Université Abou Bekr Belkaid
Tlemcen Algérie
جامعة أبي بكر بلقايد
تلمسان الجزائر

People's Democratic Republic of Algeria
University of Abou Bekr Belkaid – Tlemcen
Faculty of Sciences
Department of Computer Science

Final Year Thesis

For the fulfillment of the Master's Degree in Computer Science

Option: Artificial Intelligence (A.I)

Title

Exploring Self-Supervised Learning for the Classification of Ophthalmic Diseases from Retinal Images

Carried out by:

- **TAIBI Mohammed**
- **BELARBI Boumediene**

Presented on Juin 20, 2026 before the jury composed of:

- Mrs. ILES Nawel (President)
- Mr. SMAHI Mohammed Ismail (Supervisor)
- Mr. EL HABIB DAHO Mostafa (Supervisor)
- Mr. ABDERRAHIM Alaedine (Examiner)

Academic Year 2025/2026


بسم الله الرحمن الرحيم

Dedication

I dedicate this modest work to my dearest parents for their unconditional love, endless sacrifices, and unwavering support throughout my academic journey. Their patience, encouragement, and prayers have always been my greatest source of strength and motivation.

To my beloved family, for their constant presence, trust, and moral support, which gave me the courage to continue even in difficult moments.

To all those who are dear to me and who contributed, directly or indirectly, to the completion of this work, I express my sincere gratitude.

Acknowledgments

First and foremost, I thank God for giving me the strength, patience, and will to complete this work.

I wish to express my sincere gratitude to my supervisors for their availability, invaluable advice, and guidance throughout the writing of this thesis.

My thanks also go to all the professors who contributed to my education during my university studies.

I am deeply grateful to my parents and family for their unwavering moral support, trust, and encouragement.

Finally, I cannot conclude without expressing my sincere gratitude to my dear partner for his genuine friendship, trust, and collaborative spirit.

Belarbi Boumediene

Dedication

I dedicate this work to my parents for their sacrifices, patience, and support, which made this journey possible for me. I cannot put into words how thankful I am to you for everything you have given me.

To my brother and sister, for their support and for their presence in moments of doubt.

To everyone in my family who supported me and believed in me.

Acknowledgments

All praise and thanks belong to Allah, without whom this work could not have been completed and from whom I derive all the strength, clarity, and patience needed to complete this work.

I would like to thank my supervisors. Their expertise, guidance, and support were extremely valuable throughout this work.

Thanks also go to all my teachers and professors, from the moment I first walked through the doors of a school until today.

I also thank all my friends and colleagues for their assistance and encouragement.

Lastly, I want to express my sincere gratitude to my friend and colleague who worked on the same project with me, for their collaboration and dedication.

Taibi Mohammed

Abstract

[en]

Self-supervised learning has recently emerged as a promising approach for retinal image representation learning by leveraging large quantities of unlabelled data. Among these methods, Masked Autoencoders (MAEs) have demonstrated strong performance by learning to reconstruct missing image regions. However, conventional MAEs primarily focus on spatial-domain reconstruction and do not explicitly encourage the preservation of frequency-domain information, which may contain important retinal structures and pathological patterns.

This work proposes a Frequency-Regularised Masked Autoencoder (FR-MAE) for retinal image representation learning. The proposed approach extends the standard MAE framework by introducing a frequency-domain regularisation term based on the Fourier transform of reconstructed images. The objective is to encourage the preservation of high-frequency information during self-supervised pre-training and improve the transferability of the learnt representations.

The proposed method was pre-trained on approximately 90,000 unlabelled retinal images and subsequently fine-tuned for Diabetic Retinopathy and Glaucoma classification. Experimental evaluation was performed using ten fine-tuning seeds and included both internal test sets and cross-dataset generalisation experiments. The results showed that frequency-domain regularisation consistently improved downstream classification performance compared with the baseline MAE. In particular, the proposed approach achieved superior cross-dataset generalisation for both tasks and substantially reduced the performance gap with RETFound despite using a considerably smaller pre-training dataset.

These findings indicate that frequency-domain regularisation constitutes an effective complementary learning signal for self-supervised retinal representation learning and can improve the transferability and generalisation ability of retinal image representations.

Keywords: Self-Supervised Learning, Retinal Images, Masked Autoencoders, Frequency-Domain Regularisation, Diabetic Retinopathy, Glaucoma, Representation Learning.

Résumé

[fr]

L'apprentissage auto-supervisé est récemment apparu comme une approche prometteuse pour l'apprentissage de représentations d'images rétiniennes, en exploitant de grandes quantités de données non annotées. Parmi ces méthodes, les Masked Autoencoders (MAE) ont montré de bonnes performances en apprenant à reconstruire des régions masquées de l'image. Cependant, les MAE classiques se concentrent principalement sur la reconstruction dans le domaine spatial et n'encouragent pas explicitement la préservation des informations fréquentielles, qui peuvent contenir des structures rétiniennes importantes ainsi que des motifs pathologiques.

Ce travail propose un Masked Autoencoder à régularisation fréquentielle (FR-MAE) pour l'apprentissage de représentations d'images rétiniennes. L'approche étend le cadre standard du MAE en introduisant un terme de régularisation dans le domaine fréquentiel basé sur la transformée de Fourier des images reconstruites. L'objectif est d'encourager la préservation des informations de haute fréquence lors du pré-entraînement auto-supervisé et d'améliorer la transférabilité des représentations apprises.

La méthode proposée a été pré-entraînée sur environ 90 000 images rétiniennes non annotées, puis affinée pour les tâches de classification de la rétinopathie diabétique et du glaucome. L'évaluation expérimentale a été réalisée avec dix graines (seeds) de fine-tuning et comprend à la fois des datasets de test internes et des expériences de généralisation cross-dataset. Les résultats montrent que la régularisation dans le domaine fréquentiel améliore de manière constante les performances en classification en aval par rapport au MAE de référence. En particulier, l'approche proposée atteint une meilleure généralisation cross-dataset pour les deux tâches et réduit significativement l'écart de performance avec RETFound, malgré l'utilisation d'un ensemble de pré-entraînement considérablement plus réduit.

Ces résultats indiquent que la régularisation dans le domaine fréquentiel constitue un signal d'apprentissage complémentaire efficace pour l'apprentissage auto-supervisé de représentations rétiniennes et peut améliorer la transférabilité et la capacité de généralisation des représentations d'images rétiniennes.

Keywords: Apprentissage auto-supervisé, images rétiniennes, autoencodeurs masqués, régularisation dans le domaine fréquentiel, rétinopathie diabétique, glaucome, apprentissage de représentations.

ملخص

[ar]

لقد ظهر التعلّم ذاتي الإشراف مؤخراً كنهج واعد لتعلّم ترميزات صور الشبكية، وذلك من خلال الاستفادة من كميات كبيرة من البيانات غير المعلّمة. ومن بين هذه الأساليب، أثبتت المشفّرات التلقائية ذات المدخلات المحجوبة (Masked Autoencoders - MAEs) أداءً قوياً عبر تعلّم إعادة بناء الأجزاء المفقودة من الصورة. ومع ذلك، فإن هذه النماذج التقليدية تركز بشكل أساسي على إعادة البناء في المجال المكاني (Spatial Domain)، ولا تعزز بشكل صريح الحفاظ على المعلومات في المجال الترددي (Frequency Domain)، والذي قد يحتوي على بنى شبكية مهمة وأنماط مرضية دالة.

يقترح هذا العمل مشفراً تلقائيًا ذو مدخلات محجوبة مُقيَّدا بالتردد (Frequency-Regularised Masked Autoencoder - FR-MAE) لتعلّم تمثيلات صور الشبكية. ويقوم النهج المقترح بتوسيع إطار المشفّرات التلقائية ذات المدخلات المحجوبة التقليدية من خلال إدخال حدّ تقييد في المجال الترددي يعتمد على تحويل فورييه (Fourier Transform) للصور المعاد بناؤها. والهدف هو تشجيع الحفاظ على المعلومات عالية التردد أثناء مرحلة التدريب ذاتي الإشراف، وتحسين قابلية نقل الترميزات المتعلّمة إلى مهام لاحقة.

تم تدريب النموذج المقترح مسبقاً على حوالي 90,000 صورة شبكية غير معلّمة، ثم تم تكييفه لاحقاً لمهام تصنيف اعتلال الشبكية السكري (Diabetic Retinopathy) والزَرَق (Glaucoma). وقد أُجري التقييم التجريبي باستخدام عشر عمليات تكييف مختلفة (seeds)، وشمل مجموعات اختبار داخلية وتجارب تعميم عبر مجموعات بيانات مختلفة. أظهرت النتائج أن التّقييد في المجال الترددي حسّن بشكل ثابت أداء التصنيف في المهام اللاحقة مقارنةً بالنموذج التقليدي (Baseline MAE). وعلى وجه الخصوص، حقق النهج المقترح أداءً أفضل في التعميم عبر مجموعات البيانات لكلا المهمتين، وقلّص بشكل ملحوظ الفجوة في الأداء مع نموذج RETFound رغم استخدام مجموعة بيانات أصغر بكثير خلال مرحلة التّدريب ذاتي الإشراف.

تشير هذه النتائج إلى أن التّقييد في المجال الترددي يشكّل إشارة تعلم مكملة فعّالة في التعلم ذاتي الإشراف، ويمكنه تحسين قابلية النقل وقدرة التعميم لترميزات الصور الشبكية.

الكلمات المفتاحية: التعلم ذاتي الإشراف، صور الشبكية، المشفّرات التلقائية ذات المدخلات المحجوبة، التّقييد في المجال الترددي، اعتلال الشبكية السكري، الزَرَق، تعلّم الترميزات.

Table of Contents

List of Figures

List of Tables

List of Algorithms

List of Acronyms

MAE Masked Autoencoder

AMD Age-Related Macular Degeneration

CNN Convolutional Neural Network

ViT Vision Transformer

SSL Self-Supervised Learning

SimCLR Simple Contrastive Learning of Representations

BYOL Bootstrap Your Own Latent

FR-MAE Frequency-Regularized MAE

DR Diabetic Retinopathy

GL Glaucoma

CLS Classification

GEN Generalization

FFT Fast Fourier Transform

General Introduction

Today, medical imaging plays a vital role in the diagnosis and management of many diseases. Fundus photography is a crucial form of imaging used in ophthalmology, as it provides a direct view of the retina and blood vessels. Fundus imaging is one of the most standard methods for diagnosing and managing many ophthalmic diseases. Given the enormous number of patients and the widespread availability of digital medical data, retinal image analysis is currently a very active area of research.

In general, deep learning-based methods have proven effective in classifying and detecting ocular pathologies. However, they typically require large amounts of annotated data. This poses a particular problem in the medical field, as annotation requires expert intervention and is time-consuming. Therefore, it seems relevant to turn to self-supervised learning.

This approach involves using large collections of unlabelled images to automatically learn visual representations. Among recent approaches, the use of Masked Autoencoders has proven effective in learning robust representations by reconstructing images from partially masked versions.

However, fundus images contain crucial information located in thin retinal structures, such as microaneurysms, small hemorrhages, and drusen. These microstructures contain important clinical information for the early diagnosis of certain pathologies.

Standard Masked Auto Encoder models can sometimes struggle to accurately preserve these fine details during model building.

In this context, our work focuses on exploring Self-Supervised Learning methods for retinal images, specifically by introducing additional constraints that should better preserve the fine structures of the image. We therefore seek, on one hand, to provide better quality representations and, on the other hand, to evaluate the usefulness of these representations in tasks of classifying ocular pathologies.

The Problem Statement

Despite significant progress in deep learning methods for retinal image analysis, several challenges remain. The limited availability of annotated retinal images continues to restrict the training of supervised models, while large quantities of unannotated retinal images remain underused. Self-supervised learning has emerged as an effective solution for exploiting such data and learning transferable visual representations.

Among recent Self-Supervised Learning approaches, Masked Autoencoders have demonstrated strong performance by learning to reconstruct masked image regions. However, the reconstruction objective is primarily defined in the spatial domain and may not sufficiently emphasise fine retinal structures. Small pathological patterns such as microaneurysms, hemorrhages, and subtle vessel abnormalities often correspond to high-frequency image components that can be overlooked during reconstruction.

Since these structures are clinically relevant and can contribute to robust representation learning, it is natural to investigate whether incorporating frequency domain information into the SSL objective can improve the quality of the learnt representations.

Therefore, the main question addressed in this work can be formulated as follows:

"Can frequency-domain regularisation that emphasises high-frequency retinal information improve the transferability and generalisation performance of representations learnt through self-supervised Masked Autoencoders?"

To investigate this question, this work proposes a Frequency-Regularised Masked Autoencoder (FR-MAE), which augments the conventional reconstruction objective with a frequency-domain regularisation term designed to encourage the preservation of fine retinal details during self-supervised pre-training.

1

An Overview of Retinal Imaging and SSL

Table of Contents

1.1 Introduction

Retinal image analysis has become an important research area due to its potential to assist in the early diagnosis of ocular diseases

We begin by presenting the basic concepts related to retinal imaging (anatomy and imaging techniques). Next, we discuss methods for the automated analysis of medical images based on deep learning. Finally, reconstruction-based SSL methods, particularly Masked Autoencoders, are presented together with recent developments in retinal image analysis, forming the basis of the proposed approach.

1.2 Retinal Imaging

Ophthalmic imaging plays a crucial role in the detection and monitoring of many pathologies that may deteriorate vision or lead to blindness. Among many imaging modalities, fundus photography is one of the most widely used techniques for retinal examination, providing detailed images of important anatomical structures. These images serve as valuable clinical resources for diagnosis and can also be used in automated analysis systems based on machine learning and deep learning methods.

However, automatic classification of retinal images remains a significant challenge due to variations in imaging devices (Topcon, Nidek, Eidon, etc.), heterogeneity in patient populations, and different variables during image acquisition [1].

1.2.1 Anatomy of the Retina

The retina is located at the back of the eye, it is a very thin light sensitive membrane responsible for vision. Acting as a receptor in image formation, It functions by converting light into nerve impulses that are transmitted to the brain through the optic nerve.

Important anatomical parts of the retina are the Blood Vessels, Macula, and Fovea along with the Optic disc. Macula and Fovea are involved in centralised and high detail vision, while the Optic disc is the area where the optic nerve leaves the eye, a region devoid of photoreceptors which explains the existence of a blind spot.

Any alteration of these anatomical structures can be a sign of ocular pathologies, making their analysis essential in diagnostic support systems [2].

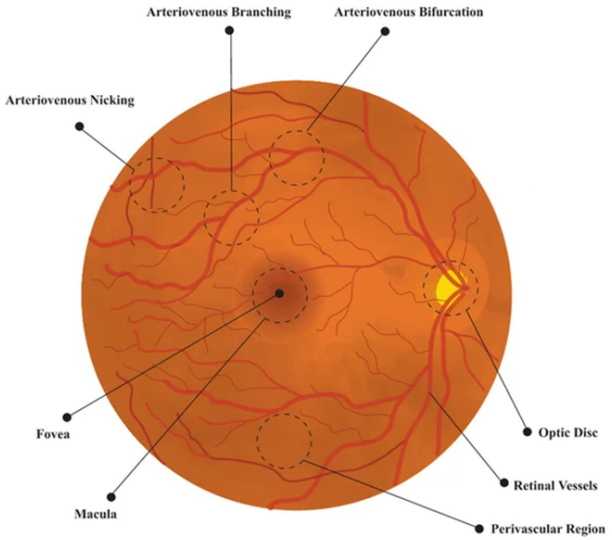

Figure 1.1: Retina Labels (macula, optic disc…)

1.2.2 Fundus Photography

Fundus photography is a diagnostic procedure used in ophthalmology to obtain detailed pictures of the back of the eye. It is well tolerated, non-invasive, provides high quality view of retinal structures (optic disc, retina, blood vessels, etc) and aids in the diagnosis, documentation, and management of many eye conditions.

A controlled light source is focused into the eye, illuminating the retina, while an image is taken via a specialised fundus camera. Depending on the available imaging technology and acquisition protocol, the resultant image may be in greyscale or colour.

The standard procedure for fundus photography includes the following:

1. Dilation of the pupil for optimal retina display.
2. Appropriate alignment of the fundus camera and the patient's eye.
3. Retinal image capture.
4. Post acquisition analysis and image evaluation.

Retinal imaging itself has its limits. Many external conditions can affect the quality of fundus images, such as light conditions, pupil dilation, type of retinal camera, operator expertise, and quality of retinal camera which then influences image analysis and clinical diagnosis [3].

1.2.3 Major Retinal Diseases

The Major retinal diseases are a set of retinal disorders, which can result in severe sight loss or blindness if not detected and managed adequately.

Common amongst these are:

- **Diabetic Retinopathy** refers to a vascular abnormality that affects the retina due to Diabetes Mellitus. It manifests itself by capillary occlusions, microaneurysms, and hemorrhages that impair vision in a degenerative manner [4].
- **Age-Related Macular Degeneration** is a degenerative disease that predominantly affects the macula. It is closely associated with ageing, although genetic predispositions and other factors also play a role. It results in a slow and insidious loss of central vision, affecting sight in daily activities such as reading, driving, and identifying faces [5].
- **Glaucoma** is a progressive optic neuropathy, with high intraocular pressure being a significant risk factor; resulting in irreparable loss of vision due to progressive loss of the optic nerve, often starting with the peripheral field of vision [6].

It is critical to achieve an early diagnosis of the diseases above and fundus photography offers an application where computer-vision algorithms can greatly help.

1.3 Automated Analysis of Medical Images

Artificial intelligence and computer vision are applied in the analysis of medical images in order to retrieve clinically relevant information. The main objective is to help physicians diagnose, treat and monitor patients, improving efficiency and reducing manual interpretation effort [7].

Recent advances in this domain are heavily driven by machine learning, specifically deep learning, which is discussed in the next section.

1.4 Deep Learning for Computer Vision

Deep learning is a sub-field of machine learning, driven by deep neural networks and capable of learning hierarchical representations of data. It has significantly advanced the field of computer vision, eliminating the need for hand-made feature descriptors by enabling automatic feature learning directly from raw images [8].

Deep learning techniques are usually deployed in supervised, unsupervised and reinforcement learning setups. Among these, supervised learning remains the most widely used for image-based tasks such as classification, object detection, segmentation, and regression [8].

1.4.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are deep learning architectures designed specifically to extract image features, utilising convolutional operators, allowing models of this architecture to learn automatically spatial image features throughout stacked layers.

The general structure of a CNN architecture includes convolution, activation function, pooling layers, then fully connected layers to calculate predictions. Whilst early CNN layers detect simple and low-level image features, like corners and edges, later layers detect more abstract and intricate visual patterns.

As CNNs efficiently manage image features, they have become widely involved in numerous computer vision tasks, such as image classification and medical image analysis.

1.4.2 Vision Transformers

Vision Transformers (ViTs) are deep learning models, based on the Transformer model, which was originally designed for natural language processing. Unlike CNNs, ViTs handle the input image by splitting it into a sequence of smaller patches.

In a ViT model, the input image is split into fixed size patches, and each patch is then mapped to a vector. These representations are fed into Transformer layers with self-attention allowing the model to learn dependencies between different parts of the images.

ViTs are effective in learning long-range dependencies from the image and have achieved good performance in different computer vision tasks, including medical image analysis [9].

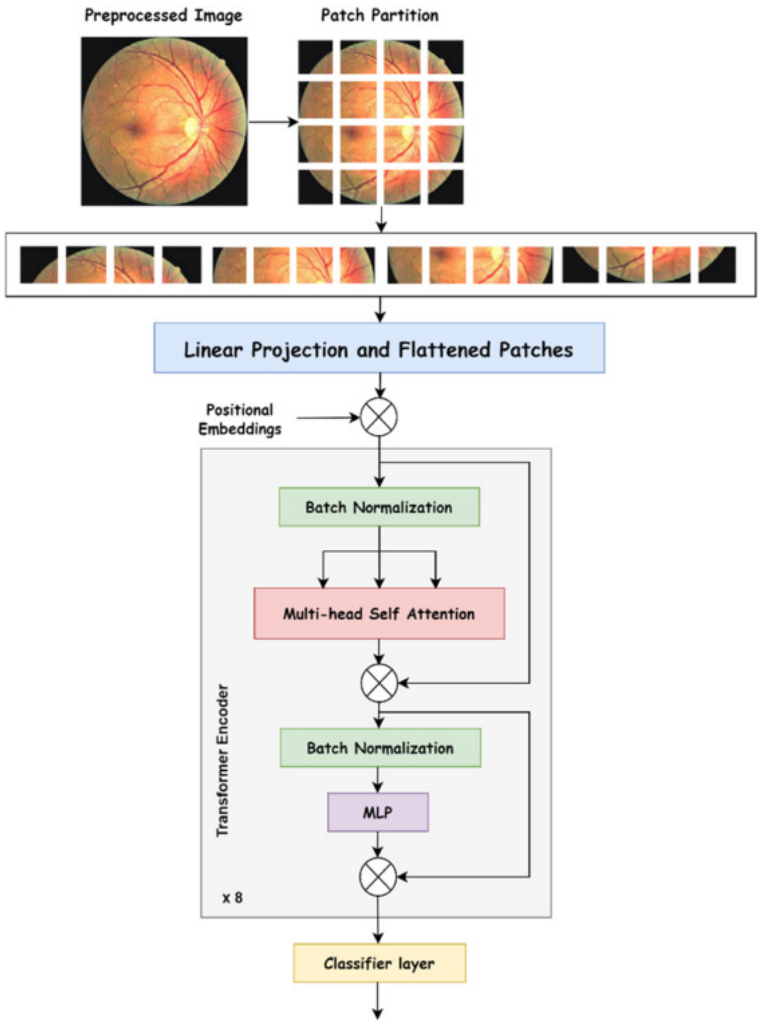

Figure 1.2: Retinal image processing pipeline with ViT (Patching and Encoding)

1.5 Self-Supervised Learning

Self-supervised learning is a machine learning method that exploits the paradigm of unsupervised learning for tasks that are traditionally tackled by supervised learning. The learning signals for SSL models are implicit and can be generated automatically from unlabelled data.

SSL is particularly useful in fields such as computer vision and natural language processing, which require large amounts of labelled data. Since such

data require costly human annotation, collecting them can be difficult. SSL approaches are therefore faster and more cost-effective, as they reduce the need for manual labelling.

Within SSL, a distinction is made between pretext tasks and downstream tasks. Pretext tasks allow the model to learn useful representations from unstructured data. These representations are then used for downstream tasks via transfer learning, reusing the whole pre-trained model or a part of it.

1.5.1 General Principle

The general principle of SSL is based on designing tasks that allow a loss function to use unlabelled input data as ground truth. In this context, pretext tasks generate "pseudo-labels" from the data itself. These tasks are not necessarily useful on their own, but they allow the model to learn representations that can be exploited for downstream tasks, hence the concept of representation learning.

The pre-trained models are then fine-tuned for specific downstream tasks, usually with supervised learning and much less labelled data.

There are different strategies in SSL, including contrastive learning, self-predictive methods, reconstruction-based approaches, or a combination of these techniques.

1.5.2 Contrastive Methods

Contrastive SSL methods provide models with multiple data samples and require them to predict the relationships between them. Models trained with these methods are generally discriminative (rather than generative).

These methods are typically based on creating data-data pairs; different views of the same instance are considered positive examples, while views from other instances serve as negative examples. These views are generated through data transformations such as cropping, rotation, flipping, or colour modification. The objective of the model is to maximise similarity between positive examples and minimise similarity with negative examples.

Among the most well-known contrastive methods:

- SimCLR relies on contrastive SSL without requiring specialised architectures or memory banks. It's objective is to maximise the agreement be-

tween augmented views of the same image (positive pairs) while minimising the agreement with views from other images (negative pairs) [10].

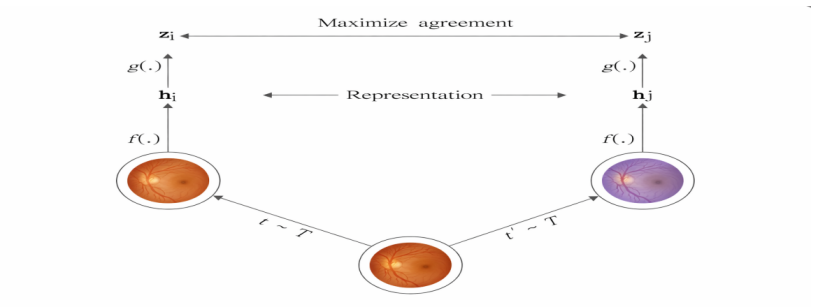

Figure 1.3: Principle of SimCLR based on maximizing the similarity between two augmented views of the same image.

- BYOL relies on an online (student) network and a target (teacher) network that interact with each other. An input (augmented view of an image) is presented to the student network and it learns to predict the target produced by the teacher network for another augmented view of the same input, eliminating the need for negative samples [11].

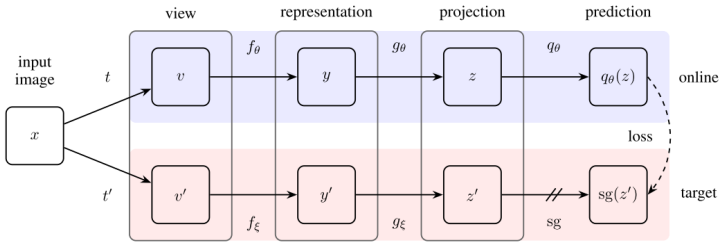

Figure 1.4: BYOL architecture showing the online and target networks with the representation, projection, and prediction stages.

1.5.3 Self-Predictive Methods

Self-predictive learning approaches train the model to predict informative representations of the input from a partial or modified view of that input. Such methods do not need an explicit comparison between positive and negative samples, as in contrastive learning. Instead, the model learns by predicting hidden or latent targets generated from the input itself, encouraging robust representation learning through internal consistency.

One of the most notable self-predictive methods is Data2Vec, an SSL framework that trains a student model to predict contextualised latent representations generated by a teacher model from the same input data, without predicting the original raw pixels or the use of a contrastive task, it predicts high-level feature representations in latent space. During training, the student receives a masked version of the input, while the teacher processes the full input and produces target representations. The objective is to minimise the difference between the student's predictions and the teacher's latent targets [12].

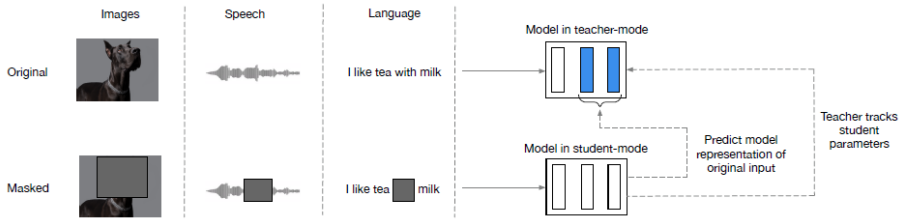

Figure 1.5: Data2Vec architecture illustrating self-supervised learning through masked input prediction and teacher–student latent representation alignment.

1.5.4 Reconstruction Methods

From many categories in SSL, reconstruction-based learning has proven to be one of the most dominant methods.

In this approach, a model is trained to recreate the input from a corrupted, compressed, or masked version. This kind of learning method helps the model capture the global structure of the data and the relations between different data parts, leading to effective data representations which can then be transferred to a wide range of downstream applications.

Reconstruction-based learning has been widely explored through approaches such as Autoencoders and, more recently, transformer-based methods. Among these approaches, Masked Autoencoders (MAEs) have demonstrated strong performance in computer vision tasks.

The main idea behind MAEs is to mask away a large fraction of the input image and then to train a model to reconstruct the masked regions from the unmasked portion of the input. By this reconstruction objective, it is possible to obtain useful visual features without supervision [13].

Despite their strong performance, MAEs primarily optimise reconstruction in the spatial domain. Consequently, fine image structures are not explicitly

emphasised during representation learning.

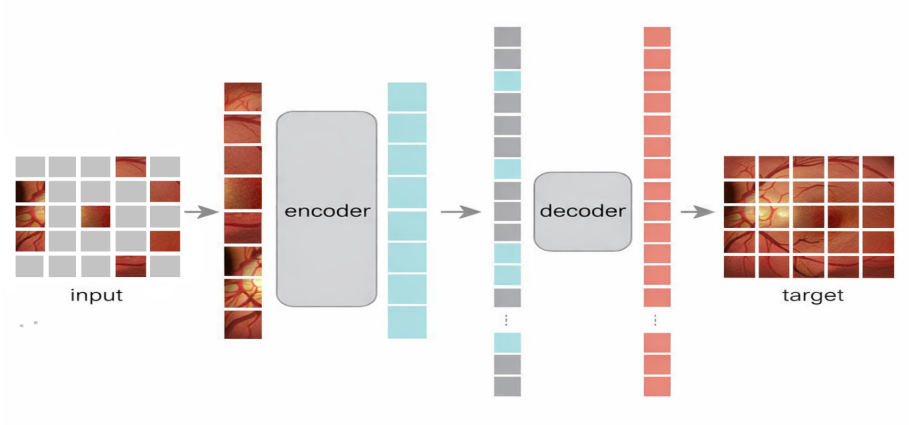

Figure 1.6: MAE architecture showing patch masking, encoder processing of visible patches, and decoder-based reconstruction of masked regions.

1.6 Recent Work on Retinal Image Analysis

In recent years, the use of deep learning in medical image analysis has evolved significantly.

In the field of ophthalmology, fundus images are very important because they allow direct observation of the condition of the retina. For this reason, several researchers have attempted to develop automatic methods capable of helping to diagnose eye diseases.

Initially, most studies were based on CNNs, models known for their ability to process images and automatically extract useful information. They have been used in several studies, particularly for the detection of diabetic retinopathy, achieving satisfactory results in many cases [14] [15].

Subsequently, more recent architectures were proposed. Transformer-based models, such as ViTs, began to be used in medical imaging. Unlike CNNs, these models focus not only on local details, but also on the image as a whole, which can improve understanding of structures present in retinal images [16].

At the same time, some studies have shifted towards SSL. MAEs are one of the most effective approaches that enable learning from unlabelled images prior to performing classification [13]. Among MAE-based approaches, RETFound has demonstrated state-of-the-art performance on a wide range of retinal image analysis tasks by leveraging large-scale self-supervised pre-training [17].

Although MAEs have demonstrated strong performance in retinal image representation learning, challenges remain regarding the preservation of fine retinal details and the generalisation of learnt representations.

1.7 Conclusion

In this chapter, the basics required to understand the proposed work were presented.

Retinal imaging and the main studied pathologies were first introduced. The different methods used for medical image analysis were then discussed, with particular emphasis on deep learning approaches such as CNNs and ViTs.

Subsequently, SSL was presented as a promising approach for exploiting unlabelled data. Particular attention was paid to reconstruction-based methods and Masked Autoencoders, which constitute the foundation of the proposed methodology.

This chapter therefore establishes the theoretical foundation for the remainder of this work, which is detailed in the following chapter.

2

Proposed Frequency-Regularised MAE Methodology

Table of Contents

2.1 Introduction

In this chapter, we describe the methodology adopted for the automatic analysis of retinal images. The objective of this work is to build a deep learning model capable of extracting relevant features from unlabelled images in order to improve classification performance.

First, we present the datasets used and the pre-processing steps applied to the images. Next, we describe the model architecture based on MAE.

We also introduce the proposed method, called Frequency-Regularised MAE (FR-MAE), which adds regularisation in the frequency domain to improve the quality of learnt representations.

Subsequently, we explain the loss function and the adopted training strategy. The model is first pre-trained in a self-supervised manner, then fine-tuned for the classification task.

Finally, we present the experimental choices and the parameters used in this work.

2.2 Presentation of the Datasets

Several public retinal fundus image datasets were used to train and evaluate the proposed model. These datasets include labelled and unlabelled images and cover different ocular conditions, mainly Diabetic Retinopathy and Glaucoma. The datasets were organised according to their specific role within the proposed experimental pipeline and subsequently divided into three distinct pools: the SSL pool, the fine-tuning pool, and the generalisation pool.

Within each pool, all contributing datasets were merged to form a single larger dataset. This organisation allowed the inclusion of images originating from different populations, acquisition devices, and imaging conditions, thereby increasing the diversity of retinal characteristics represented within each pool. In consequence, the models were subjected to a wide variety of image variations during training and evaluation.

2.2.1 SSL Dataset

Table 2.1 presents the datasets used to construct the SSL pool employed during the SSL pre-training phase. During dataset collection, all available unlabelled

retinal images were gathered and merged into a single SSL pool regardless of their dataset of origin, with the exception of datasets reserved for the generalisation experiments, whose unlabelled images were excluded to prevent data leakage. The Total row indicates the final number of unlabelled images available for SSL pre-training after merging all contributing datasets.

Table 2.1: Retinal datasets used for SSL pre-training

No.	Dataset	Count
1	EYEPACS [18]	53,573
2	Messidor 02 [19] [20]	1,748
3	IDRID [21]	81
4	RFMiD [22]	2,462
5	1000x39 [23]	856
6	ODIR [24]	3,373
7	PAPILA [25]	68
8	STARE [26] [27]	396
9	ARIA [28]	143
10	FIVES [29]	200
11	AGAR300 [30]	28
12	APTOS [31]	1,928
13	FUND-OCT [32] [33]	30
14	DiaRetDB1 [34]	89
15	DRIONS-DB [35]	110
16	E-Ophta [36]	434
17	HEI-MED [37]	169
18	HRF [38]	45
19	ROC [39]	100
20	BREST [40] [41] [42]	8,491
21	JICHI [43]	9,939
22	CHAKSU [44]	1,345
23	DR1-2 [45]	1,601
24	Cataract [46]	601
25	MMAC [47]	2,769
26	DiaRetDB0 [48]	130
Total		**90,709**

2.2.2 Fine-Tuning Dataset

Tables 2.2 and 2.3 report the labelled datasets used for fine-tuning on the Diabetic Retinopathy and Glaucoma classification tasks, respectively. For each disease,

the individual datasets were merged into a single fine-tuning dataset. The Total row therefore represents the final class distribution obtained after combining all contributing datasets.

Table 2.2: Diabetic Retinopathy datasets used for fine-tuning

No.	Dataset	0 - No DR	1 - Mild	2 - Moderate	3 - Severe	4 - Proliferative DR	Count
1	EYEPACS [18]	25,808	2,443	5,292	873	708	35,124
2	IDRID [21]	168	25	168	93	62	516
3	1000x39 [23]	38	18	49	39	0	144
4	ODIR [24]	2,816	460	745	144	21	4,186
5	APTOS [31]	1,805	370	999	193	295	3,662
6	BREST [40] [41] [42]	7,287	97	164	14	213	7,775
Total		**37,922**	**3,413**	**7,417**	**1,356**	**1,299**	**51,407**

Table 2.3: Glaucoma datasets used for fine-tuning

No.	Dataset	0 - Normal	1 - Glaucomatous	Count
1	ODIR [24]	2,816	200	3,016
2	PAPILA [25]	333	87	420
3	FIVES [29]	200	200	400
4	FUND-OCT [32] [33]	27	44	71
Total		**3,376**	**531**	**3,907**

It is important to note that both fine-tuning datasets exhibit a significant class imbalance. In the Diabetic Retinopathy dataset, the majority of images are part of the No DR class, while severe and proliferative classes are sparsely represented. In the Glaucoma dataset a similar class imbalance is observed, where normal images numerically overwhelm glaucomatous images. Class imbalance is typical in medical imaging datasets and may lead to a training biased towards the majority classes. To address this issue, specific mechanisms were incorporated into the fine-tuning procedure with the intention of reducing the impact of class imbalance and promote more balanced learning across classes. These mechanisms are described in detail later in this chapter.

Another aspect worth noting is that a small number of datasets, such as ODIR, is used in both the Diabetic Retinopathy and Glaucoma classification tasks. However, this overlap does not affect the validity of the experimental protocol, as the two tasks are treated as independent classification problems and are trained separately. Each task is fine-tuned directly from the same SSL pre-trained model rather than from a model previously trained on the other task. Consequently, the presence of a dataset in both classification pools does not introduce information leakage between the two experiments.

Furthermore, some datasets contribute to both the SSL pool and the fine-tuning pool, including datasets such as EYEPACS, ODIR, and FIVES. This overlap exists at the dataset level only and not at the image level. As discussed previously, several datasets were only partially annotated or contained labels that were not suitable for the target classification tasks. In such cases, the labelled images were assigned to the corresponding fine-tuning pool, while the remaining unlabelled images were transferred to the SSL pool. During this process, special care was taken to ensure that no image was present in both pools simultaneously. Consequently, there is no overlap between the images used in the two stages.

2.2.3 Generalisation Dataset

Tables 2.4 and 2.5 present the datasets used for cross-dataset generalisation evaluation in the Diabetic Retinopathy and Glaucoma tasks, respectively. The Total row represents the final class distribution of the merged generalisation dataset.

Table 2.4: Diabetic Retinopathy datasets used for cross-dataset generalisation evaluation

No.	Dataset	0 – No DR	1 – Mild	2 – Moderate	3 – Severe	4 – Proliferative DR	Count
1	PARAGUAY [49]	711	6	110	210	261	1,298
2	OIA-DDR [50]	6,266	630	4,477	236	913	12,522
3	SUSTech-SYSU [51]	631	24	401	87	76	1,219
Total		**7,608**	**660**	**4,988**	**533**	**1,250**	**15,039**

Table 2.5: Glaucoma datasets used for cross-dataset generalisation evaluation

No.	Dataset	0 – Normal	1 – Glaucomatous	Count
1	Drishti-GS1 [52] [53]	31	70	101
2	G1020 [54]	724	296	1,020
3	ORIGA [55]	482	168	650
4	REFUGE [56] [57]	1,720	280	2,000
Total		**2,957**	**814**	**3,771**

The generalisation pool plays a crucial role in assessing the transferability and generalisation capability of the learnt representations. For this reason, the datasets assigned to this pool were kept completely independent from both the SSL and fine-tuning pools. No overlap was allowed at either the dataset level or the image level, ensuring that the model had never encountered any of the generalisation data during training. Furthermore, any unlabelled images originating from datasets reserved for the generalisation pool were excluded rather

than transferred to the SSL pool. This precaution was taken to avoid data leakage and to guarantee that the reported results reflect true cross-dataset generalisation performance.

2.3 Image Preprocessing

Medical fundus images collected from different datasets exhibit significant variations in image resolution, acquisition devices, illumination conditions, and frame quality. Therefore, preprocessing is required before the images are provided to the proposed model.

The objective of this preprocessing stage is not only to improve consistency between datasets, but also to ensure that the model learns meaningful retinal structures rather than non-informative elements such as borders and background regions.

A deliberate choice was made to use datasets that contain complete fundus images. Several publicly available datasets were excluded when they only provided region-specific retinal views. For example, datasets such as ACRIMA were not retained because they do not consistently provide full retinal images.

This decision was motivated by the fact that the proposed approach relies on learning global retinal structures, where peripheral context such as optic disc location, macula positioning, and vascular distribution plays an important role.

2.3.1 Cleaning Images

All images are first cleaned to minimise acquisition artefacts and to increase the effective coverage of the retinal region prior to resizing and data augmentation.

Typically, fundus images consist of considerable areas of black boundaries, stemming from acquisition sensors of the camera, circular masks during the acquisition, or incorrect cropping during the database generation. These non-informative regions do not contribute to retinal structure learning and may negatively affect representation learning if retained.

To address this issue, an automatic preprocessing procedure was designed to detect and remove black borders. The method is based on intensity thresholding combined with spatial ratio analysis.

Algorithm 2.1 Black Border Removal Procedure

Require: Fundus image I, Intensity threshold t, Minimum valid ratio r

Ensure: Cropped retinal image I'

1: Convert the image I into an array of pixel intensity values
2: Create a binary mask where a pixel is valid if its max RGB intensity is greater than t
3: calculate the ratio of valid pixels per each row and per each column
4: Keep rows and columns with valid pixel ratio $\geq r$
5: Get the smallest bounding box that contains all valid rows and columns
6: Crop the image using the computed bounding box
7: **return** the cropped retinal image I'

This procedure is less sensitive to partially damaged images and varying acquisition conditions.

The majority of the images were correctly processed by this automatic pipeline. A small fraction of the images had to be pre-processed by hand, preventing additional noise from being introduced. This typically occurred when text annotations were embedded within the original image document or had a wrong colour in the background that the thresholding did not easily pick up.

2.3.2 Image Resizing

After the cleaning stage, all retinal images were resized in order to standardise the input dimensions required by the Vision Transformer architecture. Since the datasets used in this work originate from multiple sources, image resolutions and aspect ratios vary significantly.

Direct resizing to a fixed square resolution would distort anatomical retinal structures and potentially alter clinically important patterns. With that in mind, the images were resized to keep the original geometric ratios of the retina.

The length of the longest side of the image was fixed at 224 pixels, and the other side was scaled proportionally. This retains the relationship between elements of the retina.

After proportional resizing, the remaining empty regions were padded with black pixels in order to obtain a final square image of size 224×224 pixels. The padding was applied symmetrically around the shortest dimension to centre the retinal image within the frame.

The final output of this stage is therefore a square retinal image of resolution 224×224 pixels, suitable for both SSL pre-training and downstream classification.

2.3.3 SSL Data Augmentation

To increase the model robustness and enrich the training samples without altering key, clinically relevant, retinal structures, data augmentation was used in the SSL pre-training stage. Augmentation procedures were selected conservatively so that pathological information would not be destroyed, as retinal lesions like microaneurysms, exudates and hemorrhages are often small and localised.

Prior to spatial transformations, Contrast Limited Adaptive Histogram Equalisation (CLAHE) was applied to the green channel of the fundus images. The green channel is commonly preferred in retinal analysis because it provides higher contrast for vascular structures and small lesions. The CLAHE algorithm increases local contrast but reduces over-amplification of noise. Consequently, it will display finer retinal details, including vessels and microaneurysms. This technique is supported in the literature on retinal image processing and has been proven to effectively increase the visibility of lesions [58], [59].

Then a conservative random resized crop was applied with a scaling range between 0.8 and 1.0 and an aspect ratio constrained between 0.9 and 1.1. Compared to aggressive cropping used for natural images, the above parameters were specifically set up to avoid losing the overall retinal context and the deletion of salient regions. Retinal SSL and RETFound-based approaches have shown the importance of retaining the structure of lesions while learning representations [17].

Horizontal flipping with a probability of 50% was also applied. Horizontal flips remain anatomically plausible for fundus photographs, whereas vertical flips were intentionally excluded because they generate unrealistic retinal orientations.

Small random rotations limited to $\pm 15^{\circ}$ were also introduced to simulate realistic acquisition variability caused by camera tilt during retinal imaging.

Unlike conventional supervised learning pipelines, no colour jitter augmentation was used during MAE pretraining. Since MAEs rely on reconstructing original pixel values, strong colour perturbations can produce inconsistent reconstruction targets and negatively affect representation learning.

Consequently, only minimal appearance modifications were retained during pretraining, following the practices adopted in MAE-based retinal representation learning approaches [17].

Finally, all images were converted into tensors and normalised using the Im-

ageNet mean and standard deviation values:

- Mean: [0.485, 0.456, 0.406]
- Standard deviation: [0.229, 0.224, 0.225]

This normalisation approach is commonly used within Vision Transformer and RETFound-based frameworks to guarantee backward compatibility with pretrained networks and stable training [17].

Algorithm 2.2 Fundus Image Augmentation Pipeline In The SSL Phase

Require: Fundus image I
Ensure: Preprocessed image I'
1: Extract the green channel from I
2: Apply CLAHE to the green channel
3: Reinsert the enhanced green channel into the image
4: Apply a random resized crop to 224×224
5: Apply a horizontal flip with 50% probability
6: Apply a random rotation within the range $[-15°, +15°]$
7: Convert the image into a tensor
8: Normalize the image using ImageNet mean and standard deviation values
9: **return** the preprocessed image I'

2.4 MAE Model Implementation

In this work, the MAE implementation employs a ViT backbone. The model follows the standard encoder-decoder workflow of Masked Autoencoders, where the encoder takes visible image patches and the decoder reconstructs the missing parts.

The implementation uses TIMM library for ViT backbone and Lightly library for MAE related components like patch masking, encoding and image reconstruction.

2.4.1 Patch Masking Principle

As stated previously, the proposed approach adopts a ViT backbone using the vit_large_patch16_224 model architecture with a masking ratio of 75%, which means that each input image is divided into small patches of size 16×16 pixels, and only a randomly selected small portion corresponding to 25% of the patches is provided to the encoder while the rest is masked.

The indices of visible and masked patches are preserved throughout the process to enable proper reconstruction during decoding. Besides reducing computational cost by limiting the number of processed patches, this masking strategy encourages the model to learn global dependencies between different image regions in order to infer the missing content.

2.4.2 Image Reconstruction

After the encoding stage, the decoder receives the latent representation produced from the visible patches together with tokens corresponding to the masked patches, then reconstructs the missing image regions.

In this work's implementation, the reconstructed patches are subsequently reinserted into the complete set of image patches to reform a full reconstructed image, which is later used for frequency-domain regularisation.

Algorithm 2.3 MAE Image Reconstruction Process

Require: Input Image I_t, visible index set I_k, masked index set I_m
Ensure: Predicted patches P_p, reconstructed image I_r
1: Get visible patches based on I_k
2: Input visible patches into encoder
3: Obtain visible patches' latent representation
4: Fill masked positions in representation vector with mask token
5: Input the whole sequence into decoder
6: Predict the masked patches P_p
7: Extract all original patches from the image I_t
8: Replace the masked patches with the predicted patches
9: Reassemble all patches to form the reconstructed image I_r
10: **return** P_p and I_r

2.5 Proposed Model: Frequency-Regularised MAE

In this work, a Frequency-Regularised Masked Autoencoder (FR-MAE) is proposed for retinal image representation learning. Although conventional MAE models mainly optimise reconstruction in the spatial domain, they may not fully capture fine structural and textural information present in retinal images.

To overcome this limitation, we adopted a frequency domain regularisation term using Fourier representations of the reconstructed image. This approach adds spectral knowledge to the spatial reconstruction aiming to improve reconstruction quality and learn better visual features.

The resulting FR-MAE model preserves the general MAE framework while extending the training objective with frequency-based constraints.

2.5.1 Frequency Regularisation

The frequency regularisation is calculated at the level of the entire image, rather than at the patch level. This choice is motivated by the fact that applying Fourier analysis to individual patches can distort the true frequency content of the image, structures that appear as low-frequency within a patch may actually correspond to high-frequency components in the global image context. Performing the regularisation on the full image preserves the true global frequency distribution and avoids artificial frequency shifts caused by patch boundaries.

Algorithm 2.4 Frequency-Domain Loss Computation

Require: Reconstructed image I_r, Original image I, Weighting factor α
Ensure: Frequency loss L_{freq}
1: Compute Fourier transforms: $F_r \leftarrow \text{FFT2}(I_r)$, $F \leftarrow \text{FFT2}(I)$
2: Compute frequency domain error: $D(x, y) \leftarrow \left|F_r(x, y) - F(x, y)\right|^2$
3: Generate spatial frequencies: $f_x \leftarrow \text{fftfreq}(W)$, $f_y \leftarrow \text{fftfreq}(H)$
4: Construct radial frequency grid: $r(x, y) \leftarrow \sqrt{f_x(x)^2 + f_y(y)^2}$
5: Define radial weighting: $W(x, y) \leftarrow r(x, y)^p$
6: Compute weighted frequency error: $S \leftarrow \sum_{x,y} W(x, y)\, D(x, y)$
7: Normalize loss: $L \leftarrow \frac{S}{\sum_{x,y} W(x,y)}$
8: Aggregate over RGB channels
9: Compute batch mean
10: Apply weighting factor: $L_{\text{freq}} \leftarrow \alpha \cdot L$
11: **return** L_{freq}

To apply this regularisation, the torch.fft library was used. A 2D Fourier transform(fft2) was applied for the generated image and ground truth image respectively.

The frequency loss is then computed from the complex difference between the two spectra, which simultaneously accounts for both magnitude and phase errors. This is important for preserving not only spectral content but also the spatial localisation of retinal structures.

In order to emphasise fine retinal structures, a radial frequency weighting is introduced using torch.fft.fftfreq. This weighting increases the importance of high-frequency components, which are responsible for edges, textures, and small lesions. In this implementation, the weight is proportional to the spatial

frequency radius raised to a coefficient p (r^p), allowing stronger penalisation of reconstruction errors in high-frequency regions.

The weighted frequency loss is computed by averaging over spatial dimensions, summing contributions across RGB channels, then averaging over the batch and scaling by a factor α, which controls the contribution of frequency regularisation to the total loss.

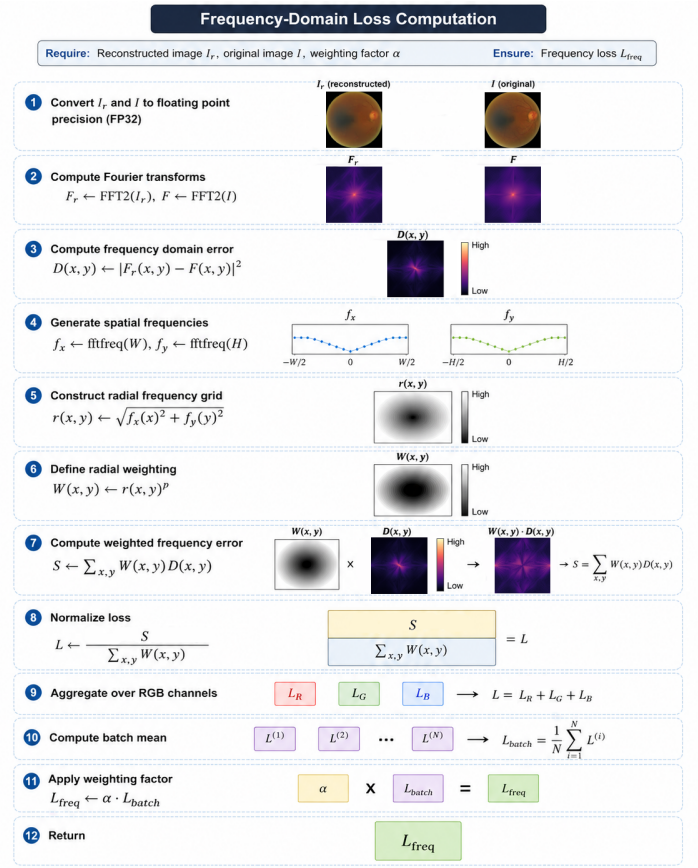

Figure 2.1: Diagram of the Weighted Frequency Loss Computation Integrating Complex Difference and Radial Weighting

2.5.2 Loss Function

The loss function combines a spatial-domain reconstruction loss with a frequency-domain regularisation.

The first component corresponds to the reconstruction loss, computed using the mean squared error (MSE) between the patches predicted by the decoder and the original masked patches.

$$L_{\text{rec}} = \text{MSE}(P_{\text{pred}}, P_{\text{target}}) \tag{2.1}$$

where P_{pred} denotes the predicted patches and P_{target} the corresponding original patches.

The second component is the frequency regularisation term (L_{freq}) detailed in the previous subsection (Subsection 2.5.1).

The final loss function of the model is defined as the sum of the two components:

$$L = L_{\text{rec}} + L_{\text{freq}} \tag{2.2}$$

This formulation allows the model to preserve not only reconstruction consistency in masked regions but also fine-grained spectral details and the spatial localisation of important structures.

2.6 SSL Pre-Training

The SSL pretraining phase is a key component of the proposed FR-MAE approach. It enables the effective use of large amounts of unlabelled retinal images, which are particularly abundant in medical datasets.

The SSL training procedure is designed to ensure stable and efficient convergence through training and validation phases. Retinal images are first preprocessed (Section 2.3), then each image is divided into patches and masked (Subsection 2.4.1). The model is then trained to reconstruct the masked patches from the visible ones (Subsection 2.4.2). This mechanism allows the encoder to learn rich visual representations that are well adapted to retinal structures.

At each iteration, the total loss combining spatial domain reconstruction loss and frequency regularisation is computed (Subsection 2.5.2). Gradients are then backpropagated to update the model parameters. In this work, different values of the weighting coefficient α were explored in order to study its impact on the quality of the learnt representations.

AdamW optimiser with a learning rate schedule compromising of warm-up and cosine decay is used during training. Mixed precision training and gradient clipping were also applied to speed up computation and reduce memory usage while keeping training stable.

A validation stage is conducted every epoch and the model achieving the

best performance based on validation loss is saved to ensure good generalisation performance.

The representations obtained after the self-supervised pretraining phase are then used as an initialisation for a downstream fine-tuning stage dedicated to retinal disease classification.

2.7 Fine-tuning for Classification

After the self-supervised pre-training phase, a supervised fine-tuning stage is performed in order to adapt the learnt representations to a specific task.

This phase aims to leverage the features extracted by the pre-trained model, focusing on Diabetic Retinopathy as a 5-class classification task and Glaucoma as a binary classification task.

Several experimental configurations are evaluated to analyse the influence of the frequency regularisation coefficient. Fine-Tuning is performed on four different pre-trained models corresponding to different values of the frequency regularisation parameter α.

For this phase, the pre-trained encoder is reused as a feature extractor and is complemented by a task-specific classification head. The head consists of a normalisation layer, a dropout layer to reduce overfitting, and a linear layer projecting the extracted representations into the class space.

Fine-tuning is carried out progressively in two stages. First, only the final classifier is trained while the parameters of the pre-trained backbone remain frozen. This strategy allows the decision layer to adapt smoothly to the new annotations. In second stage, the entire network is unfrozen to perform global weight refinement. This second phase allows deeper representations to be fine-tuned while preserving knowledge acquired during pre-training.

AdamW is used as the optimiser. Different learning rates are used for the backbone and the classification head, ensuring that pre-trained weights are updated slowly and the output layer is adapting quickly. A learning rate scheduling strategy combining a warm-up phase followed by cosine decay is employed to stabilise training.

The loss function used in this phase is cross-entropy loss. To handle class imbalance, weights are assigned to classes according to their effective occurrence in the training set. In addition, for better generalisation and to prevent overconfidence, label smoothing was employed.

The data is divided into training, validation and test sets using a stratified 70/15/15 split in order to preserve class distribution across the three subsets. After the initial split, an additional leakage-prevention procedure was applied using the stored metadata. This procedure checks whether images belonging to known eye pairs are assigned to different subsets. If one image of a pair is found in the test set, while its corresponding pair is located in the training or validation set, the paired image is moved to the test set as well. To preserve the split proportions, a randomly selected image from the test set is then moved back to the subset from which the paired image was taken. The same procedure is then applied to the validation set, ensuring that paired eye images are not shared between different subsets.

During training, data augmentation is performed by applying random cropping, horizontal flipping, and slight photometric perturbations. These transformations improve robustness to acquisition variability in retinal images. The images are then normalised before being fed into the network.

The primary performance metric used for evaluation and selection criterion is AUCROC due to the severe class imbalance. This metric is particularly appropriate in this context as it remains robust to biassed class distributions. The best model for each configuration is saved based on its validation's AUCROC performance.

Finally, an early stopping mechanism is applied to stop training when performance no longer improves over several consecutive epochs.

Once fine-tuning is completed, the best model is evaluated on an independent generalisation set, allowing an objective comparison between different frequency regularisation coefficients, and identifying the best-performing configuration for each task.

To reduce the influence of random factors during supervised training, each model was independently fine-tuned and evaluated using the same set of 10 seeds. The reported results correspond to the mean and standard deviation across the 10 runs. This procedure provides a more robust estimate of model performance and ensures that comparisons between models are not influenced by favourable or unfavourable random initialisations.

The goal of this study is to determine the contribution of frequency regularisation introduced during pre-training on the final generalisation performance.

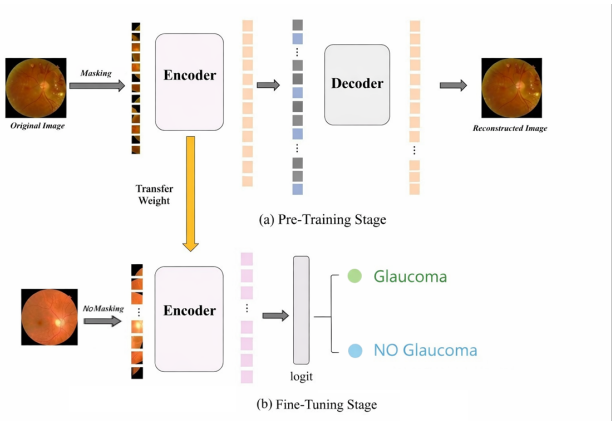

Figure 2.2: Self-Supervised Learning Framework with Pre-training and Fine-tuning for Glaucoma Classification

2.8 Conclusion

In this chapter, the methodology adopted for SSL pre-training and downstream classification task was presented. Firstly, retinal datasets used in this project were introduced along with their organisation into SSL, classification and generalisation pools.

Next, retinal images pre-processing pipeline used was discussed in detail, including image cleaning, resizing, and data augmentation techniques designed to preserve clinically relevant retinal structures while improving dataset consistency.

The implementation of the MAE model was then detailed, covering the patch masking mechanism and the image reconstruction process. Subsequently, the proposed FR-MAE model was presented, together with the introduced frequency-domain regularisation based on Fourier analysis and the associated loss function.

Finally, the SSL pre-training strategy and the supervised fine-tuning procedure for retinal disease classification were explained, including the optimisation methods, evaluation strategy, and experimental configurations adopted in this work.

This chapter therefore establishes the complete methodological framework used for the experiments and evaluations presented in the following chapter.

3

Results and Discussion

Table of Contents

3.1 Introduction

This chapter presents the experimental evaluation of the proposed FR-MAE approach for the classification of Diabetic Retinopathy and Glaucoma. The objective is to assess the contribution of self-supervised pre-training with frequency regularisation.

First, the experimental configuration is presented, including training parameters, and evaluation metrics. The behaviour of the SSL pre-training phase is then analysed, followed by an evaluation of the fine-tuning performance obtained with different values of the frequency regularisation coefficient.

The impact of frequency regularisation on downstream classification and cross-dataset generalisation performance is subsequently investigated. The obtained results are also compared with those of RETFound, which is used as the reference model throughout this study.

Finally, the limitations of the proposed method and possible directions for future improvement are discussed.

3.2 Experimental Configuration

In this section, the detailed information of experimental settings used in this work are presented.

All pre-training, fine-tuning, and evaluation experiments were performed on one NVIDIA H100 GPU with 80GB of VRAM. The large memory capacity of this hardware enabled efficient training with large batch sizes and mixed-precision computation.

To ensure a fair and repeatable comparison among all models, including RETFound, a fixed seed was set in both the SSL pre-training and fine-tuning stages to control dataset splitting, model initialisation, and other random operations, such as patch masking and data augmentations.

3.2.1 Pre-Training Parameters

Table 3.1 summarises the hyperparameters used during the SSL pre-training phase.

The number of training epochs was fixed at 400. Preliminary experiments showed that extending pre-training beyond this point did not provide additional

benefits and could negatively affect downstream fine-tuning performance, suggesting a degradation in the transferability of the learnt representations.

The frequency regularisation coefficient α was evaluated using the values $0.0, 0.1, 0.5$ and 1.0. The explored range was intentionally limited to values not exceeding 1.0 because the frequency loss and reconstruction loss operate at different scales. Larger coefficients would significantly increase the contribution of the frequency term, potentially introducing unstable optimisation dynamics and causing the training objective to be dominated by frequency-domain regularisation rather than reconstruction learning.

The decoder architecture and masking ratio were selected to remain consistent with the configuration adopted in RETFound, facilitating a direct comparison while relying on settings that have already demonstrated effectiveness for learning retinal representation.

The initial learning rate was set to 1.5×10^{-4}. Preliminary experiments indicated that larger learning rates produced noticeable loss spikes during training, even though the model was generally able to recover from them. Therefore, a lower learning rate was adopted to improve optimisation stability. In addition, gradient clipping was employed to further reduce the influence of occasional gradient explosions and stabilise the training process.

Table 3.1: SSL Pre-Training Parameters

Parameter	**Value**
Mask Ratio	`75%`
Decoder Dim	`512`
Decoder Depth	`8`
Decoder Num Heads	`16`
Batch Size	`512`
Number of Epochs	`400`
Initial Learning Rate	`1.5e-4`
Weight Decay	`0.05`
Warm-up Epochs	`20`
Minimum Learning Rate	`1e-6`
Gradient Clipping	`1.0`
Tested `Alpha` Values	`0.0`, `0.1`, `0.5`, `1.0`

3.2.2 Fine-Tuning Parameters

Table 3.2 summarises the hyperparameters used during the fine-tuning phase.

As described in the previous chapter, fine-tuning was performed in two successive stages. The first stage consisted of five epochs of linear probing with a frozen backbone, during which only the classification head was updated. This stage allows the classifier to adapt progressively to the target task before modifying the pre-trained representations. The second stage consisted of end-to-end optimisation of the entire network using a lower learning rate for the backbone, enabling gradual refinement of the learnt features while preserving the knowledge acquired during SSL pre-training.

Although a maximum number of training epochs was specified, most experiments converged earlier, with early stopping typically occurring after approximately 30 epochs.

Gradient clipping and label smoothing were used throughout fine-tuning to improve optimisation stability, reduce prediction overconfidence, and enhance generalisation performance.

Table 3.2: Fine-Tuning Parameters

Parameter	**Value**
Batch size	`16`
Epochs with frozen backbone	`5`
Number of epochs	`100`
Warm-up Epochs	`3`
Head Learning Rate	`1.2e-4`
Backbone Learning Rate	`8e-6`
Weight Decay	`0.03`
Minimum Learning Rate	`1e-7`
Gradient Clipping	`1.0`
Label Smoothing	`0.05`
Early Stopping Patience	`10`
Head Dropout	`0.30`

3.3 SSL Pre-Training Results

The SSL pre-training stage was evaluated using the reconstruction loss and the frequency-domain loss. However, the best model for each configuration was chosen based on total validation loss. Since all models were trained using the same random seed and initialisation settings, the first epoch started with the same reconstruction and frequency loss values for all α configurations. This makes the comparison between different α values more consistent, because the observed differences are mainly due to loss weighting rather than random initialisation effects.

The validation curves showed the same general trend, although they were less smooth than the training curves. This means that the models remained stable generally as there was no significant instability or divergence, and all models demonstrated consistent convergence.

3.3.1 Reconstruction Loss Behaviour During Pre-Training

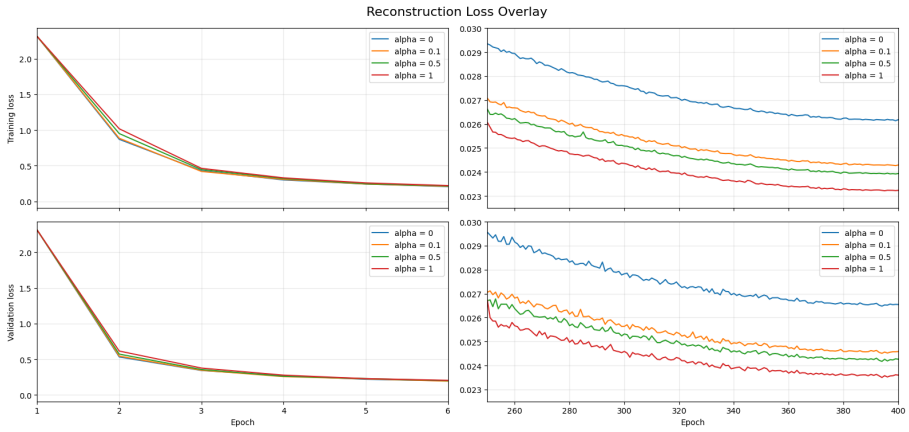

Figure 3.1: Training and Validation Reconstruction Losses with Zoomed Views of the Early (Epochs 1-6) and Late (Epochs 250-400) Training Stages

For the reconstruction loss, all models showed a rapid decrease during the first few epochs. In the first 6 epochs, the reconstruction loss dropped to below 0.25, after which the decrease became much slower and the curves gradually converged, indicating that models had already learnt the main image structure and were then refining smaller details. In the initial epochs, the smallest reconstruction loss belonged to the baseline model with $\alpha = 0$ then $\alpha = 0.1$ then $\alpha = 0.5$ and the largest was from the model with $\alpha = 1$. This shows that when frequency

loss was given more weight, the model initially focused less on pixel-level reconstruction accuracy.

However, in later epochs, the ranking of the reconstruction loss was reversed. The model with $\alpha = 1$ achieved the lowest reconstruction loss, while the base model $\alpha = 0$ had the highest reconstruction loss among the four models, indicating that, over time, the frequency loss helped the model learn more useful representations, even for the reconstruction objective itself.

3.3.2 Frequency Loss Behaviour During Pre-Training

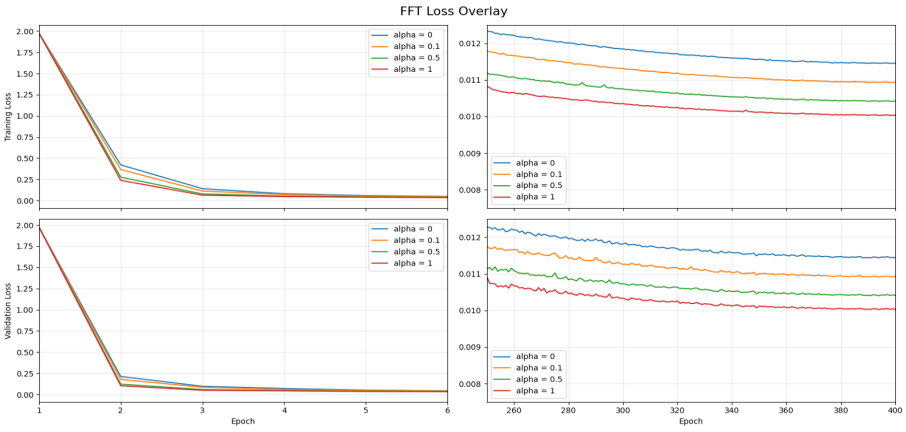

Figure 3.2: Training and Validation Frequency Losses with Zoomed Views of the Early (Epochs 1–6) and Late (Epochs 250–400) Training Stages

For frequency loss, the behaviour was similar in terms of convergence speed. The loss decreased rapidly during the first 6 epochs, then continued to improve more slowly until convergence. However, Unlike the reconstruction loss, the ranking between models remained consistent from the early epochs to the final epochs. The model with $\alpha = 1$ achieved the lowest frequency loss, followed by $\alpha = 0.5$, then $\alpha = 0.1$, while the base model with $\alpha = 0$ had the highest frequency loss. This result is expected because larger α values give more importance to the frequency-domain component during optimisation.

3.3.3 Observations

Although the difference between the losses of the models was relatively small, the zoomed views of the curves show that the choice of α still had a measurable effect. These small numerical differences are important because they correspond

to fine image details, especially in the frequency domain. Therefore, adding the frequency loss not only improved frequency preservation, but also contributed to more refined reconstruction behaviour in the later stages of SSL pre-training.

3.4 Fine-Tuning Results

As described in the previous Chapter, fine-tuning Section (Section 2.7), the SSL pre-trained models were subsequently fine-tuned for two downstream retinal disease classification tasks: Diabetic Retinopathy and Glaucoma. Additionally, to reduce the influence of random factors related to dataset splitting and model optimisation, each model was independently fine-tuned and evaluated using the same set of 10 seeds. The reported results therefore correspond to the mean and standard deviation across these runs.

This section presents the performance obtained on the internal test set, which corresponds to the 15% test partition extracted from the fine-tuning pool after applying the pair-preserving data leakage prevention procedure described previously. Although this test set originates from the same classification pool used for training, it remains unseen during optimisation and therefore provides an initial assessment of model generalisation.

Model performance is evaluated primarily using the Area Under the Receiver Operating Characteristic Curve (AUROC), while the Area Under the Precision-Recall Curve (AUPR) is reported as a complementary metric.

The objective of this analysis is to investigate how the frequency regularisation coefficient α influences downstream classification performance and to determine whether frequency-domain regularisation improves the quality and transferability of the representations learnt during SSL pre-training.

3.4.1 Diabetic Retinopathy

For Diabetic Retinopathy, all frequency-regularised models outperformed base MAE on both AUROC and AUPR, indicating that the addition of frequency-domain regularisation during pre-training positively influenced downstream classification performance. The base model achieved an AUROC of 79.43% and an AUPR of 43.10%, on the other hand, the frequency-regularised models accomplished AUROC values ranging from 81.01% to 81.60% and AUPR values ranging from 45.95% to 46.62%.

The best performance was obtained with $\alpha = 1.0$, which achieved an AUROC of 81.60% and an AUPR of 46.62%. This corresponds to improvements of 2.17% and 3.52% over the base model for AUROC and AUPR, respectively. Models with $\alpha = 0.1$ and $\alpha = 0.5$ also consistently improved both evaluation metrics, suggesting that the observed gains are not limited to a single choice of frequency regularisation coefficient.

Table 3.3: Diabetic Retinopathy classification performance on the internal test set. Results are reported as mean ± standard deviation over 10 fine-tuning runs.

Model	**AUROC (%)**	**AUPR (%)**
Base MAE	79.43 ± 0.91	43.10 ± 1.33
FR-MAE ($\alpha = 0.1$)	81.01 ± 0.58	45.95 ± 1.00
FR-MAE ($\alpha = 0.5$)	81.27 ± 0.82	46.41 ± 0.88
FR-MAE ($\alpha = 1.0$)	$\mathbf{81.60 \pm 0.59}$	$\mathbf{46.62 \pm 0.93}$

The reported standard deviations are relatively small across all models, indicating stable behaviour across the 10 fine-tuning runs. Overall, these results suggest that frequency-domain regularisation improves the quality of the representations learnt during SSL pre-training and leads to more effective transfer to the Diabetic Retinopathy classification task.

3.4.2 Glaucoma

For the Glaucoma classification task, the frequency-regularised models were again able to consistently outperform the base MAE on both AUROC and AUPR. The base model achieved an AUROC of 89.47% and an AUPR of 83.96%, while frequency-regularised models obtained AUROC values between 90.75% and 91.65% and AUPR values between 84.59% and 85.60%.

$\alpha = 1.0$ model was the best performing model, achieving an AUROC of 91.65% and an AUPR of 85.60%. Compared with the base model, this represents 2.18% increase in AUROC and 1.64% increase in AUPR. Models with $\alpha = 0.1$ and $\alpha = 0.5$ also produced better results than the base MAE on both evaluation metrics, indicating that the benefits of frequency-domain regularisation are consistent across different values of the regularisation coefficient.

Although the standard deviations are larger than those observed for the Diabetic Retinopathy task, reflecting a higher sensitivity to the fine-tuning seed, the

performance improvements remain consistent across all frequency-regularised models.

Table 3.4: Glaucoma classification performance on the internal test set.

Model	AUROC (%)	AUPR (%)
Base MAE	89.47 ± 2.63	83.96 ± 3.17
FR-MAE ($\alpha = 0.1$)	90.97 ± 2.65	85.34 ± 3.34
FR-MAE ($\alpha = 0.5$)	90.75 ± 1.43	84.59 ± 2.49
FR-MAE ($\alpha = 1.0$)	$\mathbf{91.65 \pm 2.14}$	$\mathbf{85.60 \pm 3.08}$

3.4.3 Observations

Despite the fact that some variability can be observed across the 10 fine-tuning runs, the frequency-regularised models consistently matched or exceeded the performance of the base MAE in most experiments. This behaviour was observed in both classification tasks, indicating that the improvements reported in Tables 3.3 and 3.4 are reproducible across different seeds rather than being attributable to a particular random initialisation or data split. Therefore, these results provide the first indication that frequency-domain regularisation contributes positively to downstream classification performance.

3.5 Generalisation Performance

The previous section evaluated model's performance on the internal test set. Although known image pairs were explicitly separated to reduce the risk of data leakage, complete elimination of leakage cannot be guaranteed because pair metadata was not available for all contributing datasets. Consequently, some unknown image pairs may still be distributed across different subsets. Furthermore, training and test sets originate from the same fine-tuning pool and may therefore share underlying population characteristics, acquisition protocols, and imaging devices. Therefore, to more rigorously assess the quality, transferability, and robustness of the learnt representations, a cross-dataset generalisation experiment is essential. This evaluation measures how effectively each model transfers to entirely unseen datasets collected from different populations and imaging conditions.

3.5.1 Diabetic Retinopathy

The Diabetic Retinopathy generalisation results indicate that the improvements shown on the internal test set also transfer to completely unseen datasets. The baseline MAE model achieved the lowest performance, with an AUROC of 71.59% and an AUPR of 37.14%. In contrast, all frequency-regularised models achieved substantially higher scores on both evaluation metrics, indicating that the representations learnt during SSL pre-training generalise more effectively across datasets when frequency-domain regularisation is employed.

The highest generalisation performance was obtained with $\alpha = 0.5$, which achieved an AUROC of 75.50% and an AUPR of 40.84%. The model with $\alpha = 1.0$ produced very similar results, reaching an AUROC of 75.00% and an AUPR of 40.75%, while $\alpha = 0.1$ achieved an AUROC of 74.50% and an AUPR of 39.67%. These results suggest that although stronger regularisation improved performance on the internal test set, the most transferable representations for Diabetic Retinopathy were obtained with an intermediate value of the frequency regularisation coefficient.

Table 3.5: Diabetic Retinopathy classification performance on the generalisation set.

Model	AUROC (%)	AUPR (%)
Base MAE	71.59 ± 1.45	37.14 ± 1.65
FR-MAE ($\alpha = 0.1$)	74.50 ± 1.23	39.67 ± 1.33
FR-MAE ($\alpha = 0.5$)	$\mathbf{75.50 \pm 1.79}$	$\mathbf{40.84 \pm 1.57}$
FR-MAE ($\alpha = 1.0$)	75.00 ± 1.47	40.75 ± 1.36

Compared to baseline MAE, the best performing frequency-regularised model improved AUROC from 71.59% to 75.50%, corresponding to a gain of 3.91%. AUPR was similarly improved from 37.14% to 40.84%, representing a gain of 3.70%. Furthermore, the relatively small standard deviations observed in all models indicate that these improvements are consistent across the 10 fine-tuning runs.

3.5.2 Glaucoma

The Glaucoma generalisation results further support the benefits of frequency-domain regularisation for learning transferable retinal representations. The baseline MAE achieved an AUROC of 74.32% and an AUPR of 66.82%, while all

frequency-regularised models obtained higher scores on both evaluation metrics. This indicates that the improvements introduced during SSL pre-training are maintained when the models are evaluated on entirely unseen datasets.

$\alpha = 1.0$ had the best generalisation performance, reaching an AUROC of 76.73% and an AUPR of 69.59%. The models with $\alpha = 0.5$ and $\alpha = 0.1$ also improved upon the baseline, achieving AUROC values of 76.29% and 76.14%, respectively. Unlike the Diabetic Retinopathy task, where the best generalisation performance was obtained with $\alpha = 0.5$, the Glaucoma results suggest that stronger frequency regularisation provides the most transferable representations for this classification task.

Table 3.6: Glaucoma classification performance on the generalisation set.

Model	**AUROC (%)**	**AUPR (%)**
Base MAE	74.32 ± 1.94	66.82 ± 1.65
FR-MAE ($\alpha = 0.1$)	76.14 ± 3.42	68.79 ± 2.35
FR-MAE ($\alpha = 0.5$)	76.29 ± 1.46	68.65 ± 1.91
FR-MAE ($\alpha = 1.0$)	$\mathbf{76.73 \pm 1.95}$	$\mathbf{69.59 \pm 1.65}$

Compared to baseline MAE, the best performing frequency-regularised model improved AUROC from 74.32% to 76.73%, corresponding to a gain of 2.41%. AUPR was similarly improved from 66.82% to 69.59%, representing a gain of 2.77%. Although the standard deviations are larger for some configurations, particularly $\alpha = 0.1$, the overall ranking remains consistent and all frequency-regularised models outperform the baseline on average.

3.5.3 Observations

In general, Diabetic Retinopathy and Glaucoma results provide strong evidence that incorporating frequency-domain information during SSL pre-training enhances the transferability of the learnt retinal representations.

Further observations can be made from the individual fine-tuning runs. Across the Diabetic Retinopathy experiments, baseline MAE consistently produced lower performance than frequency-regularised models, with only occasional cases where its performance approached that of the weakest regularised configuration. For the Glaucoma task, a small number of runs produced results comparable to those of the frequency-regularised models. Examination of these cases revealed signs of prolonged training and reduced effectiveness of

early stopping for some regularised configurations, suggesting occasional overfitting during fine-tuning. Nevertheless, such cases remained uncommon, and the frequency-regularised models generally maintained a performance advantage across the evaluated seeds.

These observations further support the conclusion that the improvements obtained by frequency-domain regularisation are consistent and not solely attributable to favourable random initialisations or dataset splits and confirm that the advantages of frequency-domain regularisation extend beyond the internal test set and translate into improved cross-dataset generalisation.

3.6 State-Of-The-Art

To the best of our knowledge, RETFound currently represents one of the strongest state-of-the-art models for retinal image representation learning. For this reason, it was selected as the reference model in this study.

To ensure a fair comparison, RETFound was fine-tuned using the same training protocol adopted for the proposed FR-MAE models, including identical fine-tuning parameters, evaluation metrics, and generalisation datasets.

The objective of this comparison is not to directly outperform RETFound, as the two approaches were pre-trained on substantially different amounts of data. RETFound was originally trained on approximately 1.6 million retinal images, whereas the SSL pool used in this work contains approximately 90,000 images, corresponding to about 5.6% of the pre-training data available to RETFound.

Consequently, RETFound is primarily used as a reference point for assessing the effectiveness of the proposed frequency regularisation strategy. The comparison makes it possible to evaluate the extent to which frequency-domain regularisation improves representation learning and helps reduce the performance gap associated with a considerably smaller pre-training dataset.

3.6.1 Fine-Tuning

For the Diabetic Retinopathy task, RETFound achieved the highest performance, reaching an AUROC of 87.04% and an AUPR of 57.35%. Nevertheless, introducing frequency-domain regularisation produced a clear improvement over the baseline MAE. The best-performing FR-MAE model ($\alpha = 1.0$) achieved an AUROC of 81.60% and an AUPR of 46.62%, compared with 79.43% and 43.10% for

the baseline MAE. Therefore, the AUROC gap between the baseline MAE and RETFound was reduced from 7.61% to 5.44% when frequency regularisation was introduced. Similarly, the AUPR gap was reduced from 14.25% to 10.73%.

Table 3.7: Comparison of the baseline MAE, best-performing FR-MAE model, and RETFound on the Diabetic Retinopathy internal test set.

Model	**AUROC (%)**	**AUPR (%)**
Base MAE	79.43 ± 0.91	43.10 ± 1.33
FR-MAE ($\alpha = 1.0$)	81.60 ± 0.59	46.62 ± 0.93
RETFound	$\mathbf{87.04 \pm 0.71}$	$\mathbf{57.35 \pm 2.04}$

A different behaviour was observed for the Glaucoma task. The baseline MAE achieved an AUROC of 89.47% and an AUPR of 83.96%, while the best-performing FR-MAE model ($\alpha = 1.0$) reached an AUROC of 91.65% and an AUPR of 85.60%. RETFound achieved an AUROC of 91.04% and an AUPR of 86.52%. Consequently, the best FR-MAE model obtained a higher AUROC than RETFound, while RETFound maintained the highest AUPR. More specifically, FR-MAE achieved an AUROC improvement of 0.61% over RETFound. Although RETFound retained an advantage in terms of AUPR, frequency-domain regularisation reduced the AUPR gap from 2.56% for baseline MAE to only 0.92% for FR-MAE.

Table 3.8: Comparison of the baseline MAE, best-performing FR-MAE model, and RETFound on the Glaucoma internal test set.

Model	**AUROC (%)**	**AUPR (%)**
Base MAE	89.47 ± 2.63	83.96 ± 3.17
FR-MAE ($\alpha = 1.0$)	$\mathbf{91.65 \pm 2.14}$	85.60 ± 3.08
RETFound	91.04 ± 1.86	$\mathbf{86.52 \pm 2.00}$

The comparison on the internal test set demonstrates that the performance gap between a standard MAE and a state-of-the-art retinal foundation model can be substantially reduced through frequency-domain regularisation. Although RETFound remained the strongest model for Diabetic Retinopathy, the proposed FR-MAE achieved the highest AUROC on the Glaucoma task despite being pre-trained on a substantially smaller SSL dataset.

3.6.2 Generalisation

The comparison with RETFound on the cross-dataset generalisation set provides another viewpoint into the transferability of the learnt representations. For the Diabetic Retinopathy task, RETFound achieved the highest performance, reaching an AUROC of 84.20% and an AUPR of 50.40%. The baseline MAE obtained substantially lower scores, with an AUROC of 71.59% and an AUPR of 37.14%. Introducing frequency-domain regularisation improved both metrics, with the best-performing FR-MAE model ($\alpha = 0.5$) achieving an AUROC of 75.50% and an AUPR of 40.84%. Therefore, the AUROC gap between the baseline MAE and RETFound was reduced from 12.61% to 8.70%, while the AUPR gap was reduced from 13.26% to 9.56%.

Table 3.9: Comparison of the baseline MAE, best-performing FR-MAE model, and RETFound on the Diabetic Retinopathy generalisation set.

Model	AUROC (%)	AUPR (%)
Base MAE	71.59 ± 1.45	37.14 ± 1.65
FR-MAE ($\alpha = 0.5$)	75.50 ± 1.79	40.84 ± 1.57
RETFound	**84.20 ± 1.23**	**50.40 ± 1.77**

A different trend was observed for the Glaucoma task. The baseline MAE achieved an AUROC of 74.32% and an AUPR of 66.82%, whereas the best-performing FR-MAE model ($\alpha = 1.0$) reached an AUROC of 76.73% and an AUPR of 69.59%. RETFound obtained an AUROC of 71.98% and an AUPR of 67.34%. Consequently, the best FR-MAE model outperformed both the baseline MAE and RETFound on both evaluation metrics. More precisely, FR-MAE achieved an improvement of 4.75% in AUROC over RETFound while also obtaining a higher AUPR by 2.25%.

Table 3.10: Comparison of the baseline MAE, best-performing FR-MAE model, and RETFound on the Glaucoma generalisation set.

Model	AUROC (%)	AUPR (%)
Base MAE	74.32 ± 1.94	66.82 ± 1.65
FR-MAE ($\alpha = 1.0$)	**76.73 ± 1.95**	**69.59 ± 1.65**
RETFound	71.98 ± 2.90	67.34 ± 1.98

Although RETFound remained the strongest model for Diabetic Retinopathy, the proposed FR-MAE substantially reduced the performance gap despite being

pre-trained on a much smaller SSL dataset. For Glaucoma, the best-performing FR-MAE model achieved the highest AUROC and AUPR among all evaluated approaches, surpassing even the state-of-the-art reference model. These results indicate that the benefits of frequency-domain regularisation are particularly evident when evaluating representation transferability across entirely unseen datasets.

3.7 Discussion

The results obtained throughout this study demonstrate that frequency regularisation positively affected the transferability of the representations learnt during SSL pre-training. Across both downstream classification tasks, the frequency-regularised models consistently outperformed the baseline MAE on the internal test sets and achieved superior cross-dataset generalisation performance. Furthermore, these improvements that were observed across multiple fine-tuning seeds, imply that the benefits of frequency regularisation are robust and not solely attributable to favourable random initialisations or dataset splits.

The experiments also showed that no single value of the frequency regularisation coefficient was optimal for all tasks and evaluation scenarios. Although increasing α generally strengthened the influence of frequency-domain information during SSL pre-training, the best-performing configuration depended on the downstream task and the evaluation dataset. Consequently, α should be regarded as a tunable hyperparameter whose optimal value may vary depending on the target application and the desired generalisation behaviour.

Compared with RETFound, the proposed approach demonstrated encouraging results. Despite being pre-trained on only approximately 90,000 retinal images, the frequency-regularised models were able to substantially reduce the performance gap with RETFound on the Diabetic Retinopathy task. Furthermore, for the Glaucoma classification task, the best-performing FR-MAE configuration achieved superior cross-dataset generalisation performance compared to RETFound. These findings suggest that incorporating frequency-domain information during SSL pre-training can improve the quality and transferability of learnt representations, helping to mitigate the disadvantage associated with a substantially smaller pre-training dataset.

One possible explanation for these improvements is that frequency-domain regularisation encourages the model to preserve image information that is only

weakly constrained by the standard reconstruction objective. In a conventional MAE, successful reconstruction can often be achieved by recovering the global structure of the image while neglecting certain fine details.

Another possible explanation is that frequency-domain regularisation can incorporate complementary information which may not be explicit in the spatial domain. Although certain structures may appear visually similar at the pixel level, their underlying frequency distributions can differ significantly. In many imaging applications, physical phenomena that are difficult to perceive directly often manifest themselves through characteristic spectral patterns. By constraining the model to preserve frequency-domain information, FR-MAE may therefore encourage the learning of features that are overlooked by conventional reconstruction objectives. This additional information could contribute to the improved transferability and cross-dataset generalisation observed in the experiments.

However, in retinal images, clinically relevant information is frequently contained in small lesions, vessel patterns, optic disc boundaries, and other high-frequency structures. By introducing a frequency-domain constraint during SSL pre-training, the model is encouraged to reconstruct not only the overall appearance of the image but also its spectral characteristics. This additional learning signal may lead to representations that capture more discriminative retinal features and are therefore more transferable to downstream classification tasks.

Moreover, frequency-domain regularisation may act as a form of inductive bias which has less reliance on dataset-specific visual characteristics. Although pixel-level reconstruction can encourage the model to learn acquisition-specific textures or colour distributions, frequency-domain information is more closely related to the underlying anatomical structures present in retinal images. This may explain why the benefits of frequency regularisation were particularly pronounced during cross-dataset evaluation, where robustness to changes in population characteristics, acquisition conditions, and imaging devices is essential.

In general, the results support the main hypothesis of this work:

"Regularisation in frequency domain during SSL pre-training helps to improve the transferability of retinal image representations and the generalisation performance of the model."

Further validation on larger datasets and other retinal diseases is still needed, but the proposed FR-MAE framework demonstrates that frequency regularisation can provide a useful complementary learning signal for retinal representa-

tion learning.

3.8 Limitations And Future Perspectives

Several limitations must be acknowledged. The most important limitation of this work is the size of the SSL pool. Although approximately 90,000 retinal images were used during pre-training, this remains considerably smaller than the datasets employed by large-scale SSL approaches such as RETFound. Therefore, it is plausible that additional gains could be achieved by combining the proposed frequency regularisation with substantially larger pre-training datasets.

Another limitation arises from the nature of frequency-domain information itself. High-frequency components correspond not only to fine anatomical details and pathological patterns but also to image noise and acquisition artefacts. Since the proposed frequency loss applies a radial weighting strategy that increasingly emphasises higher frequencies, it does not explicitly distinguish between clinically relevant structures and other high-frequency image content. Although the proposed regularisation improved downstream and cross-dataset performance, the specific spectral components responsible for these gains remain unclear. Consequently, further investigation is required to better understand which frequency bands contribute most strongly to representation learning and whether more selective frequency-domain weighting strategies could further improve performance.

A further restriction is related to the combination of reconstruction loss and frequency loss. The reconstruction objective is computed at the patch level, whereas the frequency loss is computed on the reconstructed image as a whole. This introduces a mismatch between the scales at which the two objectives operate. Patch-level frequency regularisation was explored during preliminary experiments, but it introduced substantial instability and degraded performance, making full-image frequency analysis the more suitable choice. Nevertheless, this difference in granularity remains a limitation of the current formulation.

An additional drawback concerns the range of downstream tasks considered in this study. The proposed approach was evaluated on Diabetic Retinopathy and Glaucoma classification tasks only. Although these tasks cover distinct retinal pathologies, further validation on additional retinal diseases and other ophthalmic imaging datasets would be necessary to determine the general applicability of frequency-domain regularisation.

Finally, the frequency loss cannot be considered entirely independent from the reconstruction loss. Since both objectives are optimised jointly, improvements in spatial reconstruction naturally influence the frequency-domain representation of the image, while frequency-domain constraints can also affect reconstruction behaviour. Indeed, the frequency-regularised models ultimately achieved lower reconstruction losses than the baseline MAE despite initially focusing less on pixel-level reconstruction accuracy. Consequently, part of the observed downstream performance gain may arise from the interaction between the two objectives rather than solely from the additional frequency-domain information. Although the results demonstrate that the combined objective improves representation learning, the individual contribution of the frequency loss remains difficult to isolate. As a result, investigating alternative formulations that more clearly decouple spatial and frequency-domain learning objectives remains an open research question.

3.9 Conclusion

In this chapter, the experimental evaluation of the proposed FR-MAE framework was presented. The first part described the experimental setup, which include the hardware environment, pre-training settings, fine-tuning parameters, and evaluation protocol used throughout the study.

The behaviour of the SSL pre-training phase was then analysed through the reconstruction and frequency-domain losses, highlighting the influence of different values of the frequency regularisation coefficient on the training dynamics and reconstruction behaviour.

The fine-tuning results obtained for Diabetic Retinopathy and Glaucoma classification were subsequently presented and compared across the different FR-MAE configurations, the baseline MAE, and the RETFound reference model. Cross-dataset generalisation experiments were then conducted to assess the transferability of the learnt representations.

Finally, the obtained results were discussed, with particular attention to the impact of frequency regularisation, the observed generalisation behaviour, and the limitations of the proposed approach.

Consequently, this chapter provides a detailed evaluation of the proposed methodology for learning retinal image representations and classifying retinal diseases.

Conclusion

Retinal image analysis has become an important research area due to its potential to assist in the early detection of diseases that can cause vision impairment or blindness. Recent advances in deep learning, particularly self-supervised learning, have demonstrated the ability to learn useful representations from large quantities of unlabelled data. However, existing reconstruction-based approaches, such as Masked Autoencoders, primarily focus on spatial-domain reconstruction and may not explicitly encourage the preservation of fine retinal details.

The objective of this work was therefore to investigate whether incorporating frequency-domain information into the SSL learning objective could improve the quality, transferability, and generalisation ability of the learnt retinal representations. To address this question, a Frequency-Regularised Masked Autoencoder (FR-MAE) was proposed. This approach extends the standard MAE framework by introducing a frequency-domain regularisation term based on Fourier analysis of the reconstructed image, with additional weighting of high-frequency components to emphasise fine retinal structures.

The proposed method was evaluated through SSL pre-training on a multi-dataset pool of approximately 90,000 unlabelled retinal images, followed by supervised fine-tuning on Diabetic Retinopathy and Glaucoma classification tasks. Experimental results showed that frequency regularisation influenced both the reconstruction behaviour during pre-training and the quality of the learnt representations. While the magnitude of the improvement was task-dependent, the frequency-regularised models consistently achieved better internal test and cross-dataset generalisation performance than the standard MAE baseline.

For Diabetic Retinopathy classification, the best-performing configuration improved generalisation AUROC from 71.59% to 75.50%, corresponding to a gain of 3.91%. For Glaucoma classification, the best frequency-regularised

model improved generalisation AUROC from 74.32% to 76.73%, corresponding to a gain of 2.41%. Furthermore, the proposed approach substantially reduced the performance gap with RETFound despite being pre-trained on only approximately 90,000 retinal images compared with the 1.6 million images used by RETFound. In the Glaucoma experiments, the best FR-MAE configuration surpassed RETFound on the cross-dataset generalisation benchmark.

The obtained results therefore support the central hypothesis of this work. Frequency-domain regularisation proved effective in improving the transferability and cross-dataset generalisation of representations learnt through self-supervised Masked Autoencoders. The findings also showed that the optimal regularisation coefficient (α) is task-dependent and should be treated as a tunable hyperparameter rather than a fixed parameter.

Nevertheless, several limitations remain. The SSL pool used for pre-training was considerably smaller than those employed by large-scale retinal foundation models, and the specific frequency components responsible for the observed improvements remain unclear. In addition, the current formulation introduces a mismatch between patch-level reconstruction and full-image frequency regularisation, while the individual contribution of the frequency loss remains difficult to isolate due to its interaction with the reconstruction objective. Finally, the proposed approach was evaluated on only two retinal disease classification tasks, and its applicability to other ophthalmic or medical imaging problems remains to be established.

As future work, the proposed approach could be evaluated on substantially larger retinal image collections, additional retinal diseases, and other medical imaging modalities. Further investigation of adaptive or frequency-selective weighting strategies may help identify the spectral components that contribute most strongly to representation learning. Alternative frequency-domain formulations that more clearly decouple spatial and frequency-domain learning objectives, as well as methods that operate consistently across multiple spatial scales, may also provide additional improvements.

In conclusion, this work demonstrates that frequency-domain regularisation constitutes a valuable complementary learning signal for self-supervised retinal image representation learning. By encouraging the preservation of fine image structures during pre-training, the proposed FR-MAE framework improves representation transferability and cross-dataset generalisation, making it a promising direction for future research in medical image analysis.

Bibliography

[1] M. Ferrara, Y. Zheng, and V. Romano, "Editorial: Imaging in ophthalmology," *Journal of Clinical Medicine*, vol. 11, no. 18, 2022. [Online]. Available: https://www.mdpi.com/2077-0383/11/18/5433.

[2] H. Kolb, "Simple anatomy of the retina," in *Webvision: The Organization of the Retina and Visual System*, H. Kolb et al. Eds., Updated 2012 Jan 31, Salt Lake City, UT: University of Utah Health Sciences Center, 2005. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK11533/.

[3] N. Panwar et al. "Fundus photography in the 21st century—a review of recent technological advances and their implications for worldwide healthcare," *Telemedicine and e-Health*, vol. 22, no. 3, pp. 198–208, 2016, Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC4790203/. eprint: https://journals.sagepub.com/doi/pdf/10.1089/tmj.2015.0068. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1089/tmj.2015.0068.

[4] U. V. Shukla and K. Tripathy, "Diabetic retinopathy," in *StatPearls [Internet]*, Updated 2023 Aug 25, Treasure Island, FL: StatPearls Publishing, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK560805/.

[5] J. H. C. Wong et al. "Exploring the pathogenesis of age-related macular degeneration: A review of the interplay between retinal pigment epithelium dysfunction and the innate immune system," *Frontiers in Neuroscience*, vol. 16, 2022. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.1009599.

[6] R. N. Weinreb, T. Aung, and F. A. Medeiros, "The pathophysiology and treatment of glaucoma: A review," *JAMA*, vol. 311, no. 18, pp. 1901–1911, May 2014. eprint: https://jamanetwork.com/journals/jama/

articlepdf/1869215/jrv140004.pdf. [Online]. Available: https://doi.org/10.1001/jama.2014.3192.

[7] A. Grzybowski et al. "Retina fundus photograph-based artificial intelligence algorithms in medicine: A systematic review," *Ophthalmology and Therapy*, vol. 13, no. 8, pp. 2125–2149, 2024. [Online]. Available: https://doi.org/10.1007/s40123-024-00981-4.

[8] D. Ravì et al. "Deep learning for health informatics," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 4–21, 2017. [Online]. Available: https://doi.org/10.1109/JBHI.2016.2636665.

[9] S. Aburass et al. "Vision transformers in medical imaging: A comprehensive review of advancements and applications across multiple diseases," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 6, pp. 3928–3971, 2025, Available at: https://www.researchgate.net/publication/390372350_Vision_Transformers_in_Medical_Imaging_a_Comprehensive_Review_of_Advancements_and_Applications_Across_Multiple_Diseases. [Online]. Available: https://doi.org/10.1007/s10278-025-01481-y.

[10] T. Chen et al. "A simple framework for contrastive learning of visual representations," *CoRR*, vol. abs/2002.05709, 2020. arXiv: 2002.05709. [Online]. Available: https://arxiv.org/abs/2002.05709.

[11] J. Grill et al. "Bootstrap your own latent: A new approach to self-supervised learning," *CoRR*, vol. abs/2006.07733, 2020. arXiv: 2006.07733. [Online]. Available: https://arxiv.org/abs/2006.07733.

[12] A. Baevski et al. *Data2vec: A general framework for self-supervised learning in speech, vision and language*, 2022. arXiv: 2202.03555 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2202.03555.

[13] K. He et al. "Masked autoencoders are scalable vision learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 979–15 988. [Online]. Available: https://arxiv.org/abs/2111.06377.

[14] V. Gulshan et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016. eprint: https://jamanetwork.com/journals/jama/articlepdf/2588763/joi160132.pdf. [Online]. Available: https://doi.org/10.1001/jama.2016.17216.

[15] D. S. W. Ting et al. "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, pp. 2211–2223, Dec. 2017. eprint: https://jamanetwork.com/journals/jama/articlepdf/2665775/jama_ting_2017_oi_170140.pdf. [Online]. Available: https://doi.org/10.1001/jama.2017.18152.

[16] W. Abdelbaki et al. "Transformer-based deep learning for population-scale retinal image screening of ophthalmic disorders," *Bioengineering*, vol. 13, no. 4, 2026. [Online]. Available: https://www.mdpi.com/2306-5354/13/4/377.

[17] Y. Zhou et al. "A foundation model for generalizable disease detection from retinal images," *Nature*, Sep. 2023. [Online]. Available: https://doi.org/10.1038/s41586-023-06555-x.

[18] J. Cuadros and G. Bresnick, "Eyepacs: An adaptable telemedicine system for diabetic retinopathy screening," *Journal of Diabetes Science and Technology*, vol. 3, no. 3, pp. 509–516, 2009, Dataset Available at: https://www.kaggle.com/c/diabetic-retinopathy-detection. eprint: https://doi.org/10.1177/193229680900300315. [Online]. Available: https://doi.org/10.1177/193229680900300315.

[19] E. Decencière et al. "Feedback on a publicly distributed image database: The messidor database," *Image Analysis and Stereology*, vol. 33, no. 3, pp. 231–234, Aug. 2014, Dataset Available at: https://www.adcis.net/en/third-party/messidor2/. [Online]. Available: https://www.ias-iss.org/ojs/IAS/article/view/1155.

[20] M. D. Abràmoff et al. "Automated analysis of retinal images for detection of referable diabetic retinopathy," *JAMA Ophthalmology*, vol. 131, no. 3, pp. 351–357, Mar. 2013. eprint: https://jamanetwork.com/journals/jamaophthalmology/articlepdf/1668203/ecs120076_351_357.pdf. [Online]. Available: https://doi.org/10.1001/jamaophthalmol.2013.1743.

[21] P. Porwal et al. "Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, 2018, Dataset Available at: https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid. [Online]. Available: https://www.mdpi.com/2306-5729/3/3/25.

[22] S. Pachade et al. "Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research," *Data*, vol. 6, no. 2, 2021, Dataset Available at: https://ieee-dataport.org/open-access/retinal-fundus-multi-disease-image-dataset-rfmid. [Online]. Available: https://www.mdpi.com/2306-5729/6/2/14.

[23] L.-P. Cen et al. "Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks," *Nature Communications*, vol. 12, no. 1, p. 4828, 2021, Dataset Available at: https://www.kaggle.com/datasets/linchundan/fundusimage1000. [Online]. Available: https://doi.org/10.1038/s41467-021-25138-w.

[24] Shanggong Medical Technology Co., Ltd., *Odir-5k: Ocular disease intelligent recognition*, Accessed via Kaggle, 2019. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k.

[25] O. Kovalyk et al. "Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment," *Scientific Data*, vol. 9, no. 1, p. 291, 2022, Dataset Available at: https://figshare.com/articles/dataset/PAPILA/14798004. [Online]. Available: https://doi.org/10.1038/s41597-022-01388-1.

[26] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical Imaging*, vol. 19, no. 3, pp. 203–210, Mar. 2000, Dataset Available at: https://cecas.clemson.edu/~ahoover/stare/. [Online]. Available: http://doi.org/10.1109/42.845178.

[27] A. Hoover and M. Goldbaum, "Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 951–958, Aug. 2003. [Online]. Available: http://doi.org/10.1109/TMI.2003.815900.

[28] P. Bankhead et al. "Fast retinal vessel detection and measurement using wavelets and edge location refinement," *PLOS ONE*, vol. 7, no. 3, pp. 1–12, Mar. 2012, Dataset Available at: https://www.researchgate.net/post/How_can_I_find_the_ARIA_Automatic_Retinal_Image_Analysis_Dataset. [Online]. Available: https://doi.org/10.1371/journal.pone.0032435.

[29] K. Jin et al. "Fives: A fundus image dataset for artificial intelligence based vessel segmentation," *Scientific Data*, vol. 9, no. 1, p. 475, 2022, Dataset Available at: https://figshare.com/articles/figure/FIVES_A_Fundus_Image_Dataset_for_AI-based_Vessel_Segmentation/19688169/1?file=34969398. [Online]. Available: https://doi.org/10.1038/s41597-022-01564-3.

[30] D. Jeba Derwin et al. "A novel automated system of discriminating microaneurysms in fundus images," *Biomedical Signal Processing and Control*, vol. 58, p. 101 839, 2020, Dataset Available at: https://ieee-dataport.org/documents/diabetic-retinopathy-fundus-image-datasetagar300. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809419304203.

[31] Karthik, Maggie, and S. Dane, *Aptos 2019 blindness detection*, https://kaggle.com/competitions/aptos2019-blindness-detection, Kaggle, 2019.

[32] T. Hassan et al. "Deep structure tensor graph search framework for automated extraction and characterization of retinal layers and fluid pathology in retinal sd-oct scans," *Computers in Biology and Medicine*, vol. 105, pp. 112–124, 2019, Dataset Available at: https://data.mendeley.com/datasets/trghs22fpg/3. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2018.12.015.

[33] T. Hassan et al. "Rag-fw: A hybrid convolutional framework for the automated extraction of retinal lesions and lesion-influenced grading of human retinal pathology," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 1, pp. 108–120, 2021. [Online]. Available: https://doi.org/10.1109/JBHI.2020.2982914.

[34] T. Kauppi et al. "The diaretdb1 diabetic retinopathy database and evaluation protocol," in *Proceedings of the British Machine Vision Conference*, Dataset Available at: https://www.kaggle.com/datasets/nguyenhung1903/diaretdb1-v21/data, BMVA Press, 2007, pp. 15.1–15.10. [Online]. Available: https://doi.org/10.5244/C.21.15.

[35] E. J. Carmona et al. "Identification of the optic nerve head with genetic algorithms," *Artificial Intelligence in Medicine*, vol. 43, no. 3, pp. 243–259, 2008, Dataset Available at: https://www.ia.uned.es/~ejcarmona/DRIONS-

DB.html. [Online]. Available: https://doi.org/10.1016/j.artmed.2008.04.005.

[36] E. Decencière et al. "Teleophta: Machine learning and image processing methods for teleophthalmology," *IRBM*, vol. 34, no. 2, pp. 196–203, 2013, Dataset Available at: https://www.adcis.net/en/third-party/e-ophtha/. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1959031813000237.

[37] L. Giancardo et al. "Exudate-based diabetic macular edema detection in fundus images using publicly available datasets," *Medical image analysis*, vol. 16, pp. 216–26, Jul. 2011, Dataset Available at: https://github.com/lgiancaUTH/HEI-MED. [Online]. Available: https://doi.org/10.1016/j.media.2011.07.004.

[38] A. Budai et al. "Robust vessel segmentation in fundus images," *International journal of biomedical imaging*, vol. 2013, p. 154 860, Jan. 2013, Dataset Available at: https://www5.cs.fau.de/research/data/fundus-images/. [Online]. Available: https://doi.org/10.1155/2013/154860.

[39] M. Niemeijer et al. "Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 185–195, 2010, Dataset Available at: http://webeye.ophth.uiowa.edu/ROC/. [Online]. Available: https://doi.org/10.1109/TMI.2009.2033909.

[40] L. F. Nakayama et al. "A Brazilian Multilabel Ophthalmological Dataset (BRSET)," *PhysioNet*, Aug. 2024, Dataset Available at: https://www.kaggle.com/datasets/tanzinabdul/fundus-patientwise-split. [Online]. Available: https://doi.org/10.13026/1pht-2b69.

[41] L. F. Nakayama et al. "Brset: A brazilian multilabel ophthalmological dataset of retina fundus photos," *PLOS Digital Health*, vol. 3, no. 7, pp. 1–16, Jul. 2024. [Online]. Available: https://doi.org/10.1371/journal.pdig.0000454.

[42] A. Goldberger et al. "Physiobank, physiotoolkit, and physionet : Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, E215–20, Jul. 2000. [Online]. Available: https://doi.org/10.1161/01.CIR.101.23.e215.

[43] H. Takahashi, *Davis grading of one and concatenated figures*, figshare. Figure, 2017. [Online]. Available: https://doi.org/10.6084/m9.figshare.4879853.v1.

[44] H. K. J. R and C. Sekhar Seelamantula, *Cháksu image: A glaucoma-specific fundus image database*, figshare. Dataset, 2023. [Online]. Available: https://doi.org/10.6084/m9.figshare.20123135.v2.

[45] R. Pires et al. *Advancing bag-of-visual-words representations for lesion classification in retinal images*, figshare. Dataset, 2016. [Online]. Available: https://doi.org/10.6084/m9.figshare.953671.v3.

[46] jr2ngb, *Cataract dataset*, 2018. [Online]. Available: https://www.kaggle.com/datasets/jr2ngb/cataractdataset.

[47] B. Qian et al. "A competition for the diagnosis of myopic maculopathy by artificial intelligence algorithms," *JAMA ophthalmology*, vol. 142, no. 11, pp. 1006–1015, 2024, Dataset Available at: https://zenodo.org/records/11025749. [Online]. Available: https://doi.org/10.1001/jamaophthalmol.2024.3707.

[48] T. Kauppi et al. "Diaretdb0: Evaluation database and methodology for diabetic retinopathy algorithms," Lappeenranta University of Technology, Tech. Rep., 2006, Dataset Available at: https://www.kaggle.com/datasets/nguyenhung1903/diaretdb1-standard-diabetic-retinopathy-database. [Online]. Available: https://www.siue.edu/~sumbaug/RetinalProjectPapers/Diabetic%20Retinopathy%20Image%20Database%20Information.pdf.

[49] K. K. Mujeeb Rahman, M. Nasor, and A. Imran, "Automatic screening of diabetic retinopathy using fundus images and machine learning algorithms," *Diagnostics*, vol. 12, no. 9, 2022, Dataset Available at: https://zenodo.org/records/4891308. [Online]. Available: https://www.mdpi.com/2075-4418/12/9/2262.

[50] T. Li et al. "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Information Sciences*, vol. 501, pp. 511–522, 2019, Dataset Available at: https://github.com/nkicsl/DDR-dataset. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025519305377.

[51] L. Lin et al. "The sustech-sysu dataset for automated exudate detection and diabetic retinopathy grading," *Scientific Data*, vol. 7, no. 1, p. 409, 2020, Dataset Available at: https://figshare.com/articles/dataset/The_SUSTech-SYSU_dataset_for_automated_exudate_detection_and_diabetic_retinopathy_grading/12570770/1. [Online]. Available: https://doi.org/10.1038/s41597-020-00755-0.

[52] J. Sivaswamy et al. "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis," *JSM Biomed Imaging Data Pap*, vol. 2, Jan. 2015. [Online]. Available: https://www.researchgate.net/publication/306939962_A_comprehensive_retinal_image_dataset_for_the_assessment_of_glaucoma_from_the_optic_nerve_head_analysis.

[53] J. Sivaswamy et al. "Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation," in *Proceedings of the 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, Dataset Available at: https://cvit.iiit.ac.in/projects/mip/drishti-gs/mip-dataset2/Home.php, IEEE, 2014, pp. 53–56. [Online]. Available: https://doi.org/10.1109/ISBI.2014.6867807.

[54] M. N. Bajwa et al. *G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection*, Dataset Available at: https://www.kaggle.com/datasets/ayush02102001/glaucoma-classification-datasets, 2020. arXiv: 2006.09158 [eess.IV]. [Online]. Available: https://arxiv.org/abs/2006.09158.

[55] Z. Zhang et al. "Origa(-light): An online retinal fundus image database for glaucoma analysis and research," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Dataset Available at: https://www.kaggle.com/datasets/ayush02102001/glaucoma-classification-datasets, 2010, pp. 3065–3068. [Online]. Available: https://doi.org/10.1109/IEMBS.2010.5626137.

[56] J. I. Orlando et al. "Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical Image Analysis*, vol. 59, p. 101 570, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841519301100.

[57] H. Fang et al. *Refuge2 challenge: A treasure trove for multi-dimension analysis and evaluation in glaucoma screening*, Dataset Available at: https://www.kaggle.com/datasets/jakkabalakrishna/refuge-2, 2022. arXiv: 2202.08994 [eess.IV]. [Online]. Available: https://arxiv.org/abs/2202.08994.

[58] A. W. Setiawan et al. "Color retinal image enhancement using clahe," in *International Conference on ICT for Smart Society*, 2013, pp. 1–3. [Online]. Available: https://doi.org/10.1109/ICTSS.2013.6588092.

[59] M. Alwazzan, M. Ismael, and A. N. Ahmed, "A hybrid algorithm to enhance colour retinal fundus images using a wiener filter and clahe," *Journal of Digital Imaging*, vol. 34, pp. 750–759, Jun. 2021. [Online]. Available: https://doi.org/10.1007/s10278-021-00447-0.

Abstract

[en]

Self-supervised learning has recently emerged as a promising approach for retinal image representation learning by leveraging large quantities of unlabelled data. Among these methods, Masked Autoencoders (MAEs) have demonstrated strong performance by learning to reconstruct missing image regions. However, conventional MAEs primarily focus on spatial-domain reconstruction and do not explicitly encourage the preservation of frequency-domain information, which may contain important retinal structures and pathological patterns.

This work proposes a Frequency-Regularised Masked Autoencoder (FR-MAE) for retinal image representation learning. The proposed approach extends the standard MAE framework by introducing a frequency-domain regularisation term based on the Fourier transform of reconstructed images. The objective is to encourage the preservation of high-frequency information during self-supervised pre-training and improve the transferability of the learnt representations.

The proposed method was pre-trained on approximately 90,000 unlabelled retinal images and subsequently fine-tuned for Diabetic Retinopathy and Glaucoma classification. Experimental evaluation was performed using ten fine-tuning seeds and included both internal test sets and cross-dataset generalisation experiments. The results showed that frequency-domain regularisation consistently improved downstream classification performance compared with the baseline MAE. In particular, the proposed approach achieved superior cross-dataset generalisation for both tasks and substantially reduced the performance gap with RETFound despite using a considerably smaller pre-training dataset.

These findings indicate that frequency-domain regularisation constitutes an effective complementary learning signal for self-supervised retinal representation learning and can improve the transferability and generalisation ability of retinal image representations.

Keywords: Self-Supervised Learning, Retinal Images, Masked Autoencoders, Frequency-Domain Regularisation, Diabetic Retinopathy, Glaucoma, Representation Learning.

Résumé

[fr]

L'apprentissage auto-supervisé est récemment apparu comme une approche prometteuse pour l'apprentissage de représentations d'images rétiniennes, en exploitant de grandes quantités de données non annotées. Parmi ces méthodes, les Masked Autoencoders (MAE) ont montré de bonnes performances en apprenant à reconstruire des régions masquées de l'image. Cependant, les MAE classiques se concentrent principalement sur la reconstruction dans le domaine spatial et n'encouragent pas explicitement la préservation des informations fréquentielles, qui peuvent contenir des structures rétiniennes importantes ainsi que des motifs pathologiques.

Ce travail propose un Masked Autoencoder à régularisation fréquentielle (FR-MAE) pour l'apprentissage de représentations d'images rétiniennes. L'approche étend le cadre standard du MAE en introduisant un terme de régularisation dans le domaine fréquentiel basé sur la transformée de Fourier des images reconstruites. L'objectif est d'encourager la préservation des informations de haute fréquence lors du pré-entraînement auto-supervisé et d'améliorer la transférabilité des représentations apprises.

La méthode proposée a été pré-entraînée sur environ 90 000 images rétiniennes non annotées, puis affinée pour les tâches de classification de la rétinopathie diabétique et du glaucome. L'évaluation expérimentale a été réalisée avec dix graines (seeds) de fine-tuning et comprend à la fois des datasets de test internes et des expériences de généralisation cross-dataset. Les résultats montrent que la régularisation dans le domaine fréquentiel améliore de manière constante les performances en classification en aval par rapport au MAE de référence. En particulier, l'approche proposée atteint une meilleure généralisation cross-dataset pour les deux tâches et réduit significativement l'écart de performance avec RETFound, malgré l'utilisation d'un ensemble de pré-entraînement considérablement plus réduit.

Ces résultats indiquent que la régularisation dans le domaine fréquentiel constitue un signal d'apprentissage complémentaire efficace pour l'apprentissage auto-supervisé de représentations rétiniennes et peut améliorer la transférabilité et la capacité de généralisation des représentations d'images rétiniennes.

Keywords: Apprentissage auto-supervisé, images rétiniennes, autoencodeurs masqués, régularisation dans le domaine fréquentiel, rétinopathie diabétique, glaucome, apprentissage de représentations.

ملخص

[ar]

لقد ظهر التعلّم ذاتي الإشراف مؤخراً كنهج واعد لتعلّم ترميزات صور الشبكية، وذلك من خلال الاستفادة من كميات كبيرة من البيانات غير المعلّمة. ومن بين هذه الأساليب، أثبتت المشفّرات التلقائية ذات المدخلات المحجوبة (Masked Autoencoders - MAEs) أداءً قوياً عبر تعلّم إعادة بناء الأجزاء المفقودة من الصورة. ومع ذلك، فإن هذه النماذج التقليدية تركز بشكل أساسي على إعادة البناء في المجال المكاني (Spatial Domain)، ولا تعزز بشكل صريح الحفاظ على المعلومات في المجال الترددي (Frequency Domain)، والذي قد يحتوي على بنى شبكية مهمة وأنماط مرضية دالة.

يقترح هذا العمل مشفرًا تلقائيًا ذو مدخلات محجوبة مُقيَّدا بالتردد (Frequency-Regularised Masked Autoencoder - FR-MAE) لتعلّم تمثيلات صور الشبكية. ويقوم النهج المقترح بتوسيع إطار المشفّرات التلقائية ذات المدخلات المحجوبة التقليدية من خلال إدخال حدّ تقييد في المجال الترددي يعتمد على تحويل فوريه (Fourier Transform) للصور المعاد بناؤها. والهدف هو تشجيع الحفاظ على المعلومات عالية التردد أثناء مرحلة التدريب ذاتي الإشراف، وتحسين قابلية نقل الترميزات المتعلّمة إلى مهام لاحقة.

تم تدريب النموذج المقترح مسبقاً على حوالي 90,000 صورة شبكية غير معلّمة، ثم تم تكييفه لاحقاً لمهام تصنيف اعتلال الشبكية السكري (Diabetic Retinopathy) والزَرَق (Glaucoma). وقد أُجري التقييم التجريبي باستخدام عشر عمليات تكييف مختلفة (seeds)، وشمل مجموعات اختبار داخلية وتجارب تعميم عبر مجموعات بيانات مختلفة. أظهرت النتائج أن التّقييد في المجال الترددي حسّن بشكل ثابت أداء التصنيف في المهام اللاحقة مقارنةً بالنموذج التقليدي (Baseline MAE). وعلى وجه الخصوص، حقق النهج المقترح أداءً أفضل في التعميم عبر مجموعات البيانات لكلا المهمتين، وقلّص بشكل ملحوظ الفجوة في الأداء مع نموذج RETFound رغم استخدام مجموعة بيانات أصغر بكثير خلال مرحلة التّدريب ذاتي الإشراف.

تشير هذه النتائج إلى أن التّقييد في المجال الترددي يشكّل إشارة تعلم مكملة فعّالة في التعلم ذاتي الإشراف، ويمكنه تحسين قابلية النقل وقدرة التعميم لترميزات الصور الشبكية.

الكلمات المفتاحية: التعلم ذاتي الإشراف، صور الشبكية، المشفّرات التلقائية ذات المدخلات المحجوبة، التّقييد في المجال الترددي، اعتلال الشبكية السكري، الزَرَق، تعلّم الترميزات.