Beatriz Rodrigues Branco de Almeida Simões

Bachelor of Science in Biomedical Engineering

Image Quality Improvement of Medical Images using Deep Learning for Computer-aided Diagnosis

Dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Science in
Biomedical Engineering

Adviser: Prof. Dr. Hugo Filipe Silveira Gamboa, Assistant Professor, NOVA University of Lisbon
Co-adviser: Dr. André Carreiro, Senior Scientist, Fraunhofer AICOS

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

March, 2021

Image Quality Improvement of Medical Images using Deep Learning for Computer-aided Diagnosis

This document was created using the (pdf)LaTeX processor, based on the "novathesis" template[1], developed at the Dep. Informática of FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

Para os meus pais, cujo amor e apoio são das maiores constantes na minha vida.

Acknowledgements

The accomplishment of this important stage of my academic and personal life would not be possible without the immense support of several people.

First of all, I would like to express my gratitude to my adviser, Professor Hugo Gamboa, for allowing me to develop this fascinating project under his guidance and for welcoming me at Fraunhofer AICOS. Your enthusiasm for science is contagious. To my co-adviser, researcher André Carreiro, a special thank you, not only for all the transmitted knowledge, brilliant advice, and continuous motivation, but also for the daily good mood and humor.

I would like to thank *Associação Fraunhofer Portugal Research* for giving me the opportunity of developing my thesis on their facilities and for the amazing environment. Even through this challenging pandemic times, their sympathy and energy made me feel included and always motivated. I am eternally grateful to Ricardo Leonardo, who closer followed me in this journey, for the immense support, for always being available to clarify my doubts and for the infinite patience. You were not only a brilliant mind but a good friend throughout this project. Thank you to everyone at the Lisbon office, particularly my master thesis colleagues who even virtually made this experience better.

To my college friends, I would also like to express my sincere gratitude. You made this five-year journey so much more fun and exciting. A special thank you to Inês, Madalena, Joana, Miguel, Lourenço and João, for being always there for me and for making me laugh every day, I will cherish all our awesome memories forever. To my long-dated friends, I can not thank you enough for all the love and companionship through all the good and bad moments. Specially to Sofia, Faria, and Pipa, you are practically family to me and I can not imagine life without you.

To my special one, Frederico, thank you for the unconditional love and support throughout the last four years. Thank you for always being there to listen (I know my stress-self is not always easy) and for believing in me even when I did not. You are my rock and best friend, and this journey would not have been the same without you.

Lastly, the biggest and most sincere thank you to all my family. I owe you all that I am. To my parents, thank you for being my home, for loving me endlessly and supporting me in all that I do. You inspire me to always be better. To my two brothers, who throughout their silliness, awkwardness and constantly English-speaking, are the friends I want close forever. And to my Wilma, who is our family not by blood but heart, I can not thank you enough for all the love, support, laughs, and of course, chocolate cake.

Abstract

Retina image analysis is an important screening tool for early detection of multiple diseases such as diabetic retinopathy which greatly impairs visual function. Image analysis and pathology detection can be accomplished both by ophthalmologists and by the use of computer-aided diagnosis systems. Advancements in hardware technology led to more portable and less expensive imaging devices for medical image acquisition. This promotes large scale remote diagnosis by clinicians as well as the implementation of computer-aided diagnosis systems for local routine disease screening. However, lower-cost equipment generally results in inferior quality images. This may jeopardize the reliability of the acquired images and thus hinder the overall performance of the diagnostic tool. To solve this open challenge, we carried out an in-depth study on using different deep learning-based frameworks for improving retina image quality while maintaining the underlying morphological information for the diagnosis. Our results demonstrate that using a Cycle Generative Adversarial Network for unpaired image-to-image translation leads to successful transformations of retina images from a low- to a high-quality domain. The visual evidence of this improvement was quantitatively affirmed by the two proposed validation methods. The first used a retina image quality classifier to confirm a significant prediction label shift towards a quality enhance. On average, a 50% increase of images being classified as high-quality was verified. The second analysed the performance modifications of a diabetic retinopathy detection algorithm upon being trained with the quality-improved images. The latter led to strong evidence that the proposed solution satisfies the requirement of maintaining the images' original information for diagnosis, and that it assures a pathology-assessment more sensitive to the presence of pathological signs. These experimental results confirm the potential effectiveness of our solution in improving retina image quality for diagnosis. Along with the addressed contributions, we analysed how the construction of the data sets representing the low-quality domain impacts the quality translation efficiency. Our findings suggest that by tackling the problem more selectively, that is, constructing data sets more homogeneous in terms of their image defects, we can obtain more accentuated quality transformations.

Keywords: Generative Models, Diabetic Retinopathy, Retina Image Quality.

Resumo

A análise da imagem da retina é um fator essencial na deteção precoce de múltiplas doenças, tais como a retinopatia diabética. A deteção de sinais patológicos em imagens da retina pode ser efetuada tanto por oftalmologistas, como pela utilização de sistemas de diagnóstico assistido por computadores. Avanços tecnológicos conduziram ao uso de dispositivos mais portáteis e economicamente acessíveis para a aquisição de imagens médicas neste contexto. Assim, é promovido o diagnóstico remoto a larga escala por especialistas, bem como a implementação de sistemas de diagnóstico assistido por computadores para rastreios locais. No entanto, os equipamentos de menor custo resultam geralmente em imagens com qualidade inferior, o que compromete a eficácia das ferramentas de diagnóstico referidas. Para ultrapassar o problema identificado, foram desenvolvidos modelos de inteligência artificial para melhorar a qualidade da imagem da retina, mantendo a informação morfológica subjacente para fins de diagnóstico. Os resultados obtidos demonstram a eficácia do modelo *Cycle Generative Adversarial Network* para a transformação de imagens de um domínio de baixa para alta qualidade. As evidências visuais desta melhoria foram quantitativamente comprovadas pelos dois métodos de validação propostos. No primeiro, foi utilizado um classificador de qualidade de imagem da retina para confirmar um deslocamento significativo das classificações no sentido de uma melhoria da qualidade. Foi verificado um aumento de imagens qualificadas como tendo alta qualidade de 50%, em média. No segundo, analisaram-se as alterações de desempenho de um algoritmo de deteção de retinopatia diabética, após ter sido treinado com as imagens de qualidade melhorada. Este método conduziu a fortes evidências de que a solução proposta satisfaz o requisito de preservar a informação original das imagens para diagnóstico, e que torna o classificador de patologia mais sensível aos sinais patológicos presentes nas imagens. Assim, confirma-se a potencial eficácia da solução proposta. Foi também analisado o impacto da utilização de conjuntos mais homogéneos, em termos dos seus defeitos de qualidade, para representar o domínio de baixa qualidade. Os resultados obtidos sugerem que são as transformações de qualidade se tornam mais acentuadas.

Palavras-chave: Modelos Generativos, Retinopatia Diabética, Qualidade Imagem Retina.

Contents

List of Figures

List of Tables

Acronyms

AHE	Adaptive Histogram Equalization
AI	Artificial Intelligence
AR-CNN	Artifacts Reduction Convolutional Neural Network
BBHE	Brightness Preserving Bi-Histogram Equalization
CADx	Computer-aided Diagnosis
CAS-CNN	Compression Artifact Suppression Convolutional Neural Network
CBDNet	Convolutional Blind Denoising Network
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
CoGAN	Coupled Generative Adversarial Network
CT	Computerized Tomography
CycleGAN	Cycle Generative Adversarial Network
DCT	Discrete Cosine Transform
DL	Deep Learning
DnCNN	Denoising Convolutional Neural Network
DR	Diabetic Retinopathy
DSLR	Digital Single-Lens Reflex
EDSR	Enhanced Deep Super-Resolution Network
ESPCN	Efficient Sub-pixel Convolutional Neural Network
EyeQ	Eye-Quality
FD-UNet	Fully Dense U-Net
FID	Frechet Inception Distance
FSRCNN	Fast Super-Resolution Convolutional Neural Network

GAN	Generative Adversarial Network
GCBD	GAN-CNN Based Blind Denoiser
HE	Histogram Equalization
IS	Inception Score
JPEG	Joint Photographic Experts Group
KL	Kullback Leibler
LLNet	Low-light Network
MCF-Net	Multiple Color-space Fusion Network
MISR	Multi-Image Super-Resolution
ML	Machine Learning
MLP	Multilayer Perceptron
MRI	Magnetic Resonance imaging
PDEs	Partial Differential Equations
RBFNN	Radial Basis Function Neural Network
RDN	residual Dense Network
ReLU	Rectified Linear Unit
ResNet	Deep Residual Network
RNN	Recurrent Neural Network
SISR	Single Image Super-Resolution
SR	Super-resolution
SRCNN	Super Resolution Convolutional Neural Network
SRDenseNet	Residual Dense Network
SRGAN	Super Resolution Wasserstein Generative Adversarial Network
VAE	Variational Autoencoder
VDSR	Very Deep Super-Resolution
VDSR	Super Resolution Generative Adversarial Network

WGAN Wasserstein Generative Adversarial Network

Chapter 1

Introduction

In this Chapter, an overview of the context and motivation behind the research are given. Moreover, the current state-of-the-art of the research topic is addressed, as well as the objective of this work.

1.1 Contextualization

Artificial Intelligence (AI) has shown great potential in several fields of research throughout the last decade. Particularly, AI models applied to Computer Vision led to successful results which surpassed human capabilities in several applications [1]. Computer-aided Diagnosis (CADx) is one of these applications, which uses medical images such as MRI, CT scans, retina images, or even microscopic analysis, for an algorithm-assisted diagnosis [2].

Technological growth allowed implementing CADx methodologies in more portable and cheaper hardware for medical assistance, which can have a global impact. This is particularly relevant in regions where healthcare facilities with advanced imaging techniques are often inaccessible [3]. Moreover, widely available CADx systems can render decentralized healthcare a reality with accuracy and performance not seen thus far [4].

In ophthalmology, CADx methods have been a relevant tool for an automatic detection of different disorders in retina images, from which we highlight Diabetic Retinopathy (DR).

DR is the most common microvascular complication of diabetes, which causes damage to the retina capillaries [5]. The damaged blood vessels might swell and leak, or close, restricting the blood flow. Additionally, abnormal new blood vessels may grow in the retina. This condition frequently leads to poor visual functioning and represents the

leading cause of blindness in working-age populations [5]. The International Diabetes Federation and IDF Diabetes Atlas estimate that in 2019, 463 million adults (20 - 79 years) had diabetes worldwide [6]. According to the IAPB Vision Atlas, the global prevalence of DR among individuals with type I and II diabetes is 35% [7]. As the number of individuals with diabetes continues to grow, the number of DR occurrences increases proportionally, as illustrated in Figure 1.1.

There is growing evidence documenting the effectiveness of routine DR screening and early treatment. Clinicians can identify the disease through the presence of lesions associated with abnormalities caused by the disease. However, this requires local expertise and equipment which are often lacking in areas with high diabetes rates [8].

The use of CADx systems allied to portable and cheaper hardware is a powerful solution for a worldwide detection and tracking of the disease. Furthermore, it contributes to the overall expansion and diversity growth of retina image databases for further algorithm development and optimization.

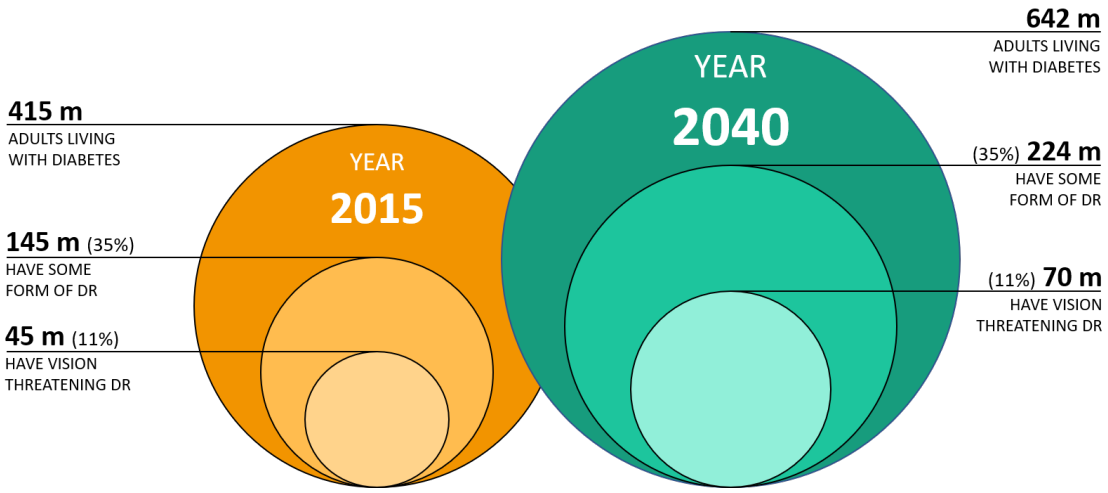


Figure 1.1: Global Prevalence of people with diabetes and Diabetic Retinopathy. Adapted from [7].

It is important to note that, aside from its impact on these algorithm-based diagnosis systems, the use of portable and low-priced imaging devices can promote healthcare decentralization by allowing large-scale remote diagnosis, even by non-experts. This further encourages telemedicine, which involves delivering healthcare and sharing medical knowledge at a distance.

1.2 Motivation

The use of portable and low-priced imaging devices for CADx systems constitutes a technological simplification with undeniable benefits. However, there is usually a trade-off between presenting low complexity and cost and assuring high accuracy and performance.

The use of lower-cost equipment generally results in inferior quality images which may hinder the overall performance of the diagnostic tool [9]. This consequence happens due to several reasons. Firstly, it becomes more difficult for experts to identify the presence and severity degree of the disease. This not only jeopardizes expert-based diagnosis but also constrains the labelling process and consequently the use of the acquired images for training disease-detection models. Secondly, even if labelling is done correctly, diagnosis models will not be as accurate because the pathology-related characteristics are shrouded by low-quality factors, such as image artifacts. Finally, if the models are trained using good-quality images only, the classification of the lower-quality examples is likely to be imprecise as the algorithms will focus not only on pathological features but also on the image quality discrepancy.

In summary, the lack of quality jeopardises the usefulness of the images for database construction and model development, as well as the accuracy of their classification for diagnosis. Thus, it compromises the overall goal of using mobile and economically inexpensive devices for automatic diagnosis: worldwide decentralized healthcare.

This project aims at mitigating the aforementioned open challenge by improving the quality of retina images, possibly acquired with portable and low-priced imaging devices. This research encompasses a solution to transform image quality while maintaining the underlying information for diagnostic purposes. The proposed solution is set to be accomplished through the development of Deep Learning (DL) methods based on generative models.

In the long-term, this solution is expected to improve the effectiveness of using more portable and cheaper hardware for CADx systems, particularly for the detection of retina pathologies such as DR. Furthermore, the implementation of such image enhancement methods will allow a more efficient image analysis by ophthalmologists, which can promote the usage of those images for remote medical assistance. This is especially relevant in areas with limited access to health facilities.

1.3 Literature Review

The aim of this dissertation is to develop a framework for image quality improvement, focusing on retina images.

Image quality improvement techniques have been applied to correct images so they can be more easily interpreted. Furthermore, enhancement methods have been used as a pre-processing step for making images more suitable for specific computer vision applications [10]. There are various methods for image quality improvement according to the type of image and application. Biomedical imaging uses these techniques to control factors such as image acquisition artifacts, fuzziness, non-homogeneity of luminance, and contrast levels [11].

Here we present the current state-of-the-art on image quality improvement techniques, addressing both conventional methods and DL approaches.

1.3.1 Conventional Methods for Image Quality Improvement

Several conventional image enhancement techniques for medical devices can be found in the literature. These include spatial-domain methods, which operate directly on pixels, and frequency-domain methods, based on the Fourier transform [12].

Homomorphic filtering has been applied in various studies to modify the illumination and reflectance components of an image. These methods involve nonlinear mapping to a frequency-domain where linear filters are used and have shown good results in overcoming non-uniform illumination [13]. Other long-established solutions from the frequency-domain include Retinex-based methods, which usually decompose an image into illumination and reflectance and then transform their information separately. These have been widely applied for enhancing non-uniform illumination images [14].

Histogram-based techniques have been extensively reported as successful in the spatial-domain. These models redistribute the grey levels within an image and are used to enhance contrast [15]. Commonly applied examples are the Adaptive Histogram Equalization (AHE) [16], the Brightness Preserving Bi-Histogram Equalization (BBHE) [17] and the Contrast Limited Adaptive Histogram Equalization (CLAHE) [18]. Despite their proven benefits, most HE techniques can cause a washed-out effect and might amplify existing noises [19]. Moreover, ambiguity can be added which increments uncertainty in the image information.

To overcome the existing solutions' limitations, fuzzy-domain approaches have been proposed which manage the imperfectness of an image modeled as its uncertainty [10]. These methods apply if-then logical rules to image processing and allow an automatic contrast enhancement. In 2012, Hasikin et al. [20] suggested a fuzzy grey-scale image enhancement technique by maximizing fuzzy measures contained in the image. Multiple combinations of HE-based and fuzzy techniques have since been proposed for various applications. In 2014, for example, Raju and his team [21] presented a novel fuzzy and histogram-based method to enhance low contrast colour images, a fast method when compared to other enhancement techniques. Nonlinear image enhancement has also proven to improve the quality of a fuzzy image as shown by Wang et al. in 2012 [22].

As previously mentioned, image acquisition artifacts may also hinder image quality. Reflection artifacts are common in photographs taken through glass windows, which is the case in some mobile device medical applications. Various studies have been carried out to solve this problem. These suggest different models for separating the transmission and reflection components of an image [23]–[25]. Researchers have also developed physics-based methods to correct dirty or partially occluded optics by estimating the attenuation and the scattering of the lens dirt and remove the artifacts [26].

1.3.2 Deep Learning Methods for Image Quality Improvement

Advances in Deep Learning introduced novel techniques in image quality improvement. One of the enhancement tasks to which neural networks have been applied is image

denoising. This process aims at recovering a clean image from a noisy observation which follows an image degradation model [27]. A common assumption is to consider the additive white noise as a Gaussian. Convolutional Neural Network (CNN) models have been employed to perform that separation. Both the Multilayer Perceptron (MLP) and auto-encoders have been successfully adopted to handle Gaussian noise removal [28], [29]. Wang et al. [30], Bae et al. [31] and Jifara et al. [32] also proposed residual learning into deeper CNN for image denoising.

CNNs have been explored to extract more useful information from noisy images with insufficient quality, which is a recurrent problem in several medical imaging fields. An example of this is the deep residual network proposed by Gholizadeh et al. [33] that encompasses dilated convolutions with enlarged receptive fields and reduced network depth for CT image denoising.

In 2017, Zhang et al. [34] proposed a novel Denoising Convolutional Neural Network (DnCNN) to extract the latent clean image in the hidden layers. The suggested model adopts the residual learning formulation by employing a single residual unit to predict the residual image. The use of batch normalization combined with the residual learning further improves training speed and stability, as well as the denoising performance. This DnCNN model proved to successfully handle Gaussian denoising with unknown noise level (i.e., blind Gaussian denoising). Other methods were proposed to tackle unknown noise, such as the usage of auto-encoder networks [35].

For handling unpaired noisy images, a model based on the Generative Adversarial Network (GAN) was suggested in 2018, which was denominated GAN-CNN Based Blind Denoiser (GCBD) [36]. This solution constitutes a novel two-step framework. First, a GAN architecture is trained to estimate the noise distribution over the noisy input images and to generate noise samples. The resulting noisy images are then utilized to construct a paired training data set. This paired data set is used to train a deep CNN for denoising. An alternative solution for real-world noisy image correction was the Convolutional Blind Denoising Network (CBDNet) proposed by Guo et al. [37]. This approach comprises two subnetworks, one for estimating the noise of the real noisy image, and the other for obtaining the latent clean image.

CNNs have also been reported as useful for enhancing low-light images and adaptively enhancing image contrast [38], [39]. The goal is to train convolutional neural network systems to take dark images as input and produce bright images as output, without disturbing the content of the image. In 2017, Lore et al. [40] proposed a stacked sparse denoising auto-encoder for simultaneous low-light enhancement and noise reduction (LLNet). This method has the disadvantage of not considering the nature of low-light pictures. To overcome this challenge, Wei et al. [41] introduced a deep Retinex-Net. The suggested method involves decomposing the input image into reflectance and illumination through a subnetwork Decom-Net. Then, an encoder-decoder based Enhance-Net corrects uneven illumination. During this step, the illumination is also adjusted from multi-scale perspectives through concatenation and the noise on the reflectance is removed. The

enhanced output image is obtained from the adjusted illumination and reflectance. A similar yet more recent and efficient solution was proposed by Wang et al. in 2019 [42], named Retinex Decomposition based GAN. The suggested framework can be divided into two subsections. The first consists of a Retinex Decomposition Network trained with a novel loss which is computed on the decomposed reflectance components of the enhanced and the reference images. The second Fusion Enhancement Network combines the input low-light image and its decomposed results to generate the final enhanced image.

Alternative deep generative structures for low-light enhancement encompass the usage of Variational Autoencoder (VAE) networks, such as proposed by Nazemi et al. [43] who used conditional VAE structures to model histogram vectors of different colour channels and parameters of image data.

Other low-light enhancement approaches might include different image decomposition methodologies. Particularly, a study carried out by Atoum et al. [44] suggested decomposing the input images into lightness and colour components using the CIE LAB space, where each component is enhanced independently. The output enhanced image corresponds to the fusion of the two previous results. This mechanism constitutes a simplification of the low-light image enhancement problem and is able to produce more realistic colours.

A recent study addressed the low-light image enhancement problem via a novel hybrid deep network. The proposed system contains a content stream to enhance the visibility of the low-light input and learn a holistic estimation of the scene content, and an edge stream devoted to edge detail learning through an improved spatially variant Recurrent Neural Network (RNN). This method was able to recover more structure details when compared to other low-light image enhancement algorithms [45].

Artifact removal is also an important factor in image quality improvement. Most state-of-the-art methods explore neural networks for compression artifact correction. These artifacts appear when lossy compression algorithms are applied and lead to detail loss or noise and small structures' addition. Multiple convolution networks have shown to be very successful at compression artifact suppression. These include the AR-CNN proposed by Dong et al. [46], the enlarged and deep CNN approach presented in [47] and the CAS-CNN with hierarchical skip connections and multi-scale loss function suggested by Cavigelli et al. [48].

Other artifact removal works relate to acquisition-induced defects that appear in photographs. Examples are the novel multi-resolution fully convolutional network trained for automatically removing moiré patterns from photos, presented by Sun et al. [49], and the cascade deep neural network introduced by Yang et al. [50] for reflection removal. The latter estimates the reflection and uses it to boost the estimation of the background image. Deep encoder-decoder networks have also proved to be successful in handling reflection removal [25]

Medical imaging-related artifacts tend to be more specific and thus require a particular assessment. The majority of these studies relate to CT artifact removal. The deep residual network structure to remove low-dose CT streak artifacts presented in [51], and the Radial Basis Function Neural Network (RBFNN) used to remove CT ring artifacts in [52] are examples thereof. Recently, Guan et al. [53] proposed using a modified CNN architecture termed Fully Dense U-Net (FD-UNet) for removing artifacts from two-dimensional photoacoustic tomography images.

Some studies reported neural networks to be useful for correcting out-of-focus images [54]. This topic has been particularly exploited in microscopy images as their lack of focus constitutes one of the biggest challenges in the field. U-Net architectures have been successfully employed for refocusing blurry out-of-focus microscopy images [55].

An additional application of CNNs is image scaling enhancement. While most scaling algorithms base their process solely on nearby pixels or interpolations, neural networks calculate how to preserve the image quality [56]. Super Resolution Convolutional Neural Network (SRCNN) algorithms are employed for that end. Since high perceptual quality images present more valuable details, SRCNNs have been widely used in many areas, such as medical imaging [57]. Super-resolution (SR) can be classified into Single Image Super-Resolution (SISR) and Multi-Image Super-Resolution (MISR), according to the number of input low-resolution images. SISR is much more popular because of its high-efficiency [57]. The SRCNN can be envisioned as an CNN that approximates the complex mapping between low/high-resolution spaces in an end-to-end manner.

In 2016, Dong et al. [58] showed that conventional sparse-coding based SR methods can be reformulated into a deep CNN. Their method starts with a bicubic interpolation as the pre-processing step. The mapping process conceptually consists of three nonlinear operations. These are patch extraction, nonlinear mapping, and image reconstruction. The loss function for optimizing the SRCNN is the mean squared error. Other relatively shallow Convolutional Neural Networks for SR include the FSRCNN [59] and the ESPCN [60].

Deeper SR networks methods have since been explored to improve accuracy, such as the VDSR inspired by VGG-net used for ImageNet classification [61]. Alternatively solutions also include GAN-based models (SRGAN). These became widely used for this application as they allow creating more perceptually satisfying and less blurry images. SR generative models use sets of high-resolution downscaled images and produce an output of SR images. The discriminator is used to distinguish original high-resolution images from SR ones. Both the generator and the discriminator are trained through back-propagation [62]. The SRGAN proposed by Ledig et al. [63] is a successful example of this methodology. This generative model employs the Deep Residual Network (ResNet) with skip-connection and uses a perceptual loss using high-level feature maps of the VGG network.

Another deep network alternative is the Enhanced Deep Super-Resolution Network (EDSR) proposed by Lim et al. [64]. They also relied on the ResNet structure for their model but removed unnecessary modules from the conventional ResNet architecture. Thus, the EDSR algorithm achieved improved results while being more compact.

While using deeper networks proved to bring benefits to representation power, these did not support optimal use of the feature information from shallow layers, which are usually associated with low-level features [65]. To overcome this challenge, several methods introducing concatenation operations between shallow layers and deep layers, as well as skipping connections, were proposed. The SRDenseNet (which presents skipped connections) [66], and the RDN (which exploits all relevant convolutional features in local and global manners) [67], are examples of this.

To access more general and realistic settings where low-/high-quality pairs of images are unavailable, image-to-image translation models have recently been explored for image quality improvement processes. Namely, for super-resolution, Yuan et al. [68] proposed a cycle-in-cycle network structure to tackle the problem within three steps. They first map the input low-resolution images to the clean and bicubic-downsampled low-resolution space using a CycleGAN. Another well-trained deep model with bicubic-downsampling assumption then up-samples the intermediate result to the desired size. The last step consists of fine-tuning the two modules in an end-to-end manner to get the high-resolution output.

A study carried out by Ignatov et al. [69] presented an end-to-end DL approach based on learning a mapping function between photos from mobile devices and a DSLR camera. This methodology uses a composite perceptual error function that combines content, colour, and texture losses. This contributed to prove the efficiency of image-to-image translation methods in image enhancement but did not fulfill the unpaired image sets condition, that is, there was still a strong supervision requirement.

Later in 2018, Ignatov et al. [70] suggested another approach using a weakly supervised photo enhancer for digital cameras. This was based on GANs and loss functions designed for accurate image quality assessment. Their model was trained to map low-quality photos into the domain of high-quality photos without requiring any correspondence, that is, using only two independent image sets representing each domain.

Chen et al. [71] suggested a two-way GAN method for image enhancement by learning from photographs. They employed an enhanced version of the U-Net as a generator by augmenting global features. Additionally, they improved the Wasserstein Generative Adversarial Network (WGAN) stability by an adaptive weighting scheme.

Recently, an adaptive weighted multi-discriminator CycleGAN was introduced by Park et al. [72] for transforming turbid underwater images into clean ones. This model was also trained with unpaired image sets.

It is important to note that, besides its recent usage in image quality improvement, generative models and image-to-image translation methods have been extensively used for multiple research ends. The work carried out by Welander et al. in [73] represents

such an example. It encompasses a comparative study of using the UNIT framework and the CycleGAN for image-to-image translation of T1- and T2-weighted MRI images.

1.3.3 Retina Image Quality Improvement

The quality factor in retina images has become a topic of great importance as it strongly influences diagnosis efficiency both by experts and AI algorithms. Several studies have been carried out for developing neural networks capable of automatically assessing the quality of retina images and thus inspect which are suitable for medical diagnosis. Examples are the EyeQual model developed by Costa et al. [74] and the MCF-Net proposed by Fu et al. [75].

Moreover, multiple enhancement techniques have been developed to meet fundus images' quality requirements for diagnostic purposes.

Retina vessels can show different states of several diseases (e.g. glaucoma, hypertension, and Diabetic Retinopathy), making the detection of vessels in retina images crucial. A spatially adaptive contrast enhancement technique for enhancing retina fundus images for blood vessel segmentation was proposed by Bandara et al. in 2017 [76]. Another study by Miri et al. [77] introduced an algorithm for retina image contrast enhancement based on the second generation of a new multi-resolution analysis tool called Curvelet Transform. This multi-scale directional transform allows an almost optimal non-adaptive sparse representation of objects with edges.

To reduce the effect of noise in colour retina images due to the acquisition process, conventional CLAHE methods have been applied and have shown positive results [78], [79].

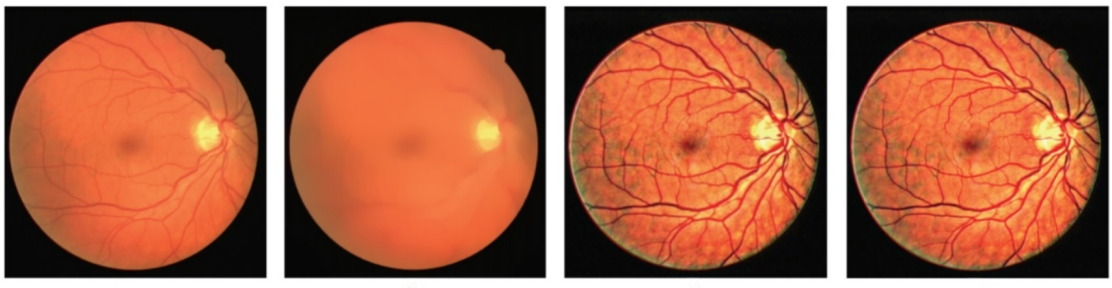


Figure 1.2: An enhancement process of retina images. (a) Original image; (b) Normalized convolution; (c) Fusion result; (d) Denoising result. Adapted from [80].

In 2016, Dai et al. [80] proposed a novel method also suitable for noise reduction in colour retina images. It used a normalized convolution algorithm with a domain transform to obtain an image containing the basic information of the background. That

image was then fused with the original one and the enhanced result went through a denoising process using the fourth-order partial differential equations and the relaxed median filter. Figure 1.2 displays an example of retina images throughout the described process.

Some studies addressing AI-based methods have been suggested. In 2016, Soomoro et al. [81] proposed a Radial Basis Function Neural Network (RBFNN) to manage the noise level of retina fundus images. The retina image colour map was converted to the green band, as it ensured better contrast, and then fed to the RBFNN filtering method to determine the denoised image. A novel stack deep convolutional autoencoder was introduced by Ghosh et al. [82] for retina image denoising. The proposed scheme has proven to restore structural details of the fundus as well as to decrease the noise level. However, there is still little research regarding the usage of DL methods for retina image enhancement, particularly for image generation with unpaired image sets.

1.3.4 Summary

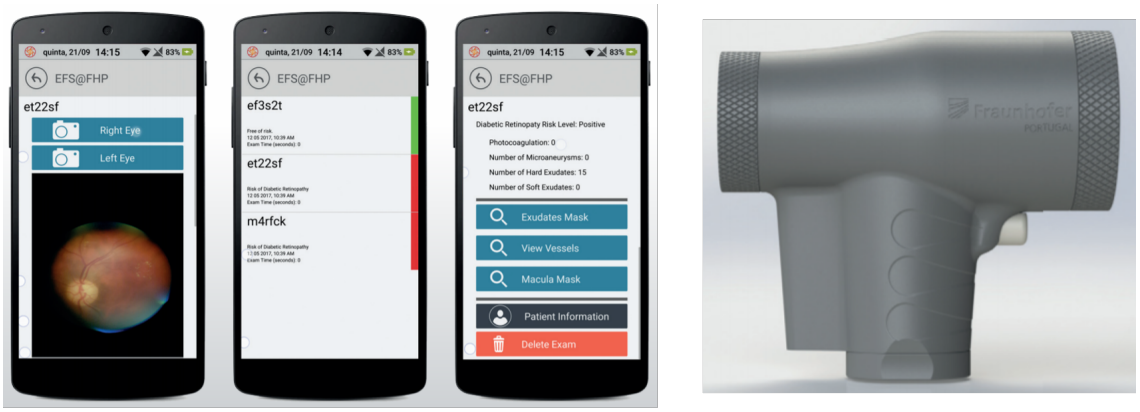
The use of more portable and cheaper hardware has gained vast interest in a medical assistance context as it allows remote expert-based diagnosis and large-scale disease screening by CADx systems. These processes depend significantly on the quality of the images. In this Chapter, we presented multiple strategies for image quality improvement based on both conventional pre-processing methods and DL algorithms.

Regarding conventional image enhancement, it is important to state that some gaps can be found in earlier research. Particularly, various enhancement frequency-domain techniques can lead to certain artifacts in the output image. Furthermore, some pixels may get lost during conversion in frequency-domain methods. Edge degradation can occur as a result of several enhancement techniques modifying them during processing. Finally, since various methods concentrate on regions within the image, illuminance imbalance may also arise [15].

Aside from the referred challenges, conventional methods have proven to be suboptimal when compared to neural network-based approaches. AI methods have supported positive results in handling different image quality problems. Recently, generative models and image-to-image translation algorithms have been reported as successful tools for many objectives, including image quality enhancement and defect correction. Additionally, several of these methods allow training unpaired images, which is crucial given the difficulty of obtaining paired medical images. However, to the best of our knowledge, these models have not yet been explored for medical image quality improvement for diagnosis purposes, particularly in regard to retina images, which is the focus of this dissertation.

1.4 Objective

The long-term goal of this project is to employ Deep Learning algorithms for enhancing the quality of retina images without low-/high-quality pairs of images (unpaired learning). This quality enhancement is set to improve the efficiency of Diabetic Retinopathy diagnosis, both by ophthalmologists and CADx systems. More precisely, this project aims to fulfill the image enhancement necessities of using portable and low-priced imaging devices for remote diagnosis and CADx applications. To this end, several steps were defined, which include the handling and processing of the image data sets, DL-based algorithm development and its optimization, as well as a final validation of the model's performance in improving retina image quality.



(a) (b)

Figure 1.3: EyeFundusScope system. (a) Android application; (b) Custom optical prototype for acquiring retina images. Adapted from [83].

In the scope of the research organization where the work is carried out, Fraunhofer AICOS, the developed algorithm is intended to be later adapted to the EyeFundusScope project. This consists of a self-contained mobile-based system capable of detecting early signs of sight-threatening DR on retina images [83]. Figure 1.3 illustrates the prototype for the referred system.

The projected algorithm would be implemented for improving the acquired fundus images and therefore, maximize the diagnostic precision of the presented CADx system.

1.5 Document Structure

The document structure is divided into five chapters as illustrated in Figure 1.4. The present chapter corresponds to the introduction, in which the study's contextualization and motivation, the literature review, and the project objectives are given. Chapter 2 regards the theoretical background required for the development of this study. Chapter 3 contains a detailed description of the implemented frameworks for retina image quality improvement. Chapter 4 addresses the main results achieved by using the proposed

frameworks. Lastly, Chapter 5 comprises the conclusions of this study and suggests future research directions.

Figure 1.4: Document structure overview.

Chapter 2

Theoretical Background

Throughout this chapter, relevant generic concepts related to the proposed dissertation are presented.

2.1 Artificial Intelligence and Deep Learning

Artificial Intelligence (AI) is an interdisciplinary field of computer science concerned with building machines capable of performing tasks that require human intelligence [84]. A considerable amount of AI algorithms are based on Machine Learning (ML), which focuses on developing systems that can access data and learn from it. The goal is to allow computer programs to learn from experience without being explicitly programmed, so they can perform tasks in an autonomous way.

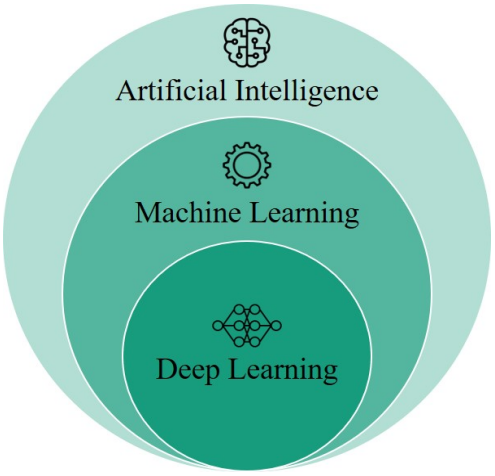


Figure 2.1: Interdependence of the AI, ML and DL concepts.

The learning process is qualified as supervised if the training data is labelled with the corresponding class. Upon training, the classifier should be capable of predicting new outputs. On the other hand, unsupervised methods aim to learn the inherent structure of the given data without using provided labels. Training with only a portion of labelled data (usually more unlabelled data than labelled) is classified as semi-supervised learning. Lastly, there is reinforcement learning, where the classifier learns by interacting with a dynamic environment and uses the gathered

experience (through actions and corresponding rewards or punishments) to optimise its performance (e.g. winning a game) [85].

Deep Learning (DL) is a subset of ML that uses deep artificial neural networks to carry out the process of machine learning. These neural networks are built using neuron nodes connected together, similarly to the networks present in the human brain. The working mechanism of these structures is presented in Section 2.3. This methodology was first theorized in the 1980s but has only recently become ubiquitous. The reason relies on its need for large amounts of data and substantial computing power. [86]. Figure 2.1 presents a scheme interconnecting the concepts of AI, ML, and DL.

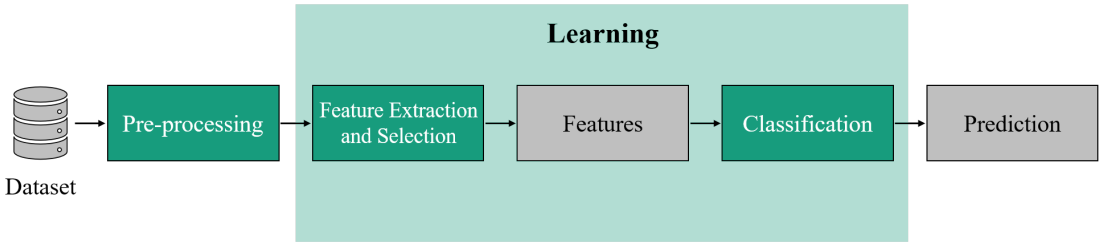


Figure 2.2: Machine learning classification pipeline. The grey objects correspond to the inputs/outputs of the processes coloured in green.

A schematic representation of a ML classification pipeline is presented in Figure 2.2 to demonstrate all steps included in the process. As can be seen, the pipeline starts with the data set. The data has to be collected and cleaned (e.g. deleting or correcting incorrect/incomplete samples). In supervised learning, data is additionally labelled according to the ground-truth.

The data is then pre-processed. This step aims at uniformizing data so that the difference between samples is mainly related to their corresponding class instead of other arbitrary characteristics. Additionally, several pre-processing methods can be employed to optimize the performance of the model. This requires performing tests under different circumstances and detecting which pre-processing techniques assure a more optimal classification performance [55]. Image processing methodologies are presented in Section 2.2.

Feature extraction and selection is then performed. This process corresponds to a dimensional reduction by which the set of data is reduced to more manageable and relevant characteristics. The selected features must be identifying for the classification task [87]. For example, in a traffic sign image classification problem, relevant features might include edge shape, colour, or area. This selection process also requires testing for reaching an optimal set of relevant features.

DL methods differ from more conventional ML algorithms in this stage, as most of them do not require feature selection prior to learning. That step is performed by the neural network itself and is a dynamic process in which the features can be successively adapted in parallel with learning [88].

The following step consists of applying the ML classifier, which is trained in a trial-and-error manner to make predictions with considerable accuracy. In DL, this algorithm corresponds to a neural network.

Upon training, the model receives the pre-processed input, extracts the previously selected features, and predicts the output automatically. The performance of the algorithm can be later validated by using a test data set that has not been previously presented to the model [89]. This is accomplished through the comparison of the predicted results from the unknown data to the ground-truth, using a loss function. Several performance parameters allow a quantitative analysis of said comparison such as Accuracy, Precision, Recall, and the F1 Score[90]. The Accuracy allows detecting the fraction of correct predictions within all samples ($\frac{TP+TN}{TP+TN+FP+FN}$ [1]). Precision gives us the proportion of positive identifications that is actually correct, for each class ($\frac{TP}{TP+FP}$ [1]). Recall permits knowing how well the model can detect a class within all ground-truth positive elements ($\frac{TP}{TP+FN}$ [1]). The F1 Score conveys, for each class, the balance between the precision and the recall ($2 \cdot \frac{precision \cdot recall}{precision+recall} = \frac{TP}{TP+0.5 \cdot (FP+FN)}$ [1]) [91].

Neural networks commonly add an intermediate step to this process. The overall data set is divided into train, test, and validation. The algorithm is validated each epoch with unknown data from the validation set. This allows monitoring the progress of the model's performance and thus detect anomalous circumstances such as overfitting. This complication happens when the model achieves a favourable fit on the training data, but it does not generalize well on new, unseen data [92]. The test set is then utilized for a final evaluation of the model's performance.

It is important to note that this pipeline exemplification is based on a classification example. However, particularly in DL, several models aim at generating data. That concept is relevant for this research project and thus it is addressed in Section 2.4.

2.2 Image Processing

Image processing is a relevant topic to consider for this work due to two reasons. First, image quality improvement or image enhancement can be accomplished through different processing mechanisms. Thus, despite this work focusing on Deep Learning approaches, it is important to contextualize the concept of quality enhancement in image processing techniques. Secondly, image pre-processing is a crucial step in preparing image data for DL models. In this Section, both circumstances are addressed.

[1] TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives

2.2.1 Generic Image Processing Concepts

There are several digital processing techniques to manipulate images. To be processed digitally, a two-dimensional image $f(x, y)$ is first sampled and transformed into a matrix of numbers. The processing algorithms operate on the elements of that matrix. Simple processing methods include matrix calculus for image rotation, shearing, and scaling [93]. More complex digital image processing methods are generally divided into enhancement, restoration, analysis, and compression [94].

Image enhancement is applied to improve the interpretability of information in images for subsequent analysis or image display. Various enhancement methods involve convolutions between kernels (filters) and an image, which replace pixels with a linear combination of its neighbours. The implementation of these filters can lead to sharpening, smoothing, and edge enhancement, for example. This type of processing also includes histogram manipulation, median filtering, image averaging, and subtraction, as well as frequency-domain filtering such as the Butterworth filter [95].

In image restoration, the goal is to recover an image from its degraded version. It relies on a statistical or mathematical description of the degradation so that it can be reverted. The objective is to obtain an approximation of the original image through that degradation function [94]. One of the most applied methods is the Wiener filter which includes both the degradation function and statistical characteristics of noise into the restoration process. Through inverse filtering, it estimates the uncorrupted image by minimizing the mean squared error [96].

Image analysis focuses on the automatic extraction of information from an image for its characterization. Several features can be extracted. These include spatial features such as edges, central moments, or entropy, as well as frequency-domain features that can be obtained through the Fourier transform. Image segmentation is frequently applied in this context for separating the images into different regions based on their properties [94].

Image compression aims to reduce the number of bits required for representing an image, maintaining the required quality. Image compression techniques can be classified as lossless if the original image can be fully retrieved, and lossy when some levels of error are allowed for obtaining larger compression rates. Transform coding lossy methods are widely used. These alter pixel specifications from spatial-domain into frequency-domain and concentrate energy in very few transform coefficients. That is, only significant coefficients are used for quantization and encoding. Discrete Cosine Transform (DCT) is the most commonly applied among these techniques [97]. An example of a DCT application is JPEG baseline compression.

2.2.2 Pre-processing Methods for AI Applications

Before feeding image data into machine learning models, it needs to be pre-processed. There are several reasons for conducting this step.

First, the acquired data is usually disorganized and derives from different sources: the acquisition equipment, capturing environment, operator collecting the image data, etc., may vary considerably. To account for all conditions in which an image is taken, there needs to be a form of normalizing the data. Employing pre-processing techniques for that end allows the model to focus more on relevant characteristics for a certain project goal than on arbitrary features. Additionally, depending on the model, some pre-processing methods may have to be necessarily applied. For example, most Convolutional Neural Networks require the input images to have a unified dimension, i.e., a uniform aspect ratio [98].

Applied methods for uniformizing an image data set include pixel scaling techniques such as pixel normalization (scale pixel values to the range 0-1), pixel centering (scale pixel values to have a zero mean), and pixel standardization (scale pixel values to have a zero mean and unit variance) [99]. An additional commonly employed method is image resizing which involves increasing or decreasing the total number of pixels in an image. This allows to obtain images with a uniform aspect ratio and is usually accomplished through interpolation algorithms such as the nearest neighbor, bilinear, or bicubic. The working mechanism of these interpolation techniques is based on estimating values at unknown points using known pixels [100]. Trimming images is also frequently applied for segmenting the area of interest and minimizing irrelevant sections. For example, in models receiving retina image data as input, images may be trimmed for obtaining a square surrounding only the retina area (and thus minimizing black sections) or the optic disc area. Therefore, even if the original images capture different area proportions around the retina, the pre-processed data will be more uniform and centered in the same anatomical structures.

An additional reason relates to optimizing the performance of the models. Applying pre-processing methods can help reduce the complexity of the problem and increase the accuracy of the applied algorithm. This can include simple image modifications such as converting colour images to grayscale to reduce computation complexity or applying algorithms for adjusting image brightness or contrast (e.g. adaptive histogram equalization). Testing the performance of the model under the usage of different pre-processing techniques allows detecting optimal conditions [98].

For some situations, more complex approaches are required. The performance of AI algorithms depends on the quantity and quality of the training data. Image quality can be affected by multiple factors including real-world noise, blur, or other quality degradation factors [101]. To overcome this problem, both conventional and DL-based image quality improvement techniques, such as the addressed in Section 1.3, may be employed.

Regarding the quantity of the training data, the performance is jeopardised when there are few and/or undiversified samples (no representativity of real-world heterogeneous settings). A pre-processing strategy that enables a significant increase in the volume and diversity of data available for training models is data augmentation. This

involves augmenting the existing data set with perturbed versions of the existing images. Image augmentation methods may include geometric transformations, colour space augmentations, and kernel filters application [98].

2.3 Neural Networks

Neural networks consist of several artificial neurons organized in layers. Each neuron is directly connected to every other neuron in the adjacent layers. The inner layers are often referred to as hidden layers. Figure 2.3 shows a representation of a common three-layered neural network. The information flows in two ways: forward- and back-propagation.

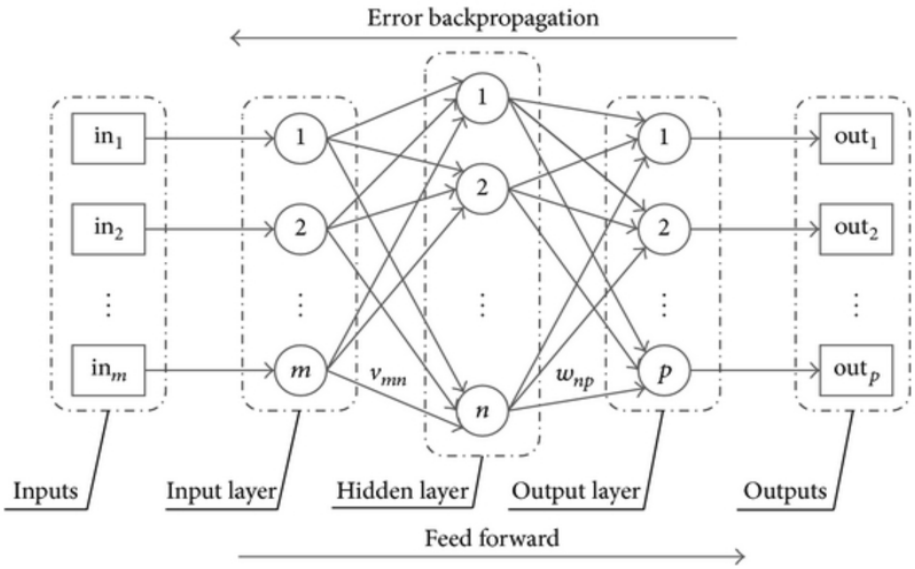


Figure 2.3: Schematic representation of a neural network [102].

When performing inference, the information enters through the input units which trigger the first hidden neurons and so on until the output neurons are reached. This design is called a feed-forward network. During the forward propagation, each neuron receives inputs from the units in the previous layer. These inputs are multiplied by the weights of the connections they travel along and are added a bias term. Throughout the process, a function is applied to each neuron to determine its level of activation, which introduces non-linear properties to the network. Rectified Linear Unit (ReLU) is the most commonly used activation function in this context [103]. The output of each neuron is given by $f(\sum_i w_i x_i + b)$ where f is the activation function which receives both the weighted sum of inputs, $\sum_i w_i x_i$, and the bias, b.

During back-propagation, there is an element of feedback that allows learning. This involves a loss function that determines the discrepancy between the algorithm's output and the given target value (in a supervised setting). Amid learning, the gradients of the loss function are calculated and the existing parameters (weights) updated accordingly.

The cycle (epoch) is repeated until reaching the minima of the loss function. By minimizing the loss, the model's accuracy is maximized. This learning process can be described by the simple equation: $W(t+1) = W(t) - \eta \nabla J(w)$ where W is the weight of the neuron, $\nabla J(w)$ the gradient of loss function- $J(\theta)$ in respect to the updated weight, and η the learning rate [104].

Higher η values lead to more abrupt weight changes which accelerate the learning process. However, the model may overshoot the minimum loss and thus fail to converge. On the other hand, lower learning rates may allow the model to learn a more optimal set of weights but may cause training to take significantly longer or get stuck on local minima [105].

In the scope of this thesis, it is also important to explore the concept of a Convolutional Neural Network (CNN). Figure 2.4 displays a generic scheme of a CNN. This DL architecture adds a set of layers to the conventional neural network architecture (fully connected network) that automatically extract relevant features from the input images. Throughout the learning process, these features are selected and adjusted dynamically.

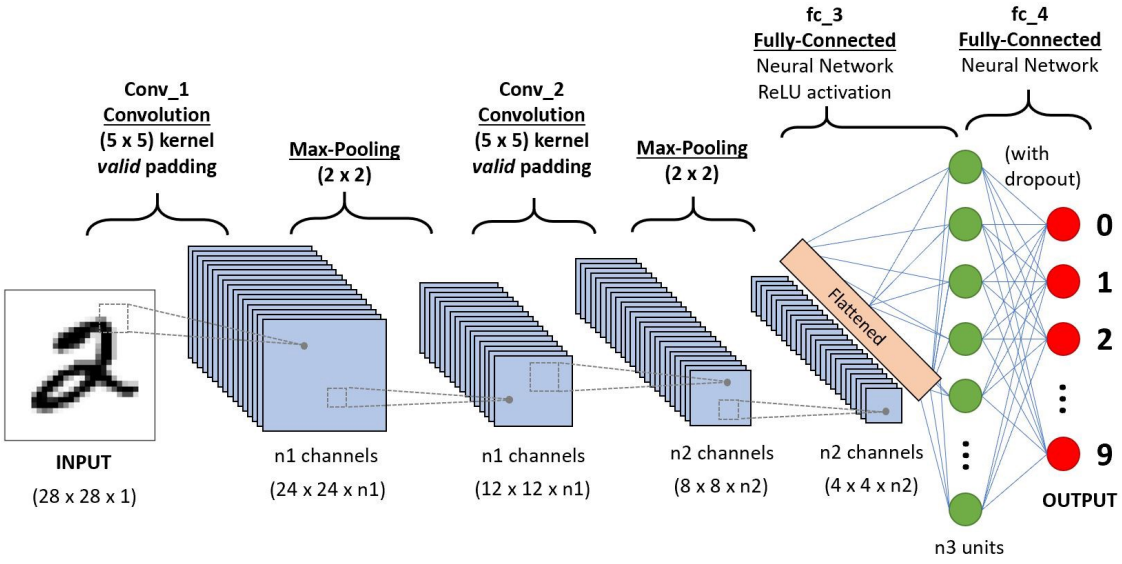


Figure 2.4: Schematic representation of a convolutional neural network [106].

The process of extracting features is done through convolutions. Each layer convolves its input with multiple filters learned by the network, usually referred to as kernels. By applying successive convolutions, it is possible to construct feature maps that characterize the image and are thus used for classification. The kernel's coefficients depend on the features to highlight. During training (back-propagation), the coefficients of the filters are updated according to the output error. Lower layer filters typically learn lower-level features such as lines and corners, whereas higher layers learn increasingly higher-level features, namely textures and whole shapes [107].

In addition to the convolution layers, pooling layers are often applied. These are

responsible for reducing the spatial size of the convolved features. Pooling techniques commonly include Max Pooling and Average Pooling. Max Pooling returns the maximum value from the portion covered by the kernel, in each patch of the feature map. On the other hand, Average Pooling returns the average of all values from the covered portion in each patch [106].

Before connecting the feature map to the final classification model, which takes place in the fully connected layers, there is a flattening process. This consists of converting the data into a 1-dimensional array. Consequently, the classification input is a single long feature vector [106].

2.4 Deep Learning in Image Generation

In addition to the conventional classification methods used to differentiate data instances, Deep Learning can also be applied for image generation. Generative models are a form of unsupervised learning which focuses on discovering the characteristics and patterns in input data to generate examples that could have plausibly been drawn from the original data set [105]. These models can be employed, for example, in image-to-image translation and reconstruction processes.

For this work, it is important to define the concepts of Variational Autoencoders, Generative Adversarial Networks, and UNIT.

2.4.1 Autoencoder Networks

An autoencoder neural network is an unsupervised learning algorithm that through back-propagation learns to generate representations ($\hat{x}$) that are as similar to the original inputs (x) as possible. This is accomplished by efficiently compressing and encoding data in a smaller hidden encoding layer, also known as the latent space.

The compression of the input data into a latent vector is carried out by the encoder while the reconstruction of the reduced encoded representation is obtained through the decoder and it is set to equal the input. The model's efficiency relies on its capacity to learn a meaningful and generalizable latent space representation, that is, to extract only the most relevant features to describe the natural data [108]. Figure 2.5 shows a generic scheme of an autoencoder network.

Regular autoencoders have the disadvantage of their latent spaces, where their encoded vectors lie, having no defined structure. This happens because the autoencoder is solely trained to encode and decode with as little loss as possible, independently of how the latent space is organised. Thus, interpolation is not as easy and overfitting is more likely to occur [109]. On the other hand, Variational Autoencoder (VAE) architectures have continuous latent spaces. This regularisation of the latent space allows easy random sampling and interpolation.

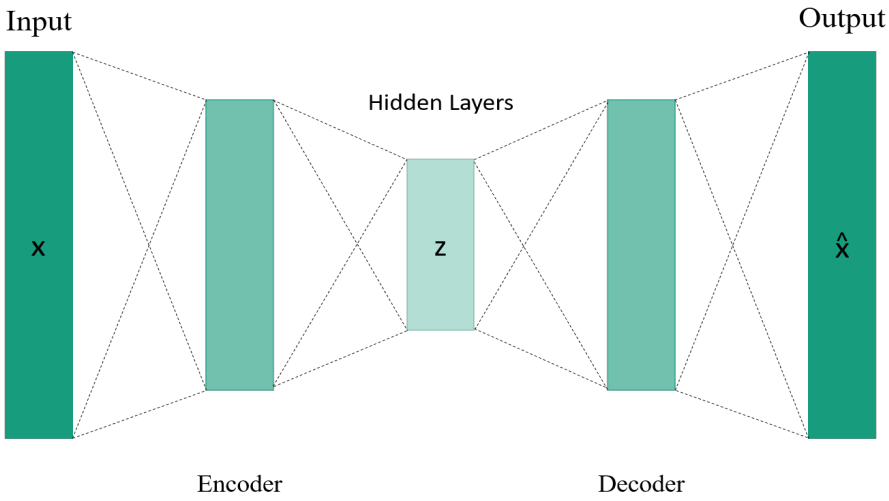


Figure 2.5: Representation of an autoencoder neural network.

VAE models assume that the source data has underlying probability distributions and attempt to reach the parameters of those distributions in the encoding process [110]. The encoded distributions are set to be normal so that the encoder can be trained to return the mean and covariance matrix that describe these Gaussians. The latent representation z is then sampled from the prior distribution and lastly, the decoder generates an output based on the latent representation.

The loss function for these networks is composed of a reconstruction term (final layer), to minimize the difference between the input and output, and a regularisation term (latent layer), to regularise the organisation of the latent space by approximating the encoded distributions to a standard normal distribution. The latter is done using the Kullback Leibler (KL) divergence metric [109].

Figure 2.6 presents a schematic representation of the steps composing both regular autoencoder and a variational autoencoder.

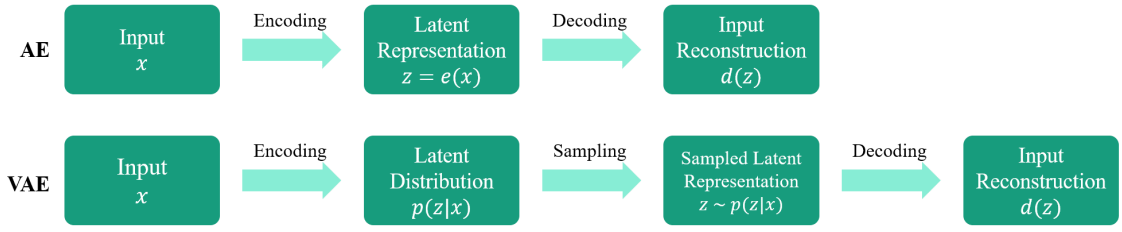


Figure 2.6: Difference between autoencoder and variational autoencoder.

2.4.2 Generative Adversarial Networks

The Generative Adversarial Network (GAN) is an unsupervised Deep Learning algorithm for training generative models through a controlled learning process.

GAN algorithms combine two neural networks: the generator, which produces new data instances, and the discriminator, that evaluates them for authenticity. The objective is to manipulate the generator into producing plausible examples [111]. Figure 2.7 shows a scheme of this mechanism. As illustrated, the generator receives random numbers from a known distribution as input and returns a synthetic image. The discriminator takes the output generated image and returns a probability of authenticity. The generator then learns to create images that "fool"the discriminator and the process is repeated, thus creating a feedback loop [112]. The discriminator is trained separately and its loss function measures the divergence between the classification outputs and the ground-truth labels (original and generated images are labelled as two binary opposites). The loss of the generator is calculated from the discriminator's classification. It decreases if the generator successfully deceives the discriminator, and gets penalized otherwise.

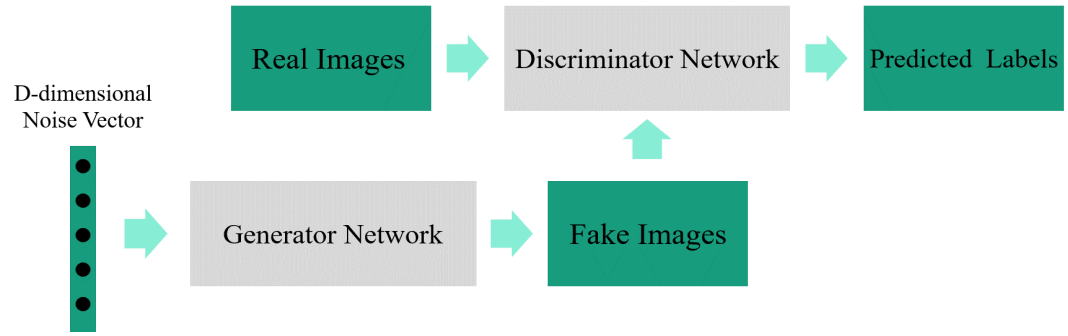


Figure 2.7: Schematic representation of a GAN model.

There are a number of different types of GANs, each presenting slightly different alterations. The Cycle Generative Adversarial Network (CycleGAN) is an extension of the GAN which allows the automatic training of image-to-image translation models without paired examples. That is, not only do they synthesize images resembling real images from the domain of interest, but they also allow transforming the domain of an image while preserving its original core characteristics. This is accomplished by adding a new generator to the conventional GAN architecture. The first generator translates the original image into the target domain. The result is then fed to the second generator which translates it back into the original domain. The output of this reconstruction should match the original image. This mechanism provides cycle consistency which is a concept of reversibility in machine translation processes. By introducing an additional loss term

for measuring the difference between the output of the second generator and the original image, consistency is assured. Consequently, the original base features are maintained [113]. An exemplification scheme of this process is displayed in Figure 2.8, where zebras are converted into horses.

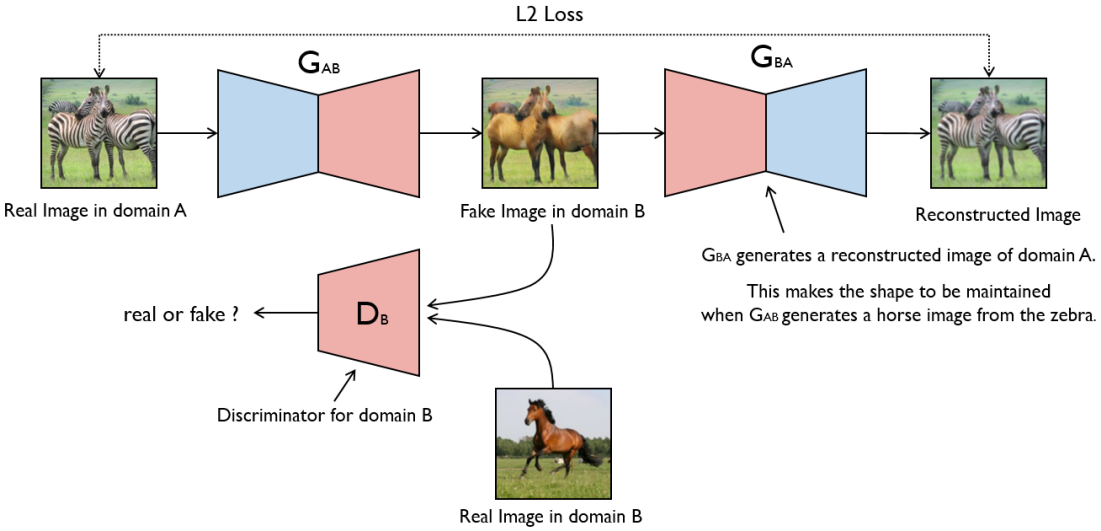


Figure 2.8: Schematic illustration of a CycleGAN model [113].

Another relevant concept based on GANs is the Coupled Generative Adversarial Network (CoGAN). As illustrated in Figure 2.9, the CoGAN consists of a tuple of GANs, each for synthesizing images resembling real images of a domain. By using shared layers, a weight-sharing constraint is imposed. This allows the CoGAN to learn a joint distribution without the existence of corresponding images in the two distinct domains. Consequently, the CoGAN can generate pairs of images that individually resemble the images in the respective domains without having paired image examples [114]. Relating to the previous example, if using two data sets to train a CoGAN, one containing images of horses and the other of zebras, it would be possible to generate synthetic pairs of corresponding images that differed only in terms of the presented animal.

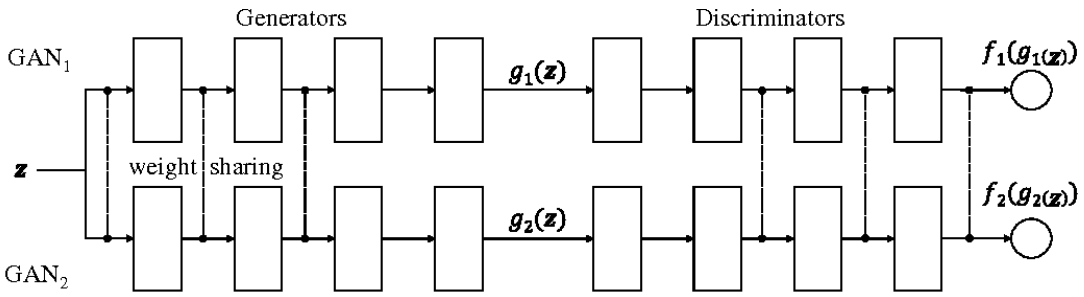


Figure 2.9: Schematic illustration of a CoGAN model [114].

2.4.3 UNIT

The term UNIT initially referred to the general concept of UNsupervised Image-to-image Translation, which includes multiple models such as the CycleGAN presented in Section 2.4.2. However, it is now highly associated with a particular framework proposed by Liu et. al [115], which is relevant to define in this context.

The UNIT framework is based on both GANs and VAEs. Each image domain is modelled through a VAE-GAN, which consists of an encoder and a decoder for the reconstruction process, and a GAN discriminator for evaluating the output's authenticity. Essentially, the VAE decoder and the GAN generator are collapsed into one. This mechanism improves the VAE reconstruction capacity by using the feature representations learned by the GAN discriminator. Consequently, element-wise errors are added to the feature-wise errors, which can improve the data distribution while offering invariance towards translation [116].

The model is trained with a weight-sharing constraint for joint distribution learning, similarly to the CoGAN architecture. This enforces a shared-latent space between the two VAE-GANs to generate corresponding images in the two domains. Moreover, the shared-latent space constraint implies the cycle-consistency constraint, which has been reported to be a successful method for image-to-image translation mechanisms.

This cycle-consistency constraint allows images in the source domain to be mapped into the target domain and the translated images in the target domain to be mapped back to the original images in the source domain, similar to the CycleGAN [115].

Upon training, the UNIT framework allows reconstruction processes if using a VAE-GAN modelling on just one of the domains, and translation processes if a domain shift is performed in the shared-latent space. The latter is accomplished by using the encoder from one domain (origin domain) and the decoder from the other domain (target domain).

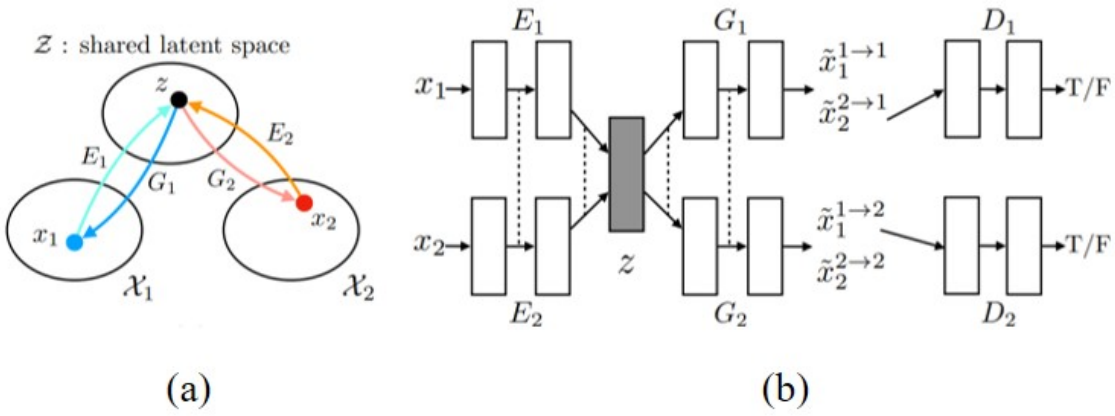


Figure 2.10: Representation of the UNIT framework. (a) Schematic illustration of the shared-latent space assumption; (b) UNIT framework.[115].

In Figure 2.10, a schematic representation of the UNIT framework is presented, where $\tilde{x}_1^{1\to1}$ and $\tilde{x}_2^{2\to2}$ correspond to the self-reconstructed images, and $\tilde{x}_1^{1\to2}$ and $\tilde{x}_2^{2\to1}$ correspond to the domain-translated images (domain shift in the shared-latent space). D_1 and D_2 are the adversarial discriminators for each respective domain. These evaluate how realistic are the translated images.

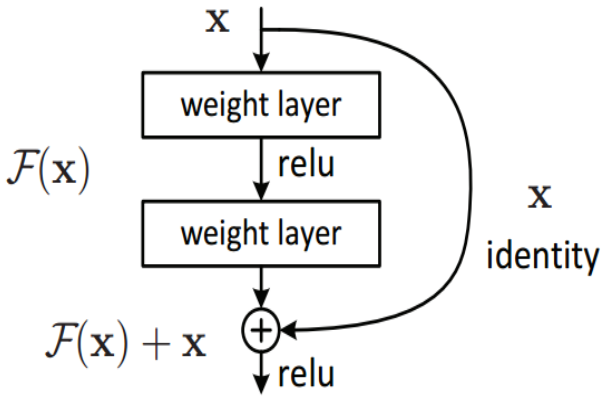


Figure 2.11: A building block of residual learning [117].

It is important to note that the UNIT architecture uses residual blocks. These were introduced by He et al. [117] and correspond to skip-connection blocks that learn residual functions regarding the layer inputs, instead of learning unreferenced functions. Figure 2.11 schematically illustrates a basic residual block.

2.5 Computer-Aided Diagnosis

The concept of Computer-aided Diagnosis was first theorized to support digital image interpretation on radiologic diagnosis [118]. Early attempts aimed at utilizing this computerized analysis of radiological images to detect abnormalities. However, these experiments were unsuccessful because computers were not sufficiently powerful, there was a lack of advanced image-processing techniques, and the access to digital images was limited [2].

Extensive and systematic research and development of several CADx systems were later restarted in the early 1980s [2]. Multiple CADx systems have since been developed to assist in the detection and diagnosis of lesions in medical images including computed tomography, ultrasound, MRI, and endoscopy. These methods have shown to be a particularly advantageous tool in early detecting multiple types of cancer through the differentiation between benign and malignant tissues. However, recent developments of CADx systems led to their usage across various other biomedical imaging fields and applications, showing relevant results in supporting diagnosis. These include detecting brain diseases such as Alzheimer in MRI images [119], vascular disorders including Diabetic Retinopathy in retina images [120], and malaria parasite detection based on microscopic blood images [121].

CADx systems have also been allied to portable and low-priced imaging devices to allow large-scale screening of certain diseases. Examples are the use of mobile devices for malignant melanoma detection [122], [123] and retinal abnormality recognition [3]. This methodology is particularly relevant in regions with limited access to healthcare facilities

with advanced imaging technology. Furthermore, this strongly encourages telemedicine, which involves delivering healthcare and sharing medical knowledge at a distance.

Technology growth and AI developments have had a great impact on CADx systems. Currently, these algorithms generally consist of several steps that may include image processing, image feature analysis, and data classification through tools such as artificial neural networks [124]. In Figure 2.12 we displays a generic scheme of a CADx system.

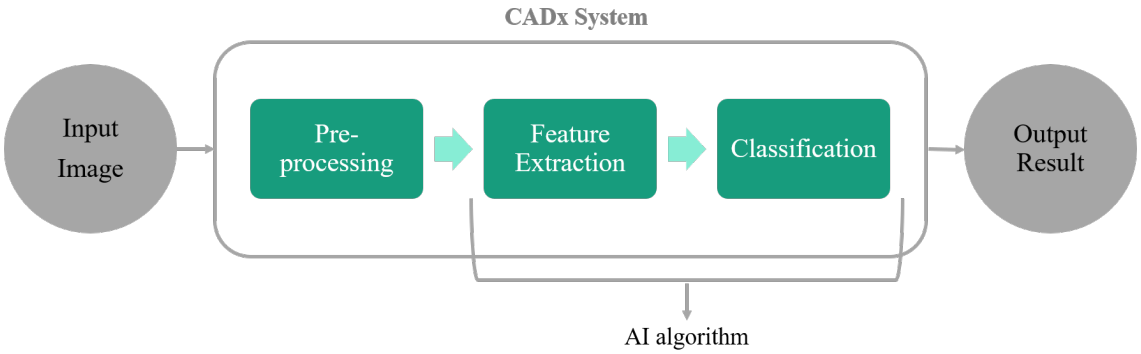


Figure 2.12: Pipeline schematic representation of a generic CADx system.

As represented, the process begins with the image data being pre-processed. This step might be employed because the model's learning process was based on images with particular characteristics (e.g. size or colour map), or because some pre-processing methods enhance the model's performance (e.g. contrast or brightness modifications). Image processing methods commonly include image enhancement to improve the interpretability of images for their analysis, and region segmentation to isolate regions of suspected abnormality [125]. A summary of commonly employed pre-processing methodologies is presented in Section 2.2.2.

The processed images are then fed into a previously trained AI algorithm. By extracting relevant pathology-related features, the model is able to provide a diagnostic prediction regarding the disease(s) of interest. As previously explained, if the model consists of a conventional ML algorithm, the considered features are previously selected for optimal classification. On the other hand, most DL algorithms have that selection and consequent extraction embedded in their network.

As aforementioned, the performance of AI algorithms for image analysis highly depends on the quantity and quality of the training data. CADx systems are required to have very high-performance levels for their output to be trustworthy in real-world medical assistance settings. This further highlights the importance of using high-quality images. Using large/diverse data sets also promotes more robust training, as well as using augmentation methods for increasing the variability of available data.

CHAPTER 3

Retina Image Quality Improvement Framework

This Chapter presents the proposed framework for retina image quality improvement. The methodology, along with the relevant processes for the implementation, is described.

The developed retina image quality enhancement system is a sequential algorithm designed to transform a baseline image with low-quality into a quality-improved image in which pathology-related characteristics are more easily identified. The low-quality image is set to simulate images acquired with a portable and inexpensive device. The image quality translation is obtained through Deep Learning methods.

The retina image quality improvement system implementation pipeline is depicted in Figure 3.1.

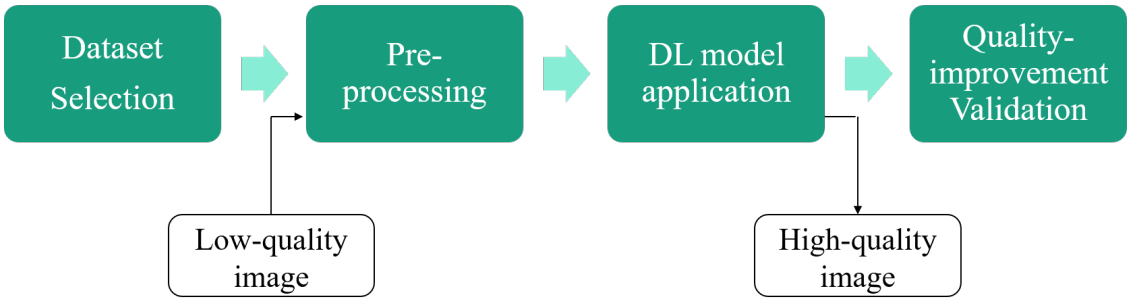


Figure 3.1: Retina image quality improvement system pipeline.

Data selection is the first stage of the process as the DL models require image data to learn how to perform the quality transformation (Section 3.1). The second step relates to the employment of pre-processing methods for preparing the selected data (Section 3.2). Three DL mechanisms for the image quality improvement process are considered for the proposed framework (Section 3.3). To validate the quality-improvement of the retina images, a last stage is employed in which two validation methods are applied. One is focused on the quality of the images themselves while the second focuses on the impact of the quality transformation on a CADx system (Section 3.4).

3.1 Data Sets

The first step to consider is the selection of retina image data sets to train and validate the algorithms. It is important to note that DL algorithms require considerably large amounts of data for training. Furthermore, we aim to transform the quality of retina images as well as to assess the resulting quality improvement and potential impact on AI-based disease-detection systems. Consequently, it is important to use a data set with images labelled according to both their quality and the severity degree of the retina-related pathology. Given these requirements, both the EyePACS and Eye-Quality data sets were selected, which are introduced in the following subsections.

3.1.1 EyePACS Data Set Overview

The first considered data set is the EyePACS, which contains over 5 million retina images of diverse populations with various degrees of DR severity. For each subject, the left and right field are provided. This large set of images was taken under a variety of imaging conditions and through different types of fundus cameras. Given the different acquisition conditions, images may contain artifacts, be out of focus, underexposed, or overexposed. This heterogeneity allows simulating the diversity that exists in real-world settings, not only in terms of retina anatomies but also regarding different image acquisition-related conditions [126].

The public set of this database is available on Kaggle, after an algorithm competition was proposed for training AI models capable of detecting signs of DR in retina images [8]. This set was used in our experiments and it is composed of 88,703 images rated by a clinician regarding the presence of the disease. The scale goes from 0 to 4, according to the following levels:

- Grade 0: No DR signs (Fig. 3.2 (a));
- Grade 1: Mild DR (Fig. 3.2 (b));
- Grade 2: Moderate DR (Fig. 3.2 (c));
- Grade 3: Severe DR (Fig. 3.2 (d));

- Grade 4: Proliferative DR (Fig. 3.2 (e)).

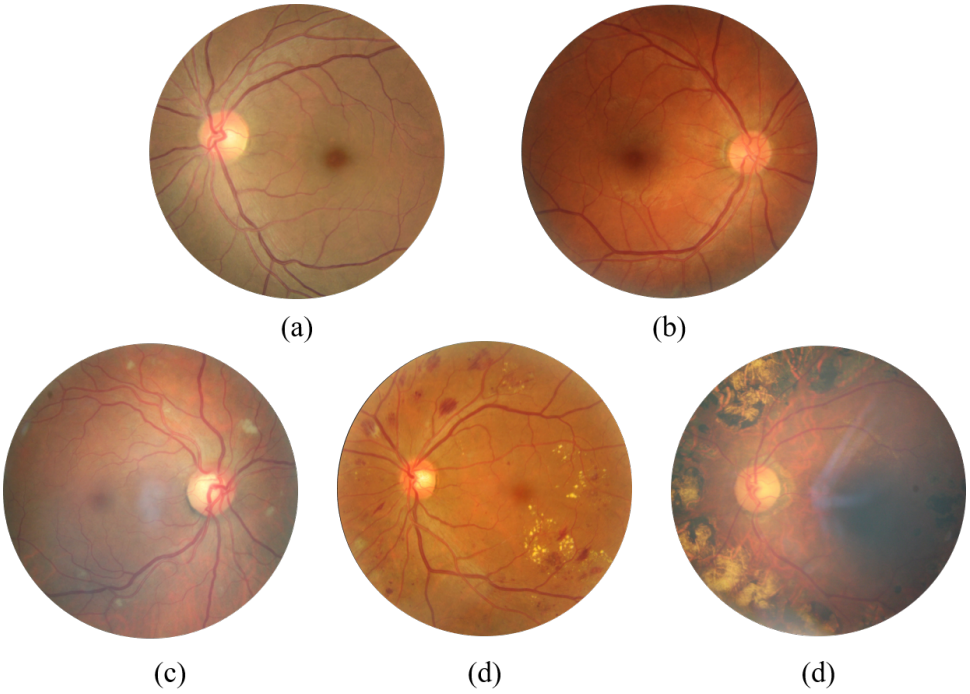
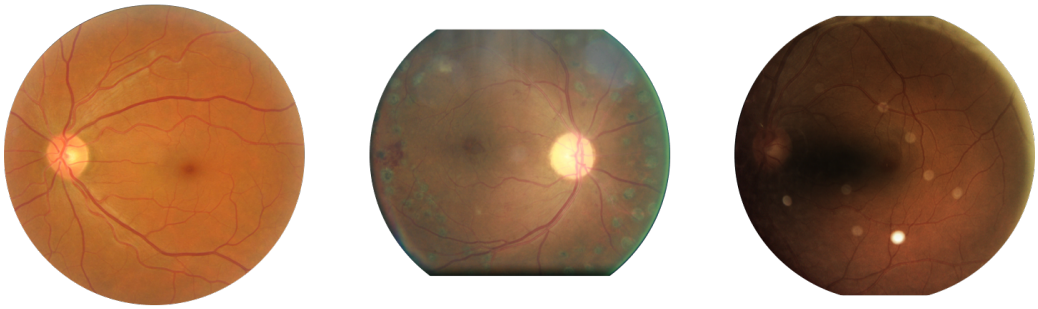


Figure 3.2: EyePACS data set retina image exemplification, with (a) no DR; (b) mild DR; (c) moderate DR; (d) severe DR; and (e) proliferative DR.

3.1.2 Eye-Quality Data Set Overview

For image quality assessment, the Eye-Quality (EyeQ) data set has been selected. It consists of a re-annotation subset from the EyePACS data set, yielding a collection of 28,792 retina images. All images were evaluated by two experts regarding their quality level for medical diagnosis reliability. The grading system considers four common quality indicators, including blurring, uneven illumination, low-contrast, and artifacts. However, the lower-quality images are not separated regarding their low-quality factors. The three proposed quality grades are defined as [75]:

- 'Good': the retina image presents no low-quality factors, and all diabetic retinopathy-related features are clearly visible (Fig. 3.3 (a));
- 'Usable': some low-quality indicators are visible, yet the lesions are clear enough to be identified by ophthalmologists (Fig. 3.3 (b));
- 'Reject': the retina image has serious quality issues, impeding a full and reliable diagnosis, even by ophthalmologists (Fig. 3.3 (c)).



(a) (b) (c)

Figure 3.3: EyeQ data set retina image exemplification, with (a) good quality; (b) usable quality; and (c) rejectable quality.

3.1.3 Low-quality Data Set Separation

The quality heterogeneity within the presented data sets allows simulating both the images acquired with mobile devices and the reference ones, acquired with advanced imaging equipment. For training the quality-translation algorithms, two retina image sets were required: a low- and a high-quality one. Initially, solely the quality level division proposed by Huazhu et al. [75] was considered. The images labelled as 'Usable' were used for the low-quality set and the ones labelled as 'Good' were considered for the high-quality set. 'Reject'-labelled images were initially discarded as their defects are much more challenging to correct, some even being irrecoverable since their morphological structures are not visible. However, after some experiments which are presented in Section 4.1.3, a further low-quality image selection was performed to construct more homogeneous data splits. This approach was accomplished by isolating each low-quality factor. Multiple data sets integrating different image defects were constructed and the models were trained separately to correct the defects contained in each data set.

To construct the low-quality sets, we first looked into the eyePACS and EyeQ retina image data sets to inspected which quality defects were more recurrent. For this selection process, we asked Fraunhofer researchers working in the EyeFundusScope project for advisory.

The most commonly observed defects were grouped into four low-quality data sets according to those which overlap the most and that are more visually similar. By grouping defects that are similar among themselves, more homogeneous low-quality data sets are obtained. Consequently, the models' learning process of correcting images with different low-quality factors is more efficient. The proposed data set division is illustrated in Figure 3.4.

The first low-quality data set contains 2,328 images presenting light low-contrast due to a hazing effect (Fig. 3.4 (a)), and with uneven illumination as a consequence of overexposure (Fig. 3.4 (b)). The second low-quality data set is composed of 2,781 images with overall dark low-contrast (Fig. 3.4 (c)), excessive darkness of the macula (Fig. 3.4

(d)), and uneven illumination due to underexposure (Fig. 3.4 (e)). The third low-quality data set contains 2,032 images presenting various types of flash-induced artifacts (Fig. 3.4 (f-g)).

Besides the acquisition defect-related image occurrences, it was also detected that in several images, the choroid vessels in the backside of the retina were clearly visible. This occurs due to a higher light penetration during the capturing of the image and it can vary with different fundus cameras and their specific wavelengths. When the backside vessels appear subtly, the phenomenon is not problematic. However, several cases in which the effect was extreme were identified and these were often associated with hyper-pigmentation, as shown in Fig. 3.4 (h). This can jeopardize the detection of lesions not only by ophthalmologists but also by AI-based systems that will be deflected from the relevant image features for diagnosis. Thus, we decided to include this characteristic as a low-quality factor and constructed a fourth data set containing 1,035 retina images.

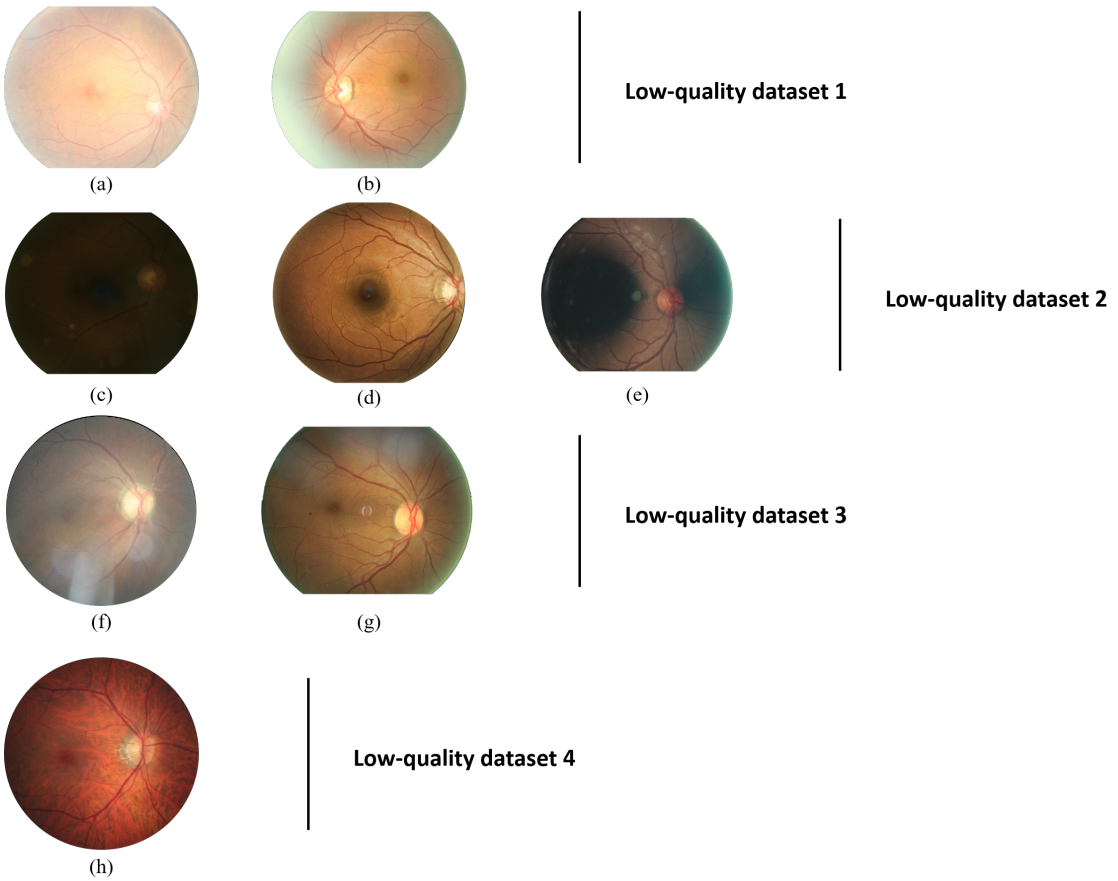


Figure 3.4: Retina Image low-quality factors exemplification and data set division, illustrating (a) light low-contrast; (b) illumination due to overexposure; (c) dark low-contrast (d) excessive darkness of the macula; (e) uneven illumination due to underexposure; (f-g) flash-induced artifacts; and (h) extreme light penetration and hyper-pigmentation.

The high-quality set is composed of 1,515 images labelled as having 'Good' quality from the EyeQ database. However, images containing signs of the above mentioned low-quality factors were discarded to ensure the algorithm did not associate the considered defects to the high-quality features.

3.2 Image Pre-processing

Before employing a DL algorithm, image data needs to be pre-processed. This guarantees uniformity across the data set and allows enhancing relevant image characteristics to optimize the performance of the models. This phase was divided into three steps which are presented and explained below.

3.2.1 Uniform Aspect Ratio

One of the first steps is to ensure that the images have the same aspect ratio, which corresponds to the ratio of their width to their height. Conventionally, neural networks assume a square shape input image and thus, an algorithm was developed to guarantee all images have a 1:1 aspect ratio. The employed algorithm checks if each image is squared. If not, it crops the image by two lateral limits corresponding to the beginning and end of the retina. These limits are established through a pixel value threshold applied to the images after converted to grayscale. The height is then adjusted to equal the width, either by clipping the surplus or by adding black pixels on the top and bottom of the image. This method ensures the cropping is centred in the middle of the retina and that no section of it is removed. Furthermore, it minimizes the black areas and consequently, its influence on the learning process. In Figure 3.5, two examples of this process are presented.

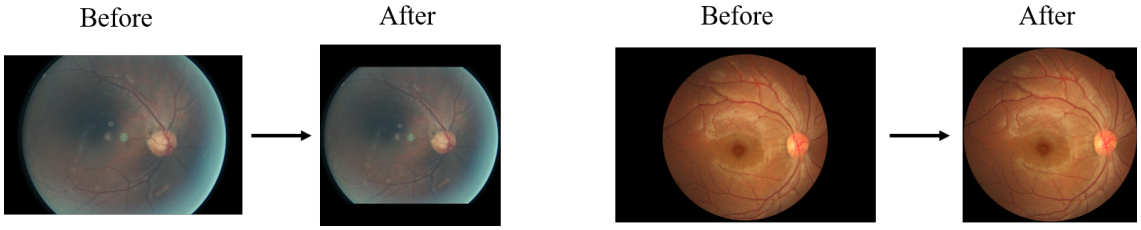


Figure 3.5: Uniform aspect ratio algorithm application examples.

3.2.2 Image Normalization

Image pixel values are often read as unsigned integers in the range between 0 and 255. Therefore, the requirement of all input parameters (pixels) having the same value distribution is already satisfied and the images could be directly fed into neural network models in their raw format. However, this value range is still considerably large which can potentiate exploding gradients and minimize the ability of the model to converge. To optimize the learning process and avoid these problems, pixel values were normalized so

that each pixel value ranges between -1 and 1. This can be achieved by simply dividing all pixel values by the largest pixel value, which is 255, then multiplying them by 2, and subtracting 1.

3.2.3 Image Resizing

Neural networks often receive inputs of the same size to guarantee uniformity across the data set and allow easier image data grouping in batches. This requires resizing the images to a fixed size before feeding them to the model. Additionally, the size of the original images coming from the acquisition device, such as a fundus camera, is always considerably larger than what is computationally accepted to use as input in a neural network. This occurs both because larger images occupy more space in memory and because these images result in more complex neural networks. Consequently, image downsampling should be done. Intuitively, this process implies performance degradation due to loss of information. Therefore, the fixed size for images constitutes a trade-off between computational efficiency and model performance.

We have established a fixed size of 224×224 because the translation methods applied in this work (presented in Section 2.4) are quite computationally heavy and using much larger images would require a demanding GPU usage. Simultaneously, some early tests were performed which allowed detecting promising quality translation performances when using a 224×224 image size with the proposed CycleGAN architecture.

For the downsampling of the images, bicubic interpolations were employed. This method uses the weighted average of the closest 4x4 neighborhood of known pixels to produce the interpolated pixel. The closer pixels are given a higher weighting in the calculation. This allows producing the best approximation of a pixel's colour and intensity based on the values of surrounding pixels [100].

3.3 Image Quality Improvement Methods

The main stage of this research corresponds to the employment of DL methods for transforming the quality of the retina images while maintaining their underlying information for diagnosis. In this Section, we present the three proposed mechanisms for translating the images from low- to high-quality.

3.3.1 Variational Autoencoder Framework

The first proposed method is based on a Variational Autoencoder. The working mechanism of these architectures is explained in Section 2.4.1.

VAEs can reconstruct their input by compressing the characteristic information in the latent dimension. Our goal is to train a VAE to encode and decode successive high-quality images so that the probability distributions composed by the latent variables, are associated with high-quality features. Hence, when inputting low-quality images, the

prediction of latent representations (inference process) would conform to high-quality distributions. As a consequence, this mechanism would allow a translation mechanism in which the underlying information of the original image is maintained while high-quality characteristics are emphasized.

Figure 3.6 illustrates the training mechanism of the model. As presented, the encoder receives as input high-quality image data that is compressed into a lower-dimensional space. This is done through latent variables that compose several Gaussian probability densities. The decoder receives as input that latent representation and generates reconstructions of the high-quality set.

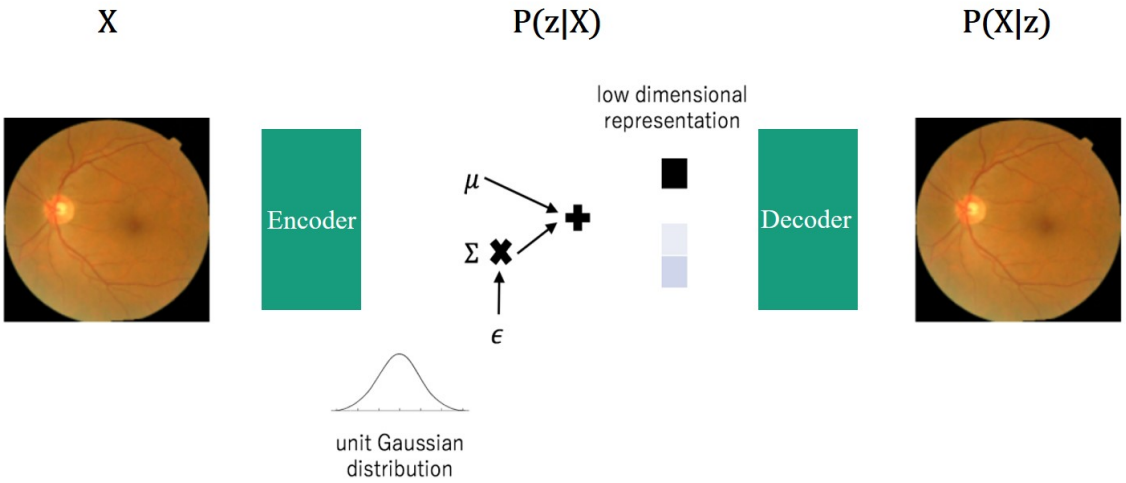


Figure 3.6: Variational Autoencoder training illustration.

To achieve this workflow, the encoder architecture is composed of successive convolution layers which lead to the parameters that characterize the latent dimension. Inversely, the decoder contains successive deconvolution layers that lead to the construction of image outputs. In Table 3.1, the architecture of the employed encoder and decoder composing the VAE model are presented in detail.

For data augmentation, vertical and horizontal flipping, and rotation in a range of 0-20 degrees were applied.

3.3.1.1 Loss Functions

The VAE loss function is a weighted sum of the Kullback Leibler (KL) loss and the reconstruction loss ($Loss_{total} = w_k Loss_k + w_r Loss_r$). The KL loss minimizes the divergence between the approximate posterior distribution P(z|X) and the true latent distribution P(z), which is a normal Gaussian, N(0,1). This loss function is given by:

$$\sum_{i=1}^{n} \sigma_i^2 + \mu_i^2 + \log(\sigma^2) - 1$$

Table 3.1: Variational Autoencoder architecture. The following abbreviations are used for ease of presentation: F=Filters, K=Kernel Size, S=Stride, and TS=Target Shape. The transposed convolutional layer is denoted by DCONV2D.

Layer	Encoder
1	CONV2D-(F64, K5, S1), BatchNormalization, LeakyReLU-(α 0.2)
2	CONV2D-(F128, K5, S2), BatchNormalization, LeakyReLU-(α 0.2)
3	CONV2D-(F256, K5, S2), BatchNormalization, LeakyReLU-(α 0.2)
4	CONV2D-(F512, K5, S2), BatchNormalization, LeakyReLU-(α 0.2)
5	CONV2D-(F1024, K5, S2), BatchNormalization, LeakyReLU-(α 0.2)
6	Flatten, Dense-(F2048)
μ	Dense-(F latent_m)
$\log(\sigma)$	Dense-(F latent_m)
Layer	**Decoder**
9	Reshape-(TS 7,7,1024)
10	DCONV2D-(F512, K5, S2), BatchNormalization,LeakyReLU-(α0.2)
11	DCONV2D-(F512, K5, S2), BatchNormalization,LeakyReLU-(α0.2)
12	DCONV2D-(F256, K5, S2), BatchNormalization,LeakyReLU-(α0.2)
13	DCONV2D-(F128, K5, S2), BatchNormalization,LeakyReLU-(α0.2)
14	DCONV2D-(F64, K5, S2), BatchNormalization,LeakyReLU-(α0.2)
15	DCONV2D-(F3, K5, S1), Activation-(tanh)

On the other hand, the reconstruction loss minimizes the difference between the input data X and the encoded-decoded data P(X|z). This enforces the encoder to generate a meaningful latent vector z.

Two reconstruction loss methods were considered for the VAE implementation. The first is the conventional mean squared error function. For this loss function, the weight losses are set to w_k=0.01 and w_r=1.

The second approach is a perceptual loss function based on a study carried out by Justin et al. [127]. This method utilizes a pre-trained network which outputs high-level feature maps, as can be seen in Figure 3.7.

The feature reconstruction loss is the Euclidean distance between each feature representation of the original and the reconstructed images. Our pre-trained model consists of a simple CNN that classifies the quality of retina images from the EyeQ data set. This allows ensuring that the pre-trained model has learned to encode the perceptual and semantic information associated with retina images that are relevant for our loss functions. The total loss function for this approach becomes

$$Loss_{total} = w_k Loss_k + \sum_{n=1}^{4} w_{r_i} Loss_{r_i}$$

where $Loss_{r_i}$ are the Euclidean distance loss functions for each feature map. The weight losses are set to w_k=1 and w_{r_i}=0.5.

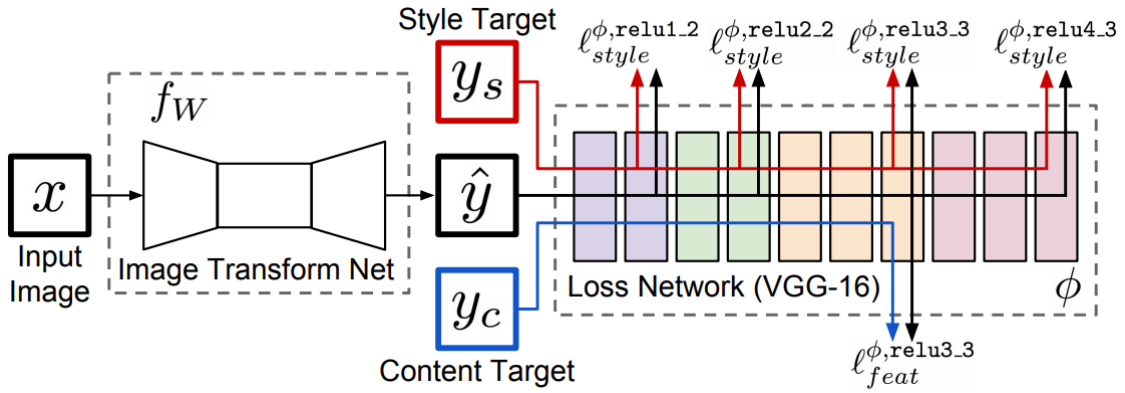


Figure 3.7: System overview of a loss network pre-trained for image classification to define perceptual loss functions that measure perceptual differences between images [127].

In Table 3.2, the architecture of the proposed pre-trained model used for the perceptual loss is presented.

Table 3.2: Loss Network architecture for the perceptual loss approach. The following abbreviations are used for ease of presentation: F=Filters, K=Kernel Size, S=Stride, R=Rate, ACT=Activation, and KR=Kernel Regulation.

Layer	Loss Pre-trained Network
1	CONV2D-(F128, K5, S1), LeakyReLU-(α~0.2), Dropout-(R0.25)
2	CONV2D-(F256, K5, S2), BatchNormalization, LeakyReLU-(α~0.2), Dropout-(R0.25)
3	CONV2D-(F512, K5, S2), BatchNormalization, LeakyReLU-(α~0.2), Dropout-(R0.25)
4	CONV2D-(F102, K5, S2), BatchNormalization, LeakyReLU-(α~0.2), Dropout-(R0.25)
5	CONV2D-(F1024, K5, S2), BatchNormalization, LeakyReLU-(α~0.2) , Dropout-(R0.25)
6	GlobalAveragePooling2D, Dense-(F3, ACT 'softmax', KR l1_2)

For optimization of both the VAE and the perceptual loss network, we employed the Adam algorithm with a learning rate of 0.0001.

3.3.2 CycleGAN Framework

The second proposed method is based on a CycleGAN architecture [128], whose working mechanism is explained in Section 2.4.2. As previously mentioned, this technique allows training unsupervised image translation models based on GAN architectures using unpaired collections of images from two different domains. In the scope of this thesis, the image domains are high- and low-quality, which will be further addressed as domain 'A' and 'B', respectively.

Figure 3.8 illustrates the training mechanism of the model. As presented, each CycleGAN architecture is composed of two generators and one discriminator. The generators carry out the translation process into the opposite domain while the discriminator evaluates the authenticity of the generated images, that is, how likely are the translated images

to belong to the target domain. The model is trained according to the presented scheme, using two CycleGANs in reversed directions.

While the generators are shared by the two architectures, the discriminators are distinct for each one, as they evaluate different domain authenticities. Hence, the system is composed of four neural networks.

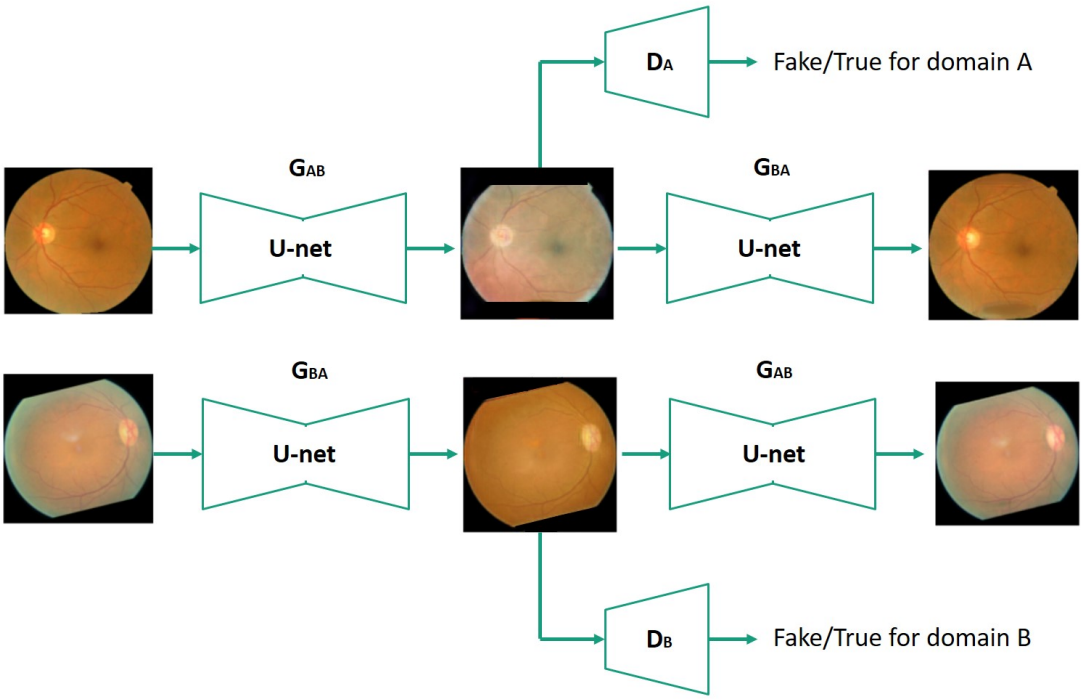


Figure 3.8: CycleGAN training illustration in which domain A and B correspond to high- and low-quality respectively.

For each CycleGAN, the process begins with the image set of one domain. These images are translated to the target domain and are then evaluated for their authenticity. Finally, the generated images are translated back to the original domain.

By employing two CycleGANs with inverse directions, the G_{BA} learns to generate not only high-quality images whose authenticity is evaluated against the ground-truth but also high-quality images that are forced to match their original form (before the translation process). Thus, this method maximizes the learning of high-quality features.

Each generator (G_{BA} and G_{AB}) is a U-Net, which consists of a CNN containing a contraction section with successive convolutions and an expansive section with successive deconvolutions. In the expansive path, each upsized image is concatenated with the corresponding image from the contracting path. This allows the network to combine both the positional and the contextual information [129]. In Table 3.3, the architecture of the employed generators is presented in detail.

Table 3.3: CycleGAN Generator Network architecture for the unsupervised image translation experiments. The following abbreviations are used for ease of presentation: F=Filters, K=Kernel Size, S=Stride, SI=Size, and ACT=Activation.

Layer	U-Net Generator	Path
1	CONV2D-(F128,K4,S2), LeakyReLU-(α 0.2), InstanceNormalization	
2	CONV2D-(F256,K4,S2), LeakyReLU-(α 0.2), InstanceNormalization	
3	CONV2D-(F512,K4,S2), LeakyReLU-(α 0.2), InstanceNormalization	Contracting Path
4	CONV2D-(F1024,K4,S2), LeakyReLU-(α 0.2), InstanceNormalization	
5	CONV2D-(F2048,K4,S2), LeakyReLU-(α 0.2), InstanceNormalization	
6	UpSampling2D-(SI2), CONV2D-(F1024,K4,S2, ACT 'LeakyRelu'), InstanceNormalization	
7	Concatenate-(out6, out4)	
8	UpSampling2D-(SI2), CONV2D-(F512,K4,S2, ACT 'LeakyRelu'), InstanceNormalization	
9	Concatenate-(out8, out3)	
10	UpSampling2D-(SI2), CONV2D-(F256,K4,S2, ACT 'LeakyRelu'), InstanceNormalization	Expansive Path
11	Concatenate-(out10, out2)	
12	UpSampling2D-(SI2), CONV2D-(F128,K4,S2, ACT 'LeakyRelu'), InstanceNormalization	
13	Concatenate-(out12, out1)	
14	UpSampling2D-(SI2), CONV2D-(F3,K4,S1, ACT 'Tanh')	

The discriminators (D_A and D_B) are PatchGAN discriminator networks which evaluate the validity of all patches composing an image, outputting a 2D array with local authenticities [130]. Thus, this method considers different sections of the image for the assessment. In Table 3.4, the architecture of the employed discriminators is presented in detail.

Table 3.4: PatchGAN Discriminator Network architecture for the unsupervised image translation experiments. The following abbreviations are used for ease of presentation: F=Filters, K=Kernel Size, S=Stride, and R=Rate.

Layer	Discriminator
1	CONV2D-(F32,K4,S2), LeakyReLU-(α 0.2), Dropout-(R0.2)
2	CONV2D-(F64,K4,S2), LeakyReLU-(α 0.2), Dropout-(R0.2), InstanceNormalization
3	CONV2D-(F128,K4,S2), LeakyReLU-(α 0.2), Dropout-(R0.2), InstanceNormalization
4	CONV2D-(F256,K4,S2), LeakyReLU-(α 0.2), Dropout-(R0.2), InstanceNormalization
5	CONV2D-(F1,K4,S1)

For data augmentation, vertical and horizontal flipping, rotation in a range of 0-20 degrees, and a slight brightness variation in a range of 0.9-1.1 were applied.

3.3.2.1 Loss Functions

In each epoch, the discriminators are trained separately from the CycleGAN itself as their goal is to validate the output of the generators. To that end, these networks learn to differentiate synthetic outputs from ground-truth samples. A mean squared error loss function was employed for training both PatchGAN discriminators. Regarding the CycleGAN, the loss function is given by

$$Loss_{total} = \sum_{n=\{A,B\}} (w_r Loss_{r_n} + w_i Loss_{i_n} + w_{id} Loss_{id_n})$$

where the $Loss_{r_n}$ corresponds to mean absolute error functions for the reconstruction of the translated images. This assures cycle-consistency which implies that the underlying

information of the image is preserved while its domain is shifted. $Loss_{i_n}$ are mean squared error loss functions for the validation of the translated images through the discriminator. The validation loss allows the generators to produce images that could have been realistically drawn from the target domain set. Lastly, the $Loss_{id_n}$ functions correspond to mean absolute error functions which force the images to equal their original form after passing through the generator. This further assures that the identity of the images is maintained. Intuitively, the weight loss of the latter needs to be considerably lower than the first two's to guarantee a successful domain translation.

The weights are set to $w_r = 10$, $w_i = 1$ and $w_{id} = 0.1$, giving a higher importance to the reconstruction loss that ensures cycle-consistency.

While training the model, it became noticeable that the losses of the discriminators converged at a high rate when compared to the ones of the generators. This phenomenon keeps the generators from learning efficiently as they cannot deceive the discriminators. Thus, a loss threshold was employed below which the discriminators are not trained. This upper limit was set to 0.26.

For optimization of both the discriminators and the CycleGANs, the Adam algorithm was implemented with a learning rate of 0.0002.

3.3.3 UNIT Framework

Lastly, we propose using the UNIT framework for translating the low-quality images into high-quality ones. The working mechanism of these architectures is presented in Section 2.4.3. As aforementioned, this framework is based on two VAE-GAN architectures, associated with each domain, that have a shared-latent space constraint. If using only the VAE-GAN related to one of the domains, image reconstruction in the respective domain is obtained. On the other hand, if a domain shift is performed in the shared-latent space (using the encoder of one domain and the decoder/generator of the second domain), the images are mapped into the opposite domain. Consequently, an image-to-image translation is achieved.

The model is trained in a cycle-consistency manner, as illustrated in Figure 3.9. As presented, each encoder is responsible for compressing the data into the shared-latent space. The generators synthesize the image outputs based on the latent distributions.

The process begins with the image sets of each domain being compressed into a latent representation. The compressed information is then reconstructed (using the generator from the same domain as the encoder) and translated (using the generator from the opposite domain of the encoder). The two translated output sets are evaluated for their authenticity regarding the target domain. Finally, they are translated back to their original domain.

For data augmentation, horizontal flipping was applied.

In Table 3.5, the architecture of the employed encoders, decoders, and discriminators is presented in detail. We employed a PatchGAN discriminator network as used in the

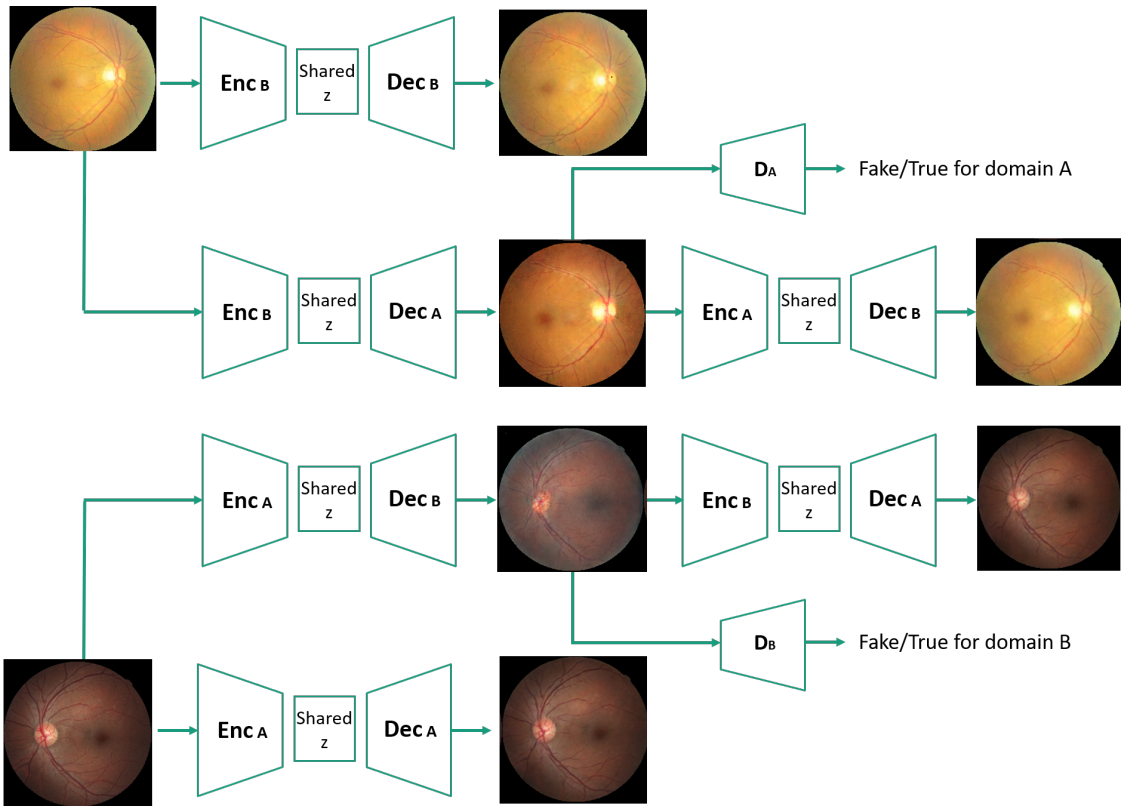


Figure 3.9: UNIT framework training illustration.

CycleGAN framework presented in Section 3.3.2.

3.3.3.1 Loss Functions

As explained for the CycleGAN architecture, the discriminators are trained separately from the UNIT itself as their goal is to validate the output of the generators upon performing translation processes. For the discriminators, a binary crossentropy loss function was implemented. The loss function of the encoder-generator models is given by

$$Loss_{total} = \sum_{n=\{A,B\}} (w_k Loss_{k_n} + w_r Loss_{r_n} + w_{id} Loss_{id_n} + w_i Loss_{i_n})$$

where $Loss_{k_n}$ correspond to Kullback Leibler loss functions (as employed in the VAE framework in Section 3.3.1), $Loss_{r_n}$ correspond to mean absolute error functions for the reconstruction error, $Loss_{id_n}$ correspond to mean absolute error functions for the identity of the translated images' reconstructions in comparison to their original form, and $Loss_{i_n}$ correspond to binary crossentropy functions for the authenticity validation of the translated images. The weights were set to $w_k = 0.1$, $w_i = 10$. For $w_r = w_{id}$, three loss weights were contemplated. These correspond to 100, 2500 and 5000.

Table 3.5: UNIT architecture for the unsupervised image translation experiments. The following abbreviations are used for ease of presentation: F=Filters, K=Kernel Size, and S=Stride. The residual basic block is denoted as RESBLK and the transposed convolutional layer is denoted by DCONV2D.

Layer	Encoders	Shared
1	CONV2D-(F64, K3, S1), LeakyReLU-(α 0.01)	no
2	CONV2D-(F64, K3, S2), LeakyReLU-(α 0.01)	no
3	CONV2D-(F64, K3, S2), LeakyReLU-(α 0.01)	no
4	RESBLK-(F256, K3, S1)	no
5	RESBLK-(F256, K3, S1)	no
μ	RESBLK-(F256, K3, S1)	yes
Layer	Decoders	Shared
7	RESBLK-(F256, K3, S1)	yes
8	RESBLK-(F256, K3, S1)	no
9	RESBLK-(F256, K3, S1)	no
10	DCONV2D-(F128, K3, S2), LeakyReLU-(α 0.01)	no
11	DCONV2D-(F64, K3, S2), LeakyReLU-(α 0.01)	no
12	DCONV2D-(F3, K3, S1), Activation-(tanh)	no
Layer	Discriminators	Shared
1	CONV2D-(F64, K3, S1), LeakyReLU-(α 0.01)	no
2	CONV2D-(F128, K3, S2), LeakyReLU-(α 0.01)	no
3	CONV2D-(F256, K3, S2), LeakyReLU-(α 0.01)	no
4	CONV2D-(F512, K3, S2), LeakyReLU-(α 0.01)	no
5	CONV2D-(F1, K1, S2), Activation-(sigmoid)	no

3.4 Validation Methods

The final developing stage of this work aims to construct methods that validate the quality improvement of retina images using the proposed DL algorithms. The two methods proposed for that end are presented in this Section. One estimates the quality label shifts of retina images using a three-level retina image quality classifier. The second assesses the quality transformation impact on the performance of a Diabetic Retinopathy severity classifier.

3.4.1 Retina Image Quality Assessment

The first validation method consists of evaluating the quality of the generated retina images in comparison to their original form.

A quality classification improvement is relevant both for ensuring the reliability of diagnoses by ophthalmologists and for promoting a more efficient classification by CADx

systems. The reason is that the lack of quality can jeopardize disease detection and certain artifacts can be mistaken as small damages to the blood vessels.

This validation system relies on the quality assessment EyeQ data set presented in Section 3.1.2. As beforehand mentioned, the images composing this data set are graded in three quality levels based on four common quality indicators, including blurring, uneven illumination, low-contrast, and artifacts. Besides the construction of the EyeQ data set, Fu et al. [75] also proposed a general Multiple Color-space Fusion Network (MCF-Net) for retina image quality classification implemented on PyTorch. Their experiments demonstrated that the MCF-Net outperforms other methods. We adapted their model into a TensorFlow implementation to use for our image quality validation [75].

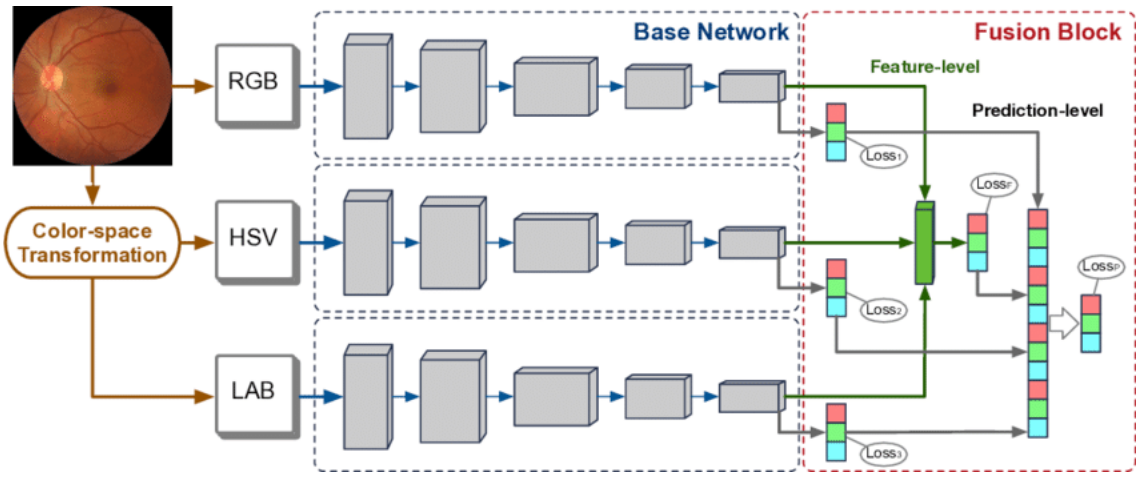


Figure 3.10: MCF-Net architecture, which contains multiple base networks for different colour-spaces [75].

As presented in Figure 3.10, the original RGB images are first transferred to HSV and LAB colour-spaces and fed into the base networks. DenseNet121 algorithms were implemented as base networks because these were reported to ensure the best results when compared to other state-of-the-art networks [75]. The feature maps generated by each base DenseNet and their corresponding predictions are combined in a fusion block. The fusion block is composed of two levels. First, the feature maps from the base networks are concatenated and fed to a fully connected layer to produce a feature-level fusion prediction. On the second level, the predictions of all the base networks and feature level fusion are concatenated and given as input to a fully connected layer. The latter generates the final prediction-level fusion result.

This solution allows the integration of the different colour-spaces while still maintaining the independence of each base network by retaining their loss functions.

The fusion loss of the last layer is given by

$$Loss_{total} = \sum_{i=1}^{3} w_i L_i + w_P L_P + w_F L_F$$

where $Loss_i$, $Loss_F$ and $Loss_P$ correspond to multi-class cross-entropy loss functions of the base networks and the two-level fusion layers, respectively. The weights are set to $w_i = 0.1$, $w_F = 0.1$ and $w_P = 0.6$, giving a higher importance to the final prediction-level fusion layer.

To train the quality assessment model, the EyeQ data set was divided into train, validation, and test sets, each composed of equal image quality grade proportions. For data augmentation, vertical and horizontal flipping, and rotation were applied. The initial weights of each base network are loaded from pre-trained models on ImageNet. For optimization, the SGD algorithm with a learning rate of 0.01 was employed.

The presented model allows evaluating the quality of the retina images. However, the interest of this validation relies on assessing alterations in image quality classification upon the quality transformation. Hence, the proposed algorithm is set to be used for evaluating the quality of the retina images before and after the quality transformation. Our goal is to assess the consequent quality classification shift.

3.4.2 Diabetic Retinopathy Computer-aided Diagnosis

The second validation method relies on an automated analysis system that detects Diabetic Retinopathy in retina images. The goal is to infer whether the performance of the model improves when receiving as input the quality-improved images generated by the proposed translation methods. If the quality enhancement is accomplished, a performance improvement of the DR classifier is theoretically expected. This is because, by eliminating or attenuating low-quality factors, their impact on the classification is minimized. Furthermore, the images become more similar among each other, and thus, the AI-based algorithm is more efficient in separating them according to only their disease sign presence instead of other external factors such as image quality.

The construction of this method is based on the EyePACS data set presented in Section 3.1.1. As previously explained, each retina image in this data set is rated on a scale of 0 to 4, according to their DR severity.

An EfficientNet-B0 was implemented as a baseline network, whose structure is illustrated in Figure 3.11. This neural network belongs to the EfficientNets family of models, developed by Tan et al. in 2019 [131]. These networks scale up CNNs dimensions, such as width, depth, and resolution with a fixed set of scaling coefficients, instead of scaling them arbitrarily as employed by conventional approaches. These models have proven to surpass state-of-the-art accuracy with up to 10x better efficiency, being smaller and faster. The B0 model was utilized because its training input has a size of 224 x 224 pixels, which matches the width and height fixed for image size in this work. The feature maps generated by the baseline network is fed to a second block whose architecture is presented in Table 3.6.

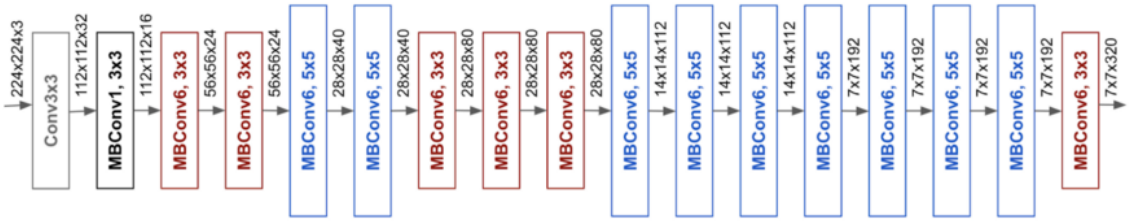


Figure 3.11: A basic block representation of the EfficientNet-B0 [132].

For the loss function, a conventional mean squared error was employed. To train the disease severity assessment model, the EyePACS data set was divided into train, validation, and test sets, each composed of equal DR severity grade proportions. To generate augmented data for training, vertical and horizontal flipping, rotation, and slight brightness shifts were applied. The initial weights of the EfficientNet-B0 are loaded from the Noisy Student pre-trained model. The Adam algorithm with a learning rate of 0.01 was employed for optimization.

Table 3.6: Second block architecture of the diabetic retinopathy severity classification model for validation. The following abbreviations are used for ease of presentation: F=Filters, R=Rate, and ACT=Activation.

Layer	DR classification model - 2nd Block
1	GlobalAveragePooling2D, Dropout -(R0.5)
2	Dense-(F5, ACT 'LeakyRelu')
3	Dense-(F1)

Cohen's Kappa score was used as the performance metric. We opted for this approach because the classification involved multiple labels with small differences between each grade interval. This approach implies that predicting a class that is adjacent to the ground-truth class is less incorrect than predicting a class with a large interval from the ground-truth. Thus, we believed this metric to be more robust as it accounts for the possibility of the agreement occurring by chance. In addition, this performance metric has been commonly and successfully applied to ordinal regression classification problems, such as the current one [133].

Given the proposed architecture for this DR severity assessment model, we will further address it as DR_EffB0-Net.

Chapter 4

Results and Discussion

This Chapter comprises the results of the proposed dissertation. The resulting efficiency of using each retina image enhancement framework, as well as the performance results of the proposed validation methods, are addressed. A critical discussion is given for each approach, where the positive and negative aspects are presented and suggested explanations for the found performance are given. To further enrich the analysis of the obtained results, we propose possible relevant usages of the methods that did not prove to be successful in achieving the goal of this research project.

4.1 Retina Image Quality Improvement Frameworks

Here, the obtained results for the three proposed frameworks for retina image quality improvement are addressed. For each methodology, a visual-based analysis of the quality improvement was performed. When this proved to be successful, a quantitative performance evaluation was employed using the proposed validation methods.

It is important to note that these results derive from the already optimized version of each proposed model. Deep Learning networks encompass multiple adjustable variables such as the number of layers, number of filters, kernel size, among others. Deciding all existing variables requires a trial-and-error process whose success will depend on the architecture of the neural network, as well as on the training data for a specific task. The chosen parameter values for our frameworks result from researching already implemented solutions of each selected DL approach (VAE, UNIT, and CycleGAN) and performing early tests with less data for a time-efficient optimization of the models' performance.

4.1.1 Variational Autoencoder Framework

One of the proposed approaches for retina image quality enhancement involved training a Variational Autoencoder to reconstruct high-quality images. The trained model would then receive low-quality images as input so these could be reconstructed in a way that high-quality features were emphasized, as explained in Section 3.3.1. Hence, the first step was to train the VAE to reconstruct high-quality images. A set of approximately 1,500 retina images from the EyeQ data set, labelled as having 'Good' quality was used for that end. The reason for not using all retina images from that class is because some high-quality samples present slight signs of the defects detected in the images with inferior quality labels. Thus, a visual-based selection of the images with a stronger demonstration of high-quality characteristics within the 'Good' label set was performed.

As explained in Section 3.3.1, two VAE architectures with distinct reconstruction loss approaches were proposed: one employing the conventional mean squared error function, and the second using a perceptual loss function based on a pre-trained network for retina image quality classification.

In Figure 4.1 and 4.2, we display the obtained results for the high-quality image reconstruction using the VAE models trained with the conventional reconstruction loss function and the perceptual loss function, respectively. For each presented sample, the input image and its corresponding reconstruction output are illustrated.

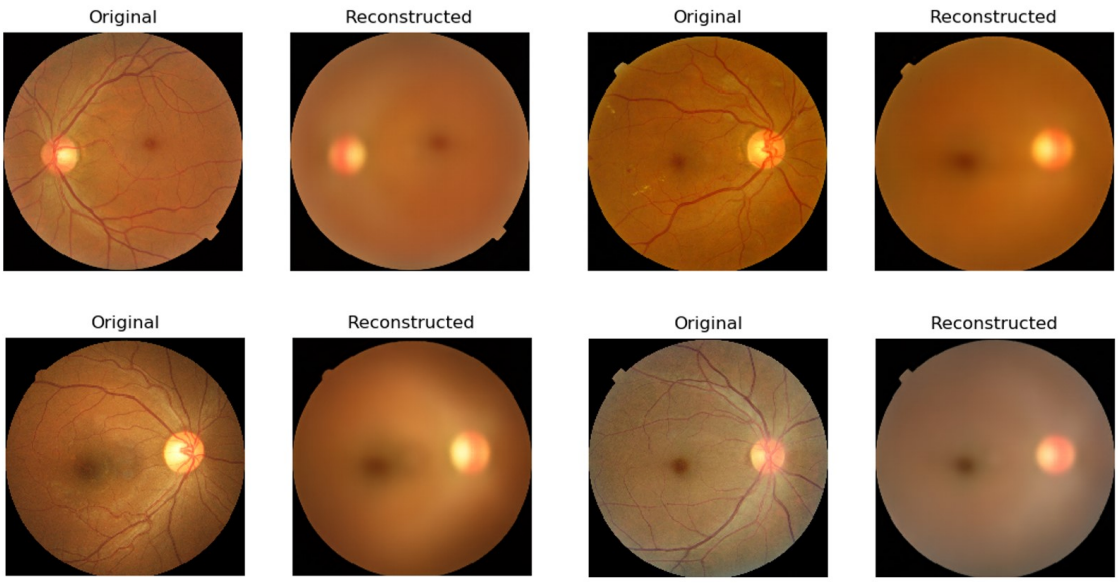


Figure 4.1: High-quality image reconstruction exemplification results through the VAE model trained with a conventional reconstruction loss function.

Based on visual analysis, it can be concluded that the reconstruction of the high-quality images using a VAE is not sufficiently effective for further use of this approach in retina image quality enhancement. As presented, the reconstructed images are only able to reproduce the background of the original images and not the relevant superficial

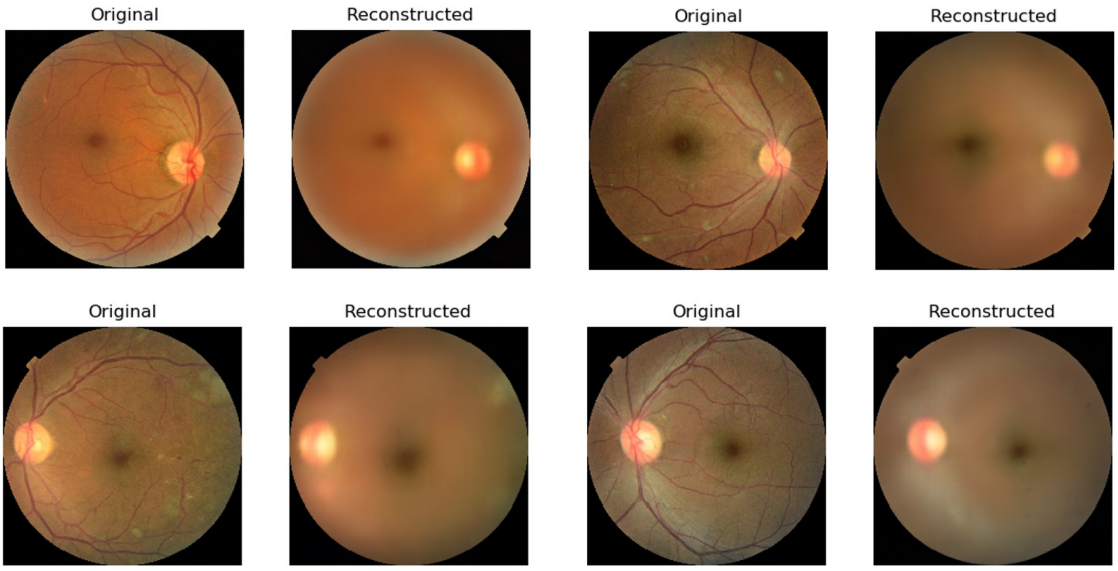


Figure 4.2: High-quality image reconstruction exemplification results through the VAE model trained with a perceptual loss function.

features, namely the blood vessels and detailed characteristics of the optic disc.

The reason for the VAEs not being able to reconstruct images with high precision and detail results from the fact that these models learn an explicit distribution by fitting the data into a multi-dimensional Gaussian distribution. This means the input data needs to be compressed in a latent vector, while gradient descent forces to minimize the difference between the input and output image. However, the compressed information might be insufficient to represent the original image or allow a successful encoder/decoder learning. Because it is not possible to parameterize a complex distribution, the reconstruction process is often oversimplified, leading to blurry output images.

Regarding the size of the latent dimension, which increase can lead to the compression of more relevant information, our settings were near the superior limit of the GPU's processing capacity. Nonetheless, the referred size could also not be excessively large because the reconstruction of the low-quality samples required an image modification margin for the emphasis of high-quality features.

When using the perceptual loss function based on a pre-trained network, a reconstruction improvement was detected. The output not only presents a colour pattern more similar to the input retina image, but also a more detailed and accurate delineation of the optic disc. The reason for this improvement relies on the fact that the perceptual loss function seeks consistency between the hidden representations of two images. That is, instead of only performing a pixel-by-pixel comparison (as used in the mean squared error approach), the perceptual loss accounts for similarities in higher-level features between the two images.

It can be deduced that using perceptual loss functions based on high-level features extracted from pre-trained networks can improve the performance of the reconstruction

process. Nevertheless, the detected reconstruction enhancement was still unsatisfying in reproducing relevant retina image details.

Because the VAE approach proved to be unsuccessful in reconstructing high-quality retina images in detail, we were not able to inspect whether feeding low-quality images as input to the already trained network would improve their quality. Despite this, it was possible to conclude that this method is able to perform accurate background reconstructions of retina images. Thus, even if the VAE alone is not useful for image quality improvement, we would like to recommend this VAE-based retina image background reconstruction mechanism for other applications.

An interesting approach would be to use the background of retina images for segmenting and isolating superficial features, such as blood vessels, optic disc details, and other pathology-related characteristics (e.g. exudates). This could be accomplished by subtracting the background image resulting from the VAE reconstruction from the original retina image.

Some studies also rely on the fusion of the background image with the original one for image quality enhancement. The work carried by Dai et al. [80] which is presented in Section 1.3.3, represents an example of this. The VAE architecture could be employed for obtaining the background component (1.2 (b)) that is later used in the fusion process for quality enhancement.

4.1.2 UNIT Framework

This approach aimed to train a UNIT framework containing two VAE-GAN models with a shared weight constraint. The goal was to allow image reconstruction when maintaining the same domain in the encoding and decoding paths, and image-to-image translation when shifting the domain between the encoding and decoding paths. The image-to-image translation path would then be used to transform the low-quality images into high-quality ones for the retina image quality enhancement, as explained in Section 3.3.3.

For training the UNIT framework, a set of low-quality images, and a set of high-quality ones was required. The low-quality data set 1, presented in Section 3.1.3, was used for that end. This data set contains 2,328 images with hazing effects and uneven illumination as a consequence of over-exposure. The high-quality set, also addressed in Section 3.1.3, contains 1,515 images labelled as having 'Good' quality in the EyePACS data set and simultaneously not presenting any signs of low-quality factors. For the testing process, 200 images of each set were randomly selected.

To allow a visual comprehension of how the outputs of the UNIT framework are generated and assessed during the cycle training, Figure 4.3 displays an exemplification of the generated images.

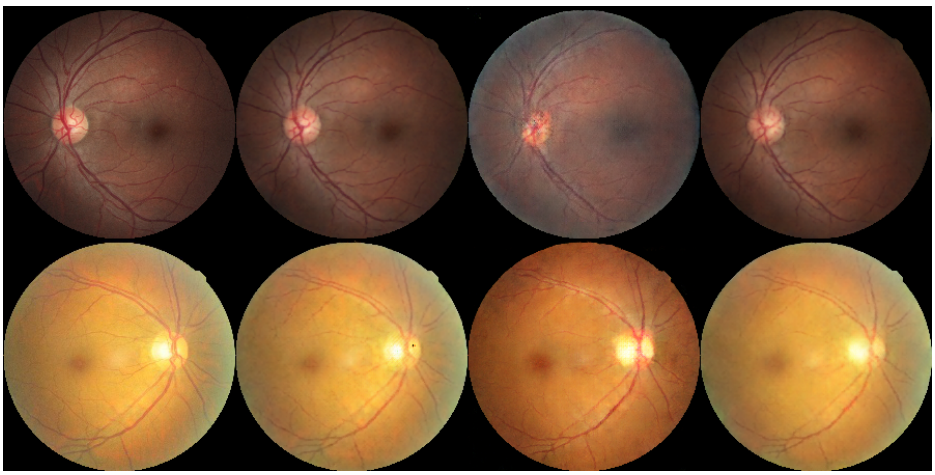


Figure 4.3: Exemplification of a generated image while training the UNIT framework. The first column corresponds to the original images in each domain, the second to the reconstructions of the original image. In the third column are the image-to-image translation outputs, and lastly, the fourth column comprises the image translations back to the original domain.

While training the model, three main indicators were considered for the performance progression assessment: the visual inspection of the output images; the losses of the discriminators; and the reconstruction losses. Other often applied metrics for judging the image outputs of generative models, such as the Inception Score (IS) and the Frechet Inception Distance (FID), were not employed. The reason is that our focus was on observing the learning impact on the correction of the different defects/artifacts, throughout its process. Hence, a visual analysis would be more sensitive to those changes and, consequently, more reliable.

To reach optimal performance, the weights of each loss function were adjusted according to the detected results. This was mainly done to find an equilibrium between an efficient image reconstruction and an efficient image-to-image translation. However, it was not possible to reach a weight loss setting that granted the fulfillment of both requirements simultaneously. We selected three different settings to demonstrate the observed phenomenon.

In Figure 4.4, 4.5, and 4.6, we present the obtained results after training the UNIT framework when using a w_r and w_{id} of 100, 2500, and 5000, respectively. These parameters correspond to the reconstruction and identity weight losses. By shifting these weight losses, the equilibrium between reconstruction efficiency, which is weighted by w_r and w_{id}, and translation efficiency, which is weighed by w_i, is adapted. In Section 3.3.3, the complete loss function is displayed. For each considered sample, we present the corresponding reconstructed image and the translated image in the opposite domain i.e. the high-quality domain.

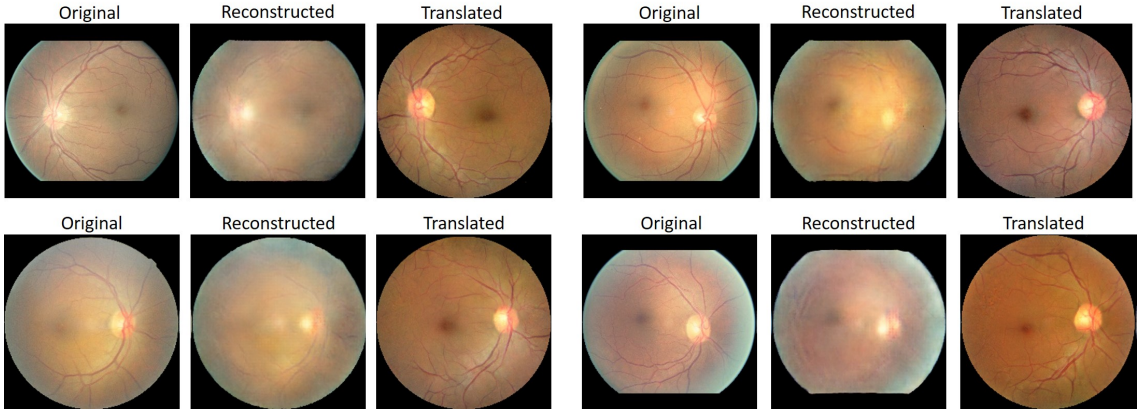


Figure 4.4: Example results of the processes of reconstruction and domain translation from low- to high-quality using the UNIT framework with the weight losses w_r and w_{id} set to 100.

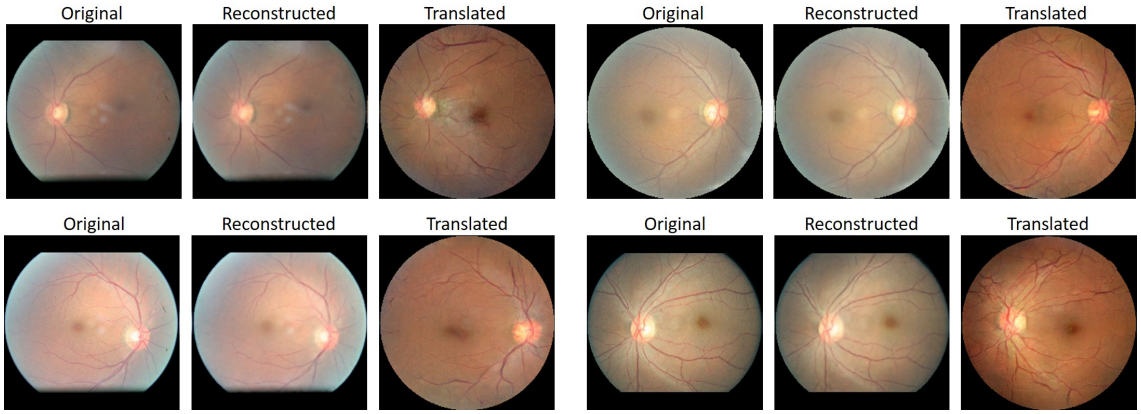


Figure 4.5: Example results of the processes of reconstruction and domain translation from low- to high-quality using the UNIT framework with the weight losses w_r and w_{id} set to 2500.

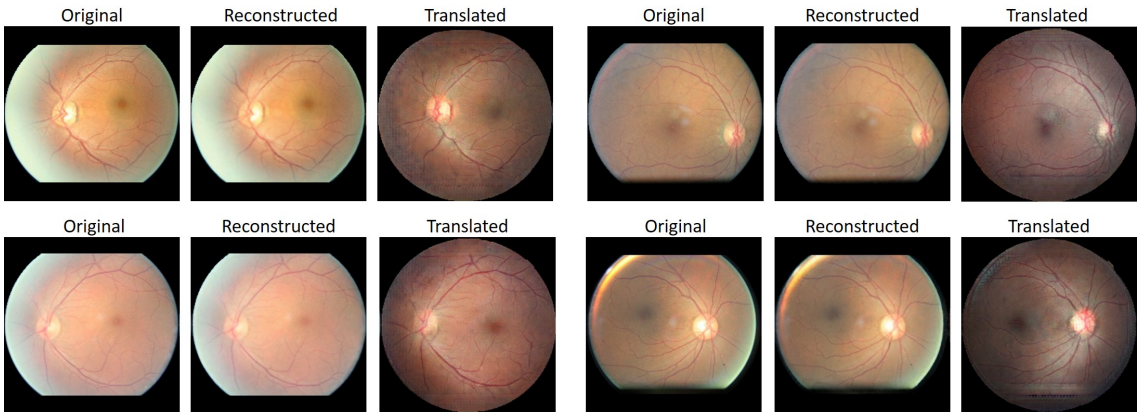


Figure 4.6: Example results of the processes of reconstruction and domain translation from low- to high-quality using the UNIT framework with the weight losses w_r and w_{id} set to 5000.

As presented, none of the displayed images assures an optimal equilibrium between an efficient image reconstruction and an efficient image-to-image translation.

When analysing the first case ($w_r = w_{id} = 100$), a poorly efficient reconstruction of the input image can be detected. The reconstructed image lacks original details related to more superficial features such as blood vessels, detailed characteristics of the optic disc, or some pathology-related elements (e.g. exudates). As a consequence of that inefficient image reconstruction, the images generated from the domain translation process are not faithful to the original inputs in terms of their underlying morphology information. One can observe that the translated images present misplaced and/or new vessels, not correctly located optic discs, imprecise delineations of some structures, etc. Thus, it can be concluded that attributing too little weight to the reconstruction process leads to translated images whose morphological information varies from the original image. On the other hand, it is also possible to visualize that images generated in the target domain are highly realistic high-quality retina images. The reason lies in the fact that by minimizing the reconstruction weight losses, we endorse the weight of authenticity validation of the translated images. Therefore, it is expected for these to be more similar to the training samples in the target domain.

For the higher values of reconstruction weight losses ($w_r = w_{id} = 5000$) the opposite situation can be observed. The reconstructed images are extremely similar to their corresponding inputs as all relevant features (both superficial elements and background characteristics) are preserved. Consequently, the translated images in the target domain maintain the original morphological characteristics of the images in the starting domain. Mainly this refers to the nonexistence of new or incorrectly located blood vessels, as well as the optic disc placement. However, the weight of the translated images' authenticity validation is minimized in comparison to the reconstruction process. As a consequence, we can detect that the translated images are often not realistic high-quality retina images. Several comprise checkerboard artifacts, blurred and not properly delineated optic discs, inaccurate generation of vessels in the optic disc, among others.

Upon detecting the two extreme opposite problems of using unbalanced weight losses, multiple attempts were made to reach an optimal equilibrium. However, we detected there was no interval in which both drawbacks were absent: either one of these was observed or they overlapped. Figure 4.5 displays an example of an intermediate weight loss setting ($w_r = w_{id} = 2500$). As can be visualized, neither the reconstruction nor the translated process is ideal. On one hand, the reconstructed images are not completely identical to their input, which leads to some morphological variance of the images generated in domain-translation. On the other hand, the translated images are not entirely realistic high-quality retina images as multiple present generation defects, especially in the optic disc.

In conclusion, we could not reach model settings with which the UNIT framework proved to be sufficiently effective in retina image quality improvement. None of the images translated from the low- to the high-quality domain represented an accurate

example of a truthful representation of a high-quality retina image while maintaining the underlying morphological information of the original input.

However, it was possible to detect that contrarily to the VAE architecture (whose results are reported in Section 4.1.1), each VAE-GAN of the UNIT framework yielded a considerably efficient reconstruction process. This was particularly noticeable when using high weight loss values for the reconstruction losses, as seen in Figure 4.6. The reason for the VAE-GAN to outperform the VAE in the reconstruction process is because we introduce learned feature representations in the GAN discriminator as a basis for the VAE reconstruction objective. Consequently, feature-wise errors are added to the element-wise errors which better captures the data distribution while offering invariance [134].

We thus decided to inspect whether using a VAE-GAN could be efficient for the proposed VAE strategy. This method consisted of inputting low-quality images to a previously trained network that reconstructs high-quality images, and detecting if the images gain high-quality features while preserving their original information. To do so, we selected the trained encoder and decoder from the high-quality domain VAE-GAN and then inputted the images from the low-quality data set 1. The resulting reconstructed images are presented in Appendix A. Relevant quality improvements were not detected as the reconstructed images were highly similar to their input and only a few presented slight uneven illumination corrections.

Despite the UNIT approach not being successful for the retina image quality enhancement objective, we consider this method to have other important potentialities. As explained in Section 2.2.2, the performance of AI algorithms highly depends on the quality and quantity of training data. Data augmentation represents a relevant tool for increasing the volume and diversity of data available for training. When attributing lower importance to reconstruction (Figure 4.4), the translated images were highly realistic. Thus, under certain settings, the UNIT model could be used to generate artificial retina images for data augmentation. This could be useful for generating retina images containing different defects to make classifiers more robust to the diversity of real-world image acquisition settings. It could be also advantageous in increasing the volume of pathological images with different disease severities as well as negative images (no pathology). Moreover, the discriminator could be further used to guarantee all produced images do not present generation defects (consider only outputs which validation by discriminator is above a certain threshold). This would enlarge the number of examples to which the network is exposed while learning. Additionally, by generating retina image examples with slight morphological divergences such as the location and form of blood vessels, data anonymity could be achieved. This topic has become more relevant with the usage increase of these images in ophthalmology biometrics for subject identification.

Different data augmentation applications would mainly depend on the employed training data.

4.1.3 CycleGAN Framework

The CycleGAN approach aimed to reach a successful image-to-image translation model capable of transforming retina images from a low-quality domain into a high-quality one, as explained in Section 3.3.2. Given this objective, the proposed CycleGAN architecture was trained using low- and high-quality image sets to represent each domain.

Five low-quality data sets were considered for the training process. First, a data set containing an ensemble of 1,817 EyePACS retina images labelled as having 'Usable' image quality was employed. The second approach involved a more strict selection: from the images labelled as 'Usable' we used only a subset of 1,690 retina images that had been classified as 'Usable' by the MCF-Net presented in Section 3.4.1 with a coefficient higher than 0.98. This conceded more uniformity among the images regarding their quality as the samples bordering the 'Good' or 'Reject' labels are discarded. Lastly, the low-quality data sets presented in Section 3.1.3 were considered. These correspond to four image sets with different low-quality factors that are commonly observed in retina images acquired with portable and inexpensive hardware.

The high-quality set is the one addressed in Section 3.1.3. It is composed of 1,515 images labelled as having 'Good' quality in the EyePACS data set from which images presenting slight signs of low-quality factors were discarded.

Figure 4.7 displays an exemplification of how the outputs of the CycleGAN model are generated and assessed while training. Similarly to the UNIT approach discussed in Section 4.1.2, the CycleGAN's performance assessment amid training was based on a visual inspection of the output images, as well as the loss values of the model. Other metrics such as the IS and the FID were discarded for optimal monitoring, following the same explanation given for the UNIT framework.

Upon training the model with the aforementioned low-quality data sets, the generators that perform the image-to-image domain translation from low- to high-quality were isolated. Each generator was applied to correct a set of images containing the low-quality factors corresponding to its training. For the two generators trained without defect isolation, the test sets are the combination of all other sets, that is, they comprise images with all the existing low-quality factors.

In Figure 4.8, 4.9 and 4.10 we present the results obtained upon training the CycleGAN architecture when using each of the five low-quality data sets. For each example, the original low-quality image and the resulting translated image in the target domain are presented.

When analysing the obtained results from using the first low-quality data set, one can see there is little or no quality improvement, as illustrated in Figure 4.8. The high diversity within the low-quality impedes the discriminator from learning what characterizes the low-quality domain and consequently, which features differentiate it from the high-quality domain. This justifies the necessity of considering a more homogeneous

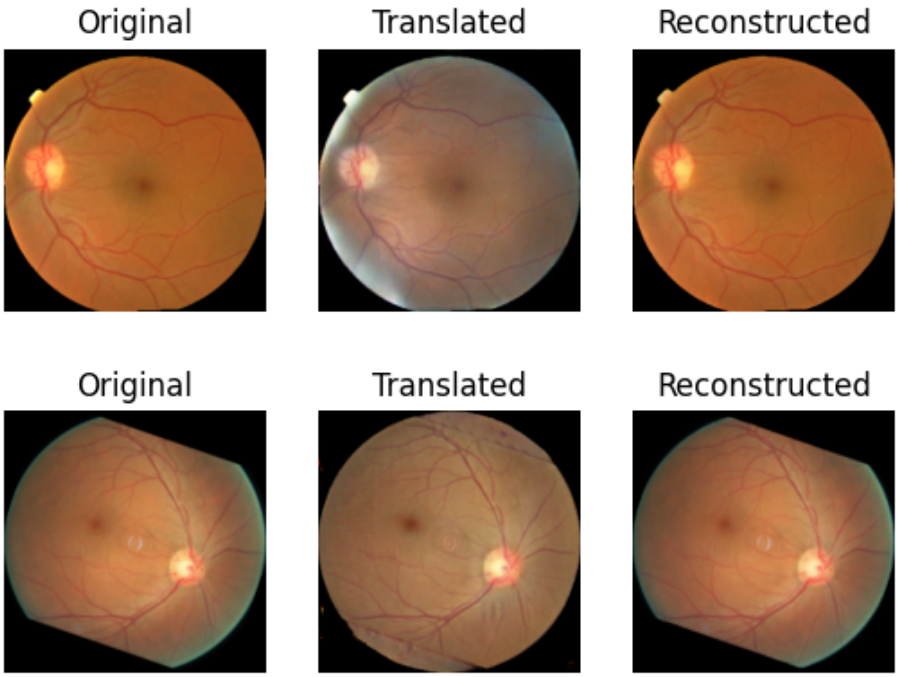


Figure 4.7: Exemplification of a generated image while training the CycleGAN framework. The first column corresponds to the original images in each domain, the second to the image-to-image translation outputs and the third column comprises the image translations back to the original domain.

low-quality data set. As presented in Figure 4.9, after performing a more strict low-quality image selection and training the model using the second low-quality data set, a significant improvement in the enhancement process can be detected. Generally, the defects observed in retina images with inferior quality are partially or completely corrected. The generated images are seemingly closer to the high-quality set and more similar to each other. Thus, it can be deduced that performing a more strict low-quality image selection and guaranteeing more uniformity within a domain set can help discriminators learning to distinguish low-quality image characteristics from high-quality ones. As a consequence, generators tend to produce better high-quality samples.

By training the model with the four low-quality data sets containing different common retina image defects, we could further tackle the problem in an even more selective and individualized manner. The presented results indicate a considerable improvement in retina image quality enhancement as each defect is more efficiently corrected (Figure 4.10). This demonstrates how isolating specific low-quality factors in medical images can help accomplish improved image-to-image translation results in transforming images with inferior quality into superior quality.

However, it is relevant to note there is a limiting factor when addressing each low-quality factor separately: when incorporating the quality improvement transformation to the automatic diagnosis pipeline, an extra classifier must be added to identify the type of defect present in each retina image.

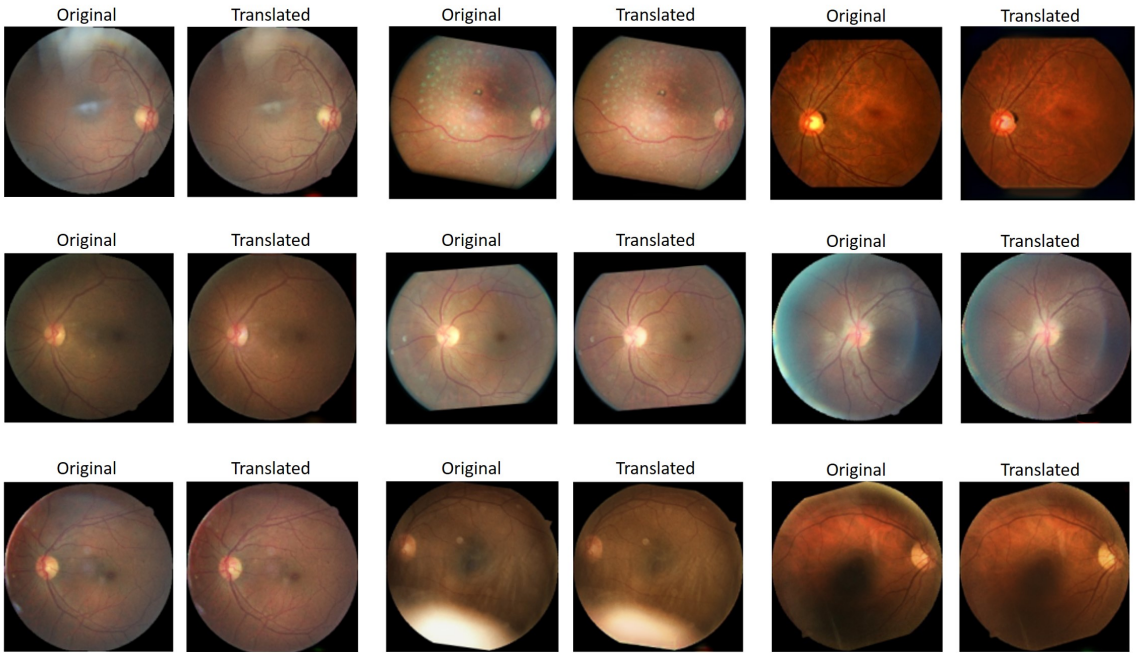


Figure 4.8: Example results of the domain translation process from low- to high-quality using the CycleGAN framework trained with a low-quality set containing 'Usable'-labelled images.

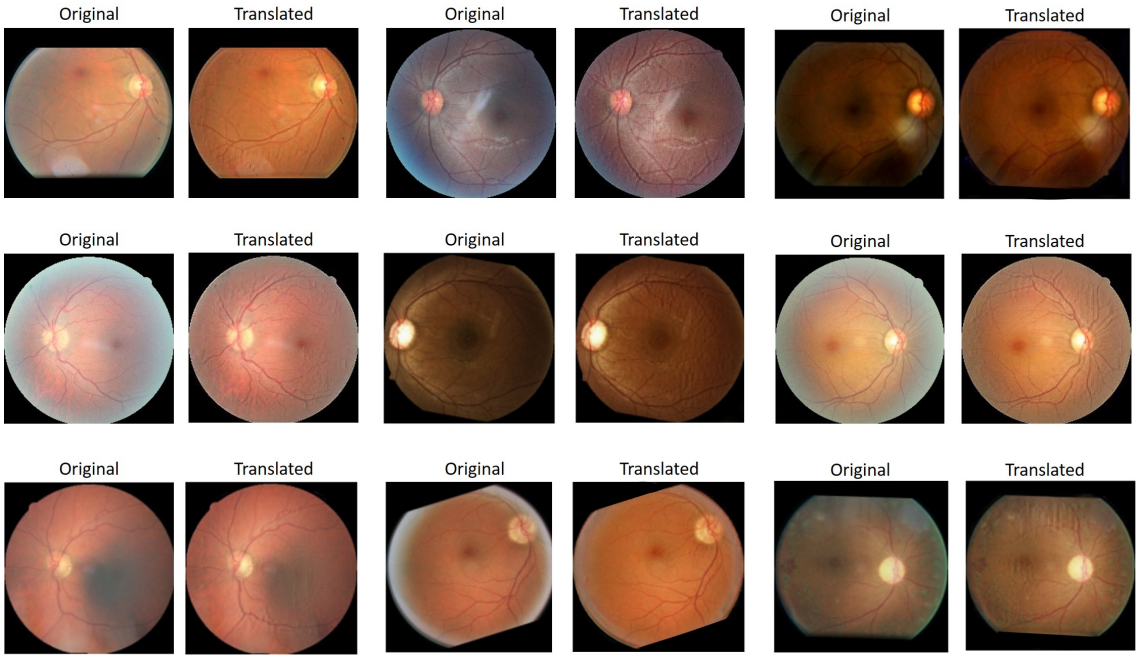


Figure 4.9: Example results of the domain translation process from low- to high-quality using the CycleGAN framework trained with a low-quality set containing 'Usable'-labelled images classified as 'Usable' by the MCF-Net with a coefficient higher than 0.98.

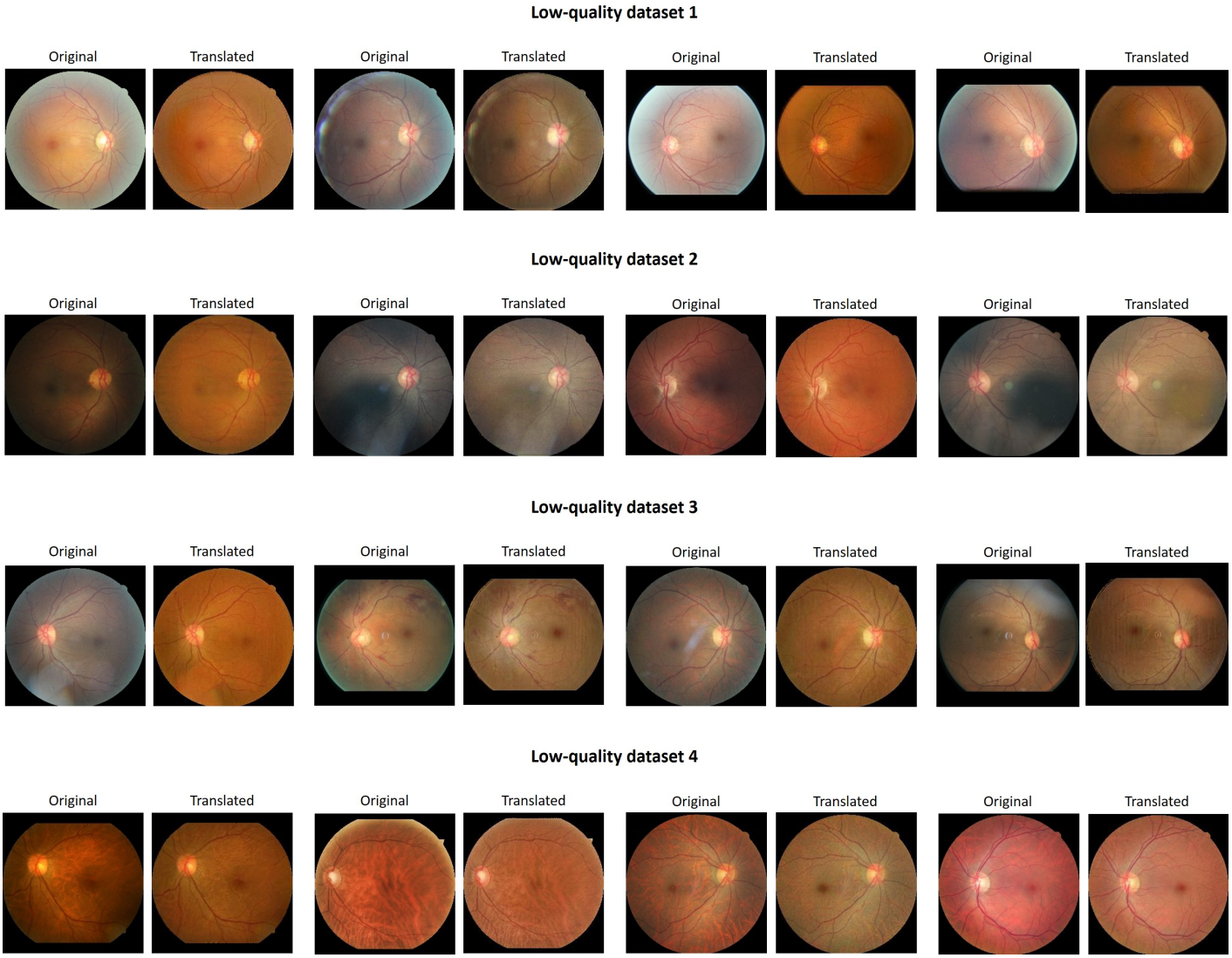


Figure 4.10: Example results of the domain translation process from low- to high-quality using CycleGAN frameworks trained with four different low-quality sets containing particular defects. Each row presents quality transformations performed by the CycleGAN generator trained to correct the corresponding low-quality factors of the mentioned data set.

From a visual-based analysis, it can be concluded the CycleGAN ensures an efficient retina image quality enhancement. The resulting images present less low-quality factors and overall upgraded quality. Additionally, the retina images became more similar to each other in the matter of their quality. This means that upon quality transformation, besides evident morphological divergences, their differences are mostly related to their pathological features.

Despite the observed quality improvement, the underlying structural information of each retina image is preserved. This is also a crucial requirement for our objective of enhancing retina image quality for diagnostic purposes and it has shown potential to be

fulfilled by the CycleGAN approach.

Also, it can be concluded that the way images presenting low-quality factors are grouped into a low-quality domain set highly impacts the efficiency of the quality transformation process. As the heterogeneity within the domain set increases, the less successful are the discriminators in differentiating the original domain from the target domain. Consequently, the identifying features that characterize the low-quality domain are less optimal. The opposite phenomenon is observed when addressing each low-quality factor individually as the discriminator is substantially more likely to learn the most relevant characteristics of the low-quality domain.

These conclusions lead us to confirm that using a CycleGAN-based image-to-image translation can be useful to transform images acquired with portable and inexpensive imaging devices into high-quality ones.

To further evaluate the quality of the generated images, we contemplated using metrics such as the IS and the FID. These are two widely employed scores for assessing the performance of GAN architectures. However, the obtained values were not plausible, incoherent with the standard range, and not aligned with our positive findings. We conjecture this occurs because the referred metrics rely on pre-trained deep networks to represent and statistically compare original and generated samples. This leads to several limitations, already reported in other studies [135]. For instance, the networks are trained to be invariant to image transformations and artifacts, which makes the evaluation method also insensitive to those distortions. Additionally, they are often trained on large-scale natural scene datasets (e.g. ImageNet) [135]. Thus, applying them to other domains, such as retina image quality, might be questionable. For the FID's metric particular case, which is more often used, it is difficult to assess the alignment between the distributions representing the original (high-quality set) and the generated images. This is because it mainly considers their marginal distributions. In the context of this work, where it is crucial to maintain information from the input domain (low-quality) to preserve the natural morphology of the images, it becomes intuitive that the synthetic samples will not approximate as much to the high-quality reference set.

Despite these metrics not being a reliable validation approach for our study, we believed it would be interesting to report their limitations. Also, this analysis further highlights the importance of developing customized validation methodologies for particular generative tasks, such as the ones proposed in this thesis.

4.2 Validation Methods

Visual image analysis allowed detecting which models are seemingly successful in enhancing retina image quality. The successful approaches were set to be quantitatively evaluated by the two validation methods presented in Section 3.4. One estimates the image quality classification shift of the retina images upon quality transformation. The

second assesses the impact of the quality transformation on the performance of a network detecting the presence of DR signs in retina images.

In this Section, the performance of both validation strategies are presented, as well as their application to the CycleGAN approach, which proved to be visually successful in enhancing retina image quality.

4.2.1 Retina Image Quality Validation

The quantitative validation of retina image quality improvement is done through the MCF-Net which grades retina image quality in 'Good', 'Usable' and 'Reject', as explained in Section 3.4.1. To train and evaluate the performance of this quality assessment network, retina image data from the EyeQ data set was used. For the training process, a subset of data comprising 9,000 retina images with equal image quality grade proportions, that is, 3,00 images corresponding to each quality label, was selected. To test the model, a different set of data containing 4,500 images was used. These were also equally divided between the three existing classes i.e. 1,500 retina images for each. For the validation set, a random balanced subset corresponding to 10% of the training data was selected.

To evaluate the performance of the employed method four evaluation metrics were considered: Accuracy, Precision, Recall, and the F1 Score. The performance metrics obtained on the test set are reported in Table 4.1. In the original paper, only the averages of the three existing classes for each performance parameter were reported. Here, the individual values for each class are also presented.

Table 4.1: Performance of our MCF-net implementation on test set in comparison to the original model.

Performance Metric	Class	Our MCF-net implementation	MCF-net proposed in [75]
Accuracy	'Good' 'Usable' 'Reject'	0.858	0.918
Precision	'Good'	0.971	-
	'Usable'	0.784	-
	'Reject'	0.835	-
	Average	0.864	0.865
Recall	'Good'	0.864	-
	'Usable'	0.800	-
	'Reject'	0.910	-
	Average	0.858	0.850
F1 - Score	'Good'	0.914	-
	'Usable'	0.792	-
	'Reject'	0.871	-
	Average	0.859	0.855

The accuracy of our model is inferior to the one presented by Fu et al. [75]. However, when we looked into the composition of their training and testing data, we detected the classes were unbalanced. They used a significantly higher set of images labelled as

'Good'. By analysing the other performance metrics, it can be inferred that this class is the most easily detected within the data set due to the higher precision and F1-Score values. This justifies the observed difference in accuracy. Moreover, when comparing the remaining metrics, a strong similarity between values can be detected and even a slight improvement for recall and F1-Score metrics. Therefore, it can be concluded that our TensorFlow implementation of the MCF-Net is equally effective in detecting the three proposed retina image quality levels.

By reporting the individual performance metric values, it can also be deduced that the 'Usable' class is the most challenging label to identify based on the lower precision, recall, and F1-Score values. This observation is coherent with our initial intuition as this class corresponds to an intermediate level between two extremes. Thus, the probability of having samples near a classification boundary increases.

To allow visualizing the performance of our classification model, the corresponding confusion matrix is presented in Figure 4.11.

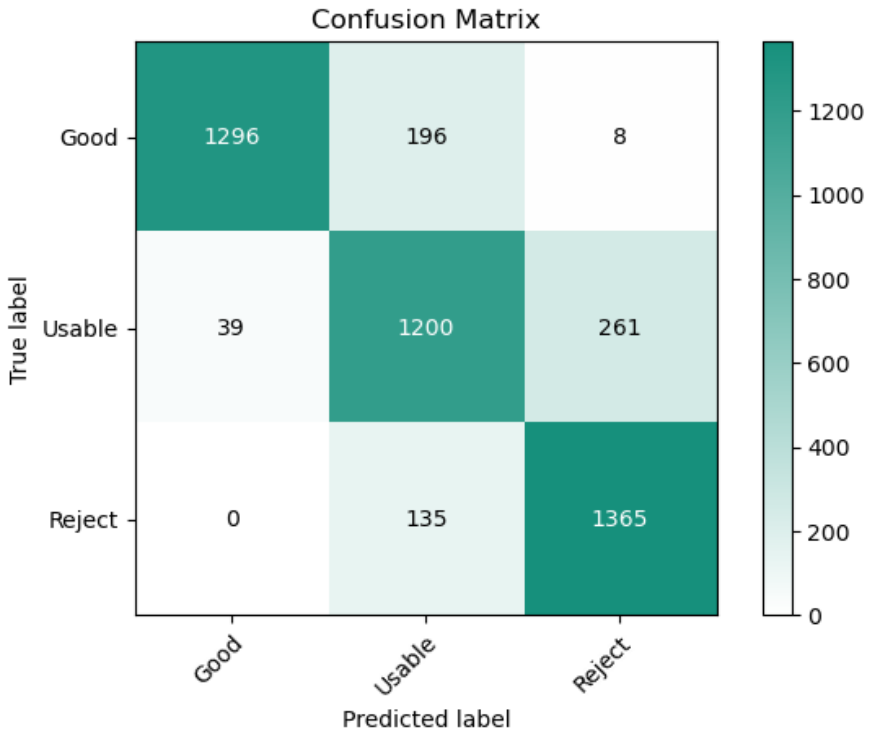


Figure 4.11: Confusion matrix for our TensorFlow implementation of the MCF-Net architecture.

As presented, most ground-truth labels are in agreement with the predicted labels. There are relatively few samples being labelled as the adjacent of their true label and almost none being labelled with two intervals from their true label. This confirms the model's efficiency in classifying retina image quality according to the considered quality grades.

Upon training the MCF-Net, it is used for evaluating changes in image quality upon image quality transformation. For that end, the retina images are classified by the network before and after the DL-based transformation. The quality label shifts are then assessed to inspect if there is an increase in the percentage of retina images being classified as having superior quality than their original form.

4.2.2 Diabetic Retinopathy Computer-aided Diagnosis Validation

The quantitative validation of the computer-aided Diabetic Retinopathy diagnosis improvement is done through the DR_EffB0-Net which grades the presence of DR in retina images in five levels of severity using as baseline an EfficientB0 network, as explained in Section 3.4.2.

To train and evaluate the performance of this DR detection network, retina image data from the EyePACS data set was used. To train the model, we selected a subset of data containing 4,805 retina images with equal DR severity grade proportions i.e. 961 retina images for each. For the testing and validation processes, a different subset of data comprising 2,342 retina images was utilized. It was separated into two sets of data containing 1,171 retina images equally divided between the five existing classes, that is, around 234 images corresponding to each DR severity label.

To evaluate the employed method, Cohen's Kappa coefficient was employed. This was also the validation metric implemented in the construction of our model. Cohen's Kappa coefficient tests inter-rater reliability, that is, it quantitatively measures the magnitude of agreement between observers [136]. For this study, the observers correspond to the expert-based classification (ground-truth) and the prediction made by the DR_EffB0-Net.

The reason for using this performance metric relates to the fact that the proposed model distinguishes five levels of DR severity that present small differences between each grade interval. This approach implies that predicting a class that is adjacent to the ground-truth class is less incorrect than predicting a class with a large interval from the ground-truth.

Table 4.2: Performance of the DR_EffB0-Net model

Cohen's Kappa coefficient	Value's Interval	Interpretability
0.729	0.61–0.80	Substantial Agreement

In Table 4.2, the resulting Cohen's Kappa coefficient obtained on test data is reported as well as its interpretation according to [137]. As presented, the obtained result suggests a significant alignment between the model's predictions and the expert-based classification. Therefore, it can be concluded that the objective of developing a multiple DR severity classifier with considerable efficiency was accomplished. Despite showing positive results, the Kappa coefficient is still not within the highest interval (0.81–1.00) which is interpreted as "almost perfect agreement"[137].

To allow visualizing the performance of our DR severity classification model, we present the corresponding confusion matrix in Figure 4.12. Considering there are several classes in this DR severity grading system, each with relatively small differences to their adjacent labels, some divergence from the main diagonal of the confusion matrix was expected to occur, as verified. Nonetheless, a significant agreement between the predicted and true labels can be detected as most predictions fall within the interval of the direct borders of the corresponding true label. This analysis further confirms a significant efficiency of the proposed model in classifying different DR severity levels. However, as previously verified by the Kappa coefficient, the prediction is not completely precise. This is not problematic as the DR_EffB0-Net is set to be used for a performance comparison when trained with retina images before and after the quality transformation. That is, we aim to enhance the reported performance when training the model with the quality-enhanced retina images.

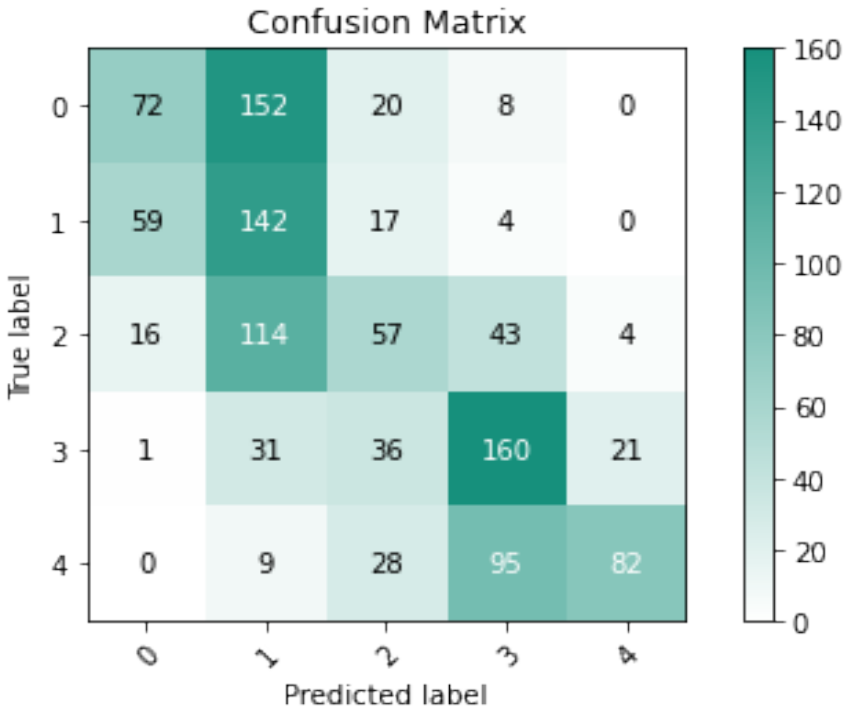


Figure 4.12: Confusion matrix for the DR severity classification model.

4.2.3 CycleGAN Results Validation

To further validate the positive results of using the CycleGAN for retina image quality improvement in a quantitative manner, the presented validation methods were applied.

4.2.3.1 Retina Image Quality Validation

For the quantitative validation of retina image quality improvement, the MCF-Net was used. This validation was performed for the images synthesized using the CycleGAN generators trained with the low-quality data sets that proved to be successful in representing the low-quality domain. These correspond to the data set composed of 'Usable'-labelled images with high classification coefficient from the MCF-Net and the four defect-isolated data sets presented in Section 3.1.3.

As previously explained, the test sets for each quality transformation process are composed of images containing the low-quality factors correspondent to the low-quality data set used for training the CycleGAN. These test sets include images with all three quality labels to detect improvement shifts in the complete grading system spectrum. Intuitively, there are considerably less 'Good'-labelled images displaying the selected defects and also few 'Reject'-labelled images with sufficient quality to be considered for the quality translation process. Consequently, the test sets used for this validation are mainly composed of images labelled as 'Usable', and all results are normalized. For the CycleGAN generator trained without defect isolation, the test set is the combination of all other sets, that is, it compresses all the existing low-quality factors.

Each trained generator was employed to perform the image-to-image quality translation and thus correct the images of the test set containing the low-quality factors corresponding to its training. The MCF-Net was used to evaluate changes in image quality in the five low-quality sets before and after applying the CycleGAN generators.

In Figures 4.13 to 4.17 we display the normalized confusion matrices representing the classification predictions of the MCF-Net before and after performing the CycleGAN-based image-to-image quality translation. The results in the first Figure (4.13) were obtained using the generator whose translation examples are presented in Figure 4.9. The remaining results were obtained using the four generators whose translation examples are presented in Figure 4.10, following the same order.

The objective is to inspect whether a shift towards the left direction is observed, that is, if the images previously predicted as having an inferior quality label are predicted as belonging to a higher quality class after the quality transformation process.

In all presented confusion matrix-based comparisons, an evident shift of predictions into higher-quality labels can be detected. This allows us to have strong quantitative evidence that the images translated into the high-quality domain through the proposed CycleGAN framework are enhanced in respect of their quality.

To quantify the amount of improvement in a numerical assessment, we report in Table 4.3 the percentage of images for which the predicted quality labels increased, as well as the overall percentage of upgrades when using each of the considered low-quality data set representing the low-quality domain. The values presented in the 'Usable' to 'Good' column were calculated by the percentage increase of samples being classified as having 'Good' quality. On the other hand, the values in the column 'Reject' to 'Usable' result from

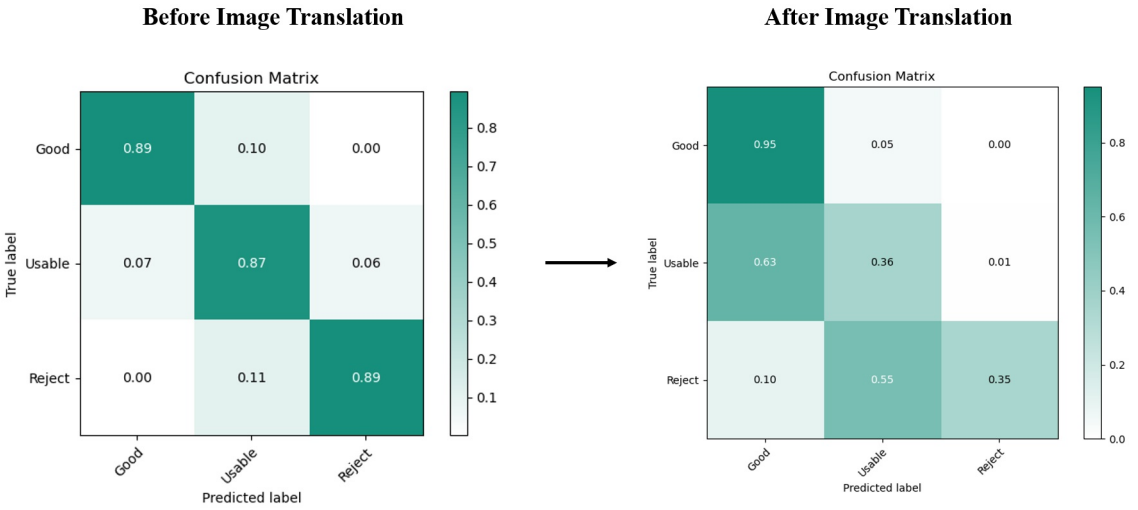


Figure 4.13: Comparison between confusion matrices before and after low-quality (with no defect isolation) retina image quality enhancement using a MCF-Net prediction. Translation examples of the employed generator are presented in Figure 4.9.

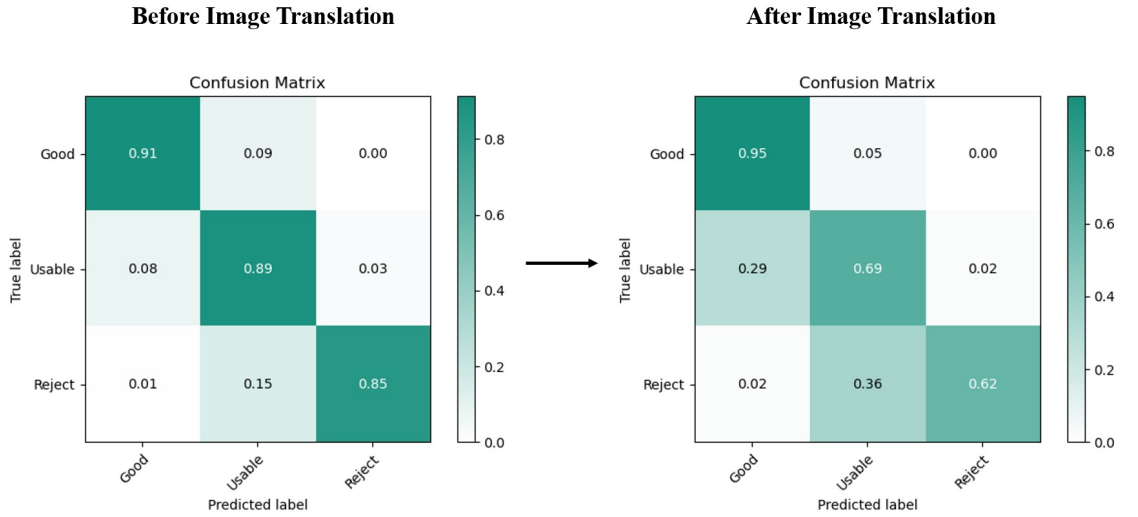


Figure 4.14: Comparison between confusion matrices before and after low-quality data set 1 retina image quality enhancement using a MCF-Net prediction. Translation examples of the employed generator are presented in the first line of Figure 4.10.

the percentage decrease of samples classified as having 'Reject' quality. Considering that most prediction shifts occur towards a quality enhance and that these often advance only in one level, class downgrades or upgrades by two levels were discarded. The calculated values are thus close approximations of the percentage of images that upgrade from 'Reject' to 'Usable' and from 'Usable' to 'Good'. Therefore, Table 4.3 allows analysing the general tendency of prediction change after the quality transformation process, as well as the significance of that change in terms of percentage.

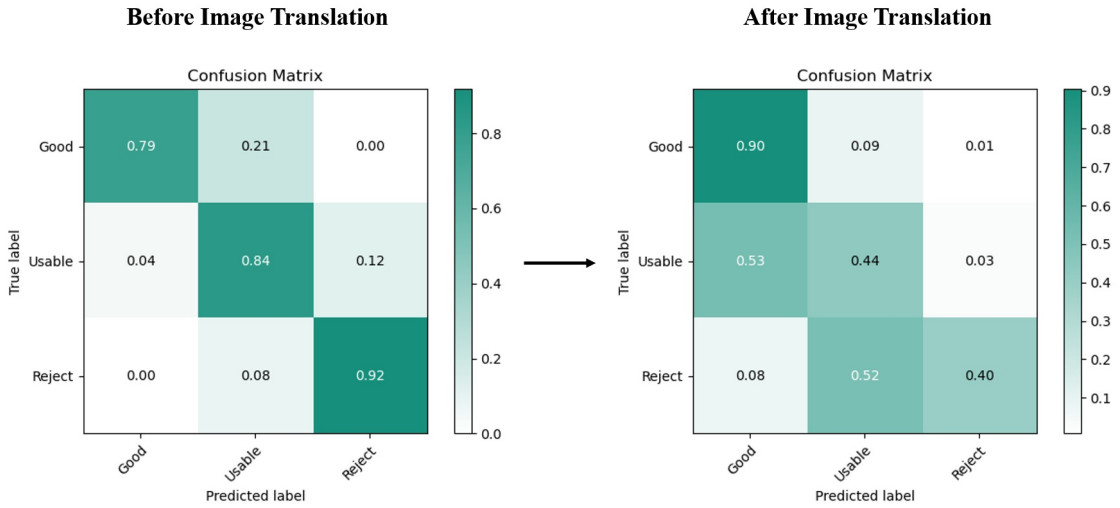


Figure 4.15: Comparison between confusion matrices before and after low-quality data set 2 retina image quality enhancement using a MCF-Net prediction. Translation examples of the employed generator are presented in the second line of Figure 4.10.

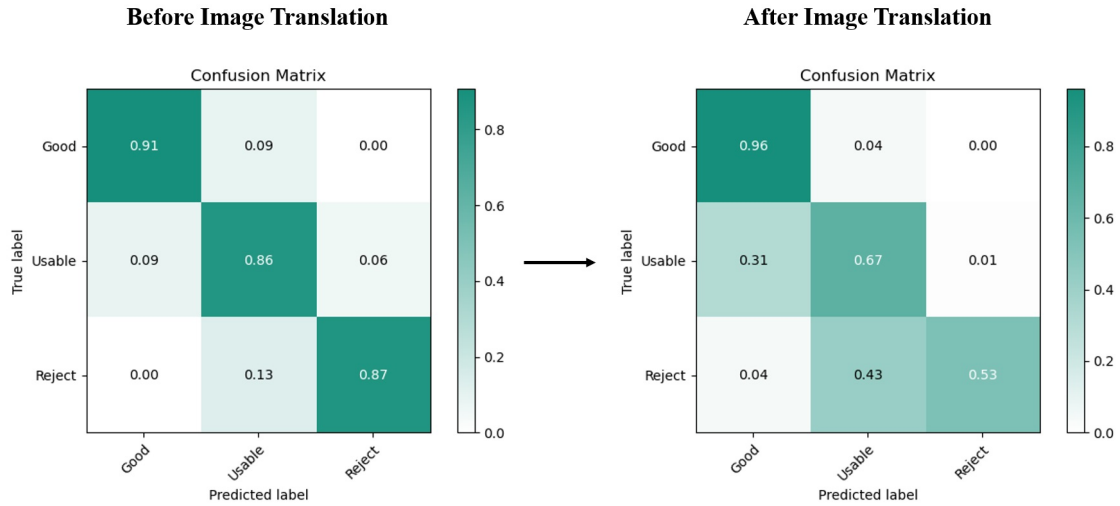


Figure 4.16: Comparison between confusion matrices before and after low-quality data set 3 retina image quality enhancement using a MCF-Net prediction. Translation examples of the employed generator are presented in the third line of Figure 4.10.

For all considered low-quality data sets, a significant quality improvement is detected with overall percentage increases above 25% and reaching maximums of around 65%. This further confirms the effectiveness of the proposed CycleGAN framework in improving retina image quality as a considerable portion of images upgraded in quality label according to our evaluator. However, some differences between the percentage increases can be noticed which justifies further analysis.

The quality transformation performed by the generator trained with the 'Usable'-labelled data set as low-quality domain assured the best performance. This is incongruous with the visual-based analysis through which a quality translation improvement is

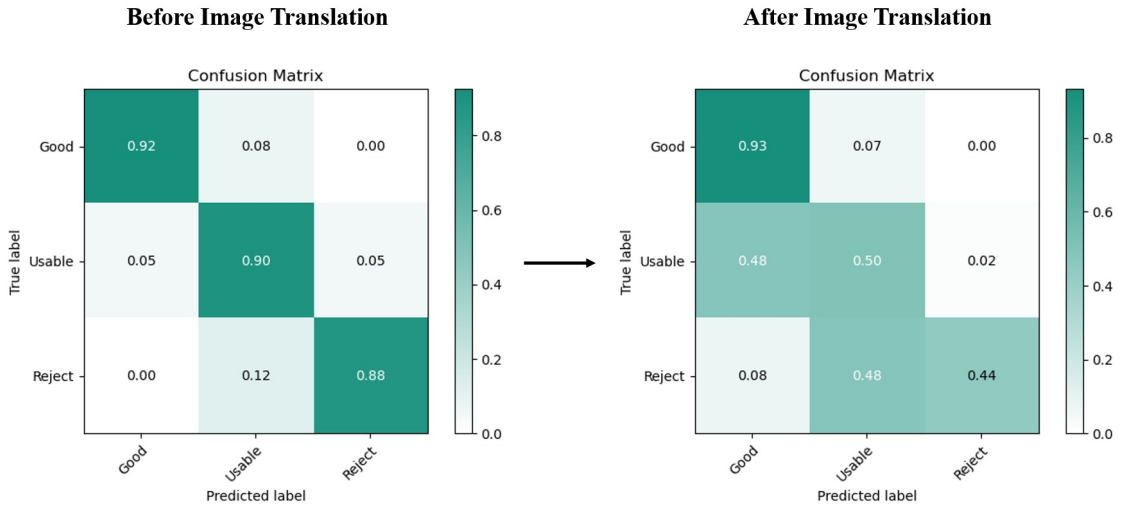


Figure 4.17: Comparison between confusion matrices before and after low-quality data set 4 retina image quality enhancement using a MCF-Net prediction. Translation examples of the employed generator are presented in the fourth line of Figure 4.10.

detected when approaching each defect separately. This higher rate of quality-class upgrades is because the used generator is trained with a low-quality set that includes only 'Usable'- labelled images. This training set considers the natural defect diverseness of the 'Usable' label and consequently, it is more representative of that class. This signifies that the image-to-image translation of low- into high-quality images is more approximated to a translation of 'Usable' into 'Good'. This explanation extends to the transformation of 'Reject' into 'Usable' as the defect of those classes is similar but with different severities. The majority of the observed defects on those label sets relate to dark low-contrast, excessive darkness of the macula, and uneven illumination due to underexposure which are the characterizing low-quality factors of the defect data set 2. Hence, we can detect a great similarity between their percentage results.

On the other hand, the appearance of choroid vessels in the backside of the retina and light low-contrast/uneven illumination due to over-exposure are the most relevant low-factors within the defect data set 3 and 1, respectively. These defects are much more present in 'Good'-labelled images from the EyeQ data set, even if in a less severe form. Consequently, the image-to-image translation of low- into high-quality images will not be such an accurate approximation of a translation of 'Usable' into 'Good'. The same extension is done for 'Reject' into 'Usable'. This explains why the percentage of images that upgraded in quality-class are inferior when using generators trained with these low-quality defect data sets.

An additional discrepancy can be detected between the results reported in the two quality transformation columns. One can observe that there is a generally higher percentage of images upgrading from 'Usable' to 'Good' than from 'Reject' to 'Usable'. In

part, this also relates to the low-quality training sets being more representative of the 'Usable' label. Despite the 'Reject' images presenting similar defect heterogeneity than the 'Usable' ones, their characterizing features are not identical. Furthermore, the defects in 'Reject'-labelled images are evidently more extreme and thus more challenging to correct while maintaining the original image identity.

All things considered, it may be concluded that the retina image quality validation through the MCF-Net quantitatively confirms a significant increase in quality-class prediction. This corroborates the quality improvement reported by the visual analysis. Moreover, it was possible to analyse how the construction of the set representing the low-quality domain used for training the CycleGAN affects the evaluation of quality improvement performance by the MCF-Net: the more similar are the domain sets to the classifier's prediction classes, the higher rate of shifts upon image-to-image translation.

Table 4.3: Quantitative assessment of retina image quality improvement.The results reported in the first row are associated to the generator which translation examples are presented in the third line of Figure 4.9. The results in the following rows are associated to the four generators which translation examples are presented in the four lines of Figure 4.10, following the same order.

	'Reject' to 'Usable' [a]	'Usable' to 'Good' [a]	Overall Improvement
'Usable'-labelled data set	0.59	0.72	0.655
Defect data set 1	0.24	0.26	0.25
Defect data set 2	0.61	0.68	0.645
Defect data set 3	0.39	0.31	0.35
Defect data set 4	0.47	0.52	0.495

[a] This constitutes an approximation as we discard sparse examples when images downgrade in quality class or when they upgrade from 'Reject' to 'Good'.

It is important to note that the classification by the MCF-Net is mostly focused on quality features. This implies that even if larger shifts indicate potential significant quality transformation, it is also positive that these are not excessive, as that would likely imply drastic changes in morphological and pathological features of the retina images.

4.2.3.2 Diabetic Retinopathy Computer-aided Diagnosis Validation

Besides validating the quality improvement of the retina images in the previous subsection, we inspect whether the quality improvement might affect the performance of CADx system using the DR_EffB0-Net proposed in Section 3.4.2.

The four defect data sets introduced in Section 3.1.3 were combined to construct a data set containing all commonly observed low-quality factors. This is similarly to what was used for testing the CycleGAN generator trained without defect isolation in the previous Section (4.2.3.1).

For the CADx validation study, we used the referred defect combination data set in three different circumstances. First in its unchanged version, where the images are only pre-processed, according to the pre-processing methods presented in Section 3.2. In the

second version, the images are translated from low- to high-quality using the CycleGAN generator trained with a low-quality domain without defect isolation. Lastly, each subset of images with a certain defect is translated from low- to high-quality using the CycleGAN generator trained with the low-quality domain correspondent to that specific defect. That is, each of the four subsets containing particular low-quality factors is transformed individually through the generator previously trained to correct the respective defects.

In summary, the DR_EffB0-Net is trained and tested using: (1) a data set with low-quality images presenting different low-quality factors; (2) the same data set but where the quality of the images is improved using the CycleGAN generator trained without defect isolation; (3) the same data set but where the images are enhanced separately using the CycleGAN generators trained to correct each of the correspondent low-quality factors.

The mentioned set contains 7,489 retina images, from which 20% are selected for validation (10%) and testing (10%). Upon that division, each set is balanced according to the DR severity grade with fewer samples, reaching around 3640 retina images for training, 440 for validation, and 440 for testing. The split in train and test is performed under a random seed condition to replicate the sequence of retina image samples in training and testing. Intuitively, this data separation and balancing are done individually for each of the three explained versions of the data set.

As aforehand mentioned, the DR_EffB0-Net model was trained and tested using each version of the considered data set. The train, validation, and test sets were split according to three random seeds for a fairer comparison. In Table 4.4, the performance of the DR_EffB0-Net in the test sets is presented through the Cohen's Kappa score.

Table 4.4: Classification performance of the DR_EffB0-Net upon training and testing before and after quality improvement.

	Cohen's Kappa coefficient		
Seed	Low-quality images	Quality-improved images ([a])	Quality-improved images ([b])
12	0.7276	0.7220	0.6666
20	0.7614	0.7666	0.6647
42	0.7051	0.7003	0.7178
μ	0.7314	0.7296	0.6830

[a] This refers to the images which quality is transformed using the generator trained without defect isolation

[b] This refers to the images which quality is transformed separately according to their low-quality factors using the generators trained for correcting each factor

As reported, the performance of the DR severity detection algorithm is very similar before and after the quality transformation process when using the CycleGAN generator that was trained with 'Usable'-labelled images with no defect isolation for composing the low-quality domain (first and second columns). For the images which qualities are transformed separately using the generators trained for correcting each corresponding defect (third column), a slight performance downgrade can be detected as the average Cohen's Kappa coefficient decreases approximately 5%.

It is believed that two reasons might justify a performance upgrade between the first and second columns not being detected.

Firstly, we will discuss the CycleGAN trained with a low-quality set that is composed of 'Usable'-labelled images with high classification coefficient from the MCF-Net. According to the explanation in [75], the mentioned label indicates that, despite the low-quality factors, the lesions in the images are clear enough to be identified. As a consequence, even if the low-quality factors are corrected and thus, the overall image quality is improved, the detection of DR signs by the AI classifier might not be greatly affected. Despite this being a possible explanation, it is incongruous with the assumption that minimizing the low-quality factors can force the algorithm to be more focused on the pathological-feature differences between the images.

Another possible and more viable explanation is that a trade-off occurs. On one hand, the quality improvement assures a more uniform data set and consequently, the algorithm becomes more fixated on detecting differences related only to DR severity characteristics. On the other hand, when looking into the EyeQ database, it can be seen that there are much fewer images with higher DR severity levels and that those often coincide with lower-quality labels. Oppositely, the images with lower levels of DR severity are seemingly more associated with higher quality labels. This apparent tendency was statistically confirmed in the EyeQ train data set upon balancing it according to the DR labels. When comparing the quality and DR level variables, a significant positive correlation is found as the obtained Spearman correlation score is 0.378 with a p-value < 0.0001. This might happen due to several reasons such as older people, who are more likely to present some level of DR, being also more probable to have cataracts or hyperopia, which can lead to certain defects in retina images. Additionally, older subjects are less likely to stay completely still during the image acquisition period.

This correlation between quality and DR severity might falsely enhance the performance of the algorithm as it associates some quality factors to particular DR levels. Therefore, while improving the images might bring more homogeneity to the image set, which theoretically should improve the pathology classification process, the augmented homogeneity can also interfere with the natural correlation between image quality and pathology signs. In practice, this can be a drawback in the performance of the DR severity classifier in the case of this data set.

Nonetheless, by maintaining the results of the performance metric and because the effect of the referred correlation is diminished/removed upon quality translation, it can be deduced that the DR_EffB0-Net algorithm becomes more sensitive to DR severity signs.

Regarding the third column, we presume that the slight performance downgrade is related to a more accentuated quality transformation. As detected in the visual-based analysis, when approaching low-quality factors in images in an individual manner (using generators from CycleGANs trained to correct specific defects), the transformed images tend to be more modified as the quality improvement is more noticeable and the defects

less perceptible. While visually the images are clearly improved in quality, in some particular cases this might transform areas where pathological signs are located. Hence, disease-related features might become hidden or less evident. We believe this might not be significant for most cases because the generators are focused on transforming low-quality regions of the retina image where, in most cases, the pathological characteristics are already not visible. This is evident e.g. for dark spots. However, in some less degrading defects, the concealing of disease signs might occur.

In addition, the positive correlation between the quality labels and DR severity levels is likely to jeopardize the maintenance of pathological features during the quality transformation. The reason for this is because there are fewer images with high-quality displaying more severe DR signs. Consequently, the CycleGAN framework might not be efficiently trained to translate low-quality images into the high-quality domain while entirely preserving disease-related characteristics (these can be perceived as low-quality factors by the network). The more accentuated is the quality translation, the more evident this phenomenon might become, and thus, less successful is the DR detection network in grading different pathology severity levels.

In Figure 4.18 we display three examples of quality image translations using the low-quality domain approach in which defect division is performed. As presented, some DR signs become less evident.

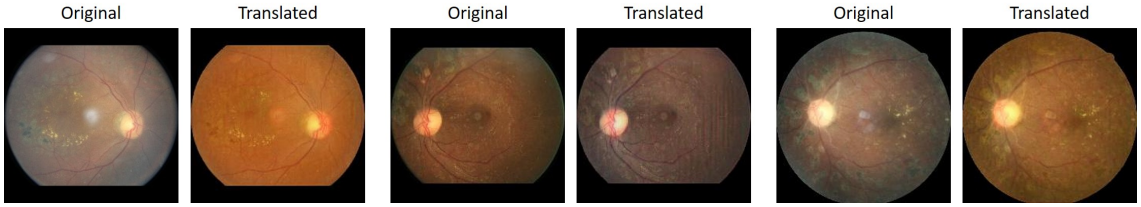


Figure 4.18: Quality translation results using the CycleGAN trained for correcting flash-induced artifacts where diabetic retinopathy signs become less clear/emphasized.

In conclusion, the objective of improving retina image quality without changing the underlying information for the diagnosis, that is, guaranteeing at minimum the same disease detection performance, can be accomplished by the CycleGAN framework.

Because the quality improvement diminishes/removes the positive correlation between DR severity and quality level, and considering that the model maintains its performance when training the CycleGAN with a low-quality domain contemplating no defect isolation, there is strong evidence that the model becomes more sensitive to DR severity signs. This implies that, for the same classification accuracy, the separation of the retina images in different DR levels is more truthful to the actual presence of DR signs. Thus, the goal of improving the efficiency of the proposed CADx system through the quality enhancement of the retina images is mostly accomplished.

Nonetheless, it can be deduced that more visual accentuated quality improvements

might jeopardize the algorithm detection of pathological signs. This occurs if the disease features become partly concealed during the quality translation process. These phenomenons are probably strongly influenced by the lack of high-quality samples with higher DR levels in contrast with the low-quality ones that frequently co-exist with the presence of more severe DR signs.

Chapter 5

Conclusion and Future Work

This Chapter summarises the developed work and the obtained results throughout this dissertation. Guidelines for future research are also proposed.

5.1 Main Conclusions

Retina image analysis is an important screening tool for early detection of multiple diseases such as Diabetic Retinopathy which greatly conditions visual function and represents the leading cause of blindness in working-age populations. Image analysis and pathology detection can be accomplished both by ophthalmologists and by Computer-aided Diagnosis systems that yield AI algorithm-based diagnosis through image data learning.

Advancements in hardware technology led to using more portable and less expensive imaging devices for medical image acquisition. This promotes large-scale remote diagnosis by clinicians as well as the implementation of CADx systems for local routine disease screening. Furthermore, it allows database expansion which further stimulates research on the detection of certain pathologies. Particularly, the development of effective AI models for diagnosis support. However, lower-cost equipment generally results in inferior quality images. This may jeopardize the reliability of the acquired images and thus hinder the overall performance of the diagnostic tool.

Throughout this research work, two main contributions were presented. The first and major part of this dissertation focused on exploring Deep Learning methodologies for transforming the quality of retina images while maintaining their underlying information for diagnosis. Three different frameworks were proposed for that end: an image reconstruction method using a VAE model and two image-to-image translation processes through the CycleGAN and UNIT architectures. The second part of the project was

focused on proposing quantitative validation methods for evaluating the quality improvement performance. Two approaches were suggested. The first consisted of employing a neural network for retina image quality evaluation before and after the quality translation and assessing the consequent quality-label prediction shifts. For the second approach, a novel DL architecture based on the EfficientNet was proposed for detecting different DR severity levels in retina images. The DR assessment model was trained and tested using both the original low-quality images and the resulting quality-improved images. Two analyses were then performed. Firstly, it was inspected if the quality improvement does not change the underlying information of the image i.e. if the DR detection performance is sustained, and secondly, if said performance can be improved due to the minimization of low-quality defects.

Experimental results demonstrate that the VAE and UNIT frameworks are not sufficiently effective in improving image quality while preserving the base morphological information. The reconstruction of high-quality images through the VAE network was not successful, as the network is only able to reproduce the background of the retina image and not all remaining relevant superficial features. Thus, its usage for reconstructing low-quality images while maximizing high-quality features was not viable. The UNIT framework, on the other hand, could not assure an optimal equilibrium between generating realistic high-quality retina images and maintaining the underlying characteristics of the original image. That is, either the outputs are realistic high-quality images but diverge from the input or they preserve the original features but fail to produce realistic retina images, particularly in the optic disc area.

Nonetheless, these methods presented promising results for other applications also relevant in this topic of research. Namely, the VAE could be employed for extracting the background of retina images, which is used in denoising processes. The UNIT framework, under certain loss weight settings, might potentially be a useful tool for data augmentation to increase the volume and diversity of training data.

It was demonstrated that using the CycleGAN framework can provide favourable results in improving retina image quality. Even so, it was shown that these results are highly influenced by the construction of the data sets representing each domain. Three different approaches for composing the low-quality set were suggested in this work. The first was using a quality separation based solely on the quality labels from the adopted database (EyeQ). The second method was similar but employed a further quality selection using the quality evaluator network for a more homogeneous set: only 'Usable'-labelled images classified as 'Usable' by the MCF-Net with a coefficient higher than 0.98 were considered. Lastly, it was proposed using four different low-quality sets containing combinations of commonly observed defects in retina images and thus correct the low-quality factors separately.

The obtained results demonstrate that when training the CycleGAN using a low-quality set with no quality selection besides the ground-truth labels, no significant image quality improvements are found. The other two approaches ensured substantial quality

improvements as defects are minimized and the images become generally closer to the high-quality domain samples. When visually comparing the two selective settings, a more accentuated improvement was detected while correcting the low-quality factors separately. This demonstrates that tackling the problem more selectively helps the CycleGAN's discriminators distinguishing low-quality image characteristics from high-quality ones.

For both successful low-quality domain construction approaches, the proposed quantitative validation methods were employed to corroborate the visual observed quality improvement, as well as its impact on a CADx system.

The predictions by the quality evaluator were considerably shifted into higher-quality labels with an average of around 48% images upgrading their quality level. The percentage increase of images evaluated as high-quality retina images after the quality translation is of 50%, on average. Hence, a significant quality improvement in the retina images is quantitatively confirmed.

Regarding the study of how the retina image quality improvement influences the performance of the DR severity detection algorithm, different results were reached for each low-quality domain approach.

The images transformed using the CycleGAN trained without low-quality factor separation led to the same performance of the CADx model as before the quality transformation. However, a natural positive correlation between the DR severity levels and the quality labels in the EyeQ data set was statistically proven through the Spearman correlation. As the quality improvement minimizes/removes this correlation and the performance accuracy is maintained, there is strong evidence that the model becomes more sensitive to the presence of DR severity signs. In other words, for the same classification accuracy, the separation of the retina images in different DR levels is more associated with the actual presence of DR signs. Therefore, it can be concluded that the quality improvement impacts the efficiency of DR detection positively.

The images transformed using the CycleGAN trained to correct each low-quality factor separately led to a slight downgrade in the performance of the CADx model. This is presumed to happen because there is a more significant modification of the images. As a consequence, some pathology characteristics located in low-quality regions might be hidden or become less evident. This phenomenon is believed to be strongly supported by the positive correlation between image quality and DR severity since few high-quality samples present severe DR levels. In contrast, the low-quality ones frequently co-exist with the presence of more severe DR signs. As a consequence, the CycleGAN is not able to learn how to optimally improve image quality while maintaining the pathological features.

In addition to the reported contributions, a parallel study using the CycleGAN proposed in this thesis was carried out. This study aimed to generate augmented data for enhancing the training process of retina disease-detection models. For that end, our

trained CycleGAN framework was employed to generate both low- and high-quality domain retina images. These were then used to construct the training dataset of a glaucoma severity classifier. The model's performance validation was done on original retina images (without transformation). The reached results demonstrate that using the proposed generative model for data augmentation leads to an improved performance of the pathology assessment network. This study resulted in a scientific paper named "Impact of GANs/synthetic data for image augmentation and image quality evaluation in Glaucoma CADx"and it is currently submitted to the IEEE Access journal.

5.2 Future Work

Despite this dissertation revealing promising results, there are still some unsolved problems and research questions that can be addressed in the future.

Firstly and most importantly, it is suggested that other retina image data sets are used in composing the low- and high-quality domains for training the CycleGAN. Particularly, image sets in which there is no significant correlation between image quality and disease presence so that both quality domains present similar pathology-label distributions. This could ensure greater preservation of the image's original pathology-related features while simultaneously allowing accentuated quality modifications. Thus, it might lead to more significant improvements in the performance of the Diabetic Retinopathy detection network.

Besides this dissertation focusing on DR diagnosis, other disease assessment algorithms could be explored such as glaucoma detection. This would allow comparing the impact of image quality improvement in the diagnosis of different retina disorders or diseases.

Regarding the obtained results, it is suggested that a subjective medical study is performed to inspect if the resulting quality improvements are corroborated by ophthalmologists. This study could allow assessing the impact of retina image quality improvement in expert-based diagnosis. The proposed evaluation was envisioned to be performed in this thesis. However, due to the current medical-collaboration constraints, it was not possible to fulfill this objective in due time.

We suggest presenting multiple pairs of retina images before and after the quality transformation to a panel of expert clinicians. The evaluators should be given a quality grading system with the three quality levels considered in this dissertation (as utilized in the EyeQ data set). Each quality level must be explained in detail regarding the presence of low-quality factors, their intensity, and the consequent impact on disease sign visualization. After all pairs of retina images being evaluated by the ophthalmologists' panel through the mentioned quality scoring system, a comparative study should be carried out. Several metrics could be extracted to assess the quality modification tendency according to the opinion of clinicians and thus inspect if a significant quality improvement is detected.

Multiple factors might influence the performance of generative neural networks. It is suggested that more experiments are performed under different settings to investigate if a more optimal CycleGAN framework and/or an improved data preparation process can be achieved. Particularly, we believe image size might have a relevant impact on the obtained results. As the images grow in size, more details are covered. Hence, the network might learn what characterizes pathology-related signs more efficiently.

Lastly, to use the proposed quality improvement tool in diagnosis systems, the construction of a fluid pipeline is suggested. The images acquired by the fundus cameras should first go through a quality assessment stage. This would allow inspecting which images should be completely rejected, which present high-quality and thus should not be modified, and the ones that need to be enhanced regarding their quality. Moreover, this quality assessment step could be useful to infer if image acquisition should be repeated.

Pre-processing methods should then be applied to prepare the retina images to be fed to neural networks. The employed modifications highly depend on the disease of interest.

If opting to perform the quality improvement process with no low-quality defect separation, a single image-to-image translation network needs to be employed. In this case, all selected retina images are modified simultaneously. On the other hand, if it is favourable to correct each specific defects individually, an additional step must be considered. This consists of applying a neural network to detect which type of low-quality factors are contained in the retina images, within all considered defects. The constructed algorithm can be trained to separate different low-quality factors which should be congruent with the ones commonly observed in the images acquired with the utilized fundus camera. In this dissertation, a low-quality image separation is suggested based on the EyeQ data set characteristics. Upon separation, the images would go through the quality transformation process using different generative networks (each trained to correct the corresponding defect).

Bibliography

[1] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification Kaiming," *Biochemical and Biophysical Research Communications*, vol. 498, no. 1, pp. 254–261, 2018, ISSN: 10902104. DOI: 10.1016/j.bbrc.2018.01.076. arXiv: 1502.01852v1.

[2] K. Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4-5, pp. 198–211, 2007, ISSN: 08956111. DOI: 10.1016/j.compmedimag.2007.02.002.

[3] S. Felgueiras, J. Costa, J. Gonçalves, and F. Soares, "Mobile-based risk assessment of diabetic retinopathy using a smartphone and adapted ophtalmoscope," *HEALTHINF 2018 - 11th International Conference on Health Informatics, Proceedings; Part of 11th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2018*, vol. 5, no. Biostec, pp. 168–175, 2018. DOI: 10.5220/0006599701680175.

[4] M. Kaur, J. Kaur, and R. Kaur, "Low cost cataract detection system using smart phone," in *Proceedings of the 2015 International Conference on Green Computing and Internet of Things, ICGCIoT 2015*, Institute of Electrical and Electronics Engineers Inc., 2016, pp. 1607–1609, ISBN: 9781467379090. DOI: 10.1109/ICGCIoT.2015.7380724.

[5] R. Cheloni, S. A. Gandolfi, C. Signorelli, and A. Odone, "Global prevalence of diabetic retinopathy: Protocol for a systematic review and meta-analysis," *BMJ Open*, vol. 9, no. 3, 2019, ISSN: 20446055. DOI: 10.1136/bmjopen-2018-022188. [Online]. Available: /pmc/articles/PMC6443069/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6443069/.

[6] *Global diabetes data report 2010 — 2045*. [Online]. Available: https://www.diabetesatlas.org/data/en/world/ (visited on 07/01/2020).

[7] N. Kourgialis, *Diabetic Retinopathy - silently blinding millions of people world-wide • IAPB Vision Atlas*. [Online]. Available: http://atlas.iapb.org/vision-trends/diabetic-retinopathy/ (visited on 06/30/2020).

[8] *Diabetic Retinopathy Detection | Kaggle*. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection (visited on 06/30/2020).

[9] B. C. S. Loh and P. H. H. Then, "Deep learning for cardiac computer-aided diagnosis: benefits, issues & solutions," *mHealth*, vol. 3, pp. 45–45, 2017. DOI: 10.21037/mhealth.2017.09.01.

[10] H. Ackar, A. A. Almisreb, and M. A. Saleh, "A Review on Image Enhancement Techniques," *Southeast Europe Journal of Soft Computing*, vol. 8, no. 1, 2019, ISSN: 2233-1859. DOI: 10.21533/scjournal.v8i1.175.

[11] M. Przytulska and J. L. Kulikowski, "Quantitative tests-based assessment of biomedical image enhancement procedures," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 205–216, 2016, ISSN: 02085216. DOI: 10.1016/j.bbe.2015.10.011.

[12] S. Kaur and P. Kaur, "Review and Analysis of Various Image Enhancement Techniques," *International Journal of Computer Applications Technology and Research*, vol. 4, no. 5, pp. 414–418, 2015. DOI: 10.7753/ijcatr0405.1016.

[13] L. Xiao, C. Li, Z. Wu, and T. Wang, "An enhancement method for X-ray image via fuzzy noise removal and homomorphic filtering," *Neurocomputing*, vol. 195, pp. 56–64, 2016, ISSN: 18728286. DOI: 10.1016/j.neucom.2015.08.113.

[14] S. Wang, J. Zheng, H. M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3538–3548, 2013, ISSN: 10577149. DOI: 10.1109/TIP.2013.2261309.

[15] S. Shashi B. Rana, "A Review of Medical Image Enhancement Techniques for Image Processing," *International Journal of Current Engineering and Technology*, vol. 5, no. 2, pp. 1282–1286, 2011, ISSN: 23475161. DOI: 10.14741/ijcet/22774106/5.2.2015.121.

[16] J. B. Zimmerman, S. M. Pizer, E. V. Staab, J. R. Perry, W. Mccartney, and B. C. Brenton, "An Evaluation of the Effectiveness of Adaptive Histogram Equalization for Contrast Enhancement," *IEEE Transactions on Medical Imaging*, vol. 7, no. 4, pp. 304–312, 1988, ISSN: 1558254X. DOI: 10.1109/42.14513.

[17] M. Moniruzzaman, M. Shafuzzaman, and M. F. Hossain, "Brightness preserving Bi-histogram equalization using edge pixels information," in *2013 International Conference on Electrical Information and Communication Technology, EICT 2013*, IEEE Computer Society, 2014, ISBN: 9781479922994. DOI: 10.1109/EICT.2014.6777872.

[18] N. M. Sasi and V. K. Jayasree, "Contrast Limited Adaptive Histogram Equalization for Qualitative Enhancement of Myocardial Perfusion Images," *Engineering*, no. 5, pp. 326–331, 2013. DOI: 10.4236/eng.2013.510B066. [Online]. Available: http://dx.http://www.scirp.org/journal/eng.

[19] Z. Y. Chen, B. R. Abidi, D. L. Page, and M. A. Abidi, "Gray-level grouping (GLG): An automatic method for optimized image contrast enhancement - Part I: The basic method," *IEEE Transactions on Image Processing*, vol. 15, no. 8, pp. 2290–2302, 2006, ISSN: 10577149. DOI: 10.1109/TIP.2006.875204.

[20] K. Hasikin and N. A. M. Isa, "Enhancement of the low contrast image using fuzzy set theory," in *Proceedings - 2012 14th International Conference on Modelling and Simulation, UKSim 2012*, 2012, pp. 371–376, ISBN: 9780769546827. DOI: 10.1109/UKSim.2012.60.

[21] G. Raju and M. S. Nair, "A fast and efficient color image enhancement method based on fuzzy-logic and histogram," *AEU - International Journal of Electronics and Communications*, vol. 68, no. 3, pp. 237–243, 2014, ISSN: 14348411. DOI: 10.1016/j.aeue.2013.08.015.

[22] L. J. Wang and Y. C. Huang, "Combined opportunity cost and image classification for non-linear image enhancement," in *Proceedings - 2012 6th International Conference on Complex, Intelligent, and Software Intensive Systems, CISIS 2012*, 2012, pp. 135–140, ISBN: 9780769546872. DOI: 10.1109/CISIS.2012.120.

[23] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman, "Reflection removal using ghosting cues," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, IEEE Computer Society, 2015, pp. 3193–3201, ISBN: 9781467369640. DOI: 10.1109/CVPR.2015.7298939.

[24] T. Li, D. P. Lun, Y. H. Chan, and Budianto, "Robust reflection removal based on light field imaging," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1798–1812, 2019, ISSN: 10577149. DOI: 10.1109/TIP.2018.2880510.

[25] Z. Chi, X. Wu, X. Shu, and J. Gu, "Single Image Reflection Removal Using Deep Encoder-Decoder Network," *Proceedings of the IEEE conference on computer vision and pattern Recognition (CVPR)*, pp. 4498–4506, 2018. arXiv: 1802.00094. [Online]. Available: http://arxiv.org/abs/1802.00094.

[26] J. Gu, P. Belhumeur, S. Nayar, and R. Ramamoorthi, "Removing Image Artifacts Due to Dirty Camera Lenses and Thin Occluders," *ACM Transactions on Graphics*, vol. 28, no. 5, pp. 1–10, 2009, ISSN: 15577368. DOI: 10.1145/1618452.1618490.

[27] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising," *CoRR*, vol. abs/1608.03981, 2016. arXiv: 1608.03981. [Online]. Available: http://arxiv.org/abs/1608.03981.

[28] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with BM3D?" In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2392–2399, ISBN: 9781467312264. DOI: 10.1109/CVPR.2012.6247952.

[29] J. Xie, L. Xu, and E. Chen, "Image Denoising and Inpainting with Deep Neural Networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 350–358.

[30] T. Wang, M. Sun, and K. Hu, "Dilated residual network for image denoising," *CoRR*, vol. abs/1708.05473, 2017. arXiv: 1708.05473. [Online]. Available: http://arxiv.org/abs/1708.05473.

[31] W. Bae, J. J. Yoo, and J. C. Ye, "Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification," *CoRR*, vol. abs/1611.06345, 2016. arXiv: 1611.06345. [Online]. Available: http://arxiv.org/abs/1611.06345.

[32] W. Jifara, F. Jiang, S. Rho, M. Cheng, and S. Liu, "Medical image denoising using convolutional neural network: A residual learning approach," *Journal of Supercomputing*, vol. 75, pp. 1–15, Jun. 2017. DOI: 10.1007/s11227-017-2080-0.

[33] M. Gholizadeh-Ansari, J. Alirezaie, and P. Babyn, "Low-dose ct denoising with dilated residual network," vol. 2018, Jul. 2018, pp. 5117–5120. DOI: 10.1109/EMBC.2018.8513453.

[34] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017, ISSN: 10577149. DOI: 10.1109/TIP.2017.2662206. arXiv: 1608.03981.

[35] A. Majumdar, "Blind denoising autoencoder," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, pp. 1–6, Jun. 2018. DOI: 10.1109/TNNLS.2018.2838679.

[36] J. Chen, J. Chen, H. Chao, and M. Yang, "Image blind denoising with generative adversarial network based noise modeling," Jun. 2018, pp. 3155–3164. DOI: 10.1109/CVPR.2018.00333.

[37] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," *CoRR*, vol. abs/1807.04686, 2018. arXiv: 1807.04686. [Online]. Available: http://arxiv.org/abs/1807.04686.

[38] L. Tao, C. Zhu, J. Song, T. Lu, H. Jia, and X. Xie, "Low-light image enhancement using CNN and bright channel prior," in *Proceedings - International Conference on Image Processing, ICIP*, vol. 2017-September, IEEE Computer Society, 2018, pp. 3215–3219, ISBN: 9781509021758. DOI: 10.1109/ICIP.2017.8296876.

[39] L. Tao, C. Zhu, G. Xiang, Y. Li, H. Jia, and X. Xie, "LLCNN: A convolutional neural network for low-light image enhancement," in *2017 IEEE Visual Communications and Image Processing, VCIP 2017*, vol. 2018-January, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 1–4, ISBN: 9781538604625. DOI: 10.1109/VCIP.2017.8305143.

[40] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, Nov. 2015. DOI: 10.1016/j.patcog.2016.06.008.

[41] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," *CoRR*, vol. abs/1808.04560, 2018. arXiv: 1808.04560. [Online]. Available: http://arxiv.org/abs/1808.04560.

[42] J. Wang, W. Tan, X. Niu, and B. Yan, "Rdgan: Retinex decomposition based adversarial learning for low-light enhancement," Jul. 2019, pp. 1186–1191. DOI: 10.1109/ICME.2019.00207.

[43] A. Nazemi, S. Kamyab, Z. Azimifar, and P. Fieguth, "Human perception-based image enhancement using a deep generative model," *Journal of Computational Vision and Imaging Systems*, vol. 4, no. 1, pp. 3–3, 2018.

[44] Y. Atoum, M. Ye, L. Ren, Y. Tai, and X. Liu, "Color-wise attention network for low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

[45] W. Ren, S. Liu, L. Ma, Q. Xu, X. Xu, X. Cao, J. Du, and M. H. Yang, "Low-Light Image Enhancement via a Deep Hybrid Network," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4364–4375, 2019, ISSN: 19410042. DOI: 10.1109/TIP.2019.2910412.

[46] C. Dong, Y. Deng, C. C. Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," *CoRR*, vol. abs/1504.06993, 2015. arXiv: 1504.06993. [Online]. Available: http://arxiv.org/abs/1504.06993.

[47] P. Svoboda, M. Hradis, D. Barina, and P. Zemcík, "Compression artifacts removal using convolutional neural networks," *CoRR*, vol. abs/1605.00366, 2016. arXiv: 1605.00366. [Online]. Available: http://arxiv.org/abs/1605.00366.

[48] L. Cavigelli, P. Hager, and L. Benini, "CAS-CNN: A deep convolutional neural network for image compression artifact suppression," *CoRR*, vol. abs/1611.07233, 2016. arXiv: 1611.07233. [Online]. Available: http://arxiv.org/abs/1611.07233.

[49] Y. Sun, Y. Yu, and W. Wang, "Moiré photo restoration using multiresolution convolutional neural networks," *CoRR*, vol. abs/1805.02996, 2018. arXiv: 1805.02996. [Online]. Available: http://arxiv.org/abs/1805.02996.

[50] J. Yang, D. Gong, L. Liu, and Q. Shi, "Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[51] H. Li and K. Mueller, "Low-dose ct streak artifacts removal using deep residual neural network," 2017.

[52] Z. Chao and H.-J. Kim, "Removal of computed tomography ring artifacts via radial basis function artificial neural networks," *Physics in Medicine and Biology*, vol. 64, Oct. 2019. DOI: 10.1088/1361-6560/ab5035.

[53] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis, "Fully dense unet for 2-d sparse photoacoustic tomography artifact removal," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 568–576, 2020.

[54] C. Khare and K. Nagwanshi, "Image restoration in neural network domain using back propagation network approach," *Image*, vol. 2, Jan. 2011.

[55] Y. Tian, *ReFocus: Making Out-of-Focus Microscopy Images In-Focus Again | by Yuan Tian | Towards Data Science*. [Online]. Available: https://towardsdatascience.com/refocus-making-out-of-focus-microscopy-images-in-focus-again-90e1fe98ead4 (visited on 08/01/2020).

[56] V. Tabora, *Applying Neural Network Image Enhancement Using CNN-based Deep-Image API | by Vince Tabora | 0xCODE | Jun, 2020 | Medium*. [Online]. Available: https://medium.com/0xcode/applying-neural-network-image-enhancement-using-cnn-based-deep-image-api-f3f43df1fbe (visited on 07/31/2020).

[57] W. Yang, X. Zhang, Y. Tian, W. Wang, and J. Xue, "Deep learning for single image super-resolution: A brief review," *CoRR*, vol. abs/1808.03344, 2018. arXiv: 1808.03344. [Online]. Available: http://arxiv.org/abs/1808.03344.

[58] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *CoRR*, vol. abs/1501.00092, 2015. arXiv: 1501.00092. [Online]. Available: http://arxiv.org/abs/1501.00092.

[59] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," *CoRR*, vol. abs/1608.00367, 2016. arXiv: 1608.00367. [Online]. Available: http://arxiv.org/abs/1608.00367.

[60] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," *CoRR*, vol. abs/1609.05158, 2016. arXiv: 1609.05158. [Online]. Available: http://arxiv.org/abs/1609.05158.

[61] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," *CoRR*, vol. abs/1511.04587, 2015. arXiv: 1511.04587. [Online]. Available: http://arxiv.org/abs/1511.04587.

[62] *Super-Resolution Deep Learning: Making the Future Clearer - MissingLink.ai*. [Online]. Available: https://missinglink.ai/guides/computer-vision/super-resolution-deep-learning-making-future-clearer/ (visited on 07/31/2020).

[63] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *CoRR*, vol. abs/1609.04802, 2016. arXiv: 1609.04802. [Online]. Available: http://arxiv.org/abs/1609.04802.

[64] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," *CoRR*, vol. abs/1707.02921, 2017. arXiv: 1707.02921. [Online]. Available: http://arxiv.org/abs/1707.02921.

[65] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, and T. S. Huang, "Wide activation for efficient and accurate image super-resolution," *CoRR*, vol. abs/1808.08718, 2018. arXiv: 1808.08718. [Online]. Available: http://arxiv.org/abs/1808.08718.

[66] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4809–4817.

[67] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[68] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," *CoRR*, vol. abs/1809.00437, 2018. arXiv: 1809.00437. [Online]. Available: http://arxiv.org/abs/1809.00437.

[69] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. V. Gool, "DSLR-Quality Photos on Mobile Devices with Deep Convolutional Networks," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 3297–3305, ISBN: 9781538610329. DOI: 10.1109/ICCV.2017.355. arXiv: 1704.02470.

[70] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool, "Wespe: Weakly supervised photo enhancer for digital cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018.

[71] Y. S. Chen, Y. C. Wang, M. H. Kao, and Y. Y. Chuang, "Deep Photo Enhancer: Unpaired Learning for Image Enhancement from Photographs with GANs," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2018, pp. 6306–6314, ISBN: 9781538664209. DOI: 10.1109/CVPR.2018.00660.

[72] J. Park, K. D. Han, and H. Ko, "Adaptive weighted multi-discriminator cyclegan for underwater image enhancement," *Journal of Marine Science and Engineering*, 2019.

[73] P. Welander, S. Karlsson, and A. Eklund, "Generative adversarial networks for image-to-image translation on multi-contrast MR images - A comparison of cyclegan and UNIT," *CoRR*, vol. abs/1806.07777, 2018. arXiv: 1806.07777. [Online]. Available: http://arxiv.org/abs/1806.07777.

[74] P. Costa, A. J. C. Campilho, B. Hooi, A. Smailagic, K. Kitani, S. Liu, C. Faloutsos, and A. Galdran, "Eyequal: Accurate, explainable, retinal image quality assessment.," in *ICMLA*, X. Chen, B. Luo, F. Luo, V. Palade, and M. A. Wani, Eds., IEEE, 2017, pp. 323–330, ISBN: 978-1-5386-1418-1. [Online]. Available: http://dblp.uni-trier.de/db/conf/icmla/icmla2017.html#0005CHSKLFG17.

[75] H. Fu, B. Wang, J. Shen, S. Cui, Y. Xu, J. Liu, and L. Shao, "Evaluation of Retinal Image Quality Assessment Networks in Different Color-spaces," *MICCAI*, 2019. DOI: 10.1007/978-3-030-32239-7_6. arXiv: 1907.05345. [Online]. Available: http://arxiv.org/abs/1907.05345http://dx.doi.org/10.1007/978-3-030-32239-7{_}6.

[76] A. M. Bandara and P. W. Giragama, "A retinal image enhancement technique for blood vessel segmentation algorithm," in *2017 IEEE International Conference on Industrial and Information Systems, ICIIS 2017 - Proceedings*, vol. 2018-January, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 1–5, ISBN: 9781538616741. DOI: 10.1109/ICIINFS.2017.8300426.

[77] M. S. Miri and A. Mahloojifar, "A comparison study to evaluate retinal image enhancement techniques," in *ICSIPA09 - 2009 IEEE International Conference on Signal and Image Processing Applications, Conference Proceedings*, 2009, pp. 90–94, ISBN: 9781424455614. DOI: 10.1109/ICSIPA.2009.5478726.

[78] A. W. Setiawan, T. R. Mengko, O. S. Santoso, and A. B. Suksmono, "Color retinal image enhancement using CLAHE," in *Proceedings - International Conference on ICT for Smart Society 2013: "Think Ecosystem Act Convergence", ICISS 2013*, 2013, pp. 215–217, ISBN: 9781479901456. DOI: 10.1109/ICTSS.2013.6588092.

[79] N. S. Datta, H. S. Dutta, M. De, and S. Mondal, "An Effective Approach: Image Quality Enhancement for Microaneurysms Detection of Non-dilated Retinal Fundus Image," *Procedia Technology*, vol. 10, pp. 731–737, 2013, ISSN: 22120173. DOI: 10.1016/j.protcy.2013.12.416.

[80] P. Dai, H. Sheng, J. Zhang, L. Li, J. Wu, and M. Fan, "Retinal Fundus Image Enhancement Using the Normalized Convolution and Noise Removing," *International Journal of Biomedical Imaging*, vol. 2016, 2016, ISSN: 16874196. DOI: 10.1155/2016/5075612.

[81] T. A. Soomro and J. Gao, "Neural network based denoised methods for retinal fundus images and mri brain images," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 1151–1157.

[82] S. Ghosh, B. Biswas, and A. Ghosh, "Sdca: A novel stack deep convolutional autoencoder - an application on retinal image denoising," *IET Image Processing*, Oct. 2019. DOI: 10.1049/iet-ipr.2018.6582.

[83] Fraunhofer Portugal, *Eye fundus scope. mobile-based risk assessment of diabetic retinopathy by image processing*, Brochure, 2018. [Online]. Available: https://www.aicos.fraunhofer.pt/content/dam/portugal/en/documents/Flyers/EyeFundusScope\%20-\%20FhP-AICOS\%20Project\%20Brochure.pdf.

[84] *What is Artificial Intelligence? How Does AI Work? - Built In*. [Online]. Available: https://builtin.com/artificial-intelligence (visited on 01/09/2020).

[85] C. M. Bishop, *Pattern Recogniton and Machine Learning*. Springer-Verlag New York, 2006, p. 738, ISBN: 978-0-387-31073-2. [Online]. Available: https://www.springer.com/gp/book/9780387310732http://users.isr.ist.utl.pt/{~}wurmd/Livros/school/Bishop-PatternRecognitionAndMachineLearning-Springer2006.pdf.

[86] *What Is Deep Learning? | How It Works, Techniques & Applications - MATLAB & Simulink*. [Online]. Available: https://www.mathworks.com/discovery/deep-learning.html (visited on 01/07/2020).

[87] J. Vergara and P. Estevez, "A review of feature selection methods based on mutual information," *Neural Computing and Applications*, vol. 24, Jan. 2014. DOI: 10.1007/s00521-013-1368-0.

[88] F. Shaheen, B. Verma, and M. Asafuddoula, "Impact of automatic feature extraction in deep learning architecture," in *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2016, pp. 1–8.

[89] J. McCaffrey, *Neural Network Train-Validate-Test Stopping – Visual Studio Magazine*. [Online]. Available: https://visualstudiomagazine.com/articles/2015/05/01/train-validate-test-stopping.aspx (visited on 01/12/2020).

[90] A. Mittal, *Cross Validation and Performance Measures in Machine Learning | by Aditi Mittal | Towards Data Science*. [Online]. Available: https://towardsdatascience.com/cross-validation-and-performance-measures-in-machine-learning-9dabdbed5459 (visited on 08/03/2020).

[91] J. Brownlee, *Classification Accuracy is Not Enough: More Performance Measures You Can Use*. [Online]. Available: https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/ (visited on 08/10/2020).

[92] B. Carremans, *Handling overfitting in deep learning models | by Bert Carremans | Towards Data Science*. [Online]. Available: https://towardsdatascience.com/handling-overfitting-in-deep-learning-models-c760ee047c6e (visited on 08/06/2020).

[93] A. Canziani, A. Paszke, and E. Culurciello, "An Analysis of Deep Neural Network Models for Practical Applications," *ArXiv*, vol. abs/1802.0, 2016. arXiv: 1605.07678. [Online]. Available: http://arxiv.org/abs/1605.07678.

[94] E. A. da Silva and G. V. Mendonca, "Digital Image Processing," in *The Electrical Engineering Handbook*, Elsevier Inc., 2005, pp. 891–910, ISBN: 9780121709600. DOI: 10.1016/B978-012170960-0/50064-5.

[95] R. M. Rangayyan, "Introduction to Enhancement," in *Handbook of Medical Imaging*, Elsevier, 2000, pp. 1–2. DOI: 10.1016/b978-012077790-7/50003-5.

[96] S. Rani, S. Jindal, and B. Kaur, "A Brief Review on Image Restoration Techniques," Tech. Rep. 12, 2016, pp. 975–8887.

[97] D. V. Rojatkar, N. D. Borkar, B. R. Naik, and R. N. Peddiwar, "Image Compression Techniques: Lossy and Lossless," *International Journal of Engineering Research and General Science*, vol. 3, no. 2, 2015, ISSN: 2091-2730. [Online]. Available: www.ijergs.org.

[98] M. Elgendy, *Deep Learning for Vision Systems*. Manning Publications, 2020, ISBN: 9781617296192. [Online]. Available: https://books.google.pt/books?id=6gkLzAEACAAJ.

[99] J. Brownlee, *How to Normalize, Center, and Standardize Image Pixels in Keras*. [Online]. Available: https://machinelearningmastery.com/how-to-normalize-center-and-standardize-images-with-the-imagedatagenerator-in-keras/ (visited on 08/04/2020).

[100] *Understanding Digital Image Interpolation*. [Online]. Available: https://www.cambridgeincolour.com/tutorials/image-interpolation.htm (visited on 07/13/2020).

[101] J. Yim and K. Sohn, "Enhancing the performance of convolutional neural networks on quality degraded datasets," in *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2017, pp. 1–8.

[102] M. Strączkiewicz and T. Barszcz, "Application of artificial neural network for damage detection in planetary gearbox of wind turbine," *Shock and Vibration*, vol. 2016, no. 4, pp. 1–12, 2016, ISSN: 10709622. DOI: 10.1155/2016/4086324.

[103] J. Torres, *Learning process of a neural network - Towards Data Science*. [Online]. Available: https://towardsdatascience.com/how-do-artificial-neural-networks-learn-773e46399fc7 (visited on 01/07/2020).

[104] Y. Ren, *A Step-by-Step Implementation of Gradient Descent and Backpropagation - Towards Data Science*. [Online]. Available: https://towardsdatascience.com/a-step-by-step-implementation-of-gradient-descent-and-backpropagation-d58bda486110 (visited on 01/07/2020).

[105] J. Brownlee, *A Gentle Introduction to Generative Adversarial Networks (GANs) - Machine Learning Mastery*. [Online]. Available: https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/ (visited on 01/08/2020).

[106] S. Saha, *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way*. [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 (visited on 01/07/2020).

[107] Z. Qin, F. Yu, C. Liu, and X. Chen, "How convolutional neural network see the world - a survey of convolutional neural network visualization methods," *Mathematical Foundations of Computing*, vol. 1, p. 149, Apr. 2018. DOI: 10.3934/mfc.2018008.

[108] *Unsupervised Feature Learning and Deep Learning Tutorial - UFLDL Tutorial*. [Online]. Available: http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/ (visited on 01/08/2020).

[109] J. Rocca, *Understanding Variational Autoencoders (VAEs) - Towards Data Science*, 2019. [Online]. Available: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73 (visited on 06/17/2020).

[110] I. Shafkat, *Intuitively Understanding Variational Autoencoders - Towards Data Science*. [Online]. Available: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf (visited on 01/08/2020).

[111] X. Liu and C. Hsieh, "From adversarial training to generative adversarial networks," *CoRR*, vol. abs/1807.10454, 2018. arXiv: 1807.10454. [Online]. Available: http://arxiv.org/abs/1807.10454.

[112] C. Nicholson, *A Beginner's Guide to Generative Adversarial Networks (GANs) - Pathmind*. [Online]. Available: https://pathmind.com/wiki/generative-adversarial-network-gan (visited on 01/08/2020).

[113] R. Vijay, *Image-to-Image Translation using CycleGAN Model - Towards Data Science*. [Online]. Available: https://towardsdatascience.com/image-to-image-translation-using-cyclegan-model-d58cfff04755 (visited on 01/08/2020).

[114] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16, Barcelona, Spain: Curran Associates Inc., 2016, 469–477, ISBN: 9781510838819.

[115] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, 700–708, ISBN: 9781510860964.

[116] A. Larsen, S. Sønderby, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," Dec. 2015.

[117] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385.

[118] K. Doi, "Overview on research and development of computer-aided diagnostic schemes," *Seminars in Ultrasound, CT and MRI*, vol. 25, no. 5, pp. 404–410, 2004, ISSN: 08872171. DOI: 10.1053/j.sult.2004.02.006.

[119] A. Goyal, M. K. Arya, R. Agrawal, D. Agrawal, G. Hossain, and R. Challoo, "Automated segmentation of gray and white matter regions in brain MRI images for computer aided diagnosis of neurodegenerative diseases," in *IMPACT 2017 - International Conference on Multimedia, Signal Processing and Communication Technologies*, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 204–208, ISBN: 9781509066742. DOI: 10.1109/MSPCT.2017.8364005.

[120] R. F. Mansour, "Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy," *Biomedical Engineering Letters*, vol. 8, no. 1, pp. 41–57, 2018, ISSN: 2093985X. DOI: 10.1007/s13534-017-0047-y.

[121] D. Kaur and G. K. Walia, "Edge detection of Malaria parasites using ant colony optimization," in *4th IEEE International Conference on Signal Processing, Computing and Control, ISPCC 2017*, vol. 2017-January, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 451–456, ISBN: 9781509058389. DOI: 10.1109/ISPCC.2017.8269721.

[122] S. L. Fernandes, B. Chakraborty, V. P. Gurupur, and P. G. Ananth, "Early Skin Cancer Detection Using Computer Aided Diagnosis Techniques," *Journal of Integrated Design and Process Science*, vol. 20, no. 1, pp. 33–43, 2016, ISSN: 10920617. DOI: 10.3233/jid-2016-0002.

[123] T. Maier, D. Kulichova, K. Schotten, R. Astrid, T. Ruzicka, C. Berking, and A. Udrea, "Accuracy of a smartphone application using fractal image analysis of pigmented moles compared to clinical diagnosis and histological result," *Journal of the European Academy of Dermatology and Venereology*, vol. 29, no. 4, pp. 663–667, 2015, ISSN: 14683083. DOI: 10.1111/jdv.12648.

[124] J. Shiraishi, Q. Li, D. Appelbaum, and K. Doi, "Computer-aided diagnosis and artificial intelligence in clinical imaging," in *Seminars in Nuclear Medicine*, 6. W.B. Saunders, 2011, vol. 41, pp. 449–462. DOI: 10.1053/j.semnuclmed.2011.06.004.

[125] H. Lee and Y. P. P. Chen, "Image based computer aided diagnosis system for cancer detection," in *Expert Systems with Applications*, 12. Elsevier Ltd, 2015, vol. 42, pp. 5356–5365. DOI: 10.1016/j.eswa.2015.02.005.

[126] *EyePACS*. [Online]. Available: http://www.eyepacs.com/data-analysis (visited on 07/08/2020).

[127] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

[128] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 2242–2251, ISBN: 9781538610329. DOI: 10.1109/ICCV.2017.244. arXiv: 1703.10593.

[129] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. arXiv: 1505.04597. [Online]. Available: http://arxiv.org/abs/1505.04597.

[130] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, vol. abs/1611.07004, 2016. arXiv: 1611.07004. [Online]. Available: http://arxiv.org/abs/1611.07004.

[131] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, Long Beach, California, USA: PMLR, 2019, pp. 6105–6114. [Online]. Available: http://proceedings.mlr.press/v97/tan19a.html.

[132] A. Kızrak, *Reviewing EfficientNet: Increasing the Accuracy and Robustness of CNNs | by Ayyüce Kızrak | Heartbeat*. [Online]. Available: https://heartbeat.fritz.ai/reviewing-efficientnet-increasing-the-accuracy-and-robustness-of-cnns-6aaf411fc81d (visited on 07/17/2020).

[133] V. Vargas Yun, P. A. Gutiérrez, and C. Martínez, "Cumulative link models for deep ordinal classification," *Neurocomputing*, Mar. 2020. DOI: 10.1016/j.neucom.2020.03.034.

[134] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16, New York, NY, USA: JMLR.org, 2016, 1558–1566.

[135] A. Borji, "Pros and cons of GAN evaluation measures," *CoRR*, vol. abs/1802.03446, 2018. arXiv: 1802.03446. [Online]. Available: http://arxiv.org/abs/1802.03446.

[136] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: The kappa statistic.," *Family Medicine*, vol. 37.5, pp. 360–363, 2005.

[137] M. McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, vol. 22, pp. 276–82, Oct. 2012. DOI: 10.11613/BM.2012.031.

Appendix A

Image improvement experiment through the VAE-GAN framework

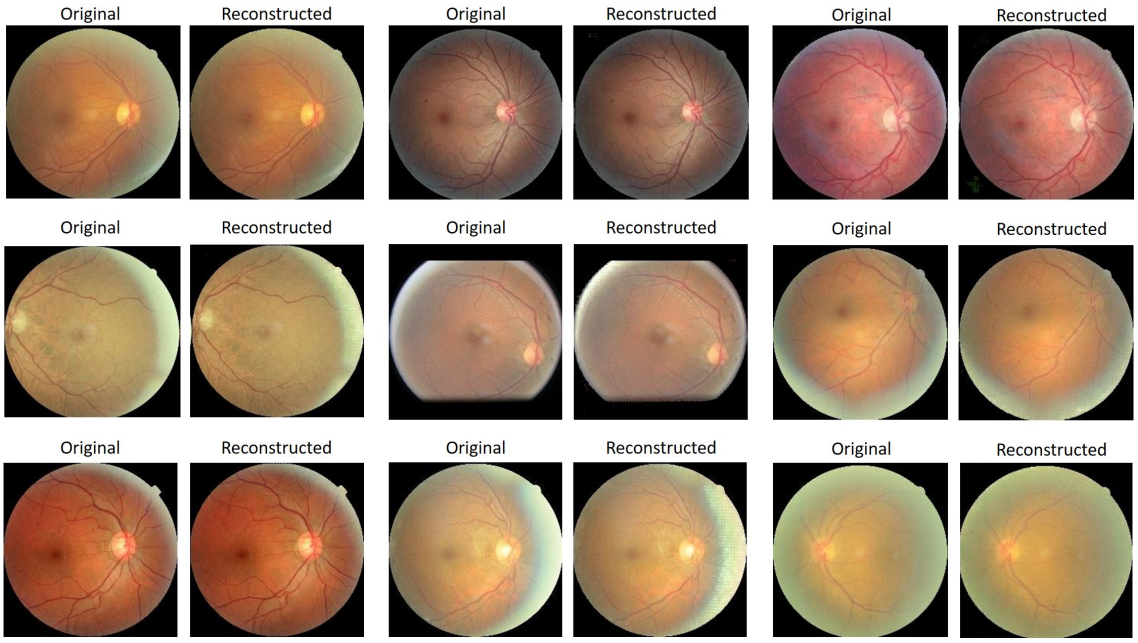


Figure A.1: Example results of low-quality image (from the low-quality data set 1) reconstruction processes using a VAE-GAN network trained for reconstructing high-quality retina images. These low-quality images integrate the low-quality dataset 1.