T.C.

BAHÇEŞEHİR UNIVERSITY

GRADUATE SCHOOL

THE DEPARTMENT OF ARTIFICIAL INTELLIGENCE ENGINEERING

A DEEP LEARNING FRAMEWORK FOR RETINAL IMAGE ANALYSIS

MASTER'S THESIS

ALNUR ALIMANOV

ISTANBUL 2024

T.C.
BAHÇEŞEHİR UNIVERSITY
GRADUATE SCHOOL
THE DEPARTMENT OF ARTIFICIAL INTELLIGENCE ENGINEERING

A DEEP LEARNING FRAMEWORK FOR RETINAL IMAGE ANALYSIS

MASTER'S THESIS
ALNUR ALIMANOV

THESIS ADVISOR
DR. MD BAHARUL ISLAM

ISTANBUL 2024

T.C.

BAHÇEŞEHİR UNIVERSITY

GRADUATE SCHOOL

MASTER THESIS APPROVAL FORM

Program name:	Artificial Intelligence Engineering
Student's Name and Surname:	Alnur Alimanov
Name Of The Thesis:	A Deep Learning Framework for Retinal Image Analysis
Thesis Defence Date:	May 23, 2024

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

. . . .

Institute Director

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

	Title/Name	**Institution**	**Signature**
Thesis Advisor's			
Member's			
Member's			

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: **Alnur Alimanov**

Signature: ....

ABSTRACT

A DEEP LEARNING FRAMEWORK FOR RETINAL IMAGE ANALYSIS

Alnur Alimanov
Master's Program in Artificial Intelligence Engineering
Thesis Advisor: Assist. Prof. Md Baharul Islam

May 2024, 73 pages

Experts use retinal images and vessel trees to detect and diagnose various eye, blood circulation, and brain-related diseases. These issues can be detected by ophthalmologists by analysing the colorful fundus images or various characteristics of vessel trees, such as length, width, curvature and shape. However, these images often happen to be of too poor quality due to device configuration and misoperations, leading to decreased efficiency, extra expenses, and even incorrect diagnoses. Additionally, experts require large high-resolution retinal images to detect tiny abnormalities, such as microaneurysms or issues of vascular branches, but these images often suffer from low image resolution. Furthermore, retinal segmentation is a challenging task due to several factors, such as the lack of extensive retinal segmentation datasets, privacy issues, and the lack of expertise in the field. Therefore, in this work we are proposing a framework for retinal image analysis, that includes retinal image quality enhancement, super-resolution, segmentation dataset generation, and retinal vessel segmentation. In particular, a Cycle-consistent Generative Adversarial Network, a hybrid super-resolution network, Denoising Diffusion Probabilistic Model, and a modified UNet are developed in this work.

Keywords: Retinal Images, Image Enhancement, Segmentation, Super-resolution, Image Generation.

ÖZ

RETİNA GÖRÜNTÜ ANALİZİ İÇİN DERİN ÖĞRENME ÇERÇEVESİ

Alnur Alimanov
Master’s Program in Artificial Intelligence Engineering
Danışman: Assist. Prof. Md Baharul Islam

Mayıs 2024, 73 sayfa

Uzmanlar, çeşitli göz, kan dolaşımı ve beyinle ilgili hastalıkları tespit etmek ve teşhis etmek için retina görüntülerini ve damar ağaçlarını kullanıyor. Bu sorunlar oftalmologlar tarafından renkli fundus görüntüleri veya damar ağaçlarının uzunluk, genişlik, eğrilik ve şekil gibi çeşitli özellikleri analiz edilerek tespit edilebilir. Ancak, cihaz konfigürasyonu ve yanlış işlemlerden dolayı bu görüntülerin kalitesi genellikle çok düşük oluyor ve bu da verimliliğin azalmasına, ekstra masraflara ve hatta yanlış teşhislere yol açıyor. Ek olarak uzmanlar, mikroanevrizmalar veya damar dalları sorunları gibi küçük anormallikleri tespit etmek için büyük, yüksek çözünürlüklü retina görüntülerine ihtiyaç duyar, ancak bu görüntüler genellikle düşük görüntü çözünürlüğünden muzdariptir. Ayrıca retina segmentasyonu, kapsamlı retina segmentasyon veri setlerinin eksikliği, gizlilik sorunları ve bu alanda uzmanlık eksikliği gibi çeşitli faktörlerden dolayı zorlu bir iştir. Bu nedenle, bu çalışmada retinal görüntü kalitesi iyileştirmesi, süper çözünürlük, segmentasyon veri seti oluşturma ve retinal damar segmentasyonunu içeren retina görüntü analizi için bir çerçeve öneriyoruz. Bu çalışmada özellikle Döngüyle tutarlı bir Üretken Çekişmeli Ağ, hibrit bir süper çözünürlüklü ağ, Gürültü Giderici Difüzyon Olasılık Modeli ve değiştirilmiş bir UNet geliştirildi.

Keywords: Retina Görüntüleri, Görüntü İyileştirme, Segmentasyon, Süper Çözünürlük, Görüntü Oluşturma.

ACKNOWLEDGMENTS

I would like to express my never-ending respect and gratitude to my family, who always supported me during my studies. I would also like to thank my thesis advisor, Dr Md Baharul Islam, for giving me an opportunity to work in his research project, supporting me throughout my work and studies. Similarly, I want to thank Dr Tarkan Aydın for assisting me with my academic and professional pursuits during my studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACPE	Adaptive Convolutional Patch Embedding
AI	Artificial Intelligence
BCE	Binary Cross-Entropy
CBAM	Convolutional Block Attention Module
CLAHE	Contrast-Limited Adaptive Histogram Equalization
CNNs	Convolutional Neural Networks
DDPM	Denoising Diffusion Probabilistic Model
DP	Deep learning
DPFR	Double Pass Fundus Reflection
ESRGAN	Enhanced Super-Resolution Generative Adversarial Network
FID	Frechet Inception Distance
GAN	Generative Adversarial Network
LSA	Locality Self-Attention
MAE	Mean Absolute Error
MCC	Matthew's Correlation Coefficient
MSE	Mean Squared Error
PSNR	Peak Signal-to-Noise Ratio
RTT	Repetitive Training Technique
SITT	Single Image Testing Time
SRGAN	Super-Resolution Generative Adversarial Network
SSIM	Structural Similarity Index Measure
ViT	Vision Transformer

Chapter 1

Introduction

1.1 General Introduction

To analyze human organs, body extremities, and other tissues, experts use medical imaging technologies such as magnetic resonance imaging, ultrasound, computed tomography, X-ray, and retinal imaging technology. The retina is a layer of nerve tissue located at the rear side of an eye, it is responsive to incoming light. This light is then converted into neural signals and passed to the brain's visual cortex for visual perception. Digital retinal imaging allows the detection of various health conditions directly related to eye diseases, e.g., diabetic retinopathy, hypertensive retinopathy, retinal tear and detachment, papilledema, optic atrophy, microaneurysms, etc. These issues may lead to more serious and severe health conditions, e.g., diabetic retinopathy is a consequence of diabetes, hypertensive retinopathy is caused by hypertension, papilledema may lead to a brain tumor, meningitis, and stroke (Das & Malathy, 2018). Segmented retinal images can also be used in diagnosing. By analysing the vessel width, curvature, and structure ophthalmologists can detect various health issues.

1.2 Statement of the Problem

However, retinal images may suffer from poor quality, usually caused by blurriness, noise, illumination problems, and image artifacts as shown in Fig. 1. It is challenging to acquire high-quality retinal images due to costly equipment and misoperations during medical imaging. According to a screening study (Philip et al., 2005), about 11.9% of the generated fundus images are of poor quality. As a result, ophthalmologists need to repeat the procedure until high-quality fundus images are produced. Overcoming these issues would help with better visual perception, understanding, and analysis. Therefore, In recent years, deep learning-based techniques (Z. Gao et al., 2018; Grassmann et al., 2018; Peng et al., 2019) are used to detect and classify diseases using retinal images. However, these methods rely on high-quality images during analysis, which is not always possible to acquire in some cases. Researchers also propose traditional methods (Shamsudeen & Raju, 2016; S. Zhang et al., 2022) to enhance these images. However, they often suffer from generalization issues, meaning they cannot be applied to all cases. For example, Zhang et al. (S. Zhang et al., 2022), took the reflective feature of the fundus camera into account. They proposed a double-pass fundus reflection model (DPFR) that improves the contrast of retinal images. However, the presence of image artifacts is noticeable. Many learning-based techniques (Alimanov & Islam, 2022a; H. Li et al., 2021; W. Li et al., 2020; Pérez et al., 2020; Wan et al., 2021; You et al., 2019) have been introduced to solve this problem. However, most of these works solved the missing paired data problem by manually degrading original high-quality fundus images. Manually degraded images are different from original low-quality retinal images. Thus, these methods show poor performance when enhancing real degradation effects. For instance, the authors of (H. Li et al., 2021) developed a restoration deep learning network that was trained using simulated cataract-like images. However, it performs worse when it comes to restoring real low-quality images.

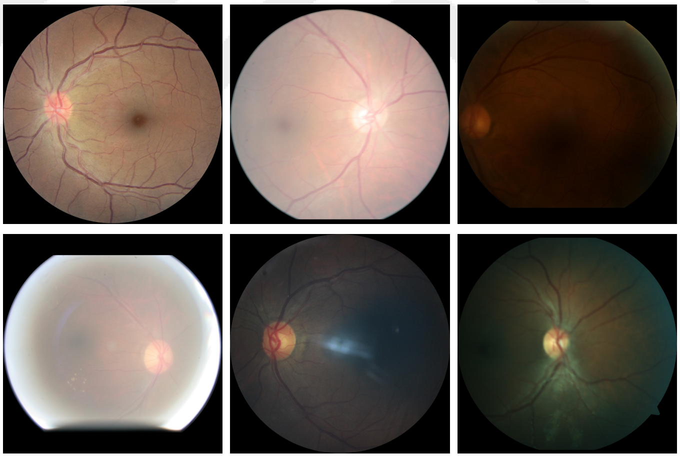

Figure 1. Retinal image qualities. Top-bottom, left-right: original high-quality image, blurring, low, high, uneven illumination, and color distortion.

In addition, acquiring high-resolution fundus images is another challenge. During imaging procedures experts may produce low-resolution images due to poor imaging device configuration. For instance, high-resolution images are used to detect microaneurysms, which are too tiny to be easily noticed; on average, the diameter of an aneurysm ranges between 43 and 266 microns with an average value of only 104 microns (H. Wang et al., 2012). Figure 2 shows the output of our model along with ground truth and low-resolution input that was up-scaled using bicubic interpolation

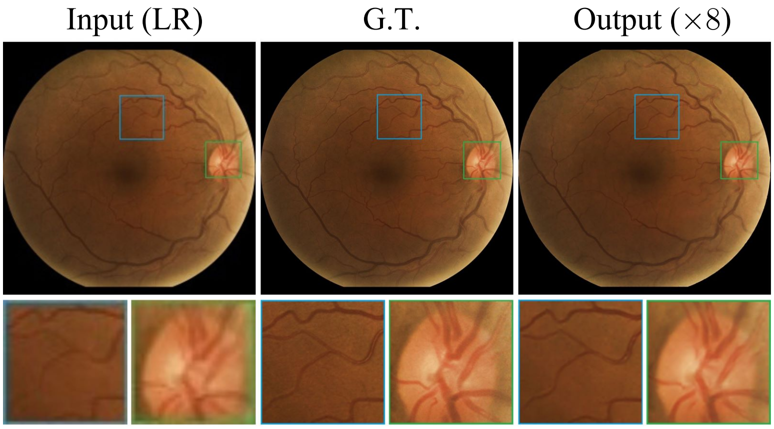

Figure 2. An example of our super-resolution method's performance. Left to right: low-resolution input, Ground Truth (G.T.), and super-resolved output of our model. The resolution was increased by a factor of 8. The bottom images zoom in on the highlighted parts by 7 times.

(Keys, 1981); as we can observe, the resulted image is not only accurate compared to the Ground Truth (G.T), but also it is not affected by the blurriness and noise. As a result, lately, vision transformer (ViT) models have taken over precedence (Dosovitskiy et al., 2020). Moreover, ViT has proven to be more superior compared to latest Convolutional Neural Networks (CNNs) in image classification tasks in the accuracy and computational efficiency (Paul & Chen, 2022). However, up-scaling natural and medical images requires different objectives. For instance, in natural image super-resolution, the objective is to generate high-resolution, realistic images. To do so, the models may generate realistic-looking results far from the ground truth. This cannot be applied to medical data, as it is crucial not to have features in the ground truth data. Additionally, the superiority of the ViT network comes at a price since ViT requires a significant amount of computational memory compared to CNN and cannot upscale the input. When ViT networks process the input, they do not change its shape; thereby, these models would require some up-scaling module, such as convolution or any interpolation method, that would reshape inputs after each ViT block. Thus, to increase the resolution of a single image standalone ViT model would need to repeat up-sampling steps several times; as the super-resolution factor increases, the number of these steps also increases, which obstructs the application of these methods in real life.

Additionally, retinal image segmentation is a challenging task due to several factors. One of the most demanding difficulties are the fundus segmentation datasets that are required for efficient training of learning-based models. The models may often produce false positive results because there is a limited difference between vessel trees and background. Three popular retinal vessel segmentation datasets are available in the literature, including DRIVE, STARE, and CHASE DB1 (Fraz et al., 2012; Hoover et al., 2000; Staal et al., 2004). These datasets have only 40, 20, and 14 image pairs, respectively. Therefore, it would be useful to generate a retinal segmentation dataset. Therefore, many deep-learning-based retinal vessel segmentation models are reported in the literature. The authors focused on increasing the size and complexity of the models, leading to lower computational efficiency. For example, IterNet (L. Li et al., 2020) is composed of one iteration of UNet (Ronneberger et al., 2015) and several iterations of mini-UNet. Several learning-based methods (Andreini et al., 2021; Guo et al., 2022; Niu et al., 2021) have been proposed to increase the training data. The authors utilized (Generative Adversarial Networks) GANs to synthesize retinal images. However, it suffers from several challenging problems, including difficulty in training as the models are susceptible to training parameters, inability to generate diverse data (mode collapse), vanishing gradient due to adversarial training, and non-convergence (H. Chen, 2021). Recently, a new class of generative models has been created, Denoising Diffusion Probabilistic Models (Sohl-Dickstein et al., 2015) (DDPM), which has proven to outperform advanced adversarial networks in image generation (Dhariwal & Nichol, 2021) as they can produce diverse results, easier to train and to converge.

1.3 Purpose of the Study

Therefore, in this work we present a retinal image restoration method based on a cycle-consistent generative adversarial network (Alimanov & Islam, 2022a), Cycle-CBAM, that relies on unpaired low-quality and high-quality retrieved images before proposing the retinal vessel segmentation method CBAM-UNet. The Cycle-CBAM does not rely on manually degraded images, making it superior to other methods. It also trains the model to preserve the structural details of fundus images. For the vessel segmentation task, we propose the CBAM-UNet that was trained and tested with three publicly available datasets, such as DRIVE (Staal et al., 2004), CHASE DB1 (Fraz et al., 2012), and STARE (Hoover et al., 2000).

In addition, we developed a novel deep learning retinal image restoration model based on a cycle-consistent generative adversarial network with modified generators (Alimanov & Islam, 2022b). In our model, we replaced the traditional CNN encoder with a vision transformer encoder, resulting in faster convergence, superior quantitative and qualitative outcomes, and better structural and color preservation than other methods.

Moreover, we have focused on solving the problems of using standalone CNN or ViT, which include lower up-sampling accuracy of CNN and high computational complexity of ViT. Therefore, we intend to develop a computationally efficient way (in terms of memory and time) of utilizing ViT for extensive and accurate image generation. We use a combination of the ViT encoder and CNN decoder. Thus, we propose a novel lightweight fixed progressive encoder-decoder model that combines ViT and CNN for single image super-resolution (Alimanov et al., 2023). The model is trained, validated, and tested in the EyeQ dataset (Fu et al., 2019). The quantitative metrics include PSNR, SSIM (Z. Wang et al., 2004), and single image testing time (SITT). For further analysis, we used three other datasets such as DRIVE (Staal et al., 2004), STARE (Hoover et al., 2000), and CHASE DB1 (Fraz et al., 2012). These datasets contain retinal images and corresponding manually segmented retinal vessels. The quantitative evaluation metrics involve the Jaccard similarity coefficient, Matthew's correlation coefficient (MCC), Cohen's kappa, F1-score, precision, recall, and accuracy and are compared with the state-of-the-art methods.

Finally, we intend to develop a computationally efficient way of using Denoising Diffusion Probabilistic Model (DDPM) for retinal image generation and vessel segmentation. To do so, we propose a novel lightweight architecture with a new training technique (Alimanov & Islam, 2023) leading to faster convergence than the original DDPM (Sohl-Dickstein et al., 2015) method. Our method has two main sub-models: (a) we perform a two-stage retinal image and corresponding vessel tree generation, (b) image super-resolution and segmentation. As shown in Fig. The model initially learns to generate vessel trees from the normal distribution before translating images from vessel trees to corresponding fundus images using our guided DDPM. In the third and fourth steps, we increase the resolution of generated images and train a biomedical segmentation UNet (Ronneberger et al., 2015) to perform the retinal image segmentation task. Therefore, we propose a novel lightweight DDPM and a retinal segmentation dataset (ReTree). The proposed dataset has been evaluated using Frechet Inception Distance (Heusel et al., 2017) (FID). The ReTree dataset's efficiency is rated by comparing training UNet with real and our data and evaluating them with real testing images. We used the Jaccard similarity coefficient, Cohen's kappa, Matthew's Correlation Coefficient (MCC), F1-score, precision, recall, and accuracy to evaluate segmentation results. Super-resolution model was tested using Structural Similarity Index Measure (Z. Wang et al., 2004) (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Binary Cross Entropy (BCE) loss.

1.4 Hypotheses/Research Questions

In this research work, we have several research questions. Firstly, does embedding Attention Modules into CNN-based GAN and UNet (Ronneberger et al., 2015) lead to more accurate image enhancement. Secondly, does replacing CNN encoder with ViT encoder in GAN produce better results in image restoration? Next, can a lightweight combination of a ViT encoder and a CNN decoder outperform the state-of-the-art image super-resolution methods in terms of computational, quantitative and qualitative results? Finally, can a lightweight DDPM outperform produce better results compared to the original DDPM?

1.5 Significance of the Study

In this work, we propose a deep learning retinal image framework for image enhancement, segmentation, super-resolution, and generation. The contributions of our first work are as follows:

- Due to the lack of paired low- and high-quality retinal images, the existing works focused on manually degrading high-quality images and used a small number of images. Therefore, we present a model for retinal image restoration and vessel segmentation without requiring the degraded pair of images.

- The proposed method eliminates the need for paired low- and high-quality images to train the network, considerably reducing the difficulty of image collection.
- We modified UNet by embedding CBAM, which led to better quantitative and qualitative results.

In our second work, we focused on replacing CNN encoder with ViT encoder in GAN architecture:

- We propose a novel CycleGAN architecture consisting of a vision transformer and convolutional neural network. To the best of our knowledge, this is the first attempt to combine CNN with a transformer in CycleGAN architecture.
- Replacing the transformer decoder with a CNN decoder reduced the computational expense compared to pure transformer CycleGAN architecture, which allowed us to use a large image size.

Next, we developed an efficient super-resolution model that comprises ViT encoder and CNN decoder:

- We propose a novel fixed progressively growing super-resolution model that combines a vision transformer encoder and a convolutional neural network decoder. To our knowledge, this is the first attempt to utilize a transformer-based technique in medical image super-resolution. This combination has led to higher computational, quantitative, and qualitative performance than standalone CNN and ViT models.
- We design a novel super-resolution training technique (Adaptive Patch Embedding Layer) that keeps the prominent architecture constant. This block has led to better quantitative and qualitative performance than progressively growing architecture.
- We adapt Locality Self-Attention instead of standard Multi-Head Self-Attention in our ViT encoder, which was designed for small- and mid-sized datasets, such as the EyeQ dataset. This has helped to achieve better performance.
- The presented hybrid model is lightweight and outperforms the state-of-the-art

models in retinal image super-resolution on diverse datasets. In addition, an extensive ablation study has been performed to demonstrate the effectiveness of ViT and Adaptive Patch Embedding Layer, including statistical analysis.

Finally, we proposed a lightweight DDPM for retinal image and corresponding vessel generation:

- Develop a new Retinal Tree ReTree dataset for more efficient retinal segmentation using Denoising Diffusion Probabilistic Model (DDPM). Our dataset has 30,000 retinal images with corresponding vessel trees.
- Propose a novel lightweight architecture of DDPM that leads to low computational cost, reducing the time and memory complexity of DDPM. In addition, it allows increasing the resolution of generated images by factors of $\times 2$ and $\times 4$.
- Design a new Repetitive Training Technique (RTT) for faster DDPM convergence. Experimental results show that RTT is capable of increasing the quantitative metrics as well as reducing the training time. In this training technique, the model gets re-trained if the loss in the current train step is higher than the global lowest train loss.
- Perform extensive quantitative and qualitative evaluation of the proposed model and dataset using three publicly available real retinal segmentation datasets.

Chapter 2

Literature Review

2.1 Medical image enhancement

These methods can be divided into traditional hand-crafted and machine-learning-based methods. Traditional methods (Datta et al., 2015; Shamsudeen & Raju, 2016; Xiong et al., 2017) are mainly based on luminosity and contrast normalization. However, these methods are challenging to generalize to all cases. For example, Shamsudeen et al. (Shamsudeen & Raju, 2016) used contrast-limited adaptive histogram equalization (CLAHE) to improve the luminosity and contrast of retinal images. However, it suffers from color distortion and additional noise. For example, Zhang et al. (S. Zhang et al., 2022), took the reflective feature of the fundus camera into account. They proposed a double-pass fundus reflection model (DPFR) that improves the contrast of retinal images. However, the presence of image artifacts is noticeable. In recent years, deep learning-based methods (Biswas et al., 2020; Cao et al., 2020; Deng et al., 2022; H. Li et al., 2022, 2021; W. Li et al., 2020; Pérez et al., 2020; Sengupta et al., 2020; Shen et al., 2020; Subramani & Veluchamy, 2020; Wan et al., 2021; Yoo et al., 2020; You et al., 2019; S. Zhang et al., 2022; Zhao et al., 2019) have been introduced to perform similar tasks with high efficiency. These methods significantly outperform the traditional methods because of their generalization ability. However, they apply manual degradation to retrieve paired low-quality and high-quality images. The manual degradation is fundamentally different from the original. For example, Deng et al. (Deng et al., 2022) developed a transformer-based Generative Adversarial Network and a paired images dataset. Li et al. (H. Li et al., 2021) proposed another CNN-based method that restores cataract-free images from simulated cataract-like retinal images. However, it performs worse when it comes to restoring real low-quality images. In some other works (Wan et al., 2021; You et al., 2019), the authors utilized cycle-consistent GAN with Convolutional Block Attention Module (CBAM) (Woo et al., 2018) for unpaired image-to-image translation, low- to high-quality. The authors in (W. Li et al., 2020; Pérez et al., 2020) focused on manually degrading high-quality images to get paired data and used conditional GAN. Subramani and Veluchamy (Subramani & Veluchamy, 2020) developed an adaptive fuzzy gray level difference histogram equalization algorithm for medical image enhancement and denoising. Cao et al. (Cao et al., 2020) developed a retinal image enhancement algorithm for deblurring and contrast improvement. Biswas et al. (Biswas et al., 2020) degraded the high-quality retinal images by adding noise and blurriness to retrieve image pairs. Then, they used a variational autoencoder to enhance retinal images. In the the double pass fundus reflection (DPFR) model (S. Zhang et al., 2022), the authors proposed a retinal image enhancement method that improves the contrast. These models have been trained to improve the overall quality of the retinal images by decreasing the effects of blurriness, poor illumination, image artifacts, and noise. However, due to the complexity of the task and models' sizes, the generated images often happen to be of low resolution, with image sizes ranging between 256×256 and 512×512 pixels; therefore, retinal image super-resolution models would greatly increase the applicability of the retinal image restoration models.

2.2 Single Image Super-Resolution Methods

A lot of works utilized CNN-based models for SISR (Behjati et al., 2022; Cai et al., 2019; Dai et al., 2019; X. He et al., 2019; Hu et al., 2021, 2019; Hui et al., 2018; Lan et al., 2020; Ledig et al., 2017; J. Liu et al., 2021; Park et al., 2018; X. Wang et al., 2018; Y. Wang et al., 2018). Liu et al. (J. Liu et al., 2021) developed a super-resolution model with the non-local self-similarity of images, cross-resolution similarity, and structural sparsity. The authors of Frequency-based Enhancement Network (FENet) (Behjati et al., 2022) proposed a deep learning-based super-resolution method focusing more on restoring the high-frequency information while forwarding the low-frequency details to the output. Hu et al. (Hu et al., 2021) developed a multi-scale information cross-fusion network (MSICF) that consists of cascaded subnetworks that infer the high-resolution features. The authors of a medium-sized dense network (MADNet) (Lan et al., 2020) have built a residual multi-scale module with an attention mechanism for better focus on more informative features. Ledig et al. (Ledig et al., 2017) was the first to utilize GAN, originally proposed by Goodfellow et al. (Goodfellow et al., 2020), in single-image super-resolution. They developed Super-Resolution GAN (SRGAN) and Super-

Resolution Residual Network (SRResNet); in addition, they used perceptual loss using pre-trained VGG19 (Simonyan & Zisserman, 2014). Later, Wang et al. (X. Wang et al., 2018) introduced Enhanced SRGAN (ESRGAN), in which they utilized Residual-in-Residual Dense Block (RRDB) as the main building block of the model. Dong et al. (Dong et al., 2015) offered a lightweight CNN-based method for SISR (SRCNN). Dai et al. (Dai et al., 2019) developed a deep CNN network with second-order channel attention (SOCA) mechanism, which uses second-order feature statistics to learn feature inter-dependencies. The authors of multi-scale GAN (MSGAN) (Daihong et al., 2022) embedded multi-scale Pyramid module into generator model to improve feature extraction.

In another work (Cai et al., 2019), the authors collected a dataset (RealSR) with real-world low-resolution images along with a Laplacian pyramid-based kernel prediction network (LP-KPN). He et al. (X. He et al., 2019) concentrated on utilizing numerical methods with CNN. They combined Leapfrog (LF-block) and Runge-Kutta (RK-block) methods with CNN architectures. Hui et al. (Hui et al., 2018) developed another CNN-based model, namely an information distillation network (IDN), that consists of feature extractors, information distillation blocks, and a reconstructor. Park et al. (Park et al., 2018) constructed a generative adversarial network that works in the feature domain with an additional discriminator. Wang et al. (Y. Wang et al., 2018) proposed another CNN-based progressively growing generative adversarial network (ProGanSR), which progressively learns to increase the resolution of images up to $\times 8$. The authors of a channel-wise and spatial feature modulation (CSFM) network (Hu et al., 2019) stacked densely connected feature modulation memory (FMM) modules for better attention. Recently, in some works, the authors utilized only ViT-based models for image super-resolution. For example, Yang et al. (Yang et al., 2020) proposed a texture learning ViT for natural image super-resolution. In their work, they trained a model to learn how to increase the resolution of images using high-resolution reference images. In other words, it transfers the relevant features of high-resolution to low-resolution images. Zamir et al. (Zamir et al., 2022) developed a ViT-based method for generating high-resolution images from low-quality noisy, blurry images. Recently, Lu et al. (Lu et al., 2022) developed an Efficient Super-Resolution Transformer (ESRT) method which consists of a Lightweight CNN Backbone (LCB) and a Lightweight Transformer Backbone (LTB); LTB also includes Efficient Multi-Head Attention (EMHA). Gao et al. (G. Gao et al., 2023) proposed an efficient CNN-Transformer Cooperation Network (CTCNet) that combines a novel Local-Global Feature Cooperation Module (LGCM) for facial image super-resolution. In another work (G. Gao et al., 2022), they proposed a Lightweight Bimodal Network (LBNet) consisting of an effective Symmetric CNN and a Recursive Transformer. However, most of these works have utilized pure CNN architectures. Only recently, most researchers proposed pure ViT networks, which require a significant amount of computational memory and are computationally inefficient.

2.3 Medical Image Super-Resolution Methods

Medical image super-resolution methods are also generally represented by CNN-based SISR models (Y. Chen et al., 2018; Lv et al., 2021; Mahapatra et al., 2019, 2017; Qiu et al., 2022, 2021; S. Zhang et al., 2018; J. Zhu et al., 2019). Zhang et al. (S. Zhang et al., 2018) developed a CNN-based fast medical image super-resolution (FMISR) method. In another work (J. Zhu et al., 2019), researchers developed a GAN-based lesion-focused SR (LFSR) method for brain tumor Magnetic resonance images (MRI) super-resolution. The authors of another research work (Y. Chen et al., 2018) developed a GAN-based multi-level densely connected super-resolution network (mDCSRN) with a generative adversarial network-guided training. Mahapatra et al. (Mahapatra et al., 2019) presented a CNN-based progressively growing GAN for retinal images super-resolution with a novel triplet loss function that compares generated images from previous stages with original high-resolution images. The authors of Multiple Improved Residual Network (MIRN) (Qiu et al., 2021) built their CNN-based model with eight improved residual blocks and up-sampling modules. The authors of the improved generative adversarial network (IGAN) (Qiu et al., 2022) proposed a GAN model that has modified residual blocks with attention layers for retinal image super-resolution. Lv et al. (Lv et al., 2021) developed dense and ReZero (residual with zero initialization) residual networks for super-resolution of retinal images (DRRN). Mahapatra et al. (Mahapatra et al., 2017) presented another CNN-based GAN model with local saliency maps to build a saliency loss function. Tian et al. (Tian et al., 2022) proposed a deep learning network based on ESRGAN with their Mixed Attention Block for retinal image super-resolution. The authors of Sparse-based domain Adaptation Super-Resolution network (Hao et al., 2022) developed a method for super-resolution of retinal optical coherence tomography angiography, proposing a novel adversarial learning-based model with a sparse edge-aware loss function. All medical-related super-resolution methods are developed using CNN architectures. In addition, most of these works have utilized adversarial training to increase the realism of generated images; however, generated medical images must be the same or similar to ground truth images. In this paper, we have built a ViT-based deep learning super-resolution method that does not rely on adversarial learning. Instead, we focused on preserving the structure of medical images rather than classifying if it is real or generated. To do so, we utilized SSIM loss that trains the network to keep the images' structure, which is crucial for medical images.

2.4 Image Generation

In several methods (Andreini et al., 2021; Appan K & Sivaswamy, 2018; Costa, Galdran, Meyer, Abramoff, et al., 2017; Costa, Galdran, Meyer, Niemeijer, et al., 2017; Diaz-Pinto et al., 2019; Guibas et al., 2017; Guo et al., 2022; Kim et al., 2022; Y.-C. Liu et al., 2019; Niu et al., 2021; Yu et al., 2019), authors developed deep learning models based on GANs (Goodfellow et al., 2020) to generate retinal images with corresponding label-maps. RetiGAN (Guo et al., 2022) developed a model with embedded Visual Geometry Group (VGG) network (Simonyan & Zisserman, 2014) to improve the generated retinal images and trained it using vessel trees acquired from segmented fundus images using UNet (Ronneberger et al., 2015). Kim et al. (Kim et al., 2022) trained a GAN with large amount of training data for retinal image synthesis. Andreini et al. (Andreini et al., 2021) proposed a two-stage Progressively Growing GAN that generates semantic label-maps initially. Later, it performs the image-to-image translation from the vessel tree to the retinal image. Patho-GAN (Niu et al., 2021) built a model that generates retinal images with symptoms related to diabetic retinopathy (DR) using vessel trees. Two pairs of images are fed into a discriminator model, and a pair of real and generated images are fed into the Perceptual network (Simonyan & Zisserman, 2014). Finally, a pair of images are passed to the DR detector to calculate the severity of it. The authors of (Y.-C. Liu et al., 2019), utilized a combination of DCGAN and Wasserstein GAN to generate retinal images with diseases for their classification. In another work (Diaz-Pinto et al., 2019), authors utilized Deep Convolutional GAN (DCGAN) to generate retinal images for glaucoma assessment. Yu et al. (Yu et al., 2019) used a pre-processing pipeline multiple-channels-multiple-landmarks (MCML) that produces images from a combination of vessel trees, optic disc, and cup images. First, they segment original fundus images and feed the output to a generator model. The generated result and a segmentation output are passed to the discriminator model with a real pair of the retinal and segmented images. On the other hand, Appan et al. (Appan K & Sivaswamy, 2018) utilized GAN to generate retinal images with lesions of various severity levels. They also developed a computer-aided diagnosis system to detect hemorrhage, which was trained using previously generated data. Costa et al. (Costa, Galdran, Meyer, Niemeijer, et al., 2017) utilized GAN architecture to generate vessel trees and then fundus images from these vessel trees. First, they built two autoencoders using adversarial learning to generate vessel trees and their corresponding retinal images. In another work (Costa, Galdran, Meyer, Abramoff, et al., 2017), they developed a retinal GAN to generate retinal images from related vessel trees. In the available literature, authors utilized only GAN-based methods for retinal image synthesis. However, GANs are proven to have such issues as vanishing gradients, mode collapse, and non-convergence. Therefore, this work focused on solving these issues by utilizing a new DDPM-based technique that has achieved state-of-the-art performance in generating retinal images.

The first original generative DDPM was proposed by Sohl-Dickstein et al. (Sohl-Dickstein et al., 2015) and Ho et al. (Ho et al., 2020). Song et al. (Song et al., 2020) proposed denoising diffusion implicit models (DDIMs), a new class of diffusion models that uses non-Markovian diffusion processes. In another work (Choi et al., 2021), the authors proposed a DDPM with a novel Iterative Latent Variable Refinement (ILVR), that conditions the generative process to synthesize high-quality images based with a given reference image. In this work, we compare the proposed model with Improved DDPM (Nichol & Dhariwal, 2021) (IDDPM), which is the current state-of-the-art method in image generation task. It utilizes Vision Transformer (ViT) encoder blocks with Multi-Head Self-Attention after each down- and up-sample block. This approach reshapes the features by a factor of 2 using 4 down- and up-sample blocks. Due to the architecture of IDDPM (Nichol & Dhariwal, 2021), it is impossible to generate images with resolution other than 64$\times$64, unlike with the proposed model. Therefore, we up-scale the generated images using Enhanced Super-Resolution GAN (ESRGAN) proposed by Wang et al. (X. Wang et al., 2018). ESRGAN (X. Wang et al., 2018) is based on SRGAN (Ledig et al., 2017), that was the first GAN used in single-image super-resolution. ESRGAN (X. Wang et al., 2018) utilizes Residual-in-Residual Dense Block (RRDB) as the main building block of the model. Recently, a new work has been proposed (Alimanov et al., 2023) for retinal image super-resolution using ViT and Convolutional Neural Network (CNN). Qiu et al. (Qiu et al., 2022) proposed Improved GAN with a novel residual attention block for a more accurate generation.

2.5 Retinal Image Segmentation

Li et al. (L. Li et al., 2020) proposed a UNet-based (Ronneberger et al., 2015) IterNet that adopts several iterations of a mini-UNet. In each iteration, the features from previous steps are shared using short and long skip connections. In Attention Guided Network (AG-Net) (S. Zhang et al., 2019), the authors proposed a method based on M-Net (Fu et al., 2018) with attention-guided filter (K. He et al., 2012) that transfers structural information from low-level feature maps to high-level. Jiang et al. (Jiang et al., 2021) proposed Multi-Scale and Multi-Branch Network (MSMB-Net) for retinal image segmentation, utilizing atrous convolutions and skip connections for more efficient feature extraction. Li et al. (Y. Li et al., 2021) in the Triple Attention Network (TA-Net) embedded a channel with a self-attention encoder and spatial attention upsampling blocks to improve the representation of target features and to train the model to pay more attention to the essential pixels. Zhang et al. (J. Zhang et al., 2021) utilized pyramid UNet with pyramid-scale aggregation blocks for feature aggregation at all levels. Recently, a few methods (Amit et al., 2021; Baranchuk et al., 2021) utilized the DDPM in image segmentation. For example, Amit et al. (Amit et al., 2021) employed DDPM for natural image segmentation. After passing them to convolutional encoders, the authors used feature addition instead of concatenation for input images.

Chapter 3

Experimental Setup and Datasets

3.1 Experimental Setup

The hardware configuration for all models includes an Intel Core i7-10700f CPU, 32 GB of RAM, and an NVIDIA GeForce RTX 2080 SUPER 8 GB.

The first retinal image restoration model was implemented using the Keras framework. The model was trained for 30 epochs. Two subsets of 2,000 low-quality and 2,000 high-quality images were randomly taken from the training set at each training epoch. To evaluate the performance of the first model, PSNR and SSIM evaluation metrics have been adopted. During retinal image restoration model training, the evaluation was performed using the validation set to see the progress in every 500 training steps. This helped to know the model's performance on images not present in the training set. The validation PSNR and SSIM values had been increasing gradually at the same rate as the training PSNR and SSIM. Total training time, including validation, took 20 hours.

The retinal vessel segmentation model was developed using the PyTorch framework. We used metrics such as Jaccard similarity coefficient, F1-score, recall, precision, and accuracy in evaluating the vessel segmentation network. The model has been set to train for 100 epochs. If the validation loss does not improve for a consecutive 5 epochs the training stops, and only the best weights are saved. As a result, the model was trained for 58 epochs.

The third ViT super-resolution model was set to train for 100 epochs for all factors, including $\times 2$, $\times 4$, and $\times 8$ with early stopping. If validation SSIM and PSNR do not increase for consecutive 4 epochs, the training with the current factor stops. As a result, the training processes with a factor of 2 took 41 epochs, a factor of 4 needed for 35 epochs, and 8 lasted for 8 epochs. The training took 10 days in total for all super-resolution factors. In our experiments, we used the Adam optimization function, batch size of 1, and learning rate of 10^{-5}. During training input images were resized to 1024×1024, 512×512, 256×256 for $\times 2$, $\times 4$, and $\times 8$ super-resolution, respectively; data augmentation techniques include random horizontal and vertical flipping. In the transformer encoder, latent vector length D is equal to 2048, the number of latent vectors N is 4096, the number of heads is 4, the dimension of each head is 64, and the MLP ratio is 4.

In our image generation model, we used the Adam optimization function, batch size of 32, and learning rate of 10^{-5}. The models have been trained for a total of 9 hours.

3.2 Datasets

In this work, we used a total of 6 retinal image datasets, such as EyeQ (Fu et al., 2019), Mendeley (Yoo et al., 2020), DRIVE (Staal et al., 2004), STARE (Hoover et al., 2000), and CHASE DB1 (Fraz et al., 2012). The EyeQ dataset contains three classes of retinal images, divided according to their quality: good (16,818 images), usable (6,434 images), and bad (5,538 images). All images are fixed resolution at 800×800. The Mendeley dataset contains two classes: artifacts (1,146 images) and non-artifacts (1,060 images) with a resolution of 350×346. DRIVE dataset includes 40 image pairs of resolution 584×565 pixels. STARE dataset comprises 20 image pairs with a resolution of 700×605 pixels. CHASE DB1 has 28 image pairs with a size of 999×960 pixels. The detailed information about theses datasets is shown Table 1.

To train both image enhancement models, we used ”Good,” ”Non-artifacts,” ”Usable,” and ”Artifacts” classes from EyeQ (Fu et al., 2019) and Mendeley (Yoo et al., 2020) datasets. The total numbers of training high-quality and low-quality images is 17,878 and 6,441, respectively. The numbers of validation and testing images that were randomly taken from both datasets are 151 and 987, respectively. To train the proposed image segmentation model, we utilized the training sets of DRIVE (Staal et al., 2004), STARE (Hoover et al., 2000), and CHASE DB1 (Fraz et al., 2012); to test the model, we used the test sets from the same datasets.

To train our super-resolution network, we used the publicly available EyeQ dataset (Fu et al., 2019). We used ”Good” and ”Usable” images for training, validation, and testing. We used 22,152 images for training, 100 for validation at each epoch, and 1,000 for quantitative and qualitative testing. EyeQ images have a size of 2048×2048 pixels. To further validate the performance of our model in the second testing stage, we used three other publicly available retinal vessel segmentation datasets, such as DRIVE (Staal et al., 2004), STARE (Hoover et al., 2000), and CHASE DB1 (Fraz et al., 2012). The low-resolution images have been up-scaled by a factor of 4 and segmented. The vessel segmentation results have been qualitatively and quantitatively compared with high-resolution manually segmented retinal images.

In the final image generation work, we utilized a total of 4 retinal datasets for image generation, super-resolution, and segmentation tasks. At first, we trained UNet (Ronneberger et al., 2015) with DRIVE (Staal et al., 2004), STARE (Hoover et al., 2000), and CHASE DB1 (Fraz et al., 2012) combined to segment retinal images. Next, we used this model to segment the EyeQ dataset (Fu et al., 2019) that comprises 28,792 512×512 pixels RGB retinal images belonging to 3 classes according to their quality. The classes include ”Good”, ”Usable”, and ”Bad”. There are 16,818” Good”, 6,434” Usable,” and 5,538 ”Bad” images; additionally, the dataset contains images with diabetic retinopathy, glaucoma, and age-related macular degeneration. In step 1, the DDPM is trained with segmented images from the EyeQ dataset that belong to the ”Good” quality class. In the next step, the same segmented images and corresponding RGB fundus images were utilized for training the second DDPM to generate retinal images from given generated vessel maps. In the next step, Enhanced Super-Resolution GAN (ESRGAN) (X. Wang et al., 2018) was trained using the same image pairs from the ”Good” quality EyeQ dataset. Using these DDPMs and ESRGAN, we generated and up-scaled pairs of images for the ReTree dataset. Finally, the segmentation UNet (Ronneberger et al., 2015) was trained with generated image pairs and tested using DRIVE, STARE, and CHASE DB1 to validate the efficiency of the proposed dataset.

Table 1

Details of datasets used for the experiments

Name	Number of Images	Image Size	Class
EyeQ (Fu et al., 2019)	16,818	2048×2048	Good
	6,434	2048×2048	Usable
	5,538	2048×2048	Bad
Mendeley (Yoo et al., 2020)	1,146	800×800	Artifacts
	1,060	800×800	Non-artifacts
DRIVE (Staal et al., 2004)	40	584×565	Images
	40	584×565	Vessel trees
STARE (Hoover et al., 2000)	20	700×605	Images
	20	700×605	Vessel trees
CHASE DB1 (Fraz et al., 2012)	28	999×960	Images
	28	999×960	Vessel trees

3.3 Evaluation Metrics

To validate and compare the proposed method results, we performed qualitative analysis and utilized quantitative evaluation metrics such as Structural Similarity Index Measure (Z. Wang et al., 2004) (SSIM), Frechet Inception Distance (Heusel et al., 2017) (FID), Peak Signal-to-Noise Ratio (PSNR), Single Image Testing Time (SITT), Jaccard similarity index measure, Matthew's Correlation Coefficient (MCC), kappa, F1-score, precision, recall, and accuracy. Jaccard, MCC, and kappa are more accurate at evaluating the classification methods since they do not suffer from sensitivity to class imbalance and asymmetricity, unlike precision, recall, F1, and accuracy. The SSIM and PSNR are calculated using 3.1 and 3.2, respectively.

The SSIM can be expressed using generated image X and original high-resolution image Y:

$$SSIM(X,Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)} \tag{3.1}$$

where μ is the mean, σ is the standard deviation, C_1 and C_2 are variables that stabilize the division.

$$PSNR(X,Y) = 10\log_{10}\left(\frac{MAX_X^2}{MSE}\right) \tag{3.2}$$

where MAX_X^2 is the maximum value of a pixel and MSE is calculated in the following formula:

$$MSE(X,Y) = \frac{1}{N}\sum_{n=1}^{N}(Y-X)^2 \tag{3.3}$$

where N is the total number of pixels.

The FID score is an evaluation metric for generative models. It uses the following formula for calculating the distance:

$$FID = |\mu_1 - \mu_2| + Tr(\sigma_1 + \sigma_2 - 2 * \sqrt{\sigma_1 * \sigma_2}) \tag{3.4}$$

where μ_1 and σ_1 refer to the mean and covariance of the generated images, while μ_1 and σ_1 refer to the mean and covariance of the real images.

The jaccard index shows the similarity between two sets and is defined in the following equation:

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{3.5}$$

Precision and recall, F1-score, and accuracy are calculated using equations 3.6; while MCC and kappa coefficient are computed using equation 3.10 and equation 3.11, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{3.6}$$
$$Recall = \frac{TP}{TP + FN} \tag{3.7}$$
$$F1 = 2\frac{Precision * Recall}{Precision + Recall} \tag{3.8}$$
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.9}$$

where TP is a true positive, TN means true negative, FP is a false positive, and FN means false negative values.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FN)(TN + FN)}} \tag{3.10}$$

$$kappa = \frac{Accuracy - Accuracy^R}{1 - Accuracy^R} \tag{3.11}$$

where random accuracy $Accuracy^R$ is calculated in the following way:

$$Accuracy^R = p_1 \times p_2 + (1 - p_1) \times (1 - p_2) \tag{3.12}$$
$$p_1 = \frac{TP + FN}{TP + TN + FP + FN} \tag{3.13}$$
$$p_2 = \frac{TP + FP}{TP + TN + FP + FN} \tag{3.14}$$

3.4 Loss Functions

To train the networks from our framework, we used several loss functions. Adversarial loss was proposed in (Goodfellow et al., 2014). Generator G_1 aims to generate fake high-quality images Ŷ from given low-quality ones X (i.e. Ŷ=G_1(X)), while G_2 tries to convert high-quality images Y to fake low-quality ones X□ (i.e. X□ = G_2(Y)).

After that, discriminator D_1 receives two images X and X□, while discriminator D_2 gets Ŷ and Y as inputs. D_1 and D_2 predict if the given patches of these image are real or fake using mean square error. The formulas for the loss function are expressed below:

$$Loss_A(D_1, X, \mathrm{X}\square) = E_X[log(D_1(X))] + E_{\mathrm{X}\square}[log(1 - D_1(\mathrm{X}\square))] \tag{3.15}$$

$$Loss_A(D_2, Y, \hat{\mathrm{Y}}) = E_Y[log(D_2(Y))] + E_{\hat{\mathrm{Y}}}[log(1 - D_2(\hat{\mathrm{Y}}))] \tag{3.16}$$

where generators are trying to minimize this value and discriminators trying to maximize it.

According to cycle consistency, image generated from one domain to another should be easily translated back to its original domain without significant changes. Translation of X to Ŷ and then back to X□ is called forward cycle consistency, whereas Y $\rightarrow$ X□ $\rightarrow$ Ŷ is called backward cycle consistency. Cycle consistency loss can be expressed in the following form:

$$Loss_C(X, \mathrm{X}\square, Y, \hat{\mathrm{Y}}) = E_X[||X - \mathrm{X}\square||_1] + E_Y[||Y - \hat{\mathrm{Y}}||_1] \tag{3.17}$$

To calculate identity loss, we use generators and images from the same domains. More specifically, H should not be modified when it is sent to G_H. The same things apply to L and G_L. To calculate this loss function, we also use the $L1$ function in the following way:

$$\begin{aligned} Loss_I(L, G_L(L)) &= E_L[||L - G_L(L)||_1] \\ Loss_I(H, G_H(H)) &= E_H[||H - G_H(H)||_1] \end{aligned} \tag{3.18}$$

The total generator and discriminator loss for the CycleGANs can be formulated as:

$$Loss(G_L, G_H, D_L, D_H) = Loss_A(D_L, L, \hat{L}) + Loss_A(D_H, H, \hat{H}) + \\ + \lambda_1 Loss_C(\acute{L}, \hat{L}) + \lambda_1 Loss_C(\acute{H}, \hat{H}) + \\ + \lambda_2 Loss_I(L, G_L(L)) + \lambda_2 Loss_I(H, G_H(H)) \tag{3.19}$$

where λ_1 and λ_2 are weights of cycle consistency and identity losses. The main problem that discriminators and generators are trying to solve can be mathematically expressed in the following equation:

$$G_L^*, G_H^* = arg \min_{G_L, G_H} \max_{D_L, D_H} Loss(G_L, G_H, D_L, D_H) \tag{3.20}$$

The SSIM loss can be expressed using generated image X and original high-resolution image Y:

$$SSIM_{Loss}(X, Y) = \frac{1 - SSIM(X, Y)}{2} \tag{3.21}$$

where $SSIM(X, Y)$ is calculated using the Formula 3.1.

The next loss function that we used is L1, or Mean Absolute Error (MAE) loss. In L1 loss, the average value of absolute difference between Y and X is derived using the following equation:

$$L1_{Loss}(X, Y) = \frac{1}{N} \sum_{n=1}^{N} |Y - X| \tag{3.22}$$

where N is the total number of pixels.

The Mean Squared Error (MSE) is shown below:

$$MSE_{Loss}(X, Y) = \frac{1}{N} \sum_{n=1}^{N} (Y - X)^2 \tag{3.23}$$

where N is the total number of image pixels, ϵ_θ is the predicted noise and z is the actual applied noise. The combined loss produced more accurate results and performance.

Binary Cross-Entropy (BCE) loss is defined with the following equation:

$$\begin{aligned} BCE_{Loss} = -\frac{1}{N}\sum_{i=1}^{N}(x_{i_{t-1}} log(x^{\theta}_{i_{t-1}}) \\ + (1 - x_{i_{t-1}}) log(1 - x^{\theta}_{i_{t-1}})) \end{aligned} \tag{3.24}$$

where N represents the number of pixels in the image.

Chapter 4

Methodology

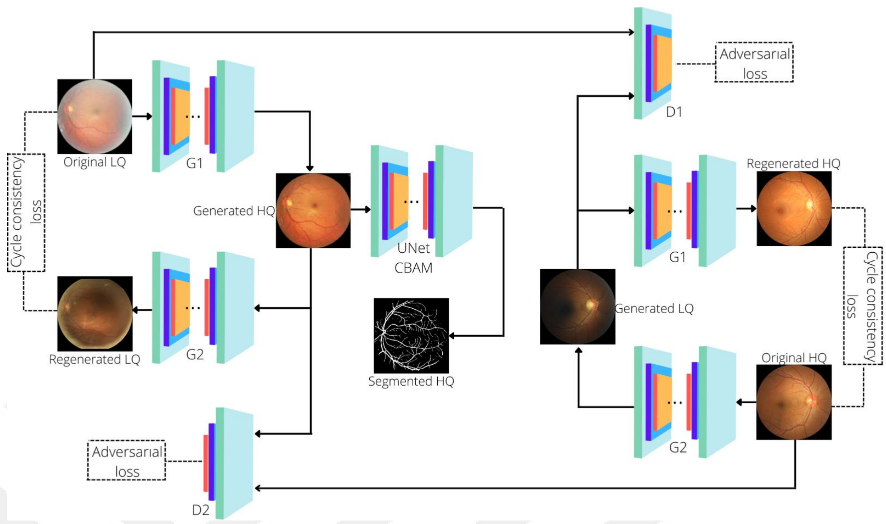

Figure 3. The workflow of the proposed method. The Original LQ image is first inputted into Generator G_1 and the output (Generated HQ) is used as input for Generator G_2 and Discriminator D_2. Next, the Original HQ image is used to make Generated LQ image using Generator G_2, then this generated image is passed to both Discriminator D_1 and Generator G_1. After these steps, cycle consistency and adversarial losses are calculated. During testing, the Generated HQ image is then segmented using our CBAM-UNet.

4.1 Cycle-Consistent Generative Adversarial Network with Convolution Block Attention Module

To enhance retinal images, we used Cycle-GAN that was firstly proposed by Zhu et al. (J.-Y. Zhu et al., 2017). The architecture of this model is shown in Fig. 3, it consists of two generators (G_1 and G_2) and two discriminators (D_1 and D_2). The goal of G_1 is to generate fake high-quality images from corresponding low-quality ones, whereas G_2 generates fake low-quality images from high-quality ones. When fake images are generated, D_1 tries to distinguish real low-quality image from fake one and D_2 chooses between real high-quality and fake high-quality. In that way, the model learns to translate low-quality images to high-quality ones and vice versa.

Generators G_1 and G_2 are fully convolutional networks, that consist of 3 down-sampling, 9 residual blocks with convolution block attention module (Res-CBAM), and

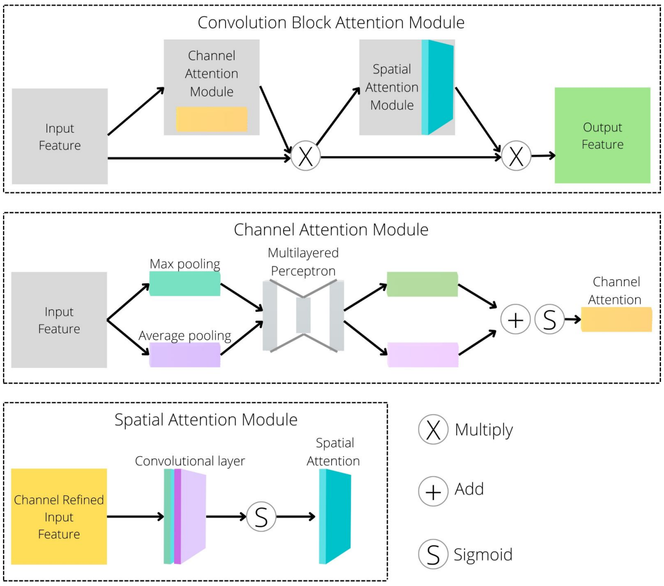

Figure 4. Convolution block attention module which is embedded into residual blocks to improve the feature extraction. Firstly, the input features are passed to channel attention module and concatenated with the output of this module. Next, the concatenated output is used as input to spatial attention module and again concatenated to get the refined output features.

3 up-sampling blocks. Each down-sampling block consists of 2D convolution, instance normalization, and ReLU activation. The numbers of filters in convolution layers are 64, 128, and 256, respectively. Kernel sizes and strides are (7, 7), (3, 3), (3, 3) and (1, 1), (2, 2), (2, 2), respectively. Each Res-CBAM block contains 2D convolutions with 256 filters having a same kernel size of (3, 3), and strides of (1, 1) and CBAM. The CBAM is capable of sequentially inferring attention maps along two channel and spatial dimensions from an intermediate feature map, and then multiplying the attention maps to the input feature map for adaptive feature refinement. Finally, there are 3 up-sampling blocks that consist of 2D transpose convolution, instance normalization, and ReLU activation for the first two blocks and tanh activation for the last block. Transpose convolution layers have 128, 64, and 3 filters and (3, 3), (3, 3), and (7, 7) kernel sizes, respectively. The padding values are (2, 2), (2, 2), and (1, 1). The architecture of CBAM is presented in Fig. 4.

Discriminators D_1 and D_2 contain 2D convolution layer with LeakyReLU activation, 4 down-sampling blocks, and an output 2D convolution layer. The filter values are 64, 128, 256, 512, 512, and 1, respectively. The kernel sizes are (4, 4) for all lay-

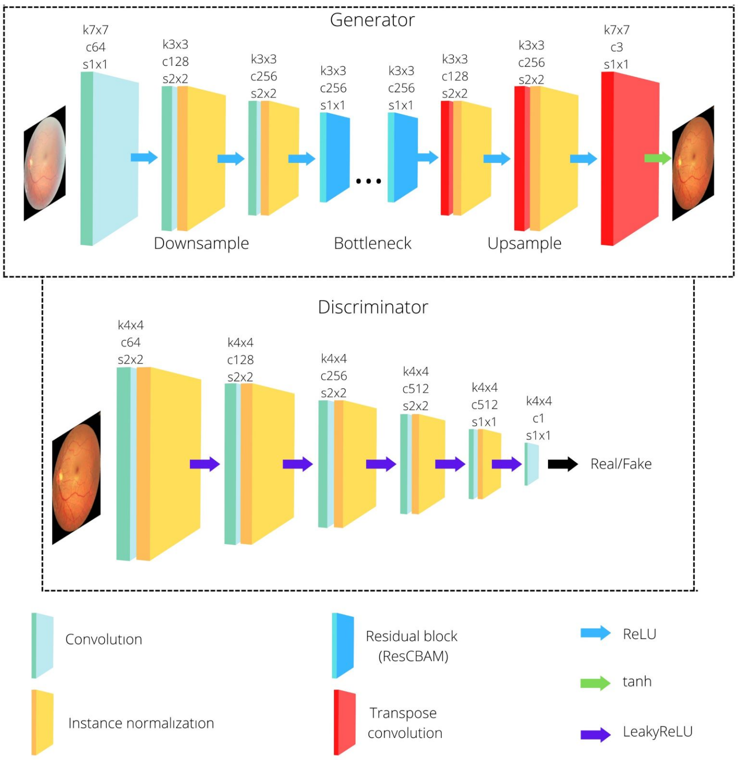

Figure 5. Architecture of generators and discriminators. Generators are deep convolutional neural networks that consist of downsample, bottleneck, and upsample layers. Discriminators are also convolutional neural networks that downsample the input to get an output that consists of 2 numbers (probabilities of two images being real).

ers and stride values are (2, 2), (2, 2), (2, 2), (2, 2), (1, 1), (1, 1), respectively. They are based on a PatchGAN network (C. Li & Wand, 2016) that aims to classify 70x70 patches of images instead of the whole image. The full architecture of generator and discriminator networks is shown in Fig. 5.

4.2 UNet with Convolution Block Attention Module

In the medical analysis of fundus images, vessels play a significant role. By analyzing these vessels' length, width, and curvature, ophthalmologists can detect various eye diseases. Therefore, the vessel segmentation task is crucial in eye care and screening. Some vessels are short and thin, so it becomes very challenging to segment them, especially if the fundus image is not of high quality. We propose a modified UNet initially developed by Ronneberger et al. (Ronneberger et al., 2015). It adopts an encoder-decoder structure, where the encoder consists of 4 down-sampling blocks with two convolution layers, instance normalization, ReLU activation function, convolution block attention module, and a max-pooling layer. The decoder contains 4 upsampling blocks with one convolution, concatenation, and two more convolution layers with ReLU activation function, instance normalization, and CBAM that is shown in Fig. 4. The model was trained and tested using three datasets with an Adam optimizer with a learning rate of 0.0001.

4.3 Vision Transformer-based Cycle-Consistent Generative Adversarial Network

Our method is based on CycleGAN, initially proposed by (J.-Y. Zhu et al., 2017). However, we replaced the CNN encoder in the generator network with a vision transformer encoder, as shown in Fig. 6. Using CNN as a decoder decreased computational expenses compared to a pure transformer network which allowed to increase the image size by a factor of 4, from 64×64 to 256×256. This network consists of a pair of generator networks (G_H and G_L) and a pair of discriminators (D_H and D_L). The task of each generator is to translate images from one domain to another and to fool the corresponding discriminator. For instance, G_H is trying to generate realistic, high-quality images from original low-quality ones, and D_H receives this generated image and one original high-quality image and calculates the probability of a given image being authentic. As a result, generators and discriminators are playing an adversarial game.

The generator consists of a vision transformer encoder and a CNN decoder. Firstly, each input image of size 256×256 is divided into patches of size 8×8 using a convolutional projection proposed by (Wu et al., 2021). It provides additional efficiency. Instead of a depth-wise convolutional layer with batch normalization, we only used a convolutional layer with 1024 output channels, kernel, and stride sizes of 1. Then we flattened the 3D output to 2D and transposed it. It reduced the computation time without affecting the accuracy of the method. Then these flattened sequences are fed into a standard transformer encoder network as shown in Fig. 6. Our network's depth and projection dimensions are 7 and 1024, respectively. The output shape of this encoder is 1024×1024. Therefore, we reshaped it into 1024×32×32 to feed into the CNN decoder. It consists of 3 upsampling blocks: transpose convolution, instance normalization, and ReLU activation function. The output channels have been decreased by 2 after each block; kernel size, stride, padding, and output padding in all blocks are 3, 2, 1, and 1, respectively. The resulted matrix of shape 128×256×256 is converted into a color image using the last convolutional layer with a kernel size of 7, the stride of 1, and padding of 3. The final output is activated using a hyperbolic tangent.

A discriminator is a CNN-based classifier that consists of such layers as 2D convolution, LeakyReLU activation function, and instance normalization. Firstly, a colorful 256×256 image is fed into the convolutional layer with output channels, kernels size, stride, and padding equal to 64, 4, 2, and 1, respectively. Then, we used the LeakyReLU activation function. The result sends to three consecutive downsampling blocks: convolutional layer, instance normalization, and LeakyReLU. Kernel size and padding are equal to 4 and 1, but output features double after each downsampling process starting from 64 and ending with 512. Stride is 1 in the first and 2 in the last two blocks. In the end, it is fed into the final convolutional layer with output channel, stride, and padding equal to 1 and kernel size being 4. As a result, we get a 30×30 patch and make predictions using the sigmoid activation function.

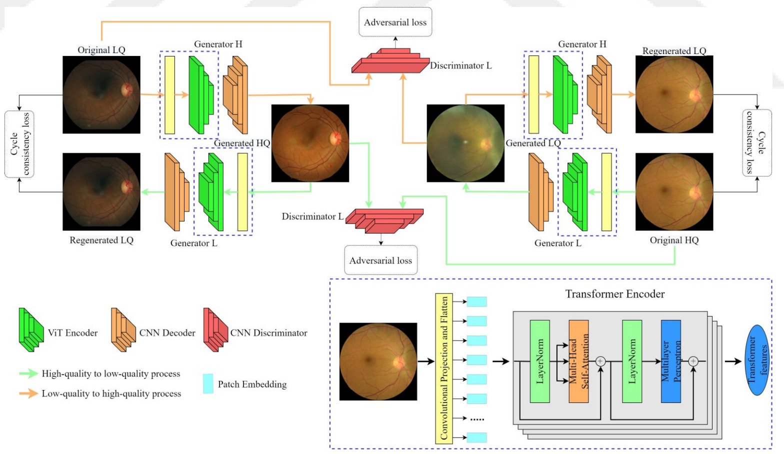

Figure 6. The workflow of the proposed method. Low-quality images go through Generator H and the result - Generated HQ - is inputted into Generator H and Discriminator H. Next, the Generated LQ image is made from high-quality image using Generator L, then this generated image is passed to both Generator H and Discriminator L. After these steps, losses are calculated.

4.4 Vision Transformer-based Super-Resolution model

The proposed method consists of a vision transformer encoder and a convolutional neural network decoder, as shown in Figure 7. In this work, we propose a novel training strategy, Adaptive Convolutional Patch Embedding (ACPE) layer, to keep the architecture of the ViT encoder and CNN decoder fixed for all up-scaling factors, starting from 2 to 8. The only part that changes is the patch embedding block that includes one 2D convolutional layer with various kernel and stride sizes for different image sizes and up-scaling factors.

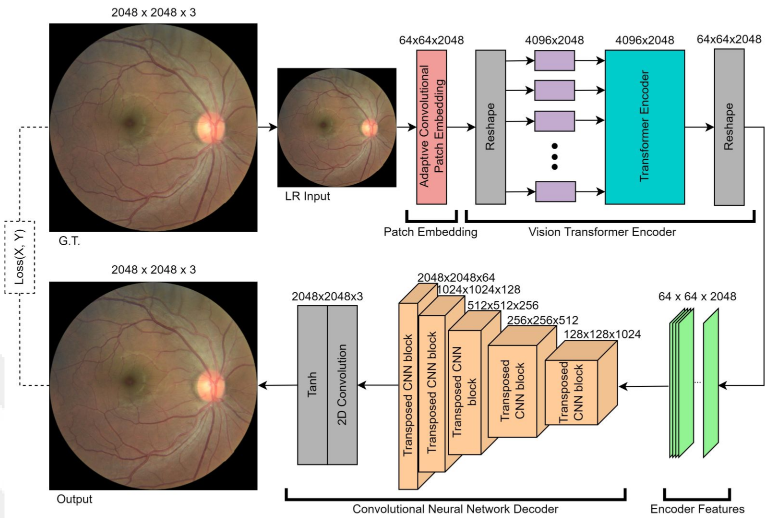

Figure 7. The proposed super-resolution model consists of a vision transformer encoder and a convolutional neural network decoder. The model remains the same for different up-scaling factors except for the patch embedding. The patch size increases as we increase the image size. For example, for input images of size 256×256, the patch size is 4×4 to perform $\times 8$ super-resolution.

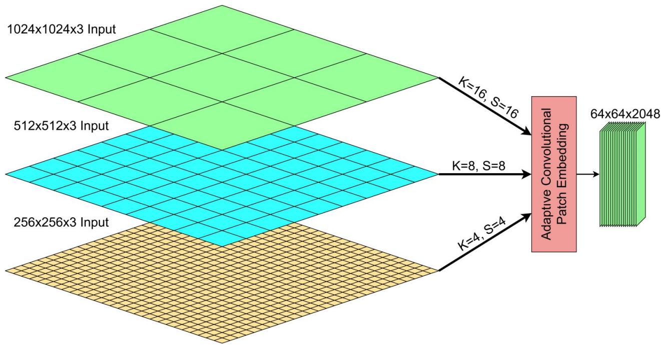

Figure 8. Adaptive Convolutional Patch Embedding (ACPE). After the input images of different sizes are split into patches by the ACPE layer, the output features have the same shape of $64 \times 64 \times 2048$. K is the kernel size, and S is the value of strides.

At first, the low-resolution image is fed into the Adaptive Convolutional Patch Embedding (ACPE) layer as shown in Figure 8; it is utilized to split the input image into non-overlapping patches. Additionally, it is used to embed the resulting image patches, as ViT models lack prior positional information of image patches. The number of input channels changes from 3 to 2048, and both the height and width of resulted features are equal to 64. ACPE has eliminated the need to add additional layers to the model while increasing the super-resolution factor, since we only need to change the patch sizes when increasing the super-resolution factors. This has led to better quantitative and qualitative results and computational performance. The output of patch embedding is fed into the transformer encoder. The output of the ViT network is up-scaled and converted back to a 2048×2048 image in the 5 transposed CNN blocks followed by 2D convolution and a Tanh activation function. To calculate the kernel size and strides, or patch size, of a given input image, we used the following formula:

$$PS = K \times S = \frac{H}{NP} \times \frac{W}{NP} \tag{4.1}$$

where PS is the patch size, NP is the number of patches equal to 64, K is kernel size, S is strides, H and W are the image height and width, respectively. For example, for image of size 1024×1024 the patch size would be $PS = \frac{1024}{64} \times \frac{1024}{64} = 16 \times 16$. NP has been derived in experiments in which we trained the model with NP being equal to 32, 64, and 128.

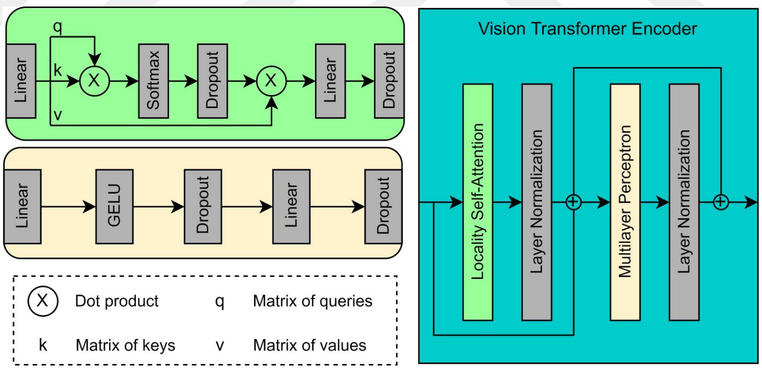

Figure 9. The architecture of the ViT encoder consists of local self-attention, layer normalization, and multilayer perceptron layers. After dividing the image into patches, the features are fed into the ViT encoder.

The encoder includes locality self-attention (Lee et al., 2021) (LSA), layer normalization (Ulyanov et al., 2016), and multilayer perceptron (MLP). LSA was initially proposed for small-sized datasets. However, it has also shown better results than multi-head self-attention when trained with medium-sized datasets thanks to learnable temperature scaling. In this work, we have decided to use LSA for our model to be efficiently trained on the EyeQ dataset, which is considered mid-sized.

The input image $(x) \in \mathbb{R}^{B\times H\times W\times C}$ with number of samples (batch size) B, height H, width W and number of channels C is divided into patches using convolutional layer, flattened across H and W, and transposed. As a result, we obtain patch sequences $(x)_p \in \mathbb{R}^{B\times N\times D}$ with $N = H \times W$ denoting the number of latent vectors and $D = 2048$ denoting the length of each latent vector. The value of D has been derived through several experiments, in which we trained the model with various values, such as 512, 1024, 2048, and 3072; the most optimal result was achieved with 2048. The input patch sequences are divided into queries, keys, and values using a linear layer, as shown in Figure 9. This layer changes the sequence length D to be equal to hidden vector length D_h, which is equal to the multiplication of the number of heads and their length ($NH \times LH$); NH and LH have been derived during experiments, and the optimal values are 64 and 4, respectively. The queries and keys matrices are multiplied to obtain $R_{i,j}$. LSA uses diagonal masking to improve the attention process. The masking is defined in equation 4.3. The diagonal elements are set to $-\infty$. Using the formula 4.4, we get attention values for R. After the dropout layer, the output is multiplied by values. The resulting outcome is sent to the final linear and dropout layers. Linear layer changes the shape of attention maps, restoring it to its original size of 4096×2048. In our model, the dropout rate is 0. The output features are normalized using layer normalization using the following formula:

$$\hat{X} = \frac{X - \mu(X)}{\sqrt{\sigma(X) + \epsilon}} * \gamma + \beta \tag{4.2}$$

where $\hat{X}$ is the normalized features, X is the input features, $\mu(X)$ is the mean, $\sigma(X)$ is the standard deviation, ϵ is a small value to avoid division by zero, γ and β are learnable parameters for affine transformations that are derived during training.

$$R_{i,j}^{Masked} = \begin{cases} R_{i,j} & i \neq j \\ -\infty & i = j \end{cases} \tag{4.3}$$

where $R_{i,j}$ represents the components of the dot product between the matrix of queries and transposed matrix of keys $R = q \times k^T$.

$$LSA = Softmax(R^{Masked}/\tau) \tag{4.4}$$

where τ is learnable temperature scaling, another feature of LSA that helps the softmax calculate its temperature during training.

The normalized features are concatenated with the encoder input and sent to MLP. The structure of MLP is presented in Figure 9, the first linear layer changes the dimension D of input features from 2048 to hidden dimension of MLP $D_h^{MLP} = 2048 \times 4$, where 4 is the MLP ratio that was derived in experiments. Next, they are sent to the Gaussian Error Linear Unit (GELU) (Hendrycks & Gimpel, 2016) activation function and dropout. Another linear layer reduces dimensions to the initial size D with dropout. Dropout rates are equal to 0. The depth of our ViT encoder is equal to 1 after testing it with depths of 2, 5, and 7.

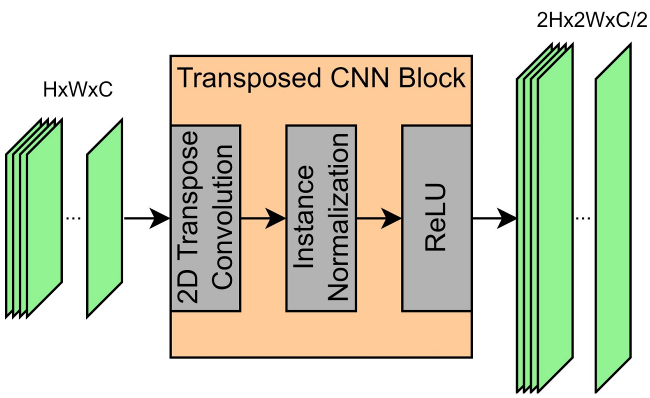

Figure 10. The structure of the CNN decoder includes 2-dimensional transpose convolution, instance normalization, and ReLU activation function layers. Each block increases the dimensions of the features (green) by a factor of 2 and reduces the number of these features by 2.

The proposed method utilizes CNN for decoding the ViT output features. The training was performed with images of different sizes, such as 256×256, 512×512, and 1024×1024. The patch embedding layer has been adjusted according to the image size. For example: for 256×256 image size, the patch size would equal 4×4 according to 4.1, and for 512×512, the patch size would be 8×8. It allowed fixing our model's encoder and decoder parts, changing only the patch embedding layer. Our convolutional decoder consists of 5 consecutive 2D transpose convolution blocks and a final 2D convolution with the tanh activation function as presented in Figure 10. The structure of all blocks is the same and includes 2D transpose convolution, instance normalization, and ReLU activation function; the number of kernels is 3, strides are 2, and padding is 1. Each CNN block up-scales the input by a factor of 2. For instance, the output shape of an encoder is $2048 \times 64 \times 64$. After the first block, it will be of shape $1024 \times 128 \times 128$, and the final block produces a $64 \times 2048 \times 2048$ matrix. In the end, it will be fed into the last layer with 2D convolution and tanh activation function, which keeps the input's height and width but reduces the channels to 3. Tanh activation function makes the input values in the range between -1 and 1.

4.5 Denoising Diffusion Probabilistic Model

The DDPM uses two stages to generate samples: forward and backward processes. In the forward diffusion process, we gradually add Gaussian noise to the input image through a series of diffusion time steps T. In the backward process, the model tries to denoise the input at the given diffusion step t. In the forward diffusion, at each diffusion step t, we sample x_t from data distribution $q(x)$ ($x_t \in q(x)$). The noise is added to the input variable x_{t-1} using the variance schedule β_t to acquire x_t with $q(x_t|x_{t-1})$. To do so, we use the following formula.

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \mu_t = \sqrt{1-\beta_t}x_{t-1}, \Sigma_t = \beta_t I) \tag{4.5}$$

where $\mathcal{N}$ is the Gaussian distribution, μ_t is the mean, Σ_t is the standard deviation, a diagonal matrix of β_t values because I is the identity matrix.

With equation 4.5, it is possible to apply noise to x_0 in a tractable way (1:T) to get x_T. This posterior probability is defined as:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{4.6}$$

However, with these equations, to sample x_T, we would have to apply $q(x_{1:T}|x_0)$ a total of T number of times repeatedly. Therefore, a reparametrization trick was used to sample x at any time step t, which is defined in the following formula:

$$\begin{aligned} x_t &= \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_{t-1} \\ &= \sqrt{\alpha_t}x_{t-2} + \sqrt{1-\alpha_t}\epsilon_{t-2} \\ &= \dots \\ &= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_0 \end{aligned} \tag{4.7}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$ and noise $\epsilon_0, \dots, \epsilon_{t-2}, \epsilon_{t-1} \sim \mathcal{N}(0, I)$.

This reparametrization trick allows us to precompute α_t and $\bar{\alpha}_t$ for any diffusion step t to sample x_t in one operation, to produce x_t we can use the following distribution:

$$x_t \sim q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \tag{4.8}$$

In our work, β_1 and β_T are 0.0001 and 0.02, respectively. In addition, we utilized cosine noise scheduler (Nichol & Dhariwal, 2021) instead of the linear one used in the original DDPM work (Sohl-Dickstein et al., 2015) as it has shown to lead to faster convergence.

In the reverse diffusion process, the model is trying to learn distribution $p_\theta(x_{t-1}|x_t)$, which is an approximation of original distribution $q(x_{t-1}|x_t)$. In this process, a deep learning method is utilized to gradually remove noise at each diffusion step from T to 1. This process can be defined as follows:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{4.9}$$

This reverse formula is applied to all diffusion steps to compute the trajectory $p_\theta(x_{0:T})$

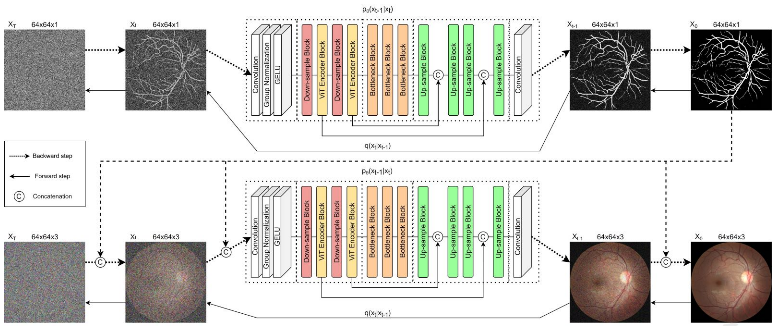

Figure 11. The general architecture of the proposed DDPMs. The first model generates semantic label maps from noise in a given number of time steps (T). At each time step t, the model predicts output from time steps $t-1$. In the second step, the model takes RGB noise and the output of the first model to generate fundus images. The image containing vessel trees is concatenated with noisy input at each diffusion step t. Each DDPM model includes initial, down-sample, ViT encoder, bottleneck, and up-sample blocks.

using equation 4.10:

$$p_\theta(x_{0:T}) = p_\theta(x_T)\prod_{t=1}^{T} q(x_{t-1}|x_t) \tag{4.10}$$

However, the output of a DDPM is not an image itself. The model takes the noisy image as input and predicts the noise. To sample an image from complete noise input x_T, we use the following formula 4.11:

$$x_{t-1} = \frac{x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} + \sqrt{\beta_t} z \tag{4.11}$$

where $\epsilon_\theta(x_t, t)$ is predicted noise by DDPM and if t is not equal to 1 then $z \sim \mathcal{N}(0, I)$, otherwise z is 0. Equation 4.11 is repeated T times, from $t = T$ until $t = 1$ to acquire x_0.

Our work focused on building a framework for retinal images with vessel tree generation, super-resolution, and segmentation. The workflow of the proposed method is shown in Fig. 12, which has four major steps. Firstly, we develop the denoising diffusion probabilistic model (DDPM) to generate semantic label maps from noise with one channel. Secondly, we applied an example-based generation technique for guided training of another DDPM to generate colorful fundus images from given vessel trees.

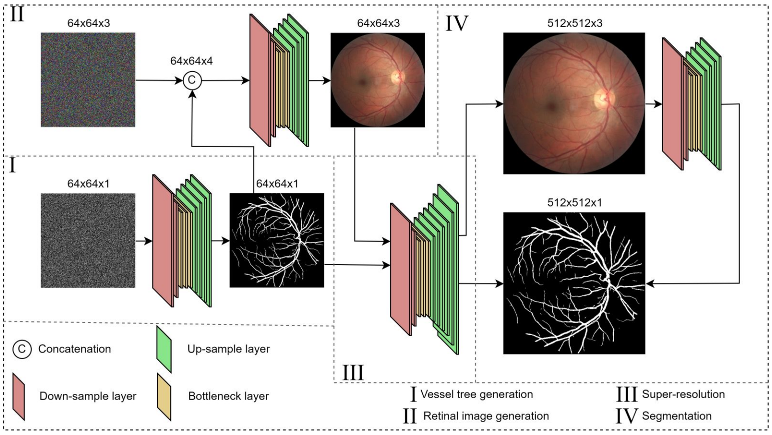

Figure 12. Workflow of the proposed framework. In the first (I) step, we train DDPM to generate vessel trees from noise. Next (II), another DDPM learns to generate retinal images from noise and vessel semantic label-map. In the third (III) step, we performed super-resolution of generated vessel trees and retinal images. Finally (IV), the biomedical segmentation model is utilized to segment generated up-sampled retinal images for validation.

In this step, the model synthesizes retinal images using noise with red, green, and blue (RGB) channels concatenated with a grayscale vessel tree; concatenation is performed at each diffusion time step t as shown in Fig. 11. After generating retinal images and vessel maps, the resulting images require up-scaling to improve the image quality for efficient image segmentation. The output of DDPMs is 64×64-pixel images.

Therefore, we performed super-resolution of retinal images up to 512×512 pixels using ESRGAN (X. Wang et al., 2018) with modifications for a specific task (RGB and binary images super-resolution). To up-sample the retinal images, we added SSIM loss in equation 3.1 during training of the ESRGAN model for preserving the image structure. To perform super-resolution of vessel maps, we had to remove the Perceptual loss computed using VGG-19 (Simonyan & Zisserman, 2014) because it requires input images with 3 channels. However, we included a BCE loss, shown in equation 3.24 to evaluate the performance of binary image super-resolution. In the final step, we performed vessel segmentation of the proposed dataset using UNet (Ronneberger et al., 2015).

Retinal image generation consists of two stages: retinal vessel generation (1) and retinal image synthesis (2). In both tasks, we utilized DDPM, the architectures of proposed DDPMs are presented in Fig. 11 and in Fig. 13. Both models comprise initial down-sample, ViT encoder, bottleneck, and up-sample blocks. The initial block includes 2D convolution with channel number, kernel size, and padding equal to 128, 3, and 1, respectively. After that, the input x_t is normalized and activated using group normalization and Gaussian Error Linear Unit (GELU) (Hendrycks & Gimpel, 2016). Next, x_t and the diffusion step are sent to one down-sample block. The input x_t is down-scaled using one max pooling layer with a window size of 4 and two convolutional blocks. Each block is built with convolution, group normalization, GELU, convolution, and group normalization layers. Also, diffusion step t is activated using Sigmoid Linear Unit (Elfwing et al., 2018) (SiLU) fed into a linear layer. The resulting embeddings are added to x_t and processed by ViT encoder block with locality self-attention (LSA) (Lee et al., 2021). The procedure with down-sampling and ViT blocks are repeated. The bottleneck is composed of three consecutive convolutional blocks. The up-scaling is performed using the bilinear up-sampling with a factor of 2 and two convolutional blocks with a residual connection. Since down-sampling was performed twice with a factor of 4, the model includes four up-sampling blocks. During up-scaling x_t, we need to add time embeddings acquired from t in the same way as in down-sampling. Finally, x_t is fed into the last convolutional layer with a kernel size of 1.

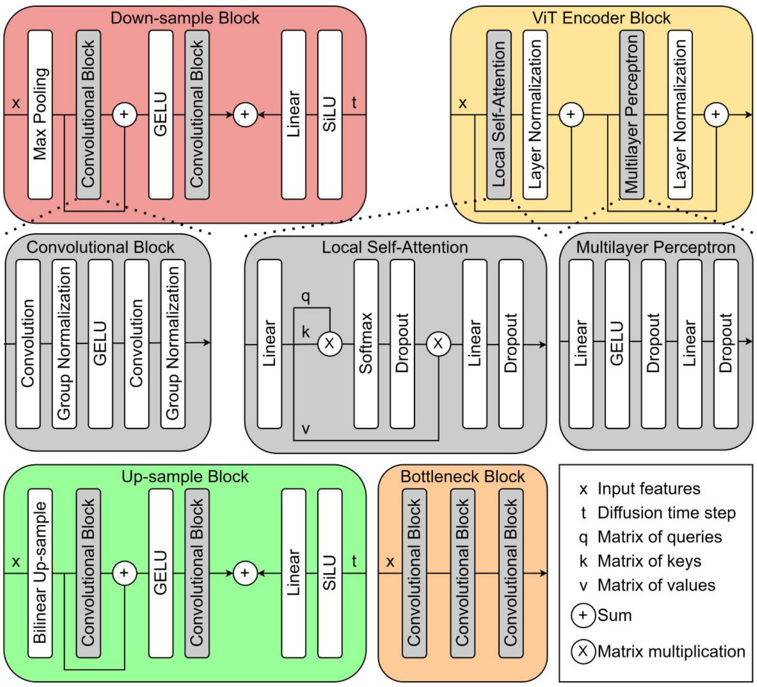

Figure 13. The detailed architecture of the proposed DDPM. It consists of down-sampling, ViT encoder, up-sampling, and bottleneck blocks.

The LSA was first proposed as a solution for small datasets by Lee et al. (Lee et al., 2021). Nonetheless, when trained on medium-sized datasets, it has also outperformed multi-head self-attention. Thanks to learnable temperature scaling. We employ LSA to train our model with the mid-sized EyeQ dataset effectively. The queries and keys matrices are multiplied in LSA to obtain $R_{i,j}$. It uses diagonal masking to improve the attention process. The masking is shown in equation 4.12. The diagonal elements are set to $-\infty$. Using equation 4.13, we get attention values for R.

$$R_{i,j}^{Masked} = \begin{cases} R_{i,j} & i \neq j \\ -\infty & i = j \end{cases} \tag{4.12}$$

where $R_{i,j}$ represents the components of the dot product between the matrix of queries and transposed matrix of keys $R = q \times k^T$.

$$LSA = Softmax(\frac{R^{Masked}}{\tau}) \tag{4.13}$$

where τ is learnable temperature scaling, another feature of LSA that helps the softmax calculate its temperature during training.

The proposed Repetitive Training Technique is a simple yet effective approach that experimentally demonstrates a reduction in training time and an improve in quantitative results. This technique aims to enhance the robustness of the proposed model against randomness and outliers. Specifically, when outlier noise is introduced, the backward diffusion process may yield unrealistic results. Therefore, re-training the model with the same data is beneficial until the loss decreases. By employing this technique, we observed a decrease in the number of epochs and total training time, along with a reduction in the generation of unrealistic images, leading to a lower FID score. In our experiments, the maximum number of repetitions is set to 5. If the current loss value surpasses the global minimum average loss across all epochs, the model is re-trained with the current data, repeating this process up to 5 times.

Chapter 5

Results and Discussions

5.1 Retinal Image Enhancement (Cycle-CBAM) and Vessel Segmentation (CBAM-UNet)

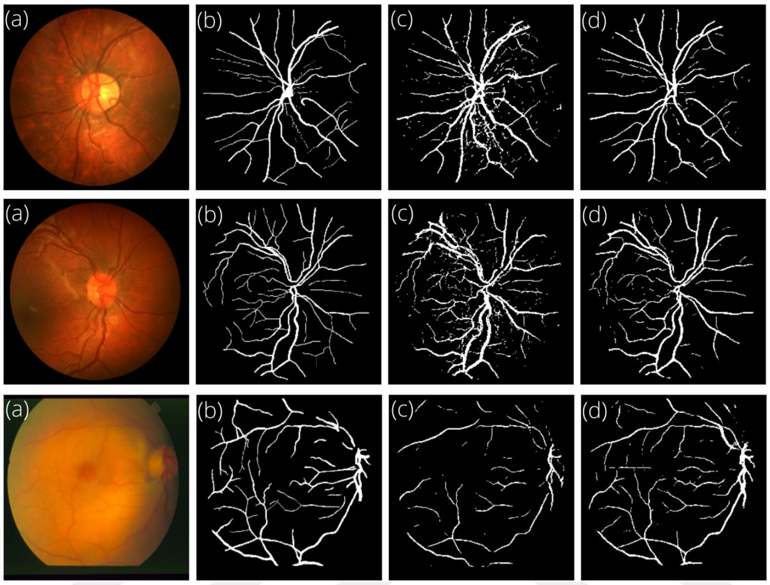

Figure 14. Qualitative analysis of proposed CBAM-UNet. Here, (a) is the original fundus image, (b) is manually segmented fundus image (ground truth), (c) is the result of UNet (Ronneberger et al., 2015), and (d) is the result of our CBAM-UNet.

In retinal images, vessels play a significant role. Therefore, it is essential to enhance the quality and preserve the structural details of these images. Vessel segmentation has been performed using our modified CBAM-UNet. Fig. 14 demonstrates the results of the original UNet (Ronneberger et al., 2015) and our model along with ground truth. As we can see, our method performs better in vessel segmentation due to the ability to segment the micro parts of the retinal vessels. In addition, our method has fewer false-positive vessels, making it look clearer with much less noise. Fig. 15 presents retinal images with various degradation types, their restored images, and the results of vessel segmentation for low-quality and restored high-quality images. The most significant difference can be noticed in low-illumination images, where blood vessels are not visible.

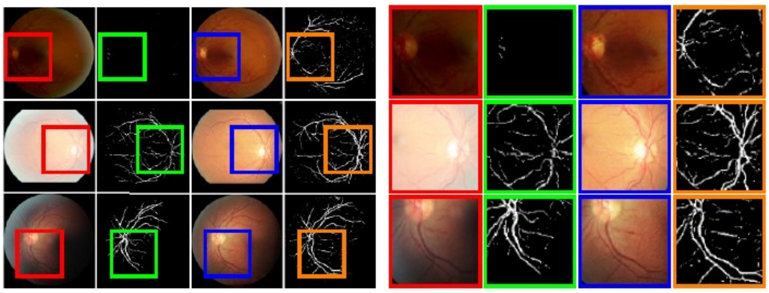

Figure 15. Original low, high, and uneven illuminations and corresponding restored high-quality images with respective vessel segmentations.

The proposed method has been qualitatively evaluated and compared with state-of-the-art methods. Fig. 16 shows the original low-quality image along with the results of the state-of-the-art restoration methods and corresponding vessel segmentation performance. Wan et al. (Wan et al., 2021) method generally improved the images. However, some retinal vessels are missed compared to the result of our method. ArcNet (H. Li et al., 2022) produces black artifacts in some of the images. On the other hand, the CLAHE (Shamsudeen & Raju, 2016) method fails to enhance low-illumination images and increases the presence of artifacts in an image. The DPFR (S. Zhang et al., 2022) makes the retinal structure more visible but at the same time significantly increases artifacts. ESRGAN (X. Wang et al., 2018) increased the resolution but could not enhance the details of a retinal image. Hence, it failed in vessel segmentation. CycleGAN (Fig. 16(g)) could enhance the images, but if we compare it to Cycle-CBAM (Fig. 16(h)), some vessels could not be segmented.

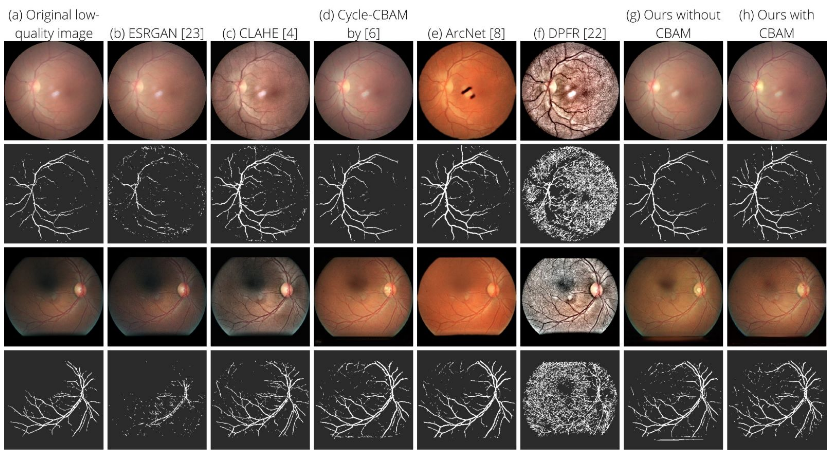

Figure 16. Qualitative analysis of various methods. (a) Original low-quality image, (b) ESRGAN (X. Wang et al., 2018), (c) CLAHE (Shamsudeen & Raju, 2016), (d) Cycle-CBAM (Wan et al., 2021), (e) ArcNet (H. Li et al., 2022), (f) DPFR (S. Zhang et al., 2022), (g) ours without CBAM, (h) ours with CBAM.

The average value of PSNR and SSIM are 26.647 dB and 0.827, respectively. Table 2 shows the PSNR and SSIM values along with time spent on a single image enhancement of our method with the state-of-the-art methods, such as Cycle-CBAM (Wan et al., 2021), ArcNet (H. Li et al., 2022), ESRGAN (X. Wang et al., 2018), CLAHE (Shamsudeen & Raju, 2016) and DPFR (S. Zhang et al., 2022). As we can see, the proposed method outperforms other methods in PSNR, and in terms of SSIM, it comes after ArcNet (H. Li et al., 2022). Regarding time spent on single image restoration, it is outperformed by CLAHE (Shamsudeen & Raju, 2016) and ResNet.

Table 2

Comparison of enhancement method with other works.

Method	PSNR	SSIM	Time
CLAHE (Shamsudeen & Raju, 2016)	25.300 dB	0.714	**0.01672s**
DPFR (S. Zhang et al., 2022)	12.446 dB	0.345	0.44903s
ArcNet (H. Li et al., 2022)	20.922 dB	**0.875**	0.18312s
ESRGAN (X. Wang et al., 2018)	**25.377 dB**	0.522	0.55390s
Cycle-CBAM (Wan et al., 2021)	24.739 dB	0.810	0.13748s
Ours without CBAM	25.212 dB	0.793	**0.12709s**
Ours with CBAM	**26.647 dB**	**0.827**	0.13780s

Table 3 provides the quantitative results of the original UNet (Ronneberger et al., 2015) and our modified CBAM-UNet. As we can see, the proposed network outperforms UNet in terms of all evaluation metrics, except for the FPS. On average, it takes 0.132 seconds to segment a single image.

Table 3

Quantitative analysis of segmentation results with DRIVE (Staal et al., 2004), STARE (Fraz et al., 2012), CHASE DB1 (Hoover et al., 2000).

Dataset	Method	Jaccard	F1	Recall	Precision	Accuracy	FPS
DRIVE	UNet	0.6520	0.7892	0.7786	0.8054	0.9639	**23.279**
	Ours	**0.6659**	**0.7992**	**0.7829**	**0.8213**	**0.9659**	19.376
CHASE DB1	UNet	0.5860	0.7388	0.7763	0.7078	0.9647	**17.211**
	Ours	**0.6077**	**0.7558**	**0.8088**	**0.7116**	**0.9665**	14.693
STARE	UNet	0.6370	0.7769	0.8005	0.7667	0.9657	**12.533**
	Ours	**0.6656**	**0.7979**	**0.8010**	**0.7997**	**0.9697**	11.417

To validate the effect of the integration of CBAM into residual blocks, we tested

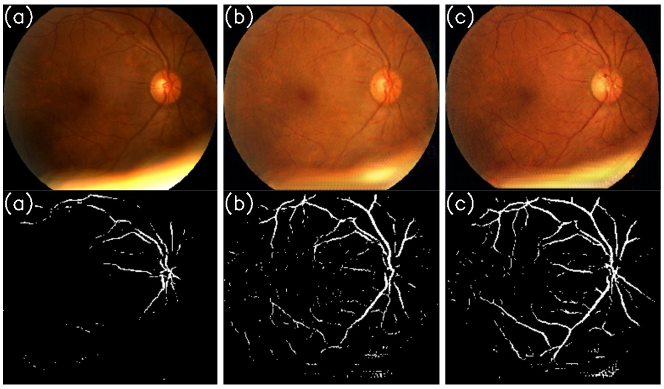

Figure 17. Ablation study. Restoration and vessel segmentation results of the input retinal image (left) without (middle) and with (right) considering the CBAM in the model.

the enhancement model with and without CBAM. The CBAM replaced ReLU activation and the convolutional layer in the residual block. As a result, the PSNR and SSIM values have been increased by 1.435 dB and 0.034, respectively. In addition, qualitative analysis has been performed to validate the efficiency of CBAM using vessel segmentation. Fig. 17 demonstrates the retinal image restoration and vessel segmentation results without (middle) and with considering the CBAM in the proposed method. The first method slightly improves the illumination problem. However, some parts of the segmented vessel are missing without considering the CBAM in the proposed method.

5.2 Retinal Image Enhancement (ViT CycleGAN)

To validate the efficiency of our method, we performed qualitative analysis for restored images using vessel segmentation with 5 state-of-the-art methods: CLAHE (Shamsudeen & Raju, 2016), CycleGAN (You et al., 2019), Cycle-CBAM (Wan et al., 2021), ArcNet (H. Li et al., 2021), DPFR (S. Zhang et al., 2022). UNet (Ronneberger et al., 2015) was trained using pairs of fundus images with corresponding manually segmented images from DRIVE dataset (Staal et al., 2004). Fig. 18 shows the restoration and vessel segmentation results. Our method enhanced the retinal image better compared to the state-of-the-art. In addition, our result has less noise compared to CycleGAN (You et al., 2019) and CycleCBAM (Wan et al., 2021). Algorithm-based methods (Shamsudeen & Raju, 2016; S. Zhang et al., 2022) could improve the contrast making vessels visible, but produced images have artifacts. ArcNet (H. Li et al., 2021) failed to restore tiny blood vessels in the dark region.

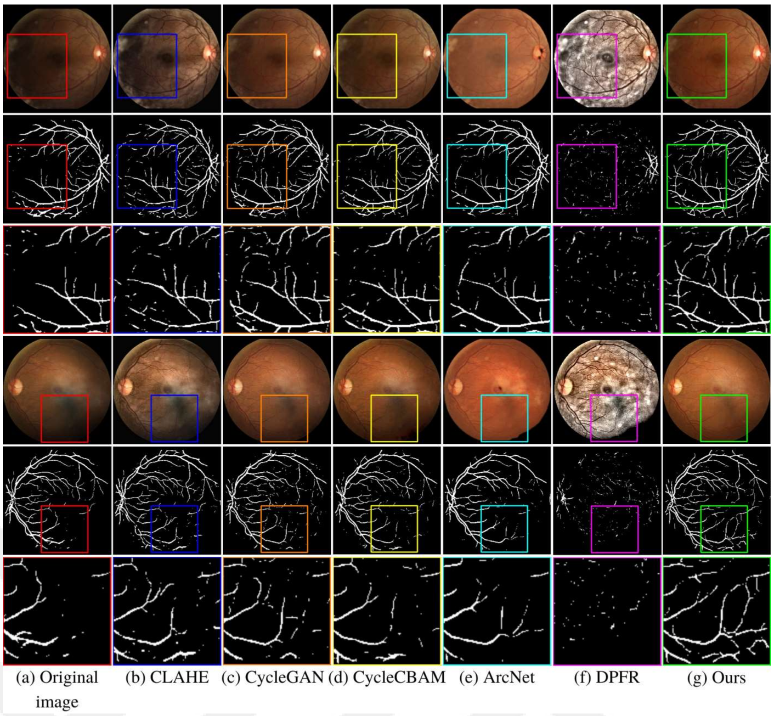

Figure 18. Comparison with state-of-the-art methods. From top to bottom are fundus image, corresponding segmentation and zoomed segmentation. (a) low-quality image, (b) CLAHE (Shamsudeen & Raju, 2016), (c) CycleGAN (You et al., 2019), (d) CycleCBAM (Wan et al., 2021), (e) ArcNet (H. Li et al., 2021), (f) DPFR (S. Zhang et al., 2022), (g) Ours.

To further evaluate our results, we conducted the quantitative analysis with the same methods. The evaluation metrics include PSNR, SSIM, and single image test time (SITT), as shown in TABLE 4. As we can see, our method significantly outperforms all other methods in terms of PSNR and SSIM with the EyeQ dataset. Our approach also shows the best results compared to the state-of-the-art for the Mendeley dataset. In terms of SITT, it shows competitive outcomes taking the third position after CLAHE (Shamsudeen & Raju, 2016) and original CycleGAN (You et al., 2019).

To conduct an ablation study of our work, we trained and tested the original fully convolutional CycleGAN model with the same datasets. Total training time for the traditional CycleGAN required 40 hours to converge, 15 hours more than ours. Additionally, testing PSNR values of our and traditional methods are 31.138 dB and 25.577 dB for the EyeQ dataset, and for the Mendeley dataset, the values are 27.798 dB and

Table 4

Comparison with state-of-the-art methods.

Dataset	Method	PSNR	SSIM	SITT
EyeQ (Fu et al., 2019)	CLAHE (Shamsudeen & Raju, 2016)	25.152 dB	0.716	**17ms**
	CycleGAN (You et al., 2019)	25.577 dB	0.882	90ms
	Cycle-CBAM (Wan et al., 2021)	26.263 dB	0.901	100ms
	ArcNet (H. Li et al., 2021)	21.073 dB	0.87	183ms
	DPFR (S. Zhang et al., 2022)	12.423 dB	0.34	449ms
	Ours	**31.138 dB**	**0.919**	97ms
Mendeley (Yoo et al., 2020)	CLAHE (Shamsudeen & Raju, 2016)	26.07 dB	0.77	**17ms**
	CycleGAN (You et al., 2019)	26.08 dB	0.9	90ms
	Cycle-CBAM (Wan et al., 2021)	25.808 dB	0.901	100ms
	ArcNet (H. Li et al., 2021)	22.17 dB	0.891	183ms
	DPFR (S. Zhang et al., 2022)	12.797 dB	0.304	449ms
	Ours	**27.798 dB**	**0.904**	97ms

26.08 dB, respectively. The SSIM values are 0.919 and 0.882 for the EyeQ and 0.904 and 0.9 for the Mendeley datasets, respectively. On average, testing one image takes 90ms for CNN CycleGAN and 97ms for our method, meaning that our modification did not add much computational expense. Fig. 19 demonstrates the qualitative comparison of these two methods. Our method is better at restoring tiny, barely visible blood vessels and preserving structural and color information of original images.

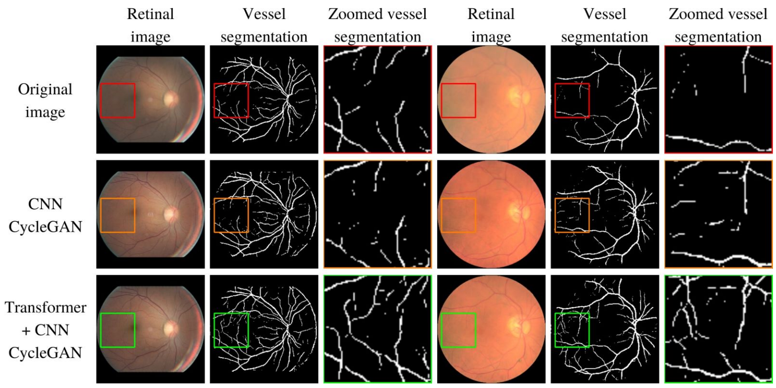

Figure 19. Qualitative comparison of tiny blood vessels restoration.

5.3 Retinal Image Super-Resolution

The proposed method has been qualitatively and quantitatively evaluated and compared with state-of-the-art methods, including SRResNet (Ledig et al., 2017), SR-GAN (Ledig et al., 2017), ESRGAN (X. Wang et al., 2018), and recent FENet (Behjati et al., 2022), and ESRT (Lu et al., 2022). These super-resolution methods have been trained and tested in the same experimental setup for a fair comparison.

Table 5

*Comparison of Single Image Testing Time of EyeQ test set (Fu et al., 2019) (SITT) in seconds, GPU Memory consumption and Number of Parameters. The **bold** and underlined numbers represent the best and second-best results.*

Factor	Method	SITT ↓	Memory ↓	Parameters
×2	SRResNet (Ledig et al., 2017)	0.484	7.8 Gb	27,400,001
	SRGAN (Ledig et al., 2017)	0.484	7.8 Gb	24,965,508
	ESRGAN (X. Wang et al., 2018)	4.874	**6.3** Gb	40,222,852
	FENet (Behjati et al., 2022)	0.273	7.9 Gb	638,086
	ESRT (Lu et al., 2022)	6.443	7.8 Gb	677,783
	Ours	**0.132**	7.0 Gb	62,399,620
×4	SRResNet (Ledig et al., 2017)	0.24	7.3 Gb	27,547,457
	SRGAN (Ledig et al., 2017)	0.246	7.1 Gb	25,113,284
	ESRGAN (X. Wang et al., 2018)	1.267	**4.7** Gb	40,259,780
	FENet (Behjati et al., 2022)	0.255	5.0 Gb	675,462
	ESRT (Lu et al., 2022)	1.801	7.8 Gb	751,767
	Ours	**0.122**	7.0 Gb	61,219,972
×8	SRResNet (Ledig et al., 2017)	0.181	6.0 Gb	27,694,913
	SRGAN (Ledig et al., 2017)	0.19	5.8 Gb	25,261,060
	ESRGAN (X. Wang et al., 2018)	0.381	**4.1** Gb	40,296,708
	FENet (Behjati et al., 2022)	0.247	5.3 Gb	712,838
	ESRT (Lu et al., 2022)	0.678	7.8 Gb	825,751
	Ours	**0.117**	7.0 Gb	60,925,060

Figure 20 demonstrates the qualitative results of given SR methods, where we can see the superiority of our model. All methods performed satisfactorily on ×2 super-resolution. However, a significant difference is noticeable in ×4 and ×8 super-resolution. For instance, SRGAN (Ledig et al., 2017), SRResNet (Ledig et al., 2017), ESRGAN (X. Wang et al., 2018), and ESRT (Lu et al., 2022) generate good-quality images making the blood vessels visible. On the other hand, if we compare these images with original high-resolution ground truth images, we can see the high presence of noise, while ESRT produces slightly over-smoothed images. FENet (Behjati et al., 2022) produces blurry images, leading to the loss of important information, especially tiny blood vessels. Our method produces images closest to the ground truth and has the highest SSIM value, which is crucial for medical images.

For qualitative comparison, the ground truth images and the images obtained

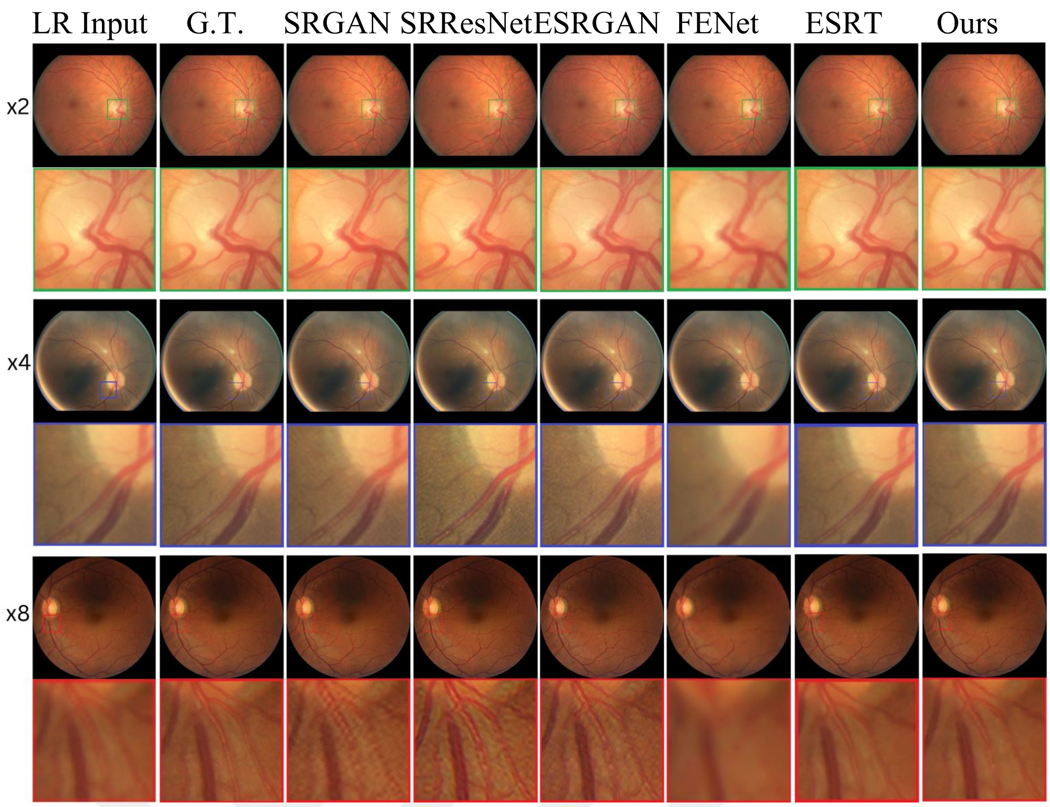

Figure 20. Qualitative comparison of Ground Truth (G.T.) with SRGAN (Ledig et al., 2017), SRResNet (Ledig et al., 2017), ESRGAN (X. Wang et al., 2018), FENet (Behjati et al., 2022), ESRT (Lu et al., 2022) and ours. For better visualization, we cropped parts of high-resolution images.

using all the other methods used for comparative analysis are given in Figure 21. As we can notice from Figure 21, the visual illustration of comparison of our model along with SRGAN, ESRGAN, and SRResNet are similar to the original (ground truth) images. However, FENet has less visible retinal blood vessels caused by blurred output. Hence, taking into account the computational efficiency in terms of time and memory usage of our proposed model that is demonstrated in Table 5, we can consider it efficient based on its quantitative and qualitative analysis results.

In Table 5, it is evident that the proposed method is the least time-consuming than other methods. Additionally, our method has the highest number of parameters due to embedding a ViT encoder that normally has more parameters than CNN. However, the proposed model maintains comparatively low GPU memory consumption for all super-resolution factors. In particular, for $\times 2$, the proposed model is second in terms of memory efficiency after ESRGAN (X. Wang et al., 2018); for $\times 4$ super-resolution,

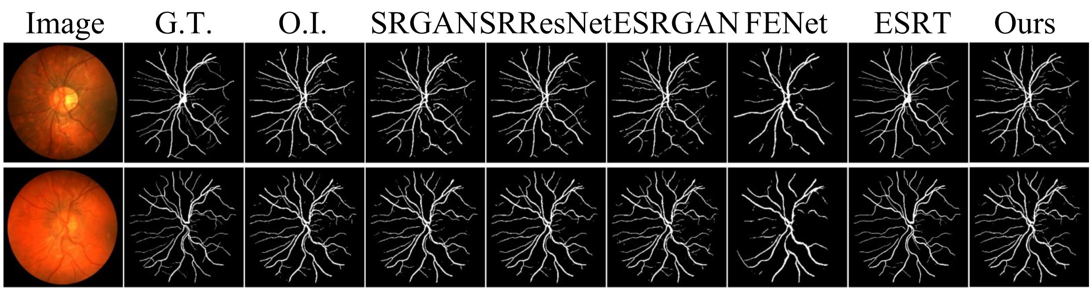

Figure 21. Qualitative comparison of segmentation results of Ground Truth (G.T.) with Original Image, SRGAN (Ledig et al., 2017), SRResNet (Ledig et al., 2017), ESRGAN (X. Wang et al., 2018), FENet (Behjati et al., 2022), ESRT (Lu et al., 2022) and ours.

it comes after ESRGAN (X. Wang et al., 2018) and FENet (Behjati et al., 2022); finally, for $\times 8$ the proposed model still outperforms another ViT-based method ESRT (Lu et al., 2022). On average, the proposed model takes testing time of 0.132, 0.122, and 0.117 seconds per image for $\times 2$, $\times 4$, and $\times 8$ up-scaling factors, respectively. Our proposed method significantly differs from the other existing methods (Behjati et al., 2022; Ledig et al., 2017; X. Wang et al., 2018). Particularly, ESRT (Lu et al., 2022) and ESRGAN (X. Wang et al., 2018) require more than 6.4 and 4.8 seconds for $\times 2$ single image super-resolution, respectively; this is almost 49 and 37 times longer than our result. Both SRGAN (Ledig et al., 2017) and SRResNet (Ledig et al., 2017) up-sample one image in about 0.49 seconds, which is almost 4 times longer than the proposed method. FENet (Behjati et al., 2022) up-scales an image by a factor of 2 in 0.273 seconds. As we can see, our model outperforms all other methods in terms of SITT for all super-resolution factors by a large margin.

In Table 6, we can observe that our model achieves the highest PSNR and SSIM values in most cases for all given up-scaling factors; only for $\times 2$ super-resolution it is slightly (0.001) outperformed by ESRT (Lu et al., 2022) in terms of SSIM. For $\times 2$, $\times 4$, $\times 8$ super-resolution, the average testing PSNR values are 46.146 dB, 43.336 dB, 38.89 dB, and SSIM values are 0.982, 0.977, 0.942, respectively. The proposed method, along with ESRT (Lu et al., 2022), which is also based on ViT, have shown very close results that are higher than those of other models. This is due to embedding ViT into the models as it has proven to outperform CNN in computer vision tasks; the experiments have proven this. The SSIM values of our model do not change dramatically when the

Table 6

*Quantitative results of EyeQ test set (Fu et al., 2019). The **bold** and underlined numbers represent the best and second-best results.*

Factor	Method	PSNR ↑	SSIM ↑	SITT ↓
×2	SRResNet (Ledig et al., 2017)	37.858 dB	0.917	0.484
	SRGAN (Ledig et al., 2017)	37.355 dB	0.933	0.484
	ESRGAN (X. Wang et al., 2018)	38.324 dB	0.95	4.874
	FENet (Behjati et al., 2022)	40.01 dB	0.96	0.273
	ESRT (Lu et al., 2022)	46.131 dB	**0.983**	6.443
	Ours	**46.146** dB	0.982	**0.132**
×4	SRResNet (Ledig et al., 2017)	34.755 dB	0.866	0.24
	SRGAN (Ledig et al., 2017)	35.392 dB	0.909	0.246
	ESRGAN (X. Wang et al., 2018)	34.706 dB	0.895	1.267
	FENet (Behjati et al., 2022)	35.607 dB	0.918	0.255
	ESRT (Lu et al., 2022)	42.376 dB	0.963	1.801
	Ours	**43.336** dB	**0.977**	**0.122**
×8	SRResNet (Ledig et al., 2017)	33.087 dB	0.879	0.181
	SRGAN (Ledig et al., 2017)	33.88 dB	0.897	0.19
	ESRGAN (X. Wang et al., 2018)	32.574 dB	0.844	0.381
	FENet (Behjati et al., 2022)	31.015 dB	0.876	0.247
	ESRT (Lu et al., 2022)	38.679 dB	0.941	0.678
	Ours	**38.89** dB	**0.942**	**0.117**

up-scaling factors increase. It is significant for medical images as the overall image structure is generally unchanged.

In addition, we performed vessel segmentation using UNet (Ronneberger et al., 2015) to validate the quantitative and qualitative results of the proposed method. Firstly, we resized the images to 512×512 pixels and up-scaled by a factor of $\times 4$ to 2048×2048. Table 7 presents the evaluation results with other methods (Behjati et al., 2022; Ledig et al., 2017; Lu et al., 2022; X. Wang et al., 2018). The presented evaluation metrics include the Jaccard similarity index, F1-score, MCC, kappa, precision, recall, and accuracy. Almost all the models performed well, and there is no significant difference in evaluation metrics. However, if we take into account the computational time from Table 5, in which our model significantly outperforms all other models in terms of SITT, we may conclude that the proposed model utilizes its performance more efficiently compared to other methods as SITT is notably lower and our evaluation results are competing. For the DRIVE dataset (Staal et al., 2004), our method has shown competitive results with the model ESRGAN (X. Wang et al., 2018) and SRResNet (Ledig et al., 2017); in particular, our method obtains the highest values of Jaccard, MCC,

Table 7

Quantitative results comparison for ×4 up-scaled images of DRIVE (Staal et al., 2004), STARE (Hoover et al., 2000), CHASE DB1 (Fraz et al., 2012) datasets. The ***bold*** *and underlined numbers represent the best and second-best results. The proposed work is compared with SRResNet (Ledig et al., 2017), SRGAN (Ledig et al., 2017), ESRGAN (X. Wang et al., 2018), FENet (Behjati et al., 2022), ESRT (Lu et al., 2022).*

Data	Method	Jaccard ↑	MCC ↑	Kappa ↑	F1-score ↑	Precision ↑	Recall ↑	Accuracy ↑
DRIVE	SRResNet	<u>0.6642</u>	0.7765	0.7764	**0.7982**	0.8080	0.7832	<u>0.9653</u>
	SRGAN	0.6489	0.7692	0.7686	0.7869	**0.8191**	0.7593	0.9643
	ESRGAN	0.6641	<u>0.7813</u>	<u>0.7811</u>	0.798	0.8171	<u>0.7835</u>	**0.9656**
	FENet	0.4935	0.6344	0.6325	0.6602	0.7134	0.6183	0.9448
	ESRT	0.6575	0.7770	0.7756	0.7942	0.7540	**0.8405**	0.9646
	Ours	**0.6643**	**0.7820**	**0.7818**	<u>0.7981</u>	<u>0.8186</u>	0.7834	**0.9656**
	Original images	0.6659	0.7833	0.7830	0.7992	0.8213	0.7829	0.9659
STARE	SRResNet	0.6501	<u>0.7709</u>	<u>0.7709</u>	<u>0.791</u>	<u>0.7813</u>	0.795	**0.9682**
	SRGAN	0.6122	0.7419	0.7418	0.7576	0.7686	0.7539	0.9643
	ESRGAN	<u>0.6509</u>	**0.7744**	**0.7744**	**0.7913**	**0.7854**	<u>0.7981</u>	<u>0.9681</u>
	FENet	0.5153	0.6560	0.6556	0.6771	0.7034	0.6595	0.9534
	ESRT	**0.6544**	0.7674	0.7673	0.7902	0.7746	**0.799**	0.9614
	Ours	0.6481	0.7692	0.7692	0.7848	0.7805	0.793	0.9676
	Original images	0.6656	0.7839	0.7839	0.7979	0.7997	0.8010	0.9697
CHASE DB1	SRResNet	<u>0.6086</u>	<u>0.7419</u>	<u>0.7399</u>	0.7565	0.7090	0.814	0.9664
	SRGAN	0.5881	0.7240	0.7228	0.7404	0.7035	0.7843	0.9647
	ESRGAN	0.6052	0.7240	0.7371	0.7538	<u>0.7108</u>	0.8055	0.9662
	FENet	0.4697	0.6158	0.6158	0.6389	0.6501	0.6307	0.9542
	ESRT	0.6096	0.7390	0.7365	<u>0.7599</u>	0.7009	<u>0.8159</u>	**0.9671**
	Ours	**0.6135**	**0.7460**	**0.7440**	**0.7603**	**0.7138**	**0.8164**	<u>0.967</u>
	Original images	0.6077	0.7409	0.7391	0.7558	0.7116	0.8088	0.9665

and kappa which demonstrates strongest positive correlation and similarity between classes. For the STARE dataset (Hoover et al., 2000), the model ESRGAN (X. Wang et al., 2018) has slightly outperformed our method with minor differences in its results while up-scaling the images. However, our proposed model performs better at up-sampling images when compared with all the other models used for CHASE DB1 (Fraz et al., 2012), having the highest Jaccard, MCC, and kappa rates. In this evaluation stage, the proposed hybrid model along with ESRGAN (X. Wang et al., 2018), SRResNet (Ledig et al., 2017), and ESRT (Lu et al., 2022) have shown favorable outcomes; the state-of-the-art methods have deeper and more complex architectures, which leads to higher computational time compared to ours.

To further validate the effectiveness of the ViT encoder, we performed an ablation study. To do so, we have built two additional models to validate the effectiveness of embedding the ViT encoder and utilizing adaptive patch embedding. As a result, we removed the ViT part from the proposed model with ViT and an adaptive patch layer (AP) to develop the first model. To build the second model, we used the proposed model and fixed the patch size to equal 16. As the super-resolution factor increases, we apply progressively growing architecture appending an additional layer to the CNN decoder part. We trained both models on the same data in the same conditions. The results are quantitatively and qualitatively compared.

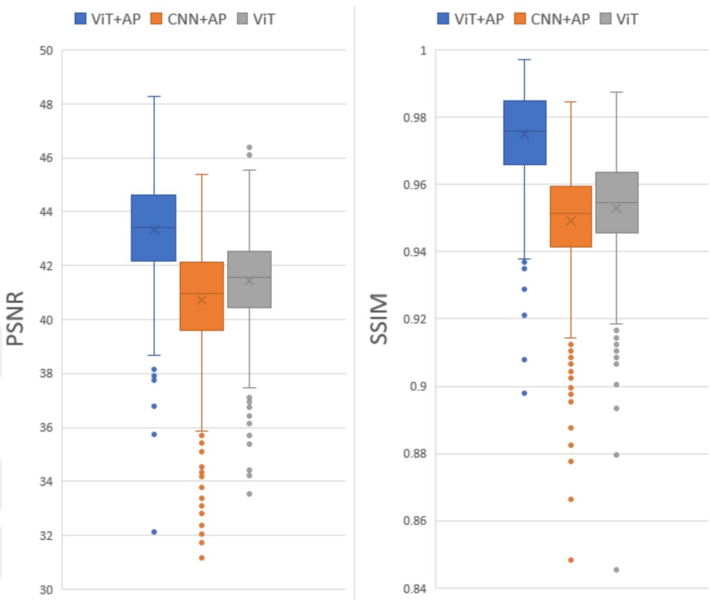

Figure 22. Comparison of the image quality of each model for $\times 4$ super-resolution. From left to right: box plots of PSNR and SSIM values.

As we can observe, the ViT-based model with AP outperforms the ViT-based model with fixed patch size and the CNN-based method in terms of PSNR and SSIM as presented in Table 8. As we can see in Table 9, the CNN-based method is superior to the ViT-based model with fixed patch size in the DRIVE (Staal et al., 2004) dataset in terms of Jaccard, MCC, and kappa. However, it is still outperformed by the ViT model with AP. When testing with STARE (Hoover et al., 2000) dataset, the proposed model has the highest Jaccard value, but it is slightly outperformed by the CNN-based method regarding MCC and kappa. Finally, the proposed method has shown superior results in CHASE DB1 (Fraz et al., 2012) dataset. The proposed method has outperformed other models of Jaccard, MCC, and kappa, proving its superior similarity and positive correlation between segmentation classes. The CNN-based method takes the least time for a single image super-resolution. The average difference between the proposed and CNN-based methods in SITT is only 0.013 seconds. Particularly, the proposed method takes 0.132, 0.122, and 0.117 seconds for each image to do $\times 2$, $\times 4$, and $\times 8$ super-resolution, respectively, which is slightly more than the CNN model that took 0.12, 0.109, 0.101 seconds, respectively. We have also performed statistical analysis, including a one-way variance analysis (ANOVA). In our experiments, a p-value less than 0.05 is considered statistically significant. The results are shown as mean $\pm$ standard deviation. As we can observe from Figure 22 that shows statistical results, utilizing the ViT encoder with AP has shown better performance. In particular, for $\times 4$ super-resolution, the values of mean and standard deviation for PSNR are 43.336 ± 1.80, 41.4 ± 1.67, 40.723 ± 2.03 for ViT+AP, ViT, and CNN+AP models, respectively. For SSIM, the values are 0.977 ± 0.01, 0.953 ± 0.02, and 0.949 ± 0.02 for ViT+AP, ViT, and CNN+AP models, respectively. In all cases, the p-value equals 0, which proves a significant difference between these models.

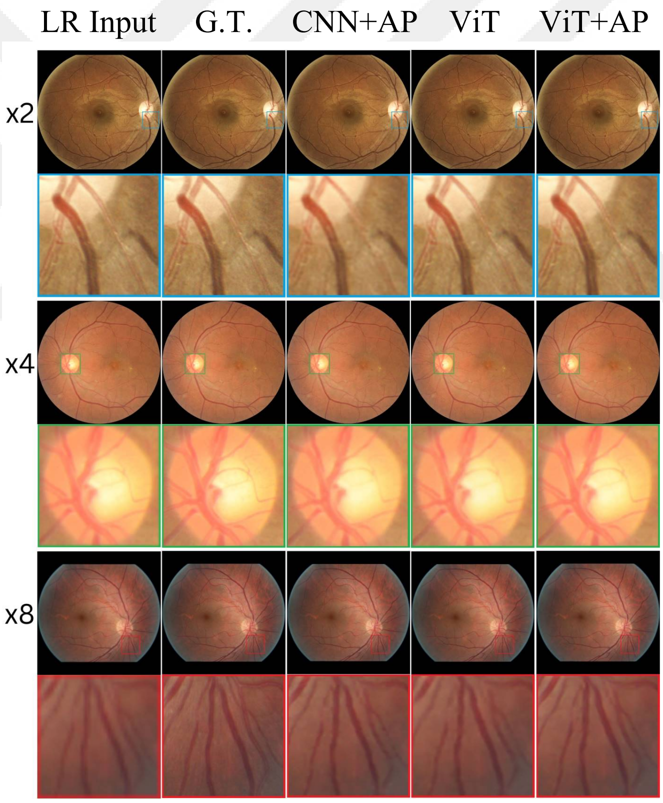

Figure 23. Qualitative comparison of ablation study. Left to right: Low-Resolution Input, Ground Truth, the model with CNN and AP, ViT encoder without AP, and the proposed method with ViT and AP. From top to bottom: $\times 2$, $\times 4$, $\times 8$ up-scaling.

As can be inferred from Figure 23, ViT-based models perform better at preserving the structures of the retinal images, particularly retinal blood vessels. The resulting images are smoother and contain less noise than the CNN model. In addition, the CNN-based model produces less sharp, blurry results, missing some tiny vessels compared to our method. Our method that includes ViT and AP has shown higher quality results than the model with ViT but with no AP since some small vessels are less visible than the proposed method's outcomes. The difference can be noticed in $\times 8$ up-scaling, where some small blood vessels are not as straightforward as in the proposed ViT-based model's output. Therefore, these results prove the effectiveness of embedding the ViT encoder and AP layer into our model.

Table 8

*Quantitative ablation study results of EyeQ test set (Fu et al., 2019). The **bold** and <u>underlined</u> numbers represent the best and second-best results.*

Factor	ViT	AP	PSNR ↑	SSIM ↑	SITT (sec) ↓
$\times 2$	✓	✓	**46.146** dB	**0.982**	<u>0.132</u>
	✓	✗	**46.146** dB	**0.982**	<u>0.132</u>
	✗	✓	<u>41.619</u> dB	<u>0.949</u>	**0.12**
$\times 4$	✓	✓	**43.336** dB	**0.977**	0.122
	✓	✗	<u>41.4</u> dB	<u>0.953</u>	<u>0.114</u>
	✗	✓	40.723 dB	0.949	**0.109**
$\times 8$	✓	✓	**38.89** dB	**0.942**	0.117
	✓	✗	<u>38.527</u> dB	0.934	<u>0.108</u>
	✗	✓	37.906 dB	<u>0.936</u>	**0.101**

Table 9

*Quantitative ablation study results $\times 4$ up-scaled images of DRIVE (Staal et al., 2004), STARE (Hoover et al., 2000), CHASE DB1 (Fraz et al., 2012) datasets. The **bold** and <u>underlined</u> numbers represent the best and second-best results.*

Dataset	ViT	AP	Jaccard ↑	MCC ↑	Kappa ↑	F1-score ↑	Precision ↑	Recall ↑	Accuracy ↑
(Staal et al., 2004)	✗	✓	<u>0.6622</u>	<u>0.7793</u>	<u>0.7791</u>	<u>0.7938</u>	<u>0.8155</u>	<u>0.7819</u>	<u>0.9649</u>
	✓	✗	0.6035	0.7095	0.7091	0.7322	0.7584	0.7110	0.9546
	✓	✓	**0.6643**	**0.7820**	**0.7818**	**0.7981**	**0.8186**	**0.7834**	**0.9656**
	Original images		0.6659	0.7832	0.7829	0.7992	0.8213	0.7829	0.9659
(Hoover et al., 2000)	✗	✓	0.645	0.7675	0.7675	0.7835	0.7798	**0.7915**	<u>0.9667</u>
	✓	✗	<u>0.6471</u>	**0.7694**	**0.7693**	<u>0.7843</u>	**0.7824**	<u>0.7915</u>	**0.9676**
	✓	✓	**0.6481**	<u>0.7692</u>	<u>0.7692</u>	**0.7848**	<u>0.7805</u>	**0.793**	**0.9676**
	Original images		0.6656	0.7839	0.7839	0.7979	0.7997	0.8010	0.9697
(Fraz et al., 2012)	✗	✓	0.6099	0.7429	0.7410	0.7576	0.7125	0.8119	0.9666
	✓	✗	<u>0.6107</u>	<u>0.7434</u>	<u>0.7416</u>	<u>0.7581</u>	<u>0.7132</u>	<u>0.8121</u>	<u>0.9667</u>
	✓	✓	**0.6135**	**0.7460**	**0.7440**	**0.7603**	**0.7138**	**0.8164**	**0.967**
	Original images		0.6077	0.7409	0.7391	0.7558	0.7116	0.8088	0.9665

5.4 Retinal Image Generation

We are proposing a retinal segmentation dataset (ReTree), which consists of images belonging to 2 classes: colored fundus images and corresponding binary vessel maps. The total number of images in each class is 30,000. The dataset is split into

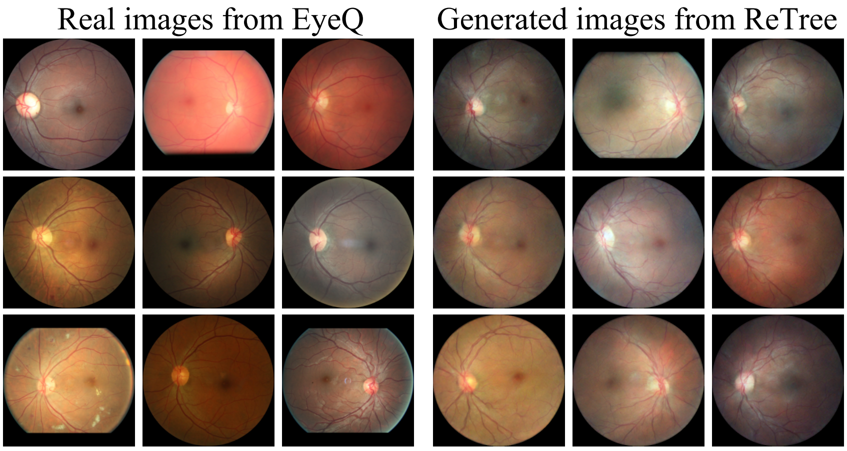

Figure 24. The generated results of the proposed method (on the right) along with real images from the EyeQ (Fu et al., 2019) dataset (on the left). The generated images by our ReTree method have similar qualities as real retinal images.

the train, validation, and test sets with 28,800, 200, and 1,000 images, respectively. The image resolution is 512×512 pixels. The proposed dataset has been extensively evaluated using quantitative and qualitative measures. To validate the effectiveness of our dataset, we trained the biomedical segmentation UNet (Ronneberger et al., 2015) model with it and tested the model with real images. The results indicate the improvement over training with solely the real images since the number of real images needs to be significantly increased to train the deep learning model efficiently. Fig. 24 presents the qualitative comparison with EyeQ (Fu et al., 2019), where we can see the similarity and realism of generated images, having similar qualities such as vessel structure, accurate disk/cup region, and realistic color.

In our experiments, we compared our method with GAN by Costa et al. (Costa, Galdran, Meyer, Niemeijer, et al., 2017), since the authors of most of the other methods that focused on retinal image generation did not publish their implementations and generated datasets. As a result, we could not compare our method with these works as we did not receive any positive reply from the corresponding authors. However, we compared the proposed method with Improved DDPM (IDDPM) (Nichol & Dhariwal, 2021) in the ablation study. In addition, the proposed model does not suffer from the mode collapse that is present in GANs; as a result, the images obtained by our model are diverse in terms of color and overall structure. Fig. 25 shows the qualitative comparison with another GAN-based method (Costa, Galdran, Meyer, Niemeijer, et al., 2017);

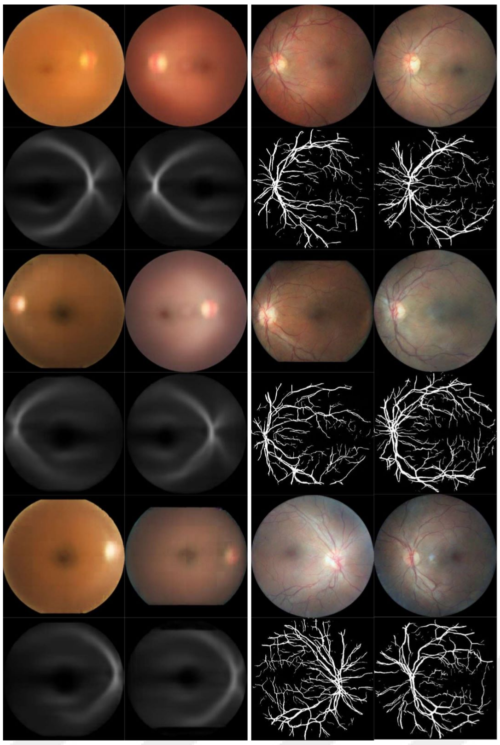

Figure 25. The qualitative comparison of the proposed method (on the right) with another GAN-based (Costa, Galdran, Meyer, Niemeijer, et al., 2017) solution (on the left). The GAN could not converge, leading to low-quality results compared to ours. Alternatively, images generated by our DDPM have an accurate vessel and disk/cup structure and realistic and diverse colors.

as we can observe, the GAN-based model has failed to converge since the generated images lack clear visual vessel structure. In addition, the generated vessel maps need more visual clarity to be used for retinal segmentation tasks.

Additionally, we validated the proposed ReTree dataset using UNet (Ronneberger et al., 2015). To do so, we trained UNet with generated images from the ReTree dataset. Then we tested it using real images from three publicly available datasets such as DRIVE (Staal et al., 2004), STARE (Hoover et al., 2000), and CHASE DB1 (Fraz et al., 2012). The qualitative results are presented in Fig. 26. As we can observe, using ReTree images and images from original datasets is best, leading to better qualitative results. As we can also observe, the results of GAN method (Costa, Galdran, Meyer, Niemeijer, et al., 2017) failed in segmentation, while GAN (Costa, Galdran, Meyer, Niemeijer, et al., 2017) + original dataset could improve the segmentation process. When training with the ReTree dataset, UNet still performs well in vessel segmenta-

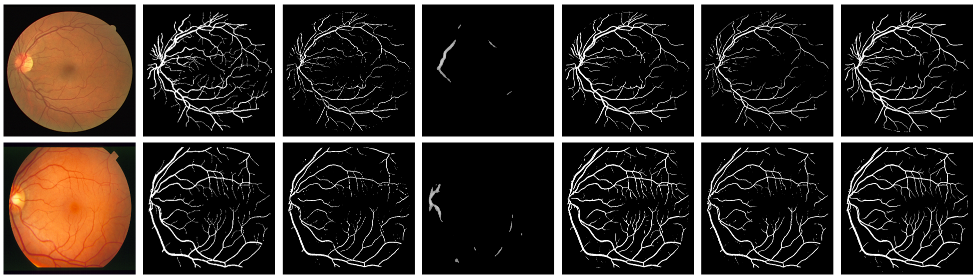

Figure 26. The qualitative comparison of the segmentation results.As we can observe, our method performs best when combined with the original dataset.

tion; however, it is outperformed by ReTree + original dataset.

Table 10 shows the FID scores along with single image generation time (SIGT) obtained by GAN-based (Costa, Galdran, Meyer, Niemeijer, et al., 2017) and our models. As we can observe, the proposed method outperforms the other techniques in having lower FID; in particular, the FI distance between ReTree and EyeQ images is more than 3 times less than that of the GAN-based method. In addition, GAN takes 6.90 seconds to generate a single image pair, while ours generates a single pair in 6.23 seconds. The proposed architecture significantly decreased the generation time by a factor of 2 compared to IDDPM (Nichol & Dhariwal, 2021).

Table 10

Quantitative comparison with GAN (Costa, Galdran, Meyer, Niemeijer, et al., 2017) in terms of Frechet Inception Distance (FID) and Single Image Generation Time (SIGT) in seconds.

Method	**FID**	**SIGT (sec)**
GAN (Costa, Galdran, Meyer, Niemeijer, et al., 2017)	162.50	6.90
Ours	**48.45**	**6.23**

Table 11 represents the second quantitative evaluation stage for the proposed ReTree dataset. In this stage, we performed retinal vessel segmentation using UNet (Ronneberger et al., 2015). First, it was trained with original images from three publicly available datasets, such as DRIVE (Staal et al., 2004), STARE (Hoover et al., 2000), and CHASE DB1 (Fraz et al., 2012). Next, it was trained with generated images only and then generated + original images from each dataset. The three UNet versions are tested with original images from three datasets during the evaluation. As we can observe from Table 11, the proposed dataset performs best when combined with original datasets.

Table 11

Quantitative comparison with GAN (Costa, Galdran, Meyer, Niemeijer, et al., 2017) for vessel segmentation using DRIVE (Staal et al., 2004), STARE (Hoover et al., 2000), CHASE DB1 (Fraz et al., 2012) datasets and UNet (Ronneberger et al., 2015).

Test data	Train data	Jaccard ↑	MCC ↑	Kappa ↑	F1-score ↑	Precision ↑	Recall ↑	Accuracy ↑
DRIVE	DRIVE + GAN	0.4419	0.6233	0.5833	0.6065	**0.9017**	0.4662	0.9496
	DRIVE + ReTree	**0.6161**	**0.7449**	**0.7414**	**0.7620**	0.8601	0.7162	**0.9723**
	GAN	0.0885	0.2252	0.1430	0.1585	0.7497	0.0928	0.9190
	ReTree	0.6023	0.7389	0.7305	0.7506	0.6714	**0.8596**	0.9621
	DRIVE	0.5042	0.6639	0.6445	0.6681	0.8517	0.5563	0.9532
STARE	STARE + GAN	0.5094	0.6570	0.6410	0.6616	**0.7912**	0.5998	0.9581
	STARE + ReTree	**0.5864**	**0.7240**	**0.7162**	**0.7355**	0.7900	**0.7303**	**0.9632**
	GAN	0.0589	0.1521	0.0928	0.1084	0.5548	0.0654	0.9238
	ReTree	0.4929	0.6324	0.6218	0.6455	0.7015	0.6324	0.9531
	STARE	0.4557	0.5918	0.5840	0.6123	0.6474	0.6068	0.9456
CHASE DB1	CHASE DB1 + GAN	0.4593	0.6052	0.6040	0.6291	0.6413	0.6210	0.9530
	CHASE DB1 + ReTree	**0.5394**	**0.6686**	**0.6655**	**0.6992**	**0.7194**	**0.7007**	**0.9649**
	GAN	0.0913	0.2277	0.1519	0.1663	0.6320	0.0972	0.9380
	ReTree	0.3580	0.5146	0.5005	0.5256	0.6700	0.4355	0.9497
	CHASE DB1	0.4319	0.5793	0.5716	0.6030	0.5308	0.7033	0.9405

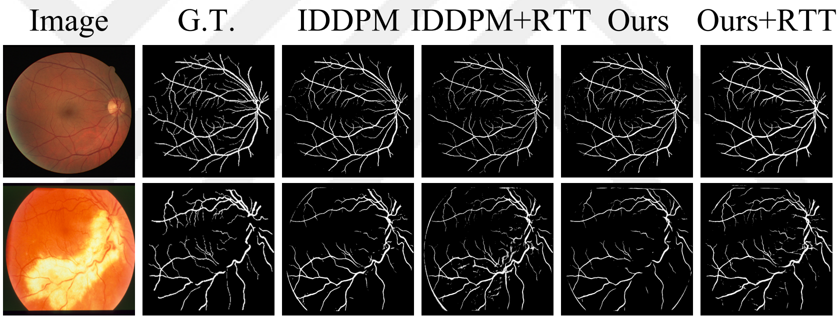

Figure 27. The qualitative comparison of the segmentation results. As we can observe, our method performs best when combined with the original dataset.

To validate the efficiency and performance of the proposed DDPM, we performed an extensive ablation study in which we compared our method with improved DDPM (Nichol & Dhariwal, 2021) that was trained in the original setup using the same two-stage generation; next, we trained this model with the proposed Repetitive Training Technique (RTT). Additionally, we trained the proposed DDPM without RTT. The training of UNet (Ronneberger et al., 2015) has been performed using generated upscaled images with resolutions of 256×256 and 512×512 pixels alone and combined with training data from original corresponding retinal segmentation datasets. Table 12 shows the FID values, SIGT, and total training time. As we can see, embedding RTT in the training process significantly reduces the total training time; for instance, RTT reduces the training time by 32.7% from 16.2 to 10.9 hours for IDDPM (Nichol & Dhariwal, 2021) and by 41.8% from 14.27 to 8.3 hours for the proposed model. Additionally, the proposed model takes less time to train and to generate a single sample; IDDPM requires 12.53 seconds which is two times higher than the proposed model. RTT also reduced the FID values for both models; however, the proposed model has the lowest FID score.

The models have also been trained and tested using segmentation datasets. Table 13 compares segmentation results using three datasets. We performed $\times 4$ and $\times 8$ super-resolution of generated images. The same models have been trained in combination with three segmentation datasets. As we can observe, the proposed model and training technique significantly outperforms the other methods in both super-resolution factors. In particular, the proposed model with RTT obtained the highest values of Jaccard similarity score, MCC, Cohen's kappa, F1-score, precision, recall, and accuracy, which means our method results in the highest similarity and positive correlation between segmentation classes. IDDPM (Nichol & Dhariwal, 2021) without RTT obtained second-best quantitative results in DRIVE (Staal et al., 2004) and STARE (Hoover et al., 2000). However, it is outperformed by the same model with RTT in CHASE DB1 (Fraz et al., 2012). Fig. 27 shows the qualitative comparison between these methods; as we can observe, utilizing our architecture along with RTT leads to better qualitative results than others. Therefore, these results present the effectiveness of the proposed DDPM architecture and RTT, which lead to higher quantitative, qualitative, and computational results.

Table 12

Quantitative comparison with original IDDPM (Nichol & Dhariwal, 2021), IDDPM + RTT, ours without RTT and ours + RTT in terms of Frechet Inception Distance (FID), Single Image Generation Time (SIGT) and training time (in hours) for two stages of generation.

Method	FID	SIGT (sec)	Train time (hours)
IDDPM (Nichol & Dhariwal, 2021) w\o RTT	68.959	6.156 + 6.375	8.8 + 7.4
IDDPM (Nichol & Dhariwal, 2021) w RTT	64.32	6.156 + 6.375	5.4 + 5.1
Ours w\o RTT	55.49	2.97 + 3.26	7.07 + 7.2
Ours w RTT	**48.45**	**2.97 + 3.26**	**4.29 + 4.01**

Furthermore, we conducted experiments with RTT to examine its impact on the training of other deep learning methods. Specifically, we applied RTT to train UNet (Ronneberger et al., 2015) for retinal image segmentation. The results are presented in Table 14. As observed, RTT has a positive effect on UNet, leading to an increase in almost all quantitative metrics. The most notable improvement can be seen in the STARE dataset (Hoover et al., 2000), where all quantitative metrics show an average increase of 3.42% while it reduces training time by an average of 28.47% across all datasets.

Table 13

*Quantitative results comparison for vessel segmentation using DRIVE (Staal et al., 2004), STARE (Hoover et al., 2000), CHASE DB1 (Fraz et al., 2012) datasets. First, we trained the super-resolution network to up-scale generated images by factors of $\times 8$ and $\times 4$ (col. 1). Then, we combined original training datasets (col. 2) with images generated by models from the ablation study (col. 3) to train UNet (Ronneberger et al., 2015). Next, we evaluated the results using test sets of corresponding original datasets and compared them with IDDPM (Nichol & Dhariwal, 2021). The **bold** and underlined numbers represent the best and second-best results.*

Resolution	Test Data	Train data	Jaccard ↑	MCC ↑	Kappa ↑	F1-score ↑	Precision ↑	Recall ↑	Accuracy ↑
512×512	DRIVE	DRIVE + IDDPM w\o RTT	0.6103	0.7417	0.7369	0.7574	0.8335	0.6981	0.9618
		DRIVE + IDDPM w RTT	0.5605	0.7077	0.6956	0.7174	0.8575	0.6207	0.9582
		IDDPM w\o RTT	0.4703	0.6073	0.6054	0.6390	0.6423	0.6427	0.9381
		IDDPM w RTT	0.4756	0.6128	0.6109	0.6437	0.6544	0.6395	0.9397
		DRIVE + Ours w\o RTT	0.5874	0.7266	0.7185	0.7395	0.8505	0.6573	0.9604
		DRIVE + Ours w RTT	**0.6161**	**0.7449**	**0.7414**	**0.7620**	**0.8601**	**0.7162**	**0.9723**
		Ours w\o RTT	0.5046	0.6402	0.6373	0.6700	0.6358	0.7145	0.9399
		Ours w RTT	0.5208	0.6565	0.6549	0.6841	0.6917	0.6824	0.9463
	STARE	STARE + IDDPM w\o RTT	0.5860	0.7226	0.7152	0.7350	0.7712	0.7272	0.9623
		STARE + IDDPM w RTT	0.5574	0.6938	0.6894	0.7122	0.7201	0.7208	0.9573
		IDDPM w\o RTT	0.2964	0.4215	0.4059	0.4376	0.5510	0.4237	0.9297
		IDDPM w RTT	0.4716	0.6171	0.6079	0.6395	0.5974	0.7145	0.9404
		STARE + Ours w\o RTT	0.5767	0.7128	0.7082	0.7296	0.7473	0.7284	0.9600
		STARE + Ours w RTT	**0.5864**	**0.7240**	**0.7162**	**0.7355**	**0.7900**	**0.7303**	**0.9632**
		Ours w\o RTT	0.3910	0.5327	0.5195	0.5546	0.5598	0.5998	0.9318
		Ours w RTT	0.4970	0.6403	0.6339	0.6630	0.6240	0.7247	0.9456
	CHASE	CHASE + IDDPM w\o RTT	0.4709	0.6196	0.6169	0.6397	0.6909	0.5985	0.9567
		CHASE + IDDPM w RTT	0.5161	0.6590	0.6577	0.6806	0.7093	0.6566	0.9572
		IDDPM w\o RTT	0.4201	0.5643	0.5559	0.5883	0.5473	0.6830	0.9401
		IDDPM w RTT	0.4484	0.5947	0.5898	0.6190	0.5597	0.6963	0.9449
		CHASE + Ours w\o RTT	0.4527	0.6003	0.5987	0.6230	0.6640	0.5880	0.9542
		CHASE + Ours w RTT	**0.5394**	**0.6686**	**0.6655**	**0.6992**	**0.7194**	**0.7007**	**0.9649**
		Ours w\o RTT	0.4218	0.5670	0.5616	0.5931	0.5328	0.6730	0.9406
		Ours w RTT	0.4625	0.6074	0.6060	0.6322	0.6124	0.6570	0.9509
256×256	DRIVE	DRIVE + IDDPM w\o RTT	0.6334	0.7572	0.7551	0.7754	**0.8151**	0.7429	0.9626
		DRIVE + IDDPM w RTT	0.6472	0.7649	0.7637	0.7830	0.8081	0.7723	**0.9635**
		IDDPM w\o RTT	0.5143	0.6531	0.6521	0.6813	0.6941	0.6917	0.9497
		IDDPM w RTT	0.5416	0.6691	0.6639	0.6995	0.7034	0.6855	0.9599
		DRIVE + Ours w\o RTT	0.6401	0.7613	0.7598	0.7803	0.7961	0.7695	0.9624
		DRIVE + Ours w RTT	**0.6486**	**0.7684**	**0.7667**	**0.7866**	0.8000	**0.7794**	0.9633
		Ours w\o RTT	0.5543	0.7062	0.6993	0.7271	0.6993	0.7528	0.9594
		Ours w RTT	0.5601	0.7085	0.7041	0.7342	0.7387	0.7464	0.9593
	STARE	STARE + IDDPM w\o RTT	0.5780	0.7105	0.7080	0.7282	0.7618	0.7037	0.9610
		STARE + IDDPM w RTT	0.5844	0.7196	0.7156	0.7355	**0.7900**	0.6957	0.9627
		IDDPM w\o RTT	0.4124	0.4511	0.4383	0.4672	0.5881	0.4617	0.9537
		IDDPM w RTT	0.5156	0.6421	0.6409	0.6691	0.6311	0.7415	0.9594
		STARE + Ours w\o RTT	0.5855	0.7161	0.7143	0.7347	0.7434	0.7325	0.9621
		STARE + Ours w RTT	**0.5869**	**0.7201**	**0.7170**	**0.7384**	0.7558	0.7322	**0.9671**
		Ours w\o RTT	0.4381	0.5687	0.5415	0.5811	0.5871	0.6251	0.9578
		Ours w RTT	0.5411	0.6716	0.6719	0.6916	0.6571	**0.7527**	0.9596
	CHASE	CHASE + IDDPM w\o RTT	0.5570	0.6962	0.6950	0.7153	0.6948	0.7400	0.9621
		CHASE + IDDPM w RTT	0.5855	0.7218	0.7191	**0.7384**	0.6929	**0.7934**	0.9639
		IDDPM w\o RTT	0.4810	0.6128	0.6101	0.6431	0.6008	0.7451	0.9581
		IDDPM w RTT	0.5095	0.6511	0.6481	0.6731	0.6251	0.7531	0.9601
		CHASE + Ours w\o RTT	0.5697	0.7076	0.7061	0.7257	0.6999	0.7567	0.9633
		CHASE + Ours w RTT	**0.5889**	**0.7267**	**0.7254**	0.7351	**0.7088**	0.7661	**0.9642**
		Ours w\o RTT	0.4891	0.6231	0.6216	0.6507	0.5931	0.7328	0.9606
		Ours w RTT	0.5310	0.6631	0.6647	0.6942	0.6785	0.7231	0.9623

Table 14

Quantitative evaluation of UNet (Ronneberger et al., 2015) with and without Repetitive Training Technique and total training time in minutes.

Dataset	RTT	Jaccard	MCC	Kappa	F1	Recall	Precision	Accuracy	Time (mins)
DRIVE (Staal et al., 2004)	✗	0.6393	0.7624	**0.7584**	**0.7793**	**0.7991**	0.7733	0.9615	17.73
	✓	**0.6417**	**0.7640**	0.7575	0.7771	0.7860	**0.7945**	**0.9628**	**13.48**
STARE (Hoover et al., 2000)	✗	0.5712	0.7142	0.6994	0.7212	0.7524	0.7446	0.9579	17.98
	✓	**0.6042**	**0.7412**	**0.7276**	**0.7464**	**0.7549**	**0.7890**	**0.9635**	**12.27**
CHASE DB1 (Fraz et al., 2012)	✗	0.5980	0.7320	0.7299	0.7482	0.7944	**0.7098**	0.9656	14.73
	✓	**0.6102**	**0.7434**	**0.7396**	**0.7577**	**0.8270**	0.7018	**0.9660**	**10.36**

Chapter 6

Conclusion

In the first work, we address the problems of low-quality retinal image restoration and vessel segmentation by proposing a cycle-consistent generative adversarial network and a retinal vessel segmentation model. The convolution block attention modules are embedded into two models to improve the feature extraction. The enhancement model does not require paired datasets because it uses unpaired low- and high-quality images. The proposed restoration and vessel segmentation methods perform better than the state-of-the-art by a large margin. Thus, it can fix the out-of-focus blurring, color distortion, and high, low, and uneven illumination problems of retinal images. However, our method may fail to restore the too low and too high luminance or uneven distribution that needs further investigation in the future.

In the second work, we combined a vision transformer and convolutional neural network, resulting in better retinal image enhancement than fully convolutional CycleGAN and other state-of-the-art techniques. In addition, our goal was to reduce the computational expense of transformer CycleGAN architecture without compromising the performance. This simple yet effective combination has led to comparatively better quantitatively, qualitatively, and computational performance. The model was tested using two datasets, such as EyeQ (Fu et al., 2019) and Mendeley dataset (Yoo et al., 2020). Testing PSNR is 31.138 dB for the first and 27.798 dB for the second dataset. Testing SSIM is 0.919 and 0.904, respectively. The ablation study demonstrates a significant advantage of our method compared to the original CycleGAN architecture.

In the third study, we focused on accurate and efficient retinal image super-resolution by proposing a model that combines the ViT encoder and CNN decoder. To perform $\times 2$, $\times 4$, and $\times 8$ super-resolution, we offer an adaptive patch embedding layer that removes the need to build additional decoding layers to increase the super-resolution factors. Our method has led to significantly greater quantitative and qualitative results than current state-of-the-art methods. It can accurately up-sample retinal images by factors of ×2, ×4, ×8 without an increase in testing time and without significant loss of structural information of retinal images. In addition, we performed an extensive ablation study. The results presented in this paper prove the effectiveness of embedding the ViT encoder and non-fixed patch size. Compared with other methods, our model outperforms them regarding PSNR, SSIM, and testing time. Our method has shown very competitive results during the evaluation of segmented images. In future works, we plan to solve this model's current limitations related to generating tiny vessels in retinal images' bright disc and cup region. We intend to solve it by normalizing the dataset, additional image pre-processing, retinal image synthesis, and more advanced methods. In the future, we intend to develop a more advanced method that is still computationally efficient, fast, and accurate for single-image super-resolution.

In our final project, our main objective was to develop a novel retinal image segmentation dataset generated with the current state-of-the-art class of generation models, namely Denoising Diffusion Probabilistic Models. Additionally, we propose a novel lightweight architecture and a training technique for DDPM that significantly improves the computational, qualitative, and quantitative performance of DDPMs. The proposed DDPM can be trained with higher resolution images, such as 128×128 and 256×256, compared to the original DDPM that generates 64×64 pixels images. This work involves four steps: retinal vessel generation (1), retinal image synthesis (2), single image super-resolution (3), and retinal vessel segmentation (4). The proposed ReTree dataset was extensively evaluated quantitatively and qualitatively. In addition, it was compared with real retinal segmentation datasets. The results prove the superiority of our dataset over other manually collected datasets.

References

Alimanov, A., & Islam, M. B. (2022a). Retinal image restoration and vessel segmentation using modified cycle-cbam and cbam-unet. In *2022 innovations in intelligent systems and applications conference (asyu)* (pp. 1–6).

Alimanov, A., & Islam, M. B. (2022b). Retinal image restoration using transformer and cycle-consistent generative adversarial network. In *2022 international symposium on intelligent signal processing and communication systems (ispacs)* (pp. 1–4).

Alimanov, A., & Islam, M. B. (2023). Denoising diffusion probabilistic model for retinal image generation and segmentation. In *2023 ieee international conference on computational photography (iccp)* (pp. 1–12).

Alimanov, A., Islam, M. B., & Abubacker, N. F. (2023). A hybrid approach for retinal image super-resolution. *Biomedical Engineering Advances*, 100099.

Amit, T., Nachmani, E., Shaharbany, T., & Wolf, L. (2021). Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*.

Andreini, P., Ciano, G., Bonechi, S., Graziani, C., Lachi, V., Mecocci, A., … Bianchini, M. (2021). A two-stage gan for high-resolution retinal image generation and segmentation. *Electronics*, *11*(1), 60.

Appan K, P., & Sivaswamy, J. (2018). Retinal image synthesis for cad development. In *International conference image analysis and recognition* (pp. 613–621).

Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., & Babenko, A. (2021). Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*.

Behjati, P., Rodriguez, P., Tena, C. F., Mehri, A., Roca, F. X., Ozawa, S., & Gonzàlez, J. (2022). Frequency-based enhancement network for efficient super-resolution. *IEEE Access*, *10*, 57383–57397.

Biswas, B., Ghosh, S. K., & Ghosh, A. (2020). Dvae: deep variational auto-encoders for denoising retinal fundus image. In *Hybrid machine intelligence for medical image analysis* (pp. 257–273). Springer.

Cai, J., Zeng, H., Yong, H., Cao, Z., & Zhang, L. (2019). Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 3086–3095).

Cao, L., Li, H., & Zhang, Y. (2020). Retinal image enhancement using low-pass filtering and α-rooting. *Signal Processing*, *170*, 107445.

Chen, H. (2021). Challenges and corresponding solutions of generative adversarial networks (gans): a survey study. In *Journal of physics: Conference series* (Vol. 1827, p. 012066).

Chen, Y., Shi, F., Christodoulou, A. G., Xie, Y., Zhou, Z., & Li, D. (2018). Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network. In *International conference on medical image computing and computer-assisted intervention* (pp. 91–99).

Choi, J., Kim, S., Jeong, Y., Gwon, Y., & Yoon, S. (2021). Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*.

Costa, P., Galdran, A., Meyer, M. I., Abramoff, M. D., Niemeijer, M., Mendonça, A. M., & Campilho, A. (2017). Towards adversarial retinal image synthesis. *arXiv preprint arXiv:1701.08974*.

Costa, P., Galdran, A., Meyer, M. I., Niemeijer, M., Abràmoff, M., Mendonça, A. M., & Campilho, A. (2017). End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, *37*(3), 781–791.

Dai, T., Cai, J., Zhang, Y., Xia, S.-T., & Zhang, L. (2019). Second-order attention network for single image super-resolution. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11065–11074).

Daihong, J., Sai, Z., Lei, D., & Yueming, D. (2022). Multi-scale generative adversarial network for image super-resolution. *Soft Computing*, *26*(8), 3631–3641.

Das, S., & Malathy, C. (2018). Survey on diagnosis of diseases from retinal images. In *Journal of physics: Conference series* (Vol. 1000, p. 012053).

Datta, N. S., Saha, P., Dutta, H. S., Sarkar, D., Biswas, S., & Sarkar, P. (2015). A new contrast enhancement method of retinal images in diabetic screening system. In *2015 ieee 2nd international conference on recent trends in information systems (retis)* (pp. 255–260).

Deng, Z., Cai, Y., Chen, L., Gong, Z., Bao, Q., Yao, X., … Ma, L. (2022). Rformer: Transformer-based generative adversarial network for real fundus image restoration on a new clinical benchmark. *IEEE Journal of Biomedical and Health Informatics*.

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, *34*, 8780–8794.

Diaz-Pinto, A., Colomer, A., Naranjo, V., Morales, S., Xu, Y., & Frangi, A. F. (2019). Retinal image synthesis and semi-supervised learning for glaucoma assessment. *IEEE transactions on medical imaging*, *38*(9), 2211–2218.

Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, *38*(2), 295–307.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Elfwing, S., Uchibe, E., & Doya, K. (2018). Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, *107*, 3–11.

Fraz, M. M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A. R., Owen, C. G., & Barman, S. A. (2012). An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, *59*(9), 2538–2548.

Fu, H., Cheng, J., Xu, Y., Wong, D. W. K., Liu, J., & Cao, X. (2018). Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE transactions on medical imaging*, *37*(7), 1597–1605.

Fu, H., Wang, B., Shen, J., Cui, S., Xu, Y., Liu, J., & Shao, L. (2019). Evaluation of retinal image quality assessment networks in different color-spaces. In *International conference on medical image computing and computer-assisted intervention* (pp. 48–56).

Gao, G., Wang, Z., Li, J., Li, W., Yu, Y., & Zeng, T. (2022). Lightweight bimodal network for single-image super-resolution via symmetric cnn and recursive transformer. *arXiv preprint arXiv:2204.13286*.

Gao, G., Xu, Z., Li, J., Yang, J., Zeng, T., & Qi, G.-J. (2023). Ctcnet: A cnn-transformer cooperation network for face image super-resolution. *IEEE Transactions on Image Processing*, *32*, 1978–1991.

Gao, Z., Li, J., Guo, J., Chen, Y., Yi, Z., & Zhong, J. (2018). Diagnosis of diabetic retinopathy using deep neural networks. *IEEE Access*, *7*, 3360–3370.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

Grassmann, F., Mengelkamp, J., Brandl, C., Harsch, S., Zimmermann, M. E., Linkohr, B., … Weber, B. H. (2018). A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology*, *125*(9), 1410–1420.

Guibas, J. T., Virdi, T. S., & Li, P. S. (2017). Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872*.

Guo, X., Lu, X., Lin, Q., Zhang, J., Hu, X., & Che, S. (2022). A novel retinal image generation model with the preservation of structural similarity and high resolution. *Biomedical Signal Processing and Control*, *78*, 104004.

Hao, H., Xu, C., Zhang, D., Yan, Q., Zhang, J., Liu, Y., & Zhao, Y. (2022). Sparse-based domain adaptation network for octa image super-resolution reconstruction. *IEEE Journal of Biomedical and Health Informatics*, *26*(9), 4402–4413.

He, K., Sun, J., & Tang, X. (2012). Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, *35*(6), 1397–1409.

He, X., Mo, Z., Wang, P., Liu, Y., Yang, M., & Cheng, J. (2019). Ode-inspired network design for single image super-resolution. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 1732–1741).

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, *30*.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Hoover, A., Kouznetsova, V., & Goldbaum, M. (2000). Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, *19*(3), 203–210.

Hu, Y., Gao, X., Li, J., Huang, Y., & Wang, H. (2021). Single image super-resolution with multi-scale information cross-fusion network. *Signal Processing*, *179*, 107831.

Hu, Y., Li, J., Huang, Y., & Gao, X. (2019). Channel-wise and spatial feature modulation network for single image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(11), 3911–3927.

Hui, Z., Wang, X., & Gao, X. (2018). Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 723–731).

Jiang, Y., Liu, W., Wu, C., & Yao, H. (2021). Multi-scale and multi-branch convolutional neural network for retinal image segmentation. *Symmetry*, *13*(3), 365.

Keys, R. (1981). Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, *29*(6), 1153–1160.

Kim, M., Kim, Y. N., Jang, M., Hwang, J., Kim, H.-K., Yoon, S. C., … Kim, N. (2022). Synthesizing realistic high-resolution retina image by style-based generative adversarial network and its utilization. *Scientific Reports*, *12*(1), 17307.

Lan, R., Sun, L., Liu, Z., Lu, H., Pang, C., & Luo, X. (2020). Madnet: a fast and lightweight network for single-image super resolution. *IEEE transactions on cybernetics*, *51*(3), 1443–1453.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., … others (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4681–4690).

Lee, S. H., Lee, S., & Song, B. C. (2021). Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*.

Li, C., & Wand, M. (2016). Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision* (pp. 702–716).

Li, H., Liu, H., Hu, Y., Fu, H., Zhao, Y., Miao, H., & Liu, J. (2022). An annotation-free restoration network for cataractous fundus images. *IEEE Transactions on Medical Imaging*.

Li, H., Liu, H., Hu, Y., Higashita, R., Zhao, Y., Qi, H., & Liu, J. (2021). Restoration of cataract fundus images via unsupervised domain adaptation. In *2021 ieee 18th international symposium on biomedical imaging (isbi)* (pp. 516–520).

Li, L., Verma, M., Nakashima, Y., Nagahara, H., & Kawasaki, R. (2020). Iternet: Retinal image segmentation utilizing structural redundancy in vessel networks. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 3656–3665).

Li, W., Liu, G., He, Y., Wang, J., Kong, W., & Shi, G. (2020). Quality improvement of adaptive optics retinal images using conditional adversarial networks. *Biomedical optics express*, *11*(2), 831–849.

Li, Y., Yang, J., Ni, J., Elazab, A., & Wu, J. (2021). Ta-net: Triple attention network for medical image segmentation. *Computers in Biology and Medicine*, *137*, 104836.

Liu, J., Liu, Y., Wu, H., Wang, J., Li, X., & Zhang, C. (2021). Single image super-resolution using feature adaptive learning and global structure sparsity. *Signal Processing*, *188*, 108184.

Liu, Y.-C., Yang, H.-H., Huck Yang, C.-H., Huang, J.-H., Tian, M., Morikawa, H., … Tegner, J. (2019). Synthesizing new retinal symptom images by multiple generative models. In *Computer vision–accv 2018 workshops: 14th asian conference on computer vision, perth, australia, december 2–6, 2018, revised selected papers 14* (pp. 235–250).

Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., & Zeng, T. (2022). Transformer for single image super-resolution. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 457–466).

Lv, Y., Ma, H., Li, J., & Liu, S. (2021). Fusing dense and rezero residual networks for super-resolution of retinal images. *Pattern Recognition Letters*, *149*, 120–129.

Mahapatra, D., Bozorgtabar, B., & Garnavi, R. (2019). Image super-resolution using progressive generative adversarial networks for medical image analysis. *Computerized Medical Imaging and Graphics*, *71*, 30–39.

Mahapatra, D., Bozorgtabar, B., Hewavitharanage, S., & Garnavi, R. (2017). Image super resolution using generative adversarial networks and local saliency maps for retinal image analysis. In *International conference on medical image computing and computer-assisted intervention* (pp. 382–390).

Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International conference on machine learning* (pp. 8162–8171).

Niu, Y., Gu, L., Zhao, Y., & Lu, F. (2021). Explainable diabetic retinopathy detection and retinal image generation. *IEEE Journal of Biomedical and Health Informatics*, *26*(1), 44–55.

Park, S.-J., Son, H., Cho, S., Hong, K.-S., & Lee, S. (2018). Srfeat: Single image super-resolution with feature discrimination. In *Proceedings of the european conference on computer vision (eccv)* (pp. 439–455).

Paul, S., & Chen, P.-Y. (2022). Vision transformers are robust learners. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 36, pp. 2071–2081).

Peng, Y., Dharssi, S., Chen, Q., Keenan, T. D., Agrón, E., Wong, W. T., … Lu, Z. (2019). Deepseenet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*, *126*(4), 565–575.

Pérez, A. D., Perdomo, O., Rios, H., Rodríguez, F., & González, F. A. (2020). A conditional generative adversarial network-based method for eye fundus image quality enhancement. In *International workshop on ophthalmic medical image analysis* (pp. 185–194).

Philip, S., Cowie, L., & Olson, J. (2005). The impact of the health technology board for scotland's grading model on referrals to ophthalmology services. *British Journal of Ophthalmology*, *89*(7), 891–896.

Qiu, D., Cheng, Y., & Wang, X. (2022). Improved generative adversarial network for retinal image super-resolution. *Computer Methods and Programs in Biomedicine*, 106995.

Qiu, D., Zheng, L., Zhu, J., & Huang, D. (2021). Multiple improved residual networks for medical image super-resolution. *Future Generation Computer Systems*, *116*, 200–208.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Sengupta, S., Wong, A., Singh, A., Zelek, J., & Lakshminarayanan, V. (2020). Desupgan: multi-scale feature averaging generative adversarial network for simultaneous de-blurring and super-resolution of retinal fundus images. In *International workshop on ophthalmic medical image analysis* (pp. 32–41).

Shamsudeen, F. M., & Raju, G. (2016). Enhancement of fundus imagery. In *2016 international conference on next generation intelligent systems (icngis)* (pp. 1–5).

Shen, Z., Fu, H., Shen, J., & Shao, L. (2020). Modeling and enhancing low-quality retinal fundus images. *IEEE transactions on medical imaging*, *40*(3), 996–1006.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning* (pp. 2256–2265).

Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Staal, J., Abràmoff, M. D., Niemeijer, M., Viergever, M. A., & Van Ginneken, B. (2004). Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, *23*(4), 501–509.

Subramani, B., & Veluchamy, M. (2020). Fuzzy gray level difference histogram equalization for medical image enhancement. *Journal of medical systems*, *44*, 1–10.

Tian, C., Yang, J., Li, P., Zhang, S., & Mi, S. (2022). Retinal fundus image super-resolution generated by optical coherence tomography based on a realistic mixed attention gan. *Medical Physics*, *49*(5), 3185–3198.

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Wan, C., Zhou, X., You, Q., Sun, J., Shen, J., Zhu, S., … Yang, W. (2021). Retinal image enhancement using cycle-constraint adversarial network. *Frontiers in Medicine*, *8*.

Wang, H., Chhablani, J., Freeman, W. R., Chan, C. K., Kozak, I., Bartsch, D.-U., & Cheng, L. (2012). Characterization of diabetic microaneurysms by simultaneous fluorescein angiography and spectral-domain optical coherence tomography. *American journal of ophthalmology*, *153*(5), 861–867.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., … Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the european conference on computer vision (eccv) workshops* (pp. 0–0).

Wang, Y., Perazzi, F., McWilliams, B., Sorkine-Hornung, A., Sorkine-Hornung, O., & Schroers, C. (2018). A fully progressive approach to single-image super-resolution. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 864–873).

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600–612.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the european conference on computer vision (eccv)* (pp. 3–19).

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 22–31).

Xiong, L., Li, H., & Xu, L. (2017). An enhancement method for color retinal images based on image formation model. *Computer methods and programs in biomedicine*, *143*, 137–150.

Yang, F., Yang, H., Fu, J., Lu, H., & Guo, B. (2020). Learning texture transformer network for image super-resolution. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5791–5800).

Yoo, T. K., Choi, J. Y., & Kim, H. K. (2020). Cyclegan-based deep learning technique for artifact reduction in fundus photography. *Graefe's Archive for Clinical and Experimental Ophthalmology*, *258*(8), 1631–1637.

You, Q., Wan, C., Sun, J., Shen, J., Ye, H., & Yu, Q. (2019). Fundus image enhancement method based on cyclegan. In *2019 41st annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 4500–4503).

Yu, Z., Xiang, Q., Meng, J., Kou, C., Ren, Q., & Lu, Y. (2019). Retinal image synthesis from multiple-landmarks input with generative adversarial networks. *Biomedical engineering online*, *18*(1), 1–15.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M.-H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5728–5739).

Zhang, J., Zhang, Y., & Xu, X. (2021). Pyramid u-net for retinal vessel segmentation. In *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1125–1129).

Zhang, S., Fu, H., Yan, Y., Zhang, Y., Wu, Q., Yang, M., … Xu, Y. (2019). Attention guided network for retinal image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 797–805).

Zhang, S., Liang, G., Pan, S., & Zheng, L. (2018). A fast medical image super resolution method based on deep learning network. *IEEE Access*, *7*, 12319–12327.

Zhang, S., Webers, C. A., & Berendschot, T. T. (2022). A double-pass fundus reflection model for efficient single retinal image enhancement. *Signal Processing*, *192*, 108400.

Zhao, H., Yang, B., Cao, L., & Li, H. (2019). Data-driven enhancement of blurry retinal images via generative adversarial networks. In *International conference on medical image computing and computer-assisted intervention* (pp. 75–83).

Zhu, J., Yang, G., & Lio, P. (2019). How can we make gan perform better in single medical image super-resolution? a lesion focused multi-scale approach. In *2019 ieee 16th international symposium on biomedical imaging (isbi 2019)* (pp. 1669–1673).

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the ieee international conference on computer vision* (pp. 2223–2232).