Figure 5.3: Discriminator ESRGAN

The discriminator is composed of a series of convolutional layers with increasing filter sizes. These layers are interleaved with batch normalization and LeakyReLU activations to ensure stable training and effective feature extraction.

It employs the PatchGAN strategy, which classifies image patches rather than the entire image. This approach encourages the generator to produce realistic high-frequency details, as the discriminator evaluates the authenticity of smaller

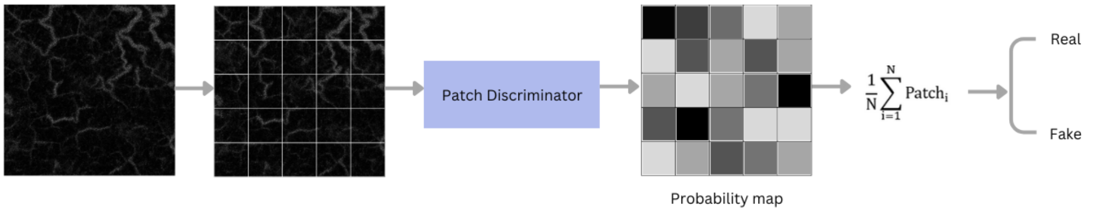


Figure 5.4: Visual explanation of the functioning of Patch Discriminator

regions within the image. In this case, the patches are sized based on the dimensions of the input images, specifically dividing both the height and width by 16.

For both generator and discriminator, the weights of the networks are initialized using the default initialization which relies on Kaiming (or He) initialization. The Kaiming initialization method relies on a Gaussian probability distribution (G) with a mean of 0.0 and a standard deviation of sqrt(2/n), where n is the number of inputs to the node.

5.1.2 Parameters Tuning

The tuning process involved adjusting several key hyperparameters to achieve optimal results in terms of both quantitative metrics and visual quality of generated images.

The following parameters were selected based on empirical experimentation and validation results:

- Number of Epochs: A total of 50 epochs were chosen to train the model, allowing sufficient time for the network to learn complex image features and converge towards optimal performance
- Batch Size: Each training batch consisted of 1 image, balancing computational efficiency and the granularity of gradient updates during backpropagation
- Warm-up Batches: The model was trained with pixel-wise loss only for the initial 200 batches, facilitating stable training before introducing adversarial and content losses
- Number of Filters: The generator architecture was configured with 64 filters, which determined the depth and complexity of feature extraction and enhancement processes
- Number of RRDB: The generator incorporated 8 residual blocks, designed to capture and refine image details across multiple scales while minimizing vanishing gradient issues

- Optimizer Parameters:
  - Learning Rate: A value of 0.0001 was set for both the generator and discriminator Adam optimizers, controlling the step size during parameter updates to balance training speed and stability
  - Adam Optimizer Beta Parameters: Values of 0.9 and 0.999 were chosen respectively for the decay rates of the first and second moments of the gradient, optimizing the convergence and robustness of the training process
  - Decay epoch: To prevent training divergence, the optimizer was configured to initiate learning rate decay beginning from the second epoch

This setting was chosen for images with 512 as input size and the limitations imposed by the hardware. In case of training with image size of 256, the configuration allowed to increase the number of RRDB blocks to 16.

5.1.3 Loss Functions

Three primary loss functions were selected to guide the training process:

1. Adversarial Loss – it is formulated using the Binary Cross-Entropy (BCE) loss with logits. This loss function combines the sigmoid activation and the BCE loss in a numerically stable manner, making it well-suited for training generative adversarial networks. It can be described as:

   $$L_{BCEwithLogits} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot log(\sigma(p_i)) + (1 - y_i) \cdot log(1 - \sigma(p_i))] \tag{5.1}$$

   where:

   - N is the batch size
   - $\mathrm{y_i}$ is the true label for the i-th sample
   - σ is the sigmoid
   - $\mathrm{p_i}$ is the predicted probability that the i-th sample belongs to the positive class

   Practically, it accepts two tensors, one being the raw generator outputs (prediction), the other being the true class labels (target), then wraps the first using a sigmoid. Then calculates Cross Entropy loss for each pair and reduces it to mean.

2. Content Loss – it is computed using the mean absolute error (L1) between the generated and ground truth image features extracted from the feature extractor network. It can be expressed as:

$$L1_{MAE} = \sum_{i=1}^{n} |y_{true} - y_{predicted}| \tag{5.2}$$

3. Pixel-wise Loss – it is measured again by the mean absolute error difference but this time between corresponding pixels of the generated and ground truth images. It emphasizes the similarity of individual pixel values.

These functions are integrated to compute the overall **generator loss**:

$$L_{generator} = L_{content} + \lambda_{adv} \cdot L_{adversial} + \lambda_{pix} \cdot L_{pixel-wise} \tag{5.3}$$

λ_{adv} and λ_{pix} are two constants chosen to balance the relative importance of each loss component during training:

- λ_{adv} is the weight assigned to the adversarial loss and set to 5x10^{-3}
- λ_{pix} is the weight assigned to the pixel-wise loss and set to 0.1

The **discriminator overall loss** is calculated as the mean of the adversarial losses calculated between the predictions for real and fake data and the respective labels:

$$L_{discriminator} = \frac{1}{2}(L_{BCEwithLogits}(D(y_R), valid) + L_{BCEwithLogits}(D(y_g), fake)) \tag{5.4}$$

where:

- D is the discriminator network
- $\mathrm{y_R}$ is the real sample
- $\mathrm{y_g}$ is the generated sample
- valid and fake are labels indicating real (1) and fake (0) sample, respectively.

5.2 Pixel2Pixel GAN

The chosen network architecture for enhancing image quality using a Pixel2Pixel approach involves an Attention U-Net combined with a Discriminator for adversarial training. This architecture is particularly effective for tasks such as image super-resolution or denoising, where high-quality outputs are desired. The U-net makes it also a suitable approach to OCTA reconstruction [22].

5.2.1 Network Architecture

Generator

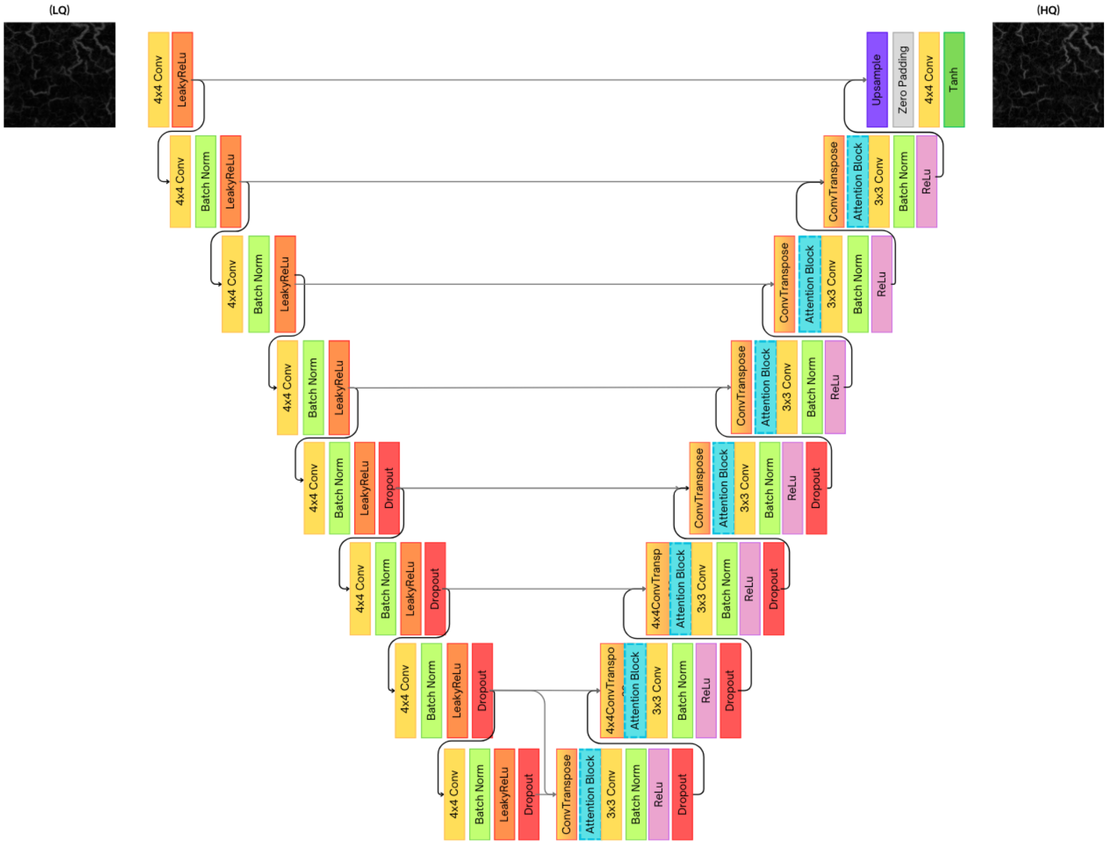


Figure 5.5: Representation of the AttentionU-Net in the Pix2Pix architecture

The attention U-Net architecture consists of an encoder-decoder structure with skip connections and attention mechanisms to selectively highlight relevant features.

The **Encoder** (Downsampling path) is composed of UNetDown blocks. Each block performs 4x4 filters convolution with stride 2 and padding 1, followed by batch normalization (except the first layer) and Leaky ReLU activation. When specified, dropout is applied. In these blocks the number of channels progressively increases while downsampling the spatial dimension.

Attention mechanisms are implemented through the attention block to enhance feature selection based on interdependencies between feature maps [23]. It combines convolutional layers (1x1 kernel, stride 1) and a sigmoid activation.

The **Decoder** (Upsampling path) concatenates UNetUp blocks. Each block consists of a transposed convolution layer (kernel size 4, stride 2, padding 1) to upsample the feature map, attention block followed by a convolutional layer (kernel size 3, stride 1, padding 1) with batch normalization and ReLU activation.

The encoder and decoder are assembled in the UNet with skip connections from corresponding blocks as required by the U.Net architecture. A final layer concludes the network consisting of a convolutional layer with Tanh activation to generate the final enhanced image output.

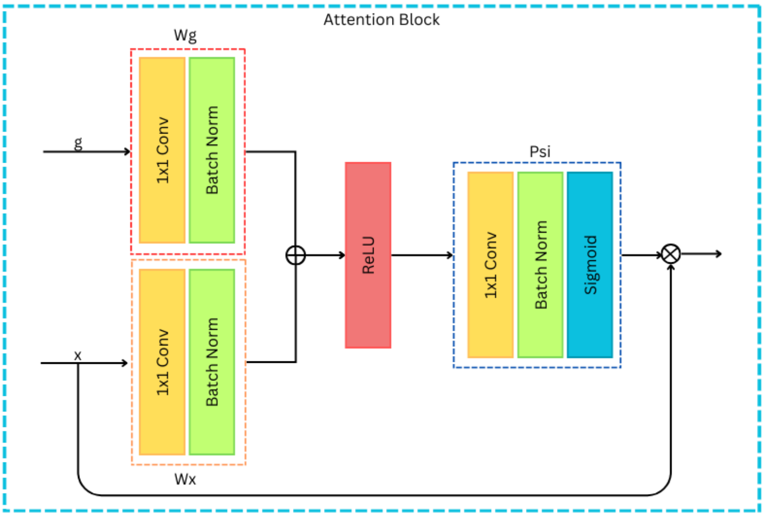


Figure 5.6: Attention Block: g is the output from the previous layer in the upsampling path and represents the global features; x is the feature map from the corresponding layer in the contracting path of the U-Net, connected via a skip connection and represents the local features.

Discriminator

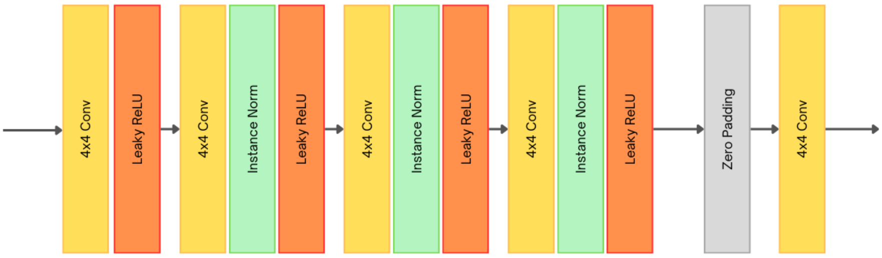


Figure 5.7: Discriminator Pix2Pix

The discriminator concatenates pairs of real and generated images and processes them through multiple discriminator blocks, progressively downsampling the spatial dimension.

Each discriminator block presents convolutional layers (kernel size 4, stride 2, padding 1) with Leaky ReLU activation and optionally instance normalization. A final convolutional layer with sigmoid activation produces the single output that indicates the probability of the input being real or generated.

Similarly to ESRGAN, a patchGAN strategy was implemented. This involved concatenating pairs of patches extracted from the entire images, each patch having dimensions 1/16 of the original size.

In both generator and discriminator, the weights of the networks are initialized to follow a normal distribution. Specifically, the weights of the convolutional layers use a normal distribution with mean of 0 and a standard deviation of 0.02, meanwhile the weights of the batch normalization layers follow a normal distribution with mean 1 and standard deviation of 0.02. The biases are initialized to 0.

5.2.2 Parameters Tuning

The following parameters were selected based on empirical experimentation and validation results:

- Number of Epochs: 50
- Batch Size: 1
- Number of Filters: The generator architecture was configured with 64 filters as input
- Number of Downsamplings: The architecture was tailored to accommodate the input image dimensions. Specifically, when the dimensions are set to 512, 8 Down blocks are incorporated.
- Optimizer Parameters:
  - Learning Rate: 0.0001
  - Adam Optimizer Beta Parameters: Values of 0.9 and 0.999 were chosen respectively for the decay rates of the first and second moments of the gradient
  - Decay epoch: 30. Given that the model was more stable in this case, the decay epoch was introduced at a later stage.

5.2.3 Loss Functions

For the pixel2pixel network two main losses are selected:

1. Adversarial Loss – It is defined as the mean square error between the generated sample and the true class labels. MSE can be expressed as:

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_{true} - y_{predicted})^2 \tag{5.5}$$

2. Pixel-wise Loss- it is measured as the mean absolute error (L1) between corresponding pixel values in the generated and target images. The formulation is the previous shown in equation 5.2.

The total generator loss combines the adversarial and pixel-wise loss scaled by λ_{pixel}:

$$L_{generator} = L_{adv} + \lambda_{pix} \cdot L_{pixel-wise} \tag{5.6}$$

λ_{pixel} is set to 10.

The discriminator loss aggregates the real and fake losses computed during its training same as in equation 5.4.

5.3 Software and Hardware

In this thesis, the training of Generative Adversarial Networks (GANs) was conducted using a combination of advanced software tools and high-performance hardware resources.

The primary software framework utilized was PyTorch, a widely-used deep learning library known for its ease of use in defining and training custom made neural networks. To support the GAN training process, additional libraries such as torchvision for data handling, and numpy for numerical operations were employed.

The computational demands of training GANs necessitated the use of a high-performance scientific computing cluster managed by the Slurm workload manager. The cluster was made available by Center for Medical Physics and Biomedical Engineering (Medical University of Vienna). This cluster leverages NVIDIA GPUs, specifically multiple NVIDIA A100-PCIE-40GB GPUs, each with 40 GB GPU-RAM. For each training the power of 1 GPU was needed. These GPUs provided the necessary computational power and memory bandwidth to handle large-scale data processing and model training tasks efficiently.

Furthermore, the training environment was encapsulated within a container from the NVIDIA NGC catalog, the pytorch ngc container, which provided a pre-configured and optimized environment for deep learning tasks. This container included all necessary dependencies and libraries, ensuring consistency and reproducibility across different runs and experiments.

Chapter 6

Metrics Evaluation

Defining and evaluating the goodness of fit of GANs is one of the most challenging aspects. It's hard to determine when an output is satisfactory even for comparison. For generative networks, relying on the loss function during training is not effective. The loss function doesn't tell us whether the GAN is producing high-quality outputs; it only indicates that the network has stopped learning. In other words, a low loss function value may simply mean that the training process has plateaued, not that the generated images are actually good or realistic. This makes it difficult to assess the performance and success of the GAN based solely on training loss.

This chapter outlines all the evaluation processes used in this work, including the validation and testing phases.

6.1 Metrics

To evaluate our GAN-based network, we used three key metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Multiscale Structural Similarity Index Measure (MS-SSIM). These metrics provide a comprehensive picture of image quality and structural accuracy.

Peak Signal-to-Noise Ratio (PSNR) measures the quality of the reconstructed image by comparing it to the original image. It quantifies how much noise is in the reconstructed image. Mathematically, PSNR is calculated using the mean squared error (MSE) between the original and reconstructed images and can be expressed as:

$$PSNR = 10log_{10}(\frac{MAX_I^2}{MSE}) \tag{6.1}$$

where $\mathrm{MAX_I}$ is the maximum possible pixel value of the original image. MSE is given by :

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1} N[I(i,j) - K(i,j)]^2 \tag{6.2}$$

Here, I is the reference image, K is the generated image, M and N are the dimensions of the image.

Higher PSNR values indicate better quality, as they signify less noise. Essentially, PSNR tells us how close the reconstructed image is to the original in terms of pixel accuracy.

Structural Similarity Index Measure (SSIM) assesses the quality of the image by considering changes in structural information, luminance, and contrast.[24] Its formulation is :

$$SSIM(I,K) = \frac{(2\mu_I\mu_K + C_1)(2\sigma_{IK} + C_2)}{(\mu_I^2 + \mu_K^2 + C_1)(\sigma_I^2 + \sigma_K^2 + C_2)} \tag{6.3}$$

where:

- μ_I and μ_K are the mean intensity
- ${\sigma_\mathrm{I}}^2$,${\sigma_\mathrm{K}}^2$ are the variances
- σ_IK is the covariance of the images I and K
- $\mathrm{C_1}$ and $\mathrm{C_2}$ are constants to stabilize the division.

Unlike PSNR, which focuses on pixel differences, SSIM evaluates how well the reconstructed image preserves the structural integrity of the original image. This metric is more aligned with how humans perceive image quality. SSIM values range from -1 to 1, with 1 indicating perfect similarity and thus better quality.

Multiscale Structural Similarity Index Measure (MS-SSIM) extends SSIM by evaluating the image at multiple scales, providing a more detailed assessment of quality. It captures image details at various levels of resolution, ensuring that both fine and coarse structures are accurately reconstructed [24]. It is defined as:

$$MSSSIM(I,K) = [SSIM_M(I,K)]^{\alpha M}\prod_{j=1}^{M-1}[SSIM_j(I,K)]^{\alpha j} \tag{6.4}$$

where:

- $\mathrm{SSIM_j(I,K)}$ represents the SSIM at the j-th scale
- α_j is the weight at the j-th scale

- M is the number of scales. Typically, M is chosen based on the desired level of detail and complexity in the multiscale analysis. Common choices for M range from 3 to 5. In this implementation 5 scales are used capturing all the details at various resolutions.

These scales represent hierarchical levels of image decomposition, where each scale captures different frequency components or spatial details of the image.

Higher MS-SSIM values indicate better overall image fidelity across different scales, which is important for capturing detailed vascular structures in OCTA images.

By using PSNR, SSIM, and MS-SSIM, thorough evaluation of both pixel-level accuracy and perceptual quality is ensured. This comprehensive approach helps confirm that a model not only produces accurate reconstructions but also preserves the essential details and structures in the images, which is crucial for practical applications in medical imaging.

6.2 Validation

Besides the loss functions, the metrics described are used to measure the performance of the model in determining the correctness of the predictions and tracking its training.

In the training phase, we store the model at every 5 epochs of training phase, so that after the complete training is done, we can evaluate the performance of the model at these various intervals. This checkpointing strategy allows us to perform a validation of the model periodically to identify the best epoch based on validation metrics.

This validation process is then divided in two steps:

1. First, the assessment is performed using the dataset containing enface angiograms of 5 slices. This is the input data of the training. The metrics are calculated for both training and validation sets. If the model shows good results in this set of data, further analysis is performed.

2. The second step involves investigating the performance on a recombined dataset which includes the Median Intensity Projection enface angiograms of the two main plexuses of the dermal layer. To do so, the images generated from the networks are again projected to recreate this more comprehensive images.

The best-fit epoch is then determined by finding the epoch with the highest Structural Similarity Index within the validation set at step 2. This best performing epoch is used again to compute the metrics on the test set to guarantee that the computed score is based on the best trained model.

Part III

Results and Discussion

Chapter 7

Results on enface angiograms of 5 slices

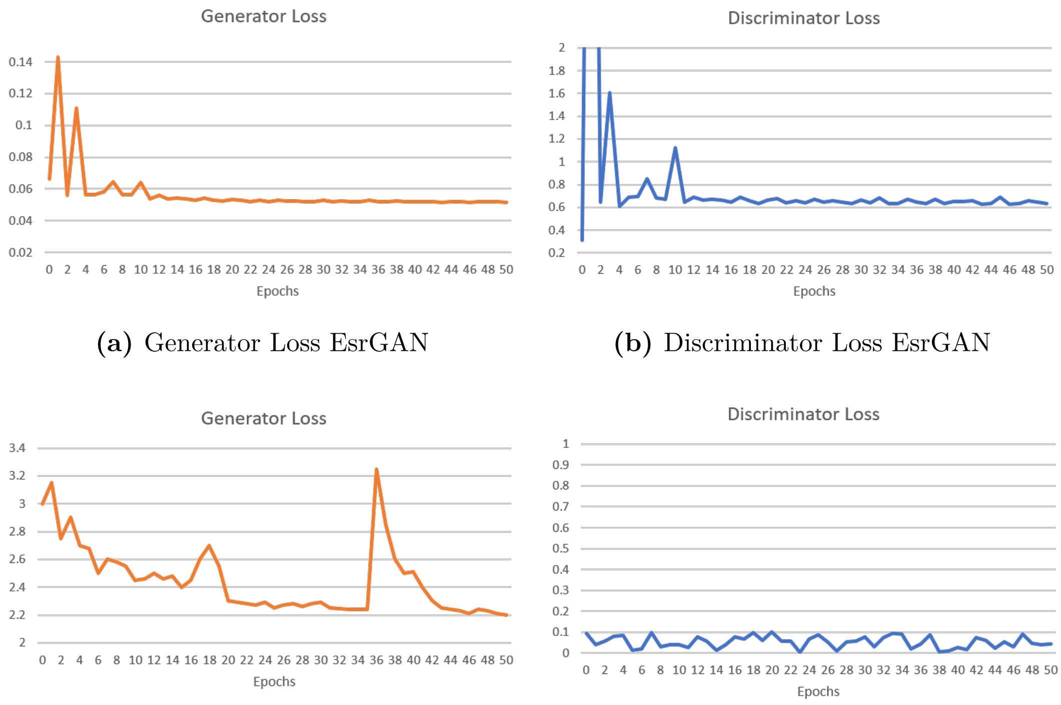
In this section, we provide a detailed presentation of the evaluation results for our GAN models, with a particular focus on the initial outputs, which consist of angiograms generated from 5 distinct slices.

7.1 Evolution of Loss Functions During Training

Before moving on to the evaluation phase of the metrics, it is helpful to review the behavior of the loss functions during training.

In the EsrGAN model (7.1 (a-b)), the generator's loss starts high, exhibiting some oscillations before gradually decreasing and reaching a plateau around epoch 12, stabilizing at approximately 0.05. Similarly, the discriminator's loss shows a very high peak at epoch 2, which then diminishes, achieving relative stability after epoch 12 at around 0.6. Both the generator and discriminator exhibit a stable and balanced training process after the initial epochs. The generator's loss plateauing around 0.05 and the discriminator stabilizing around 0.6 suggest that the GAN has reached a state where the generator can produce relatively realistic samples, and the discriminator maintains a good level of challenge without overpowering the generator. This indicates that the EsrGAN has achieved a reasonable equilibrium between the generator and discriminator. The discriminator's loss aligns closely with the rule of thumb that suggests it should remain around 0.5, confirming a balanced training process.

In contrast, the Pix2Pix GAN shows a different pattern 7.1 (c-d). The generator's loss starts high at around 3, gradually decreasing, but it experiences a significant



(a) Generator Loss EsrGAN

(b) Discriminator Loss EsrGAN

(c) Generator Loss Pix2Pix GAN

(d) Discriminator Loss Pix2Pix GAN

Figure 7.1: Behaviour of the loss functions for both EsrGAN (a-b) and Pix2Pix GAN (c-d) along epochs

spike at epoch 38 before settling at 2.2 by the end of the training. On the other hand, the discriminator's loss remains relatively stable between 0 and 0.1, suggesting that it has no difficulty distinguishing between real and fake samples. This indicates potential issues with the generator's ability to produce convincing samples, as the discriminator is consistently successful in identifying fakes. This can mean that the discriminator is underfitting.

7.2 Training and Validation Set

Here are the plots illustrating the changes in PSNR, SSIM, and MSSSIM values for the training and validation sets throughout the training phase. All these metrics are measured and saved at each checkpoint; the checkpoint is set to every five epochs. It allows evaluating the model and its evolution process in order to make necessary enhancements if needed.

The metric values are reported, and the results are supported by standard errors' deviations to display the variability and robustness of the result. Secondly, to add more credit to outcomes and increase the confidence level, an additional criterion – the 10th percentile of the measurements is incorporated. It assists in sensing and analyzing the set of potential outliers, which provides a deeper comprehension of the model's effectiveness.

7.2.1 EsrGAN

The SSIM metric for the training set in fig. 7.2 (a-b) exhibits a fluctuating trend around 0.3, characterized by discernible peaks and dips across successive epochs. The error bars underscore significant variability, indicating inconsistent performance across different subsets of data. Despite this variability, the 10th percentile remains consistently low, aligning closely with the overall metric trend. It could indicate that there may be a specific class of images, such as those with many motion artifacts, that the model finds more challenging to process, impacting the overall results and producing an asymmetric distribution in the values.

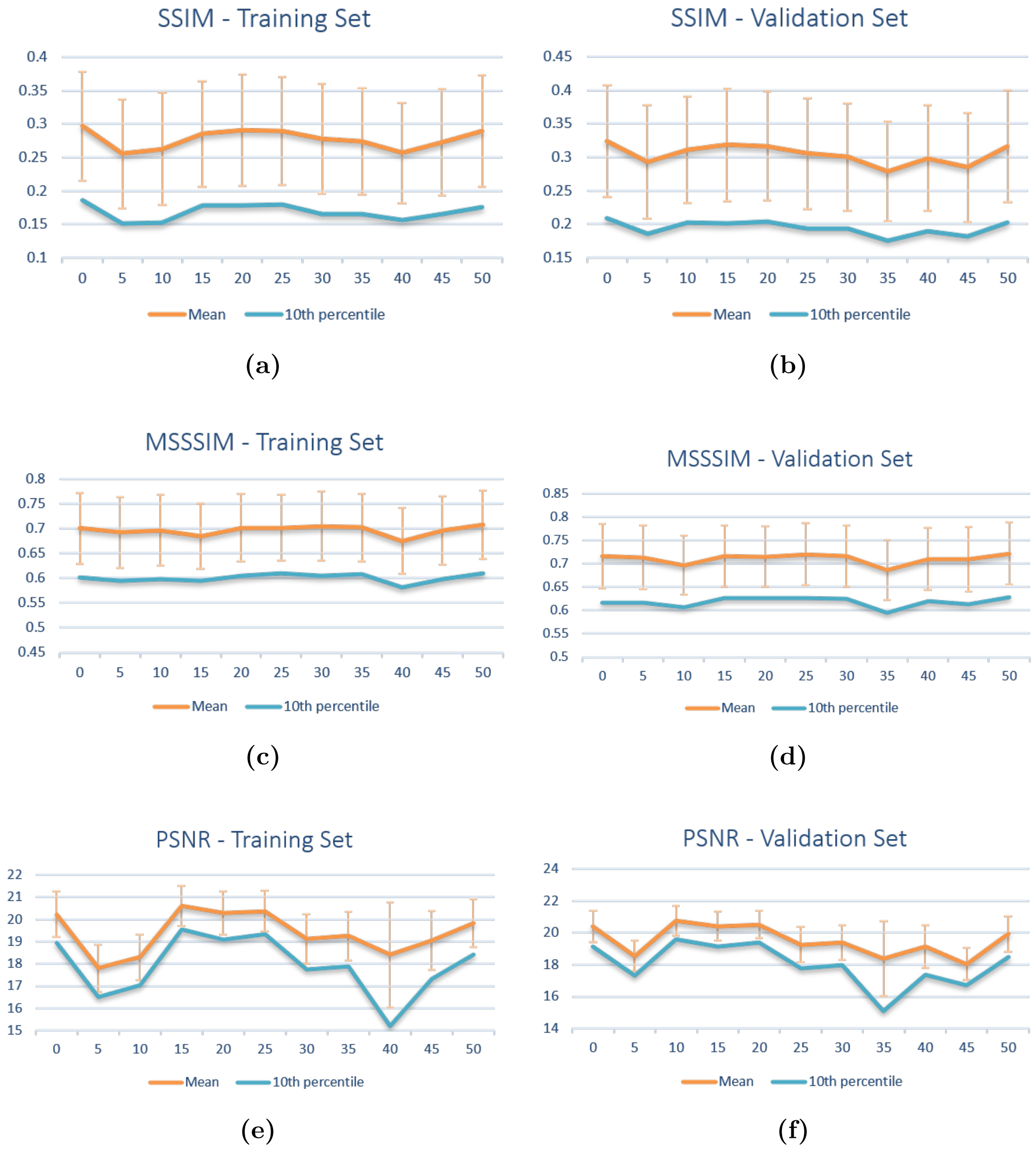
Conversely, the validation set demonstrates a similar pattern, where mean values show slight improvements over time but with an increase in variability. Notably, these fluctuations do not suggest overfitting, as the metric's behaviour remains relatively stable.

In contrast, the MSSSIM metric, fig. 7.2 (c-d) , portrays a more stable scenario on the training set, maintaining a mean value between 0.7 to 0.75 with minor fluctuations. Shorter error bars in comparison to SSIM imply more consistent model performance. While the validation set exhibits a slight decline in mean values, hovering around 0.7, there is an observable enhancement in variability, indicating robust model generalization.

Turning to PSNR, its trajectory on the training set fluctuates approximately between 18 and 21 dB, punctuated by notable peaks around epoch 15 and 30, but with a significant dip by epoch 40. The presence of substantial error bars highlights considerable variability in PSNR values across epochs, with the 10th percentile consistently trailing the mean values at lower dB levels (around 17 to 19 dB).

On the validation set, PSNR starts higher, approximately 21 dB, but experiences a decline to about 17 dB before a modest recovery to 18 dB towards the conclusion of training. Despite noticeable variability, error bars suggest comparatively more consistent performance on the validation set. The 10th percentile mirrors the mean trend closely, reflecting similar performance trends across epochs.

Each metric provides unique insights into different aspects of the model's performance. Overall, despite some fluctuations the ESRGAN model shows stability across epochs. The validation sets show similar trends to the training sets, indicating that the model's performance generalizes reasonably well to unseen data without severe overfitting.



(a) **(b)**

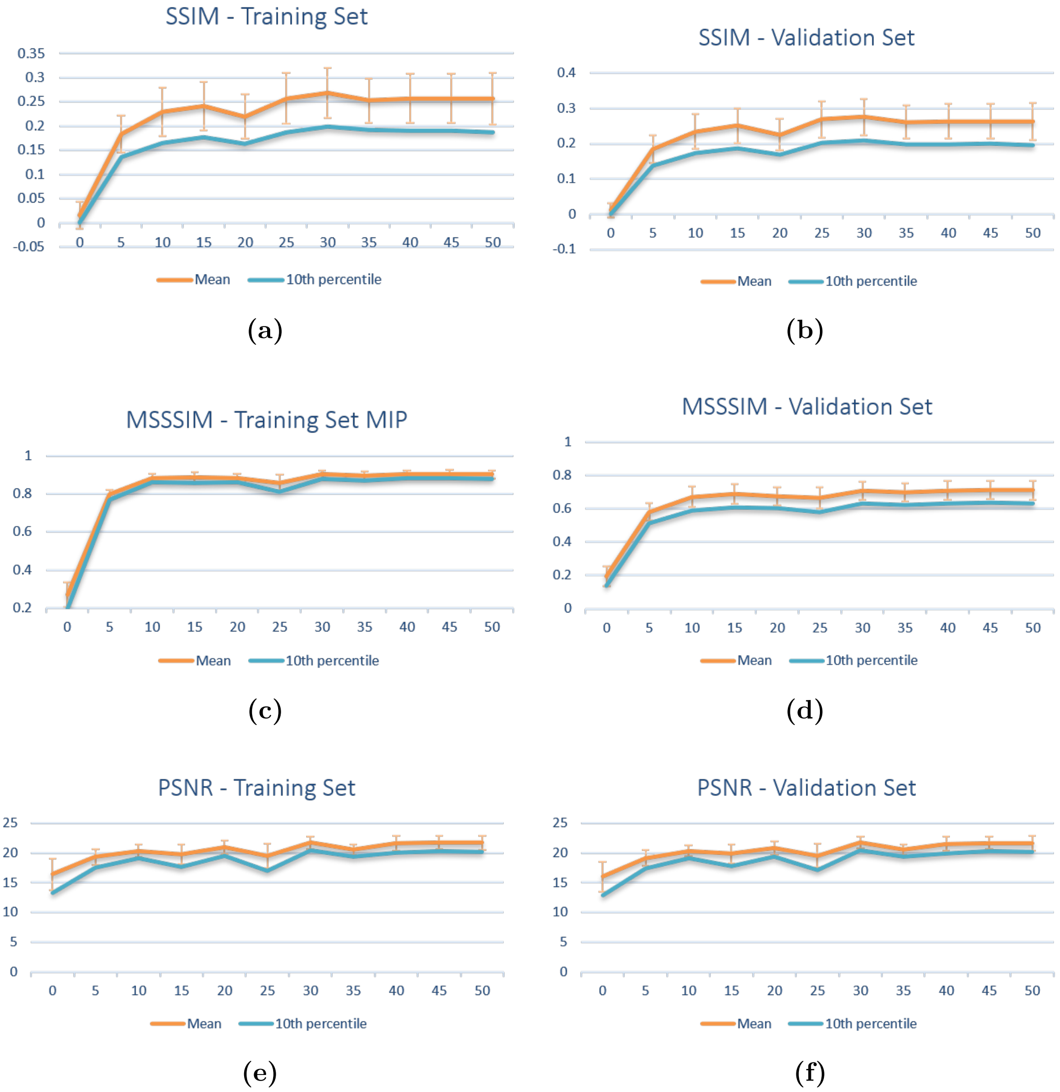
(c) **(d)**

(e) **(f)**

Figure 7.2: Metrics evaluation along training epochs for the angiograms of 5 slices with the EsrGAN

7.2.2 Pix2PixGAN

The Pix2Pix GAN shows a different scenario.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 7.3: Metrics evaluation along training epochs for the angiograms of 5 slices with the Pix2Pix

In Figure 7.3 (a-b) , the SSIM metric on the training set exhibits a trend of gradual increase, stabilizing around 0.25 from epoch 35 onwards. The error bars indicate moderate variability, suggesting consistent performance. The 10th percentile closely mirrors the metric's trend without significant outliers. Similarly, the validation set displays a comparable trend, with mean values showing slight improvements over time while maintaining stable variability patterns.

In the case of MSSSIM, a similar progression is observed. The mean value gradually increases to 0.7, maintaining consistent variability throughout. This trend is also reflected in the stability of the 10th percentile across various data subsets within the training set. Similarly, the validation set exhibits analogous behavior across all evaluated aspects.

The PSNR metric follows a similar progression but with a slower rate of increase and noticeable fluctuations. However, both the training and validation sets eventually stabilize around 20 dB in mean value. Throughout the epochs, variability in PSNR fluctuates, however during periods of stability it remains relatively small.

Considering the described trends in SSIM, MSSSIM, and PSNR metrics, The Pix2Pix GAN demonstrates a coherent learning process where it steadily enhances structural and multi-scale similarities with the target images over time. The variability observed in PSNR suggests ongoing adjustments during training, but ultimately, the model achieves stable performance metrics across different evaluation aspects.

7.3 Comparative analysis

The evaluations presented have highlighted distinct characteristics of the two models. The EsrGAN displayed fluctuations in metrics like SSIM and PSNR, indicating varying performance over epochs. In contrast, the Pix2Pix GAN shows a trend towards stability and consistent improvement in these metrics. This suggests that the Pix2Pix GAN maintains more reliable performance over time.

Both models demonstrate the ability to generalize to validation data, but the PIX2PIX GAN exhibits more stable and improved performance metrics across different evaluation aspects. Specifically, the Pix2Pix GAN shows gradual improvements in SSIM and MSSSIM metrics, maintaining comparable mean values with less variability. Similarly, the PSNR metric in Pix2Pix GAN stabilizes around a higher average dB value, indicating better image quality preservation compared to the EsrGAN, which showed more fluctuations.

However, as depicted in 7.4 , both SSIM and MSSSIM metrics indicate superior performance in the outputs generated by EsrGAN. Furthermore, comparing the results of both GANs with the values obtained for the low-quality (LQ) input, EsrGAN has demonstrated incremental improvements across all metrics. In contrast, while Pix2Pix significantly enhances PSNR, it does not show improvement in SSIM and MSSSIM metrics.

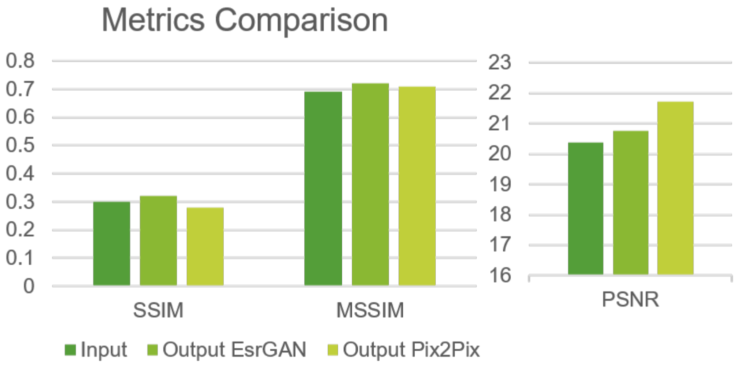


Figure 7.4: Comparison of the best performance metrics on the validation set among input vs. target (baseline), ESRGAN output vs. target, and Pix2Pix output vs. target

Upon qualitative comparison, as illustrated in the example in Figure 7.5 , the same distinct differences emerge. The output generated by EsrGAN appears noticeably more accurate than both Pix2Pix output and the LQ input and closely resembles the target image. Although the numerical improvement may be marginal, the visual enhancement is pretty evident.

Conversely, Pix2Pix exhibits less contrast, resulting in less sharp details in morphology. However, it appears smoother in appearance, which aligns with its higher PSNR metric.

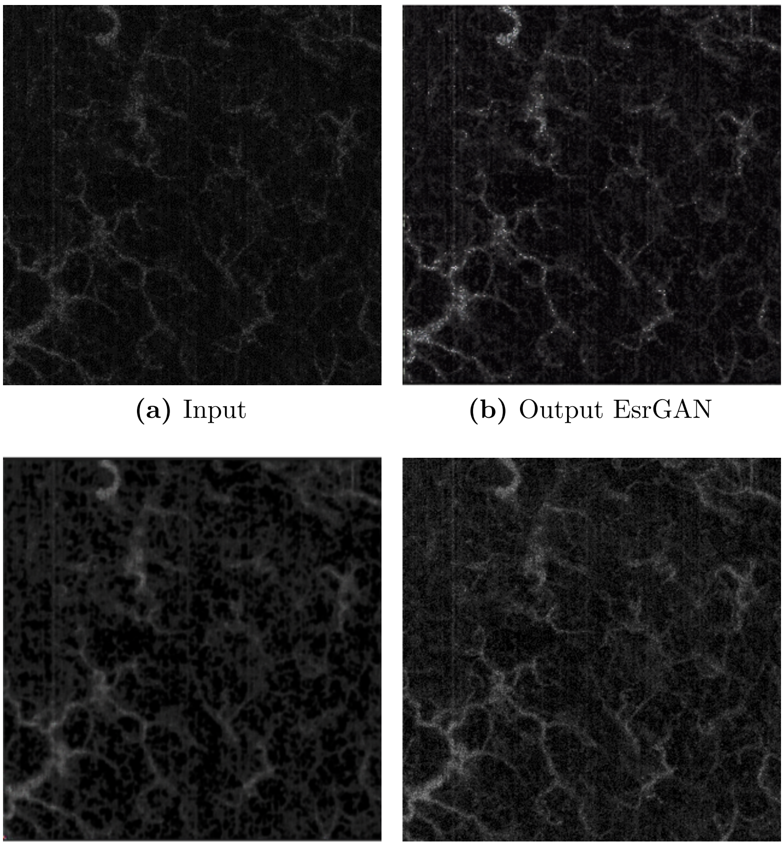


Figure 7.5: Example of input image and the outputs obtained from the two models

Chapter 8

Results on enface MIP

Once the methods are proven to achieve good performances, a further assessment on a dataset consisting of MIPs is performed. Visualizing the skin volume with 2 MIPs helps identify specific differences between the two plexuses that characterize the dermis. This allows for the distinction of abnormalities specific to the microcirculation in CVI.

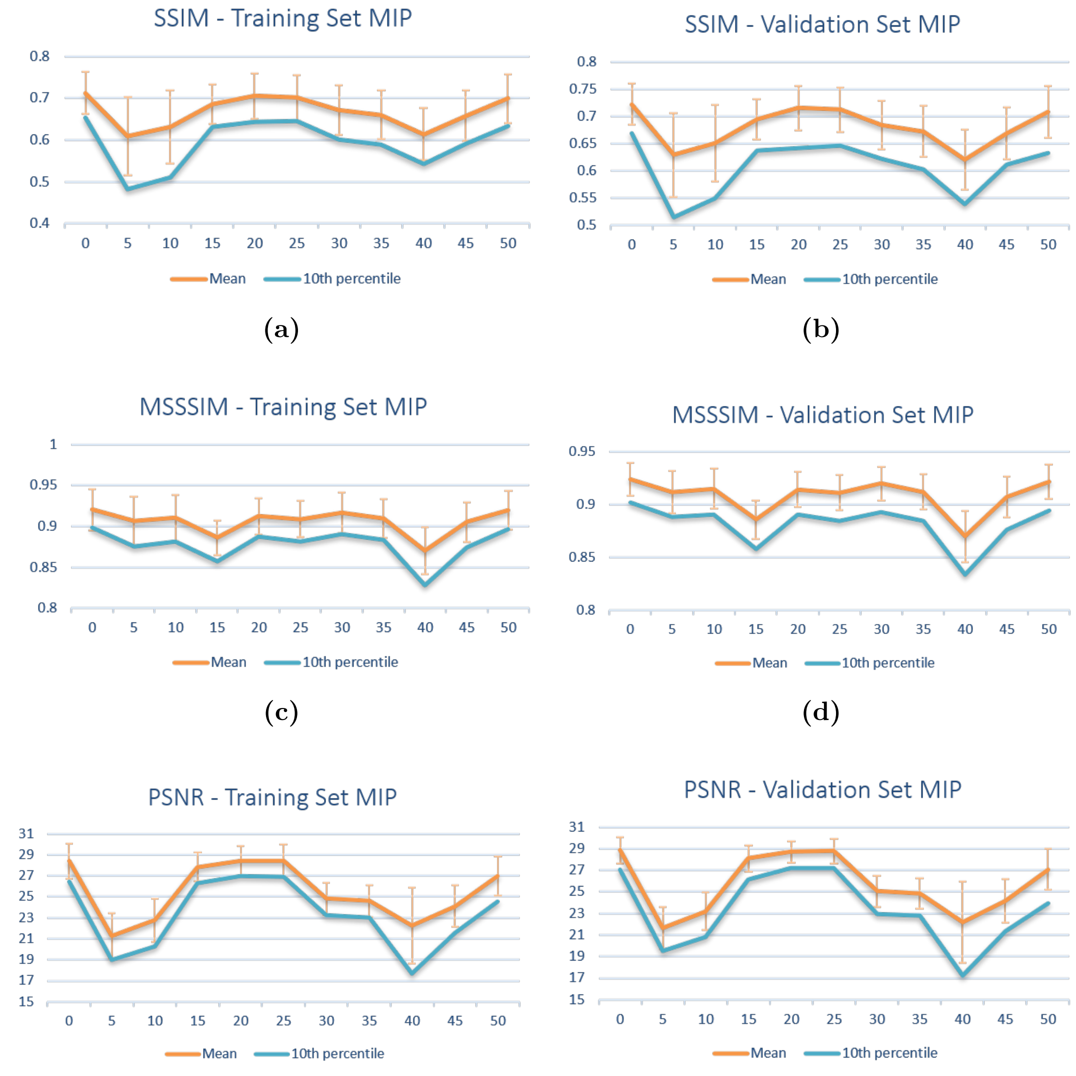
8.1 Training and Validation Set

The same procedure as before is performed for this dataset. The aim of this phase is to identify the best epoch and finally test the models.

8.1.1 EsrGAN

The SSIM for the training set begins at approximately 0.7. It exhibits a slight downward trend with some fluctuations. This metric reaches its lowest point around epoch 10. Then it stabilizes back to approximately 0.7. The 10th percentile values of SSIM display more significant fluctuations. They initially drop sharply then gradually rise. This pattern suggests considerable variability in the model's worst-case performance during the early epochs. Similarly the validation set SSIM mirrors the training set trend. It also shows initial significant fluctuations. It drops markedly at the beginning before gradually stabilizing around 0.7. The variability in the validation set is prominent in the initial epochs. It stabilizes after epoch 10.

The mean MSSSIM for both the training and validation sets remains relatively stable around 0.9. There are only minor fluctuations. The 10th percentile values of MSSSIM exhibit minor fluctuations. This indicates consistent performance across different scales. This stability in MSSSIM values suggests the model maintains



(a) **(b)**

(c) **(d)**

(e) **(f)**

Figure 8.1: Metrics evaluation along training epochs for the reconstructed MIPs with the EsrGAN

high structural similarity across various resolutions both in training and validation phases.

The mean PSNR begins at roughly 30 dB. By epoch 10 it sharply declines to about 20 dB. Subsequently, it stabilizes between 25 and 30 dB. The 10th percentile values of PSNR exhibit a similar initial drop. These values then recover and stabilize around 20 dB. This pattern is evident in the training set. It is also evident in the validation set indicating similar behavior.

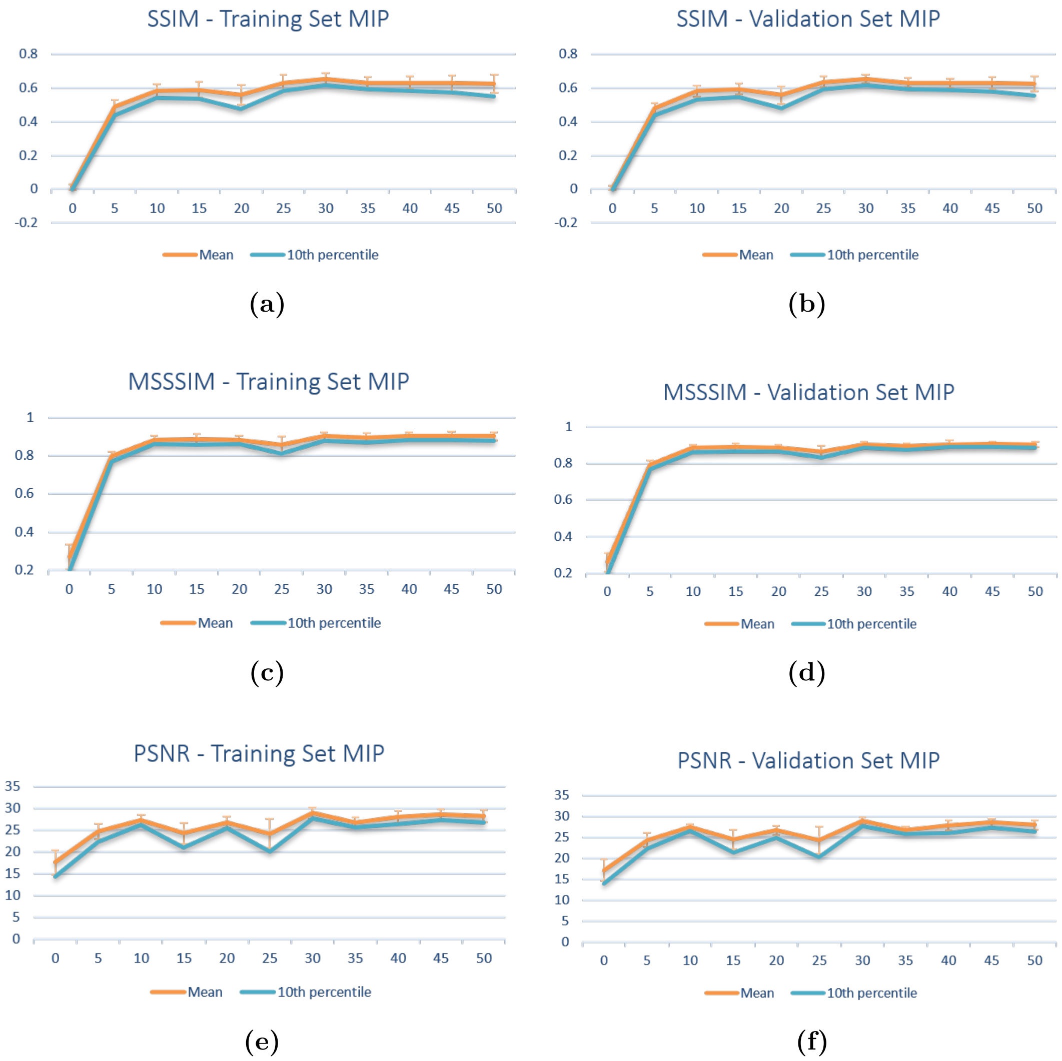
All the metrics demonstrate significant fluctuations during initial epochs. This observation suggests a phase of instability or adjustments as the model undergoes early training. After this phase, metrics stabilize. Mean values reflect consistent performance. However, the 10th percentile values indicate some variability. This highlights differences in the model's performance in lower-quality cases. Overall, the training and validation sets exhibit similar trends across all metrics. This suggests that the model generalizes well from training data to unseen validation data. This generalization is crucial for the robustness of model. Performance improvements during training are effectively translated to new data.

8.1.2 Pix2Pix GAN

For training set the Pix2Pix GAN exhibits a SSIM beginning at a low value. It experiences a rapid increase during the initial 10 epochs. It reaches approximately 0.6. Beyond the first 10 epochs, the SSIM stabilizes with a slight continued increase. It ultimately peaks at around 0.7. The 10th percentile of SSIM follows a similar trajectory. It remains consistently lower than the mean. This indicates variability in model's performance across different samples. This suggests that while model performs well on average, certain samples do not achieve the same level of reconstruction quality. For validation set the SSIM demonstrates a pattern very similar to the training set. The SSIM values increase quickly in the initial stages of training then stabilize. They maintain values between 0.6 and 0.7. This parallel trend between training and validation sets indicates that model generalizes well. It does not overfit the training data.

In the training set mean MSSSIM exhibits a rapid increase within the first 10 epochs. It stabilizes around 0.9. The 10th percentile of MSSSIM follows a similar pattern. However, it remains consistently lower than the mean. This indicates some variance in performance. This suggests that while the model generally captures structural details well there is still some variability in how well different samples are reconstructed. The validation set mirrors this trend. The MSSSIM rapidly increases in the initial epochs and stabilizes around 0.9. The close alignment between the training and validation sets for MSSSIM suggests robust performance.

For both training and validation sets the mean PSNR increases quickly within first 10 epochs. It reaches values around 25-30 dB. There is noticeable fluctuation in the PSNR values after the initial increase. This indicates variability in the noise level of the reconstructed images. The 10th percentile of PSNR also shows a similar trend. These values are slightly lower. This highlights variability in reconstruction quality.



(a) **(b)**

(c) **(d)**

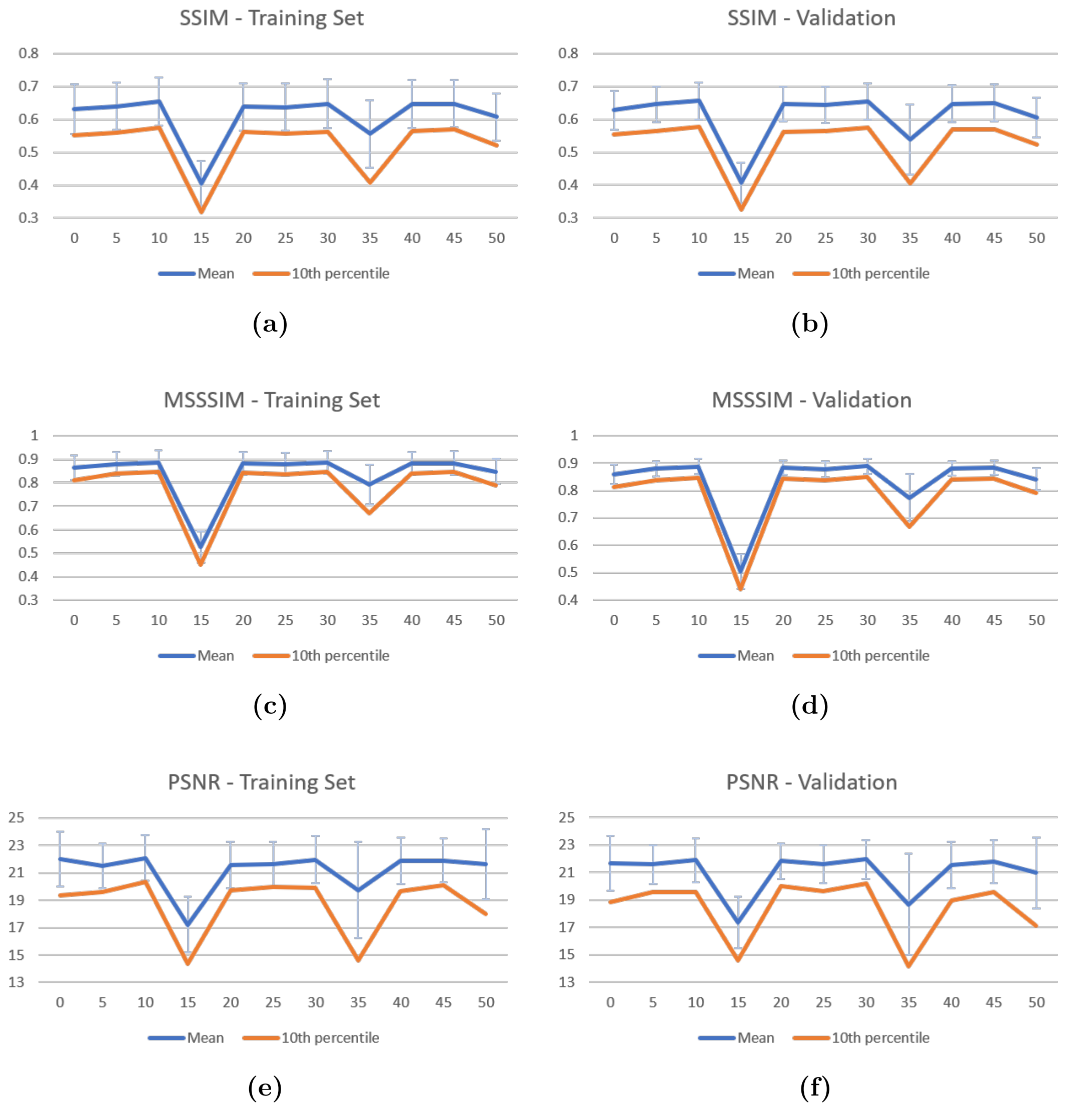
(e) **(f)**

Figure 8.2: Metrics evaluation along training epochs for the reconstructed MIPs with the Pix2Pix GAN

Overall both SSIM and MSSSIM show rapid improvement during the initial epochs (0-10). They then stabilize, indicating that the model quickly learns structural features of images. The stabilization of these metrics suggests that the model achieves good level of reconstruction quality early in the training process. In contrast PSNR shows more fluctuation compared to SSIM and MSSSIM. This suggests that while the model can reconstruct images with good structural similarity. There is more variability in the noise level of the reconstructions. The consistency between the training and validation sets across all metrics— SSIM, MSSSIM and

PSNR —indicates that the model is not overfitting. The validation metrics closely follow the training metrics. This demonstrates that the model's performance is robust and generalizes well to unseen data. This consistency is a positive indicator of the model's reliability and effectiveness in reconstructing OCTA images.

8.1.3 Training the Networks with MIP Images



(a) (b) (c) (d) (e) (f)

Figure 8.3: Metrics evaluation along training epochs of EsrGAN trained with MIPs

In addition to training the GANs with angiograms of 5 slices of the volume and evaluating their performance on reconstructed MIPs, direct training with MIPs was also explored to provide a comparative analysis. Networks architecture, losses and parameters are maintained the same.

The EsrGAN model (fig. 8.3) shows a similar scenario as before. Throughout the training process mean PSNR values showed moderate fluctuations between approximately 17 and 23. This suggested that the model's ability to reconstruct high-quality images improved over time. Occasionally, there were dips. Similarly mean SSIM values ranged from 0.55 to 0.7. This indicated consistent structural similarity between the generated and original images. There was slight downward trend around certain epochs. MSSSIM, which remained high between 0.75 and 0.95 confirmed that model maintained good perceptual quality throughout training.

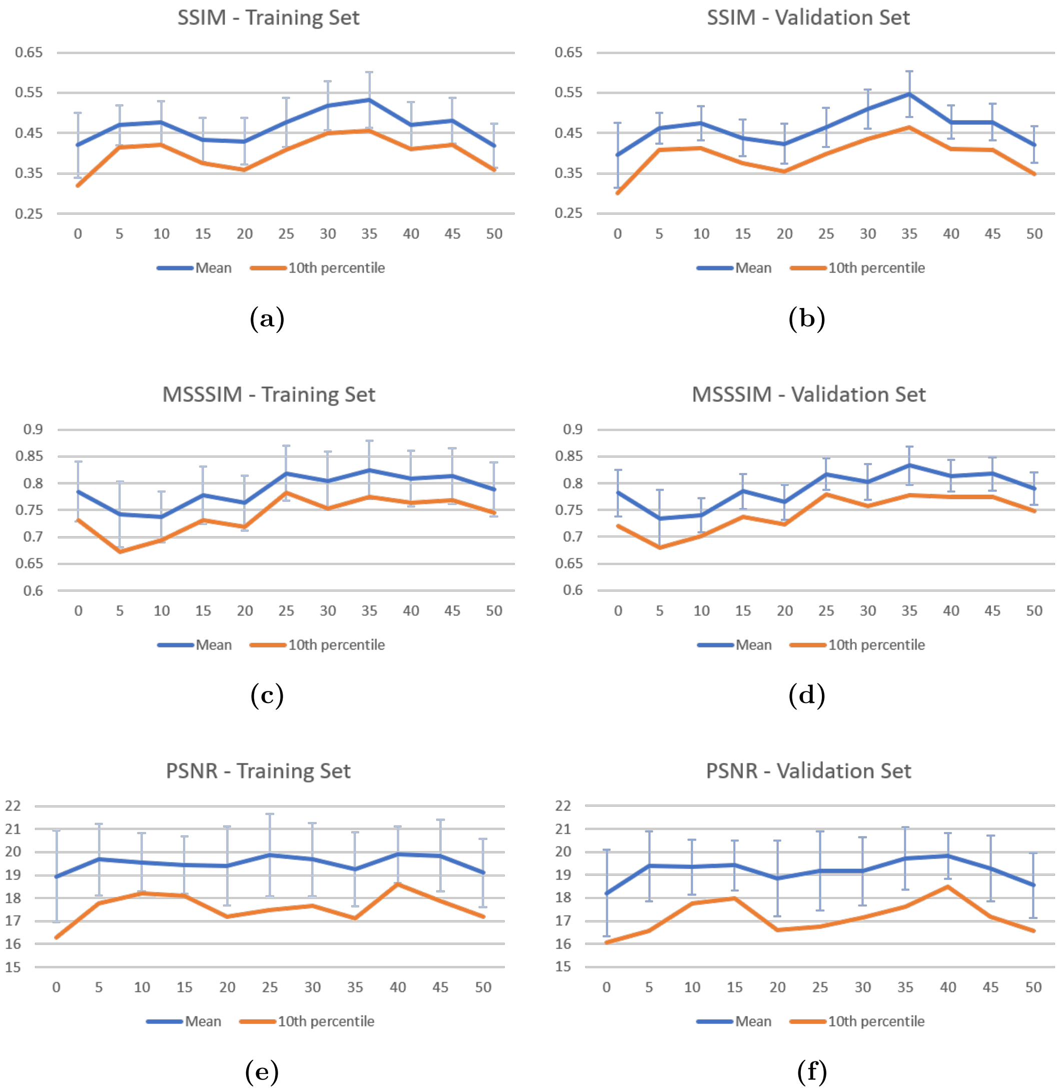
For the validation set, mean PSNR values were slightly higher. They ranged from 19 to 23 reflecting the model's generalization capability. SSIM values were also higher, fluctuating between 0.6 and 0.75. This indicated robust performance in maintaining structural similarity on unseen data. MSSSIM scores remained consistently high from 0.85 to 0.97. This underscored the model's strong perceptual quality.

The disparity between mean values and the 10th percentile in all metrics suggests that while most outputs were of high quality a small fraction experienced lower performance. This could be attributed to particularly challenging images in the datasets.

For the Pix2Pix model (8.4), the PSNR values suggest that while the model can generate images with relatively high detail there are fluctuations that might indicate instability in the learning process or the influence of noise at different stages. The increasing mean values over epochs demonstrate improvement. The variability shown by the error bars indicates not all generated images consistently achieve high quality. SSIM values both mean and 10th percentile, indicate moderate structural similarity. This suggests that while the model can replicate structural information there is room for improvement in capturing finer details and textures. The steady increase over epochs is promising, but the plateauing effect observed towards the later epochs might suggest the model is reaching capacity in terms of structural fidelity under the current configuration.

MSSIM provides more favorable assessment. Higher values indicate better perceived image quality overall. The consistent improvement and less variability compared to PSNR and SSIM highlight the model performs well in maintaining overall structural similarity. Pixel-level accuracy as indicated by PSNR, is more variable.

Considering these metrics together model demonstrates clear learning trend and improvement over time. Yet the variability and eventual plateau suggest potential



(a) **(b)**

(c) **(d)**

(e) **(f)**

Figure 8.4: Metrics evaluation along training epochs of Pix2Pix GAN trained with MIPs

areas for refinement.

In general the obtained results can be observed to present similar issues but have slightly lower values. This can be attributed to several factors. For example, dataset size is reduced: Training with MIP limits the number of images available for training the network. Using only 2 images per volume instead of the 16 used previously reduces the dataset by 1/8. This significant reduction in data can impact the model's ability to generalize. This leads to lower performance metrics.

Detail loss in MIP is another factor. Although MIP images are visually sharper and more contrasted they can hide details and lead to information loss. This occlusion of finer details can cause lower PSNR, SSIM and MSSIM values may also suffer. The model might miss subtle features that contribute to higher similarity scores.

In essence training the network with 5-slice angiograms helps in retaining even minor details. This consequently leads to higher metric values. This suggests more detailed and varied training dataset, even if it involves more computational complexity is beneficial for enhancing the overall performance of the GAN model. It results in generating high-quality images. Additionally exploring alternative training strategies could improve the metrics. Incorporating hybrid approaches that balance the sharpness of MIP images with the detail retention of multi-slice images also could be effective.

8.2 Test Set

When evaluating and picking out the best epoch for GAN-based model to enhance OCTA images, it is very important to rely on strong and meaningful metrics. Of the three metrics calculated we chose SSIM as the prime metric which would help determine the optimal epoch. Numerous key factors informed this selection.

SSIM has been developed specifically in order to mimic human visual system sensitivity to structural changes in images. Unlike PSNR, a pixel-wise difference measure, SSIM assesses changes in luminance, contrast and structure thus providing a more perceptually aligned appraisal of image quality. Moreover, MSSSIM, though robust, introduces additional complexity in interpretation and computation. For the sake of simplicity and direct relevance to structural integrity, SSIM is preferred.

EsrGAN

For the EsrGAN referring to the fig. 8.1 , the higher SSIM is reached at epoch 20. Therefore this was chosen as best model. According to this the performaces reached are:

PSNR	SSIM	MSSSIM
28.64	0.70	0.92

Pix2Pix GAN

For the Pix2Pix GAN the best performances in the validation set are obtained at epoch 30. The final metrics obtained are:

PSNR	SSIM	MSSSIM
29.12	0.65	0.91

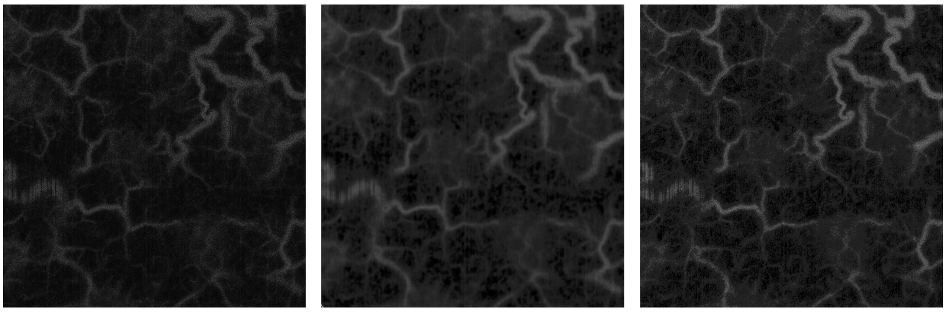
8.3 Comparative Analysis

Both ESRGAN and Pix2Pix GAN enhance OCTA images with similar final metrics. Aside the similarity in efficacy between the two, the training dynamics and stability for ESRGAN and Pix2Pix GAN are different.

1. Pix2Pix GAN: this model rapidly stabilizes and maintains more consistent performance across samples. Even though the metrics reveal a better scenario, from the loss curves 7.1 (c-d) a possible problem of underfitting of the discriminator is presented. It could explain the qualitative differences between the generated images and the target. This also shows the inability of the metrics to describe the real quality of this images, creating the need of new metrics.

2. EsrGAN: EsrGAN demonstrates variable stability and performance early in the performances, but eventually gains a comparable performance metric likely due to the network's ability to adapt over time. The loss curves anyway show the stable training process that produces images more similar to the desired ones.

Anyway, the use of MIPs significantly enhances the performance metrics for both EsrGAN and Pix2Pix GAN by leveraging the combined information from multiple images, thus providing a more reliable and accurate assessment of image quality.



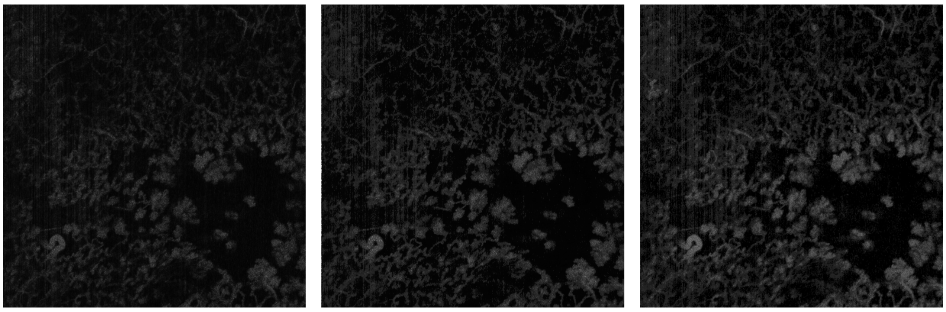
(a) MIP inputs **(b)** MIP Pix2PixGAN **(c)** MIP Target

Figure 8.5: Example of a MIP obtained from inputs, outputs and targets for the Pix2Pix GAN

When evaluating the quality of images, visual inspection carries greater significance than relying solely on metrics or measurements.

In figure 8.5 it is possible to observe a MIP recreated with the outputs of the Pix2Pix compared with the 2-volume octa (input) and the 4-volume octa (target). The structural clarity is significantly improved over the input MIP, with features becoming more distinct and easier to recognize. The texture quality is also enhanced, giving the image a smoother and more detailed look. Features are more defined and stand out better against the background. There is a marked improvement in overall image quality, with reduced noise and enhanced detail visibility.

The images show improvements achieved with the Pix2Pix GAN while also underlining the remaining gap between the enhanced image and the high-quality target image.



(a) MIP inputs **(b)** MIP EsrGAN **(c)** MIP Target

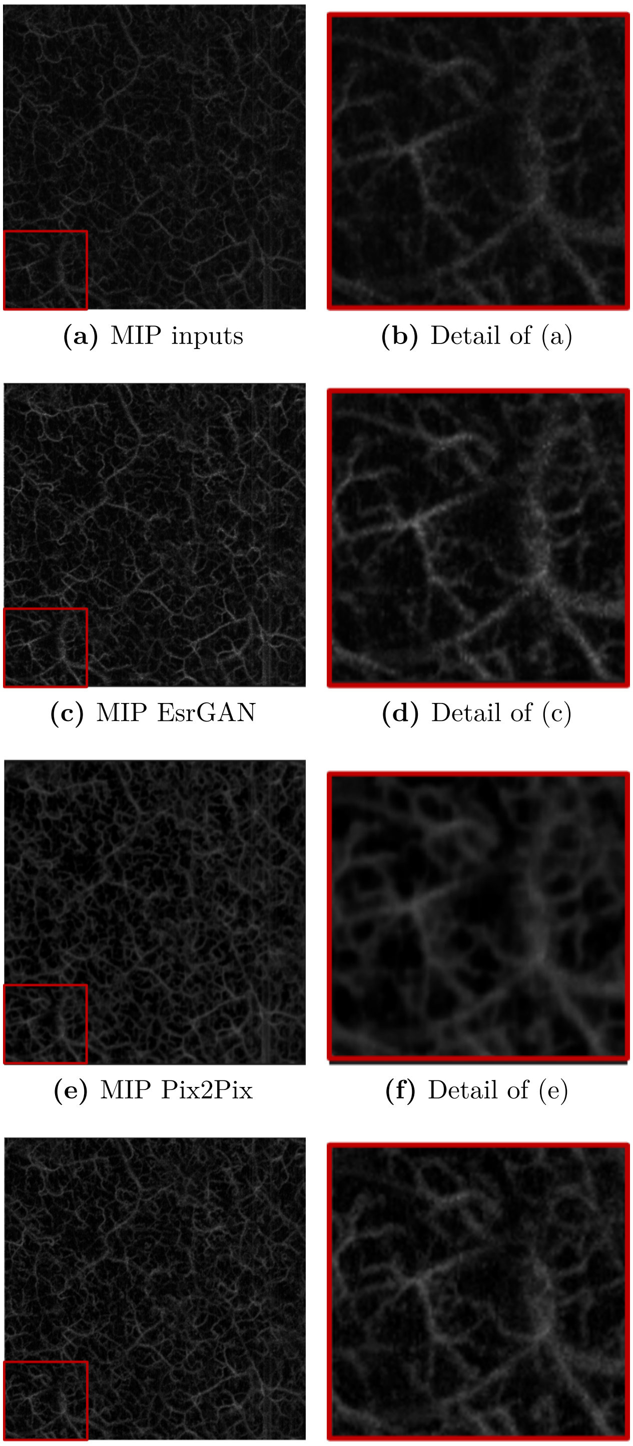
Figure 8.6: Example of a MIP obtained from inputs, outputs and targets for the EsrGAN

Turning to the EsrGAN, in figure 8.6 are shown three images as (a) MIP inputs, (b) MIP EsrGAN, and (c) MIP Target. The contrast in the input image is relatively low which makes it hard to distinguish between different structures and details. In the generated image the contrast is noticeably improved and this helps to better differentiate between various elements within the image. The areas that were previously blended now stand out more clearly. They get very close to the target.

There is also a significant reduction in noise in this image compared to input. This reduction helps in bringing out more detail: it provides a cleaner look, although some noise artifacts may still be present. Details are much more pronounced in the EsrGAN-enhanced image, edges are sharper and the texture is more defined. This helps in recognizing and analyzing specific areas.

The overall visual clarity is greatly improved. This makes the image more useful for analysis as the structures are more distinguishable. The image appears more visually appealing.

A further example of a MIP generated by both networks is shown in the figure 8.7.



(a) MIP inputs **(b)** Detail of (a)

(c) MIP EsrGAN **(d)** Detail of (c)

(e) MIP Pix2Pix **(f)** Detail of (e)

(g) MIP target **(h)** Detail of (g)

Figure 8.7: Example of a MIP obtained from inputs, outputs and targets. Side by side a magnified image of the detail in the red square

The comparison in qualitative visual assessment reveals that the image generated by EsrGAN is significantly more contrasted compared to the LQ input version. This increased contrast makes it much easier to visualize all the smaller vessels located in deeper layers, which would otherwise be lost (as shown in the detail below). It effectively resembles the target image.

In contrast, while the Pix2Pix GAN achieves comparable quantitative results, the difference is visually evident. Despite the increased contrast and apparent highlighting of all vessels, the final effect is far from the target. It appears to lose a lot of sharpness, making the image less clear. Hence, there is a need for additional metrics that can more accurately describe the quality of the MIP compared to the target.

In the end, from both quantitative and qualitative point of view, EsrGAN offers a more suitable solution to the purpose of this thesis.

Part IV

Conclusion

This thesis has explored the application of GANs to enhance the quality of OCTA images. More in detail the proposed method aims to reduce the number of OCT volumes needed to reconstruct a single OCTA volume while maintaining high image quality. Going from 4 to 2 OCT volumes means indeed halving scan time, which means also reducing the possible artifacts caused by patient motion or laser instability, and cutting in half the data weight, which means reducing saving time and session duration for the patient.

Two solutions, EsrGAN and Pix2Pix, were evaluated using MIP projections of the volumes. The results demonstrated that ESRGAN more than the other provided a promising solution in achieving the desired image quality.

However, even with these improvements, there were challenges related to the quality of the initial OCTA data. The inherent variability and the sometimes limited quality of the initial data show difficulty with GANs, which capture minute details for high-quality image reconstruction.

New metrics are needed to accurately assess the performance of this models in tasks like enhancing images quality, as existing metrics often do not adequately measure fidelity to real-world data and preservation of crucial details as shown in the Pix2Pix GAN case.

Moreover, architectural constraints on the software and hardware side bound us to work with a smaller batch size, impacting computational efficiency in the training and inference phases. This may be improved by the enhancement of hardware capabilities or optimization of the software framework for large batch sizes without compromising performance.

Looking forward, research can be extended toward including training strategies for complete 3D OCTA volumes instead of only 2D projections. This strategy is very promising for a more complete perception of spatial and volumetric information, with potential benefit that lies in an increment of accuracy and detail in skin vasculature reconstructions.

The use of integrated phase information, not wholly employed in this work, is of great promise for further improved accuracy and interpretability of OCTA images. Phase-sensitive OCTA techniques could hold improved differentiation capability of tissue layers and flow characteristics for enhanced diagnostic capabilities [25].

Diffusion models could be another promising avenue [26]. Latent diffusion models offer several benefits, such as their capability to synthesize high-resolution images while preserving intricate details and quality. They are designed to be memory-efficient, enabling the generation of high-resolution images even under resource constraints [27].

In summary, despite the noted challenges, this thesis provides a foundation for further advancements in leveraging GANs to optimize OCTA imaging protocols.

Bibliography

[1] A. F. Fercher, W. Drexler, C. K. Hitzenberger, and T. Lasser. «Optical coherence tomography - principles and applications». In: *Reports on Progress in Physics* 66 (2003) (cit. on p. 4).

[2] W. Drexler, M. Liu, A. Kumar, T. Kamali, A. Unterhuber, and R. A. Leitgeb. «Optical coherence tomography today: speed, contrast, and multimodality». In: *Journal of Biomedical Optics* 19.10 (2014) (cit. on p. 5).

[3] M. Ali and R. Parnapalli. «Signal Processing Overview of Optical Coherence Tomography Systems for Medical Imaging». In: *Texas Instrum.* (2010) (cit. on p. 5).

[4] Pete Tomlins, Jordi Lopez, Karl Alvarez, Rob Donnan, and Adina Michael-Titus. «Heart rate sensitive optical coherence angiography». In: *Proceedings of SPIE*. Vol. 70. 2018. DOI: `10.1117/12.2317602` (cit. on p. 6).

[5] Z. Chen, M. Liu, M. Minneman, L. Ginner, E. Hoover, H. Sattmann, M. Bonesi, W. Drexler, and R. A. Leitgeb. «Phase-stable swept source OCT angiography in human skin using an akinetic source». In: *Biomed. Opt. Express* 7 (2016) (cit. on pp. 8, 22).

[6] J. Wang, M. Zhang, T. S. Hwang, S. T. Bailey, D. Huang, D. J. Wilson, and Y. Jia. «Reflectance-based projection-resolved optical coherence tomography angiography». In: *Biomed. Opt. Express* 8.3 (2017) (cit. on p. 9).

[7] T. Cammarota, F. Pinto, A. Magliaro, and A. Sarno. «Current uses of diagnostic high-frequency US in dermatology». In: *European Journal of Radiology* 27 (1998) (cit. on p. 10).

[8] A. J. Deegan and R. K. Wang. «Microvascular imaging of the skin». In: *Physics in Medicine & Biology* 64 (2019) (cit. on p. 10).

[9] R. T. Eberhardt and J. D. Raffetto. «Chronic venous insufficiency». In: *Circulation* 130.4 (2014) (cit. on p. 11).

[10] Ian J. Goodfellow, Jean Pouget Abadie, Mehdi Mirza, Bing Xu, David Warde Farley, and Sherjil Ozair. «Generative Adversarial Nets». In: *Advances in Neural Information Processing Systems (NIPS)*. 2014 (cit. on pp. 12–14).

[11] Martin Arjovsky, Soumith Chintala, and Léon Bottou. «Towards Principled Methods for Training Generative Adversarial Networks». In: *International Conference on Learning Representations (ICLR).* 2017 (cit. on p. 14).

[12] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. «Unrolled Generative Adversarial Network». In: *International Conference on Learning Representations (ICLR).* 2017 (cit. on p. 15).

[13] Christian Ledig et al. «Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2017 (cit. on p. 18).

[14] Xing Yuan et al. «Image enhancement of wide-field retinal optical coherence tomography angiography by super-resolution angiogram reconstruction generative adversarial network». In: *Biomedical Signal Processing and Control* 78 (2022), p. 103957 (cit. on p. 18).

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. «Image-to-Image Translation with Conditional Adversarial Networks». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2017 (cit. on p. 19).

[16] J. Zhu, C. W. Merkle, M. T. Bernucci, S. P. Chong, and V. J. Srinivasan. «Can OCT angiography be made a quantitative blood measurement tool?» In: *Applied Sciences* 17.7 (2017) (cit. on p. 22).

[17] G. Rottunno. «Optical coherence tomography angiography and automatic vascular analysis: seeing skin lesions from a different perspective». Master's thesis. Politecnico di Torino, 2023 (cit. on p. 23).

[18] Jianwei Liao, Shiyu Yang, Chao Li, Ting Zhang, and Zhenghua Huang. «Fast optical coherence tomography angiography image acquisition and reconstruction pipeline for skin application». In: *Biomedical Optics Express* 14.8 (2023), pp. 3899–3913 (cit. on p. 24).

[19] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. «Residual Dense Network for Image Super-Resolution». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* 2018 (cit. on p. 27).

[20] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. «ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks». In: *Proceedings of the European Conference on Computer Vision (ECCV).* 2018 (cit. on p. 27).

[21] Mohit Bansal, Manish Kumar, Mamta Sachdeva, et al. «Transfer learning for image classification using VGG19: Caltech-101 image data set». In: *Journal of Ambient Intelligence and Humanized Computing* 14.5 (2023), pp. 3609–3620. DOI: `10.1007/s12652-021-03488-z`. URL: `https://doi.org/10.1007/s12652-021-03488-z` (cit. on p. 27).

[22] Zhe Jiang et al. «Comparative study of deep learning models for optical coherence tomography angiography». In: *Biomedical Optics Express* 11 (2020), pp. 1580–1597 (cit. on p. 32).

[23] Steven McDonagh et al. «Attention U-Net: Learning Where to Look for the Pancreas». In: *arXiv preprint arXiv:1904.03449* (2019) (cit. on p. 33).

[24] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. «Multiscale structural similarity for image quality assessment». In: *Proceedings of the Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*. Vol. 2. 2003, pp. 1398–1402 (cit. on p. 39).

[25] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. «MS-GAN: multiscale GAN with parallel class activation maps for image reconstruction». In: *The Visual Computer* 36.8 (2018), pp. 179–196 (cit. on p. 64).

[26] Darwon Rashid et al. «Using latent diffusion models to generate synthetic OCTA images». In: *ARVO Annual Meeting Abstract*. June 2024 (cit. on p. 64).

[27] Kun Huang, Xiao Ma, Yuhan Zhang, Na Su, Songtao Yuan, Yong Liu, Qiang Chen, and Huazhu Fu. «Memory-efficient High-resolution OCT Volume Synthesis with Cascaded Amortized Latent Diffusion Models». In: *arXiv preprint arXiv:2405.16516* (2023). arXiv: `2405.16516 [cs.CV]` (cit. on p. 64).