POLITECNICO DI TORINO

Master's Degree in Artificial Intelligence & Data Analytics

Politecnico di Torino

Master Thesis

Evaluating Cross-Domain Adaptation Strategies for Foundation Models: A Study on Fluorescein Angiography

Supervisors
Prof. Filippo Molinari
Prof. Massimo Salvi
Dr. André Anjos
Dr. Oscar Jiménez del Toro

Candidate
Roberto Pulvirenti

Academic Year 2024–2025
April 2025

Acknowledgements

Il percorso raccontato in questo elaborato è solo un piccolo frammento del lungo e appassionato viaggio che ho intrapreso in questi mesi. Un viaggio che, a dir la verità, è iniziato ben prima della decisione di svolgere la mia tesi magistrale all'estero, fatto di esperienze ma, soprattutto, di persone che mi hanno reso ciò che sono oggi. A tutte loro vorrei dedicare questa prima pagina, prendendomi il tempo che meritano per ringraziarle sinceramente.

Per prima cosa, un sentito ringraziamento a coloro che hanno contribuito direttamente alla realizzazione di questo lavoro di tesi. Al mio relatore, Filippo Molinari, e al mio correlatore, Massimo Salvi, per la costante disponibilità e gentilezza dimostrata lungo tutto il percorso. Ad André Anjos e Oscar Jiménez del Toro, che mi hanno offerto l'incredibile opportunità di lavorare a un progetto così entusiasmante, supportandomi e guidandomi in ogni fase di questa tesi. A voi devo un'esperienza di vita che porterò sempre nel cuore.

Non basterebbe un libro per contenere tutti i ringraziamenti che vorrei rivolgere ai miei genitori. Mamma e Papà, grazie di tutto: per l'amore incondizionato, per il sostegno mai mancato e per aver sempre creduto in me. Se oggi sono arrivato fin qui, lo devo soprattutto a voi e ai vostri insegnamenti.

Un ringraziamento di cuore ai tanti amici incontrati tra le mura del Politecnico, con cui ho condiviso ansie, paure, ma soprattutto traguardi.

Un pensiero speciale va a BEST, quella meravigliosa e un po' folle associazione che mi ha profondamente cambiato, insegnandomi quanto sia importante mettersi in gioco, credere in qualcosa e impegnarsi per realizzarlo. È proprio lì che ho conosciuto Pierpaolo, una parte di me stesso che porterò sempre con me e che si riflette nei miei legami più autentici.

Un grazie particolare va all'amico di sempre, Federico, l'unico che mi accompagna sin dall'infanzia. Nonostante la distanza e il tempo che può passare tra un incontro e l'altro, con lui è sempre come se ci fossimo visti il giorno prima. La nostra amicizia è una costante silenziosa ma profondamente presente della mia vita.

Infine, il mio ultimo ma più profondo Grazie va a Giada, che mi è stata accanto fin dal principio di questa avventura all'estero. Sapevamo entrambi cosa avrebbe significato questa scelta: il peso della lontananza, i mesi trascorsi separati e i giorni difficili. Eppure, non hai mai esitato a sostenermi, con amore, forza e una generosità rara, anteponendo il mio bene ai tuoi dispiaceri. A te, amore mio, grazie dal profondo del cuore: per le videochiamate che mi hanno fatto sorridere, per le risate che mi hanno salvato nei momenti più duri, per avermi ascoltato con pazienza e per esserci stata, sempre e comunque.

Abstract

Large-scale artificial general intelligence models have achieved unprecedented success in various general-domain tasks in recent years. However, their direct application to specialized fields like medical imaging remains challenging due to the field's inherent complexities. Unlike natural images, which feature everyday objects recognizable through common sense, medical images demand deep domain expertise for accurate interpretation. This makes the annotation process particularly challenging, as producing a large volume of high-quality labeled medical data requires the involvement of multiple experts, a resource that is often limited and difficult to scale. Moreover, while some medical imaging modalities, such as Color Fundus Photography (CFP), benefit from extensive annotated datasets, others, like Fluorescein Angiography (FA), suffer from a severe lack of data due to their niche and less frequently used nature. This discrepancy in data availability across modalities highlights the necessity of cross-domain adaptation, leveraging knowledge from well-established imaging modalities to improve performance on underrepresented ones.

Vision Transformers (ViTs) have shown significant promise in medical image analysis tasks such as reconstruction, segmentation, and classification, owing to their strong representational and generalization capabilities. Yet, their dependency on large, annotated datasets renders training from scratch impractical for medical applications. To overcome this hurdle, transfer learning from models pre-trained on large-scale datasets, such as ImageNet, has become a widely adopted approach. The success of these pre-trained ViTs, often referred to as visual foundation models, has inspired further research into adapting these architectures for more powerful, domain-specific applications. In ophthalmology, this progress has led to the development of RETFound, a large ViT model pre-trained on nearly one million CFP images using the Masked Autoencoder technique, and its smaller variant, RETFound Green, which was pre-trained on only 75K CFP images using a novel self-supervised technique called Token Reconstruction.

The objective of this thesis is to systematically evaluate and compare different cross-domain adaptation techniques for vision models within the medical imaging domain of fluorescein angiography. Specifically, it examines strategies for transferring knowledge to FA by comparing models pre-trained on CFP images (e.g., RETFound and RETFound Green) with those pre-trained on natural images (e.g., ImageNet-trained ViTs) using three primary adaptation approaches: full fine-tuning, self-supervised pre-training, and parameter-efficient fine-tuning. By evaluating the capabilities of these foundation models in adapting to FA tasks, this study assesses their robustness, efficiency, and clinical applicability across eleven distinct tasks using two datasets: the publicly available *3rd Aptos Competition dataset* and the private *HOJG dataset* provided by the Hospital Ophthalmic Jules-Gonin in Lausanne.

Particular emphasis is placed on evaluating the trade-offs between full fine-tuning, widely regarded as the de facto standard for adapting foundation models to downstream tasks despite its high computational cost, and parameter-efficient methods such as LoRA,

which substantially reduce the number of trainable parameters while maintaining competitive performance. Additionally, the study explores the role of self-supervised learning in mitigating domain shifts and enhancing model generalization.

Through a comprehensive analysis of cross-domain adaptation strategies, this thesis contributes to a deeper understanding of how foundation models can be effectively leveraged in medical imaging applications, supporting the development of more efficient and clinically relevant AI-assisted diagnostic systems.

Contents

Chapter 1

Introduction

1.1 Artificial Intelligence

Artificial Intelligence (AI) is a research field within computer science devoted to enabling computational systems to perform tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, perception, and decision-making. Its goal is to create rational agents, systems capable of perceiving their environment, setting goals, and taking informed actions to maximize their chance of success [1].

A key milestone in the history of AI is Alan Turing's seminal paper "*Computing Machinery and Intelligence*", published in 1950 in Mind [2]. Turing, widely regarded as one of the founding figures of computer science and artificial intelligence, posed the fundamental question: "Can machines think?" To approach this question, he proposed an indirect method known as the Turing Test. In this test, a machine is considered intelligent if a human Interrogator, after engaging in a written conversation, cannot distinguish the machine's written responses from those of a human. Passing the test implies that the machine demonstrates intelligence indistinguishable from that of a person.

Successfully fooling the Interrogator requires a variety of intelligent capabilities. For instance, the ability to perceive, recognize and interact with objects in the environment, known as perceptual intelligence, is tackled through fields like *Computer Vision* and *Robotics*. The ability to acquire and store what has been perceived, called cognitive intelligence, involves *Knowledge Representation*. High-level abilities like pattern extraction and recognition to make decisions and adapt to the circumstances are addressed by *Machine Learning* and *Automated Reasoning*. Finally, the capacity to produce coherent and contextually appropriate responses relies on *Natural Language Processing*. Together, these six subfields form the core disciplines of modern AI [3].

Despite various criticisms and limitations, Turing's test remains a foundational concept in AI, continuing to influence discussions about machine intelligence even 75 years later.

1.2 Machine Learning

Learning is the process of acquiring new understanding or knowledge. In the context of computer science, Machine Learning (ML) refers to a branch of artificial intelligence that enables systems to learn from data and improve their performance on a specific task without being explicitly programmed to do so. More formally, a computer program is said to learn from experience E for some class of tasks T and performance measure P if its performance on task T, as measured by P, improves with E [4]. Rather than relying on hard-coded rules, ML automates the construction of analytical models by using algorithms that iteratively learn from data. This allows systems to generalize from past observations, uncover hidden patterns, and make informed predictions or decisions in new situations.

Machine learning encompasses a wide variety of algorithms, commonly referred to as models, each differing in how they represent candidate solutions and how they search for optimal ones. These models can take many forms, including decision trees, mathematical functions, or even general-purpose programming languages. The process of learning involves searching through a space of possible programs to identify the one that best maps input data to the desired output. The choice of model depends not only on the characteristics of the data and the nature of the task but also on how the algorithm performs this search. Some methods use optimization techniques with well-understood convergence properties, such as gradient-based approaches, while others rely on evolutionary strategies that explore the solution space through successive generations of randomly mutated candidates. This diversity in representation and search strategies allows ML to be applied effectively across a wide range of problem domains especially those involving large-scale datasets and high-dimensional spaces. As a result, it has become a core technology across many sectors, including healthcare, where it is used for diagnostics, personalized treatment, and predictive analytics.

1.3 Deep Learning

Deep learning is a subset of machine learning that uses artificial neural networks (ANNs) to process and analyze information without human intervention.

1.3.1 Artificial Neural Network

Drawing inspiration from the mechanisms of information processing in biological systems, ANNs are composed of interconnected computational units, known as artificial neurons, which are defined through mathematical models.

As illustrated in Fig. 1.1, given an input signal encoded as a vector $\mathbf{x} = [x_1, x_2, ..., x_n]$, the output of a single neuron is computed through a weighted linear combination of its values, followed by a non-linear activation function. Mathematically, this is expressed as follows:

$$z = \sum_{i=1}^{n} w_i x_i + b \tag{1.1}$$

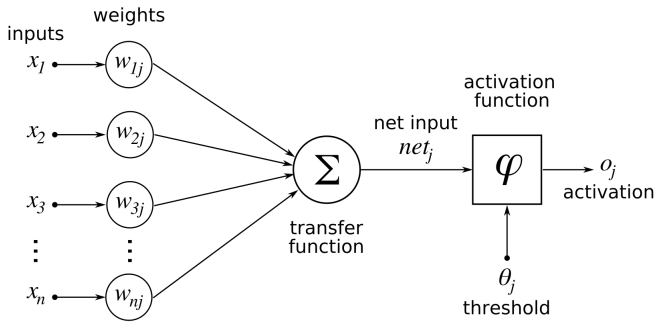

Figure 1.1: Schematic representation of an artificial neuron.

$$o = \varphi(z) \tag{1.2}$$

where w_i denotes the weight associated with input x_i, b is a bias term, and $\varphi(\cdot)$ is a non-linear activation function, such as the sigmoid, hyperbolic tangent, or Rectified Linear Unit, whose output o is passed onto the next layer of the network. Equations (1.1) and (1.2) enable the neuron to model complex relationships in the data and serve as the core of deep learning models. Neurons are typically organized into layers: an input layer that receives raw data, an output layer that produces the desired outcome, and one or more hidden layers in between that allow the network to learn non-linear mappings from inputs to outputs [5].

Similarly to synapses in a biological brain, the weights modulate the influence of each input and are adjusted during training to minimize the error in the network's predictions through a backpropagation algorithm.

1.3.2 Backpropagation

Backpropagation, short for *backward propagation of error*, is a key algorithm used to train artificial neural networks. As discussed previously, the layers of a neural network can be viewed as a composition of nested mathematical functions, where each layer transforms its input and passes the result to the next. During training, this composition becomes part of a larger function, known as the *loss function*, which quantifies the discrepancy between the network's prediction and the true target, called the *ground truth*, for a given input. To determine how each individual neuron influences the overall error, the chain rule, a principle that enables the differentiation of composite functions, is applied. This approach allows to compute how changes in any parameter, specifically the weights and biases, affect the total loss. By understanding these contributions, the model can adjust its parameters to reduce the error, thereby improving its predictions over time.

To provide a clear and structured understanding of the overall process, in the following a step-by-step breakdown accompanied by its mathematical formulation is presented.

1. The network first performs a *forward pass*, in which the input data $\mathbf{x}$ is propagated through the network until it reaches the output layer, where it is transformed into a probability prediction.

2. The final prediction $\hat{y}$ is then compared with the ground truth y using a *loss function* $\mathcal{L}(y, \hat{y})$, which measures the error between the predicted and actual values.

3. To improve the performance of the model, backpropagation computes how each weight and bias in the network contributes to this error by applying the *chain rule* of calculus in reverse, from the output back to the input layer. Specifically, it calculates the *gradient* of the loss function with respect to each parameter. For a given weight $w_{ij}^{(l)}$, which connects neuron i in layer $l-1$ to neuron j in layer l, the gradient is given by (1.3):

$$\frac{\partial \mathcal{L}}{\partial w_{ij}^{(l)}} = \delta_j^{(l)} \cdot o_i^{(l-1)} \tag{1.3}$$

where $o_i^{(l-1)}$ is the activation of neuron i in the previous layer, following (1.2) and $\delta_j^{(l)}$ is the error term of neuron j in the current layer, computed as:

$$\delta_j^{(l)} = \frac{\partial \mathcal{L}}{\partial o_j^{(l)}} \cdot \varphi'\left(z_j^{(l)}\right) \tag{1.4}$$

In (1.4) $z_j^{(l)}$ is the weighted input to neuron j, and φ' denotes the derivative of the activation function. The gradient with respect to the bias term $b_j^{(l)}$ is simply:

$$\frac{\partial \mathcal{L}}{\partial b_j^{(l)}} = \delta_j^{(l)} \tag{1.5}$$

4. Once these gradients are computed for all weights and biases in the network, they are used to update the parameters using an optimization algorithm, typically *gradient descent*. In its simplest form, each parameter is updated by moving it in the direction opposite to the gradient:

$$w_{ij}^{(l)} \leftarrow w_{ij}^{(l)} - \eta \cdot \frac{\partial \mathcal{L}}{\partial w_{ij}^{(l)}} \tag{1.6}$$

$$b_j^{(l)} \leftarrow b_j^{(l)} - \eta \cdot \frac{\partial \mathcal{L}}{\partial b_j^{(l)}} \tag{1.7}$$

where η is the *learning rate*, a hyperparameter that controls the step size of the update.

By iteratively performing the above steps, the network gradually improves its predictions to learn from the data by minimizing the loss function.

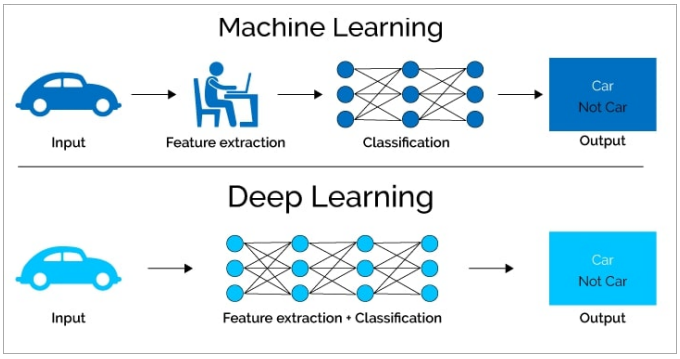

Figure 1.2: Comparison between Machine Learning and Deep Learning workflows. In traditional Machine Learning, a manual feature extraction step is required to convert raw data into structured representations that are then used by the learning algorithm. In contrast, Deep Learning models learn these representations automatically from raw data, eliminating the need for handcrafted features. Image reproduced from [6]

1.3.3 The main differences with ML

Now that the mechanisms underlying the potential of neural networks have been outlined, it is important to clarify the key differences between ML and DL.

As illustrated in Figure 1.2, a fundamental distinction lies in the way these approaches handle input data and feature extraction. Traditional ML methods typically require manual feature engineering, an intermediate step in which domain knowledge is used to extract relevant features from raw data. These features are then used by algorithms such as decision trees, support vector machines, or linear regression to learn the mapping between input and output. In contrast, DL models automatically learn hierarchical representations of data through their layered architectures. This end-to-end learning process allows deep neural networks to operate directly on raw data such as images, text, or audio, identifying complex patterns without requiring handcrafted input representations.

Thanks to this ability, DL has achieved state-of-the-art performance in several challenging tasks, including image recognition, speech processing, and natural language understanding, domains where traditional ML techniques often fall short due to their limited capacity to model unstructured and high-dimensional data. However, this expressive power comes at a cost: DL models typically require vast amounts of labeled data to outperform traditional ML methods and demand significantly higher computational resources during training and inference.

1.4 AI in Medicine

Medical imaging has become a cornerstone of modern clinical practice, playing a critical role in both diagnosis and treatment planning. Its growing importance is reflected in the continuous rise in imaging utilization across healthcare systems worldwide. As a result, clinicians are facing an ever-growing workload, often requiring them to interpret images at a pace that imposes an unsustainable cognitive burden. This pressure not only affects

efficiency, but also increases the risk of diagnostic delays and errors. These challenges underscore the urgent need for automated systems that can assist medical professionals by improving both the speed and accuracy of image interpretation. AI, particularly deep learning, has achieved expert-level performance in many medical image interpretation tasks and could represent a valid tool to be integrated into clinical workflows to support physicians in diagnostic decision making [7].

Ophthalmology, while not the most studied area in medical AI research, is uniquely positioned for the application of deep learning due to its strong dependence on high-resolution medical imaging such as color fundus photography (CFP), optical coherence tomography (OCT), fluorescein angiography (FA), and indocyanine green angiography (ICGA). Deep learning approaches have been applied to detect and classify a range of ocular diseases, including diabetic retinopathy and age-related macular degeneration, demonstrating encouraging performance in these tasks [8] [9]. Beyond performance metrics, researchers have begun to evaluate the real-world impact of these systems. For instance, semi-automated and fully automated AI screening tools have shown potential for cost savings in specific contexts, for example, the detection of diabetic retinopathy [10]. This shows that AI can also improve accessibility and efficiency of healthcare delivery.

Despite these advances, current AI success in medical imaging is mainly driven by *supervised learning*, a training approach in which models learn to associate inputs with corresponding labeled outputs. However, producing such labeled datasets in the medical domain is particularly challenging due to the need for expert annotation, making it both time-consuming and expensive. Moreover, models trained in this way often struggle to generalize to external clinical environments or different tasks, as they tend to learn label-specific features rather than robust, transferable representations. Addressing this limitation requires a paradigm shift in the way medical AI systems are trained.

1.4.1 Foundation models

In this context, the development of Foundation Models (FMs) offers a promising path forward. These are large-scale models pre-trained on vast and diverse datasets, often using self-supervised learning techniques that do not rely on manual annotation [11]. Once pre-trained, FMs can be adapted to a wide range of downstream tasks through fine-tuning, even when only limited labeled data is available. This flexible, task-agnostic training paradigm enhances the model's representational power, resulting in improved generalizability across different applications.

The growing enthusiasm towards FMs has spurred interest in developing domain-specific applications tailored to medical contexts. Unlike models trained on natural images, medical foundation models must learn to identify subtle and complex features that are highly specific. In the field of ophthalmology, this has led to the emergence of retinal foundation models. A notable example are *RETFound* and *RETFound Green*, both pre-trained on a wide number of unlabeled CFP images and subsequently fine-tuned for a variety of ocular disease tasks. These models have demonstrated the potential to significantly reduce the need for large, labeled datasets while maintaining high adaptability and performance across diverse ophthalmic applications. As a result, foundation models represent a promising advancement in the development of efficient, generalizable, and

scalable AI solutions for medical imaging.

1.5 Thesis objective

Building on recent advancements in AI-driven medical imaging in ophthalmology, there is increasing interest in applying these technologies to support the diagnosis and treatment of complex retinal diseases. Among these, uveitis stands out as a challenging and potentially sight-threatening condition where imaging plays a central role. *Uveitis* refers to a heterogeneous group of intraocular inflammatory disorders that affect the uvea and the retina. With an estimated annual incidence ranging from 17 to 52 cases per 100,000 individuals, this debilitating condition can cause significant vision loss and continues to be a major contributor to blindness in industrialized countries [12].

Unlike other imaging modalities, *Fluorescein Angiography* (FA) uniquely allows the assessment of the integrity and function of the blood-retinal barrier in vivo, making it the gold standard for evaluating inflammatory conditions. The procedure involves intravenous injection of fluorescein dye followed by sequential fundus photography to capture the transit of the dye through the retinal vessels. This provides information on vessel permeability over time and allows clinicians to identify active sites of inflammation, including the macula, optic disc, and peripheral retina, structures often implicated in disease severity.

The ability to detect earlier vascular abnormalities is critical to guide therapeutic decisions, particularly the initiation or adjustment of immunomodulatory treatment. Despite its diagnostic value, FA is rarely considered in clinical trials due to the many difficulties it has to face. In addition to being an invasive examination, its accurate interpretation is time-consuming and inherently complex, as it is influenced by the dynamic nature of dye diffusion and the presence of confounding factors. Furthermore, current approaches to identifying fluorescein leakage rely on manual assessment by experts, which, even with many years of clinical experience, introduces subjectivity and variability into the evaluation process.

1.5.1 Scientific question

This thesis investigates how foundation models can be effectively adapted for the automated interpretation of FA, a retinal imaging modality with a significant lack of available data due to its specialized nature and less frequent use. Specifically, it addresses the scientific question: "*How can knowledge learned from data-rich modalities, such as color fundus photography, be transferred to improve the performance of AI models on underrepresented modalities like fluorescein angiography?*"

1.5.2 Major contributions

To answer this, the thesis systematically evaluates cross-domain adaptation strategies using foundation models. The key contributions of this study are as follows:

- **Systematic comparison of foundation models**. A rigorous evaluation of four different vision foundation models, including RETFound, RETFound Green, and

their respective baselines pre-trained on natural images, is conducted across 11 downstream FA classification tasks. This offers one of the most comprehensive cross-domain assessments in the retinal imaging literature to date.

- **Comprehensive evaluation of adaptation strategies**. A complete comparison of three major adaptation strategies, full fine-tuning, self-supervised continual pre-training, and parameter-efficient fine-tuning with LoRA, is carried out. The analysis goes beyond standard performance metrics by also evaluating model reliability and confidence, key factors for deployment in clinical settings, providing a well-rounded perspective on the strengths and limitations of each approach.
- **Investigation of the value of self-supervised learning in cross-domain setting**. Self-supervised strategies, including Masked Autoencoder and Token Reconstruction, are examined as continual pre-training methods to assess their potential for improving generalization in the FA domain under a data-scarce condition.

1.6 Thesis structure

The structure of this thesis is designed to guide the reader through a logical progression from foundational knowledge to methodological implementation, experimental validation, and final conclusions.

- **Chapter 2** introduces the *background and related work* that inform the research presented in this thesis.
- **Chapter 3** details the *experimental design* and *methodological framework* used to address the thesis question. This includes a description of the datasets, adaptation strategies, and implementation details, ensuring transparency and reproducibility of the research.
- **Chapter 4** presents the experimental results and evaluates them in light of the objectives and hypothesis clearly outlined at the end of Chapter 2.
- **Chapter 5** concludes the thesis by summarizing the main contributions, discussing the limitations of the current work, and offering directions for future research.

Chapter 2

Background and related work

This chapter presents the foundational concepts and prior work that underpin the methods explored in this thesis. It is organized into two main sections: the first provides the background on Vision Transformers (ViTs) and their adaptations, with a focus on architectural components, transfer learning strategies, and efficient fine-tuning methods such as adapters. The second section surveys related work, with particular attention to how deep learning models have been applied in the context of ophthalmology, including the development of retinal foundation models like RETFound and RETFoundGreen, as well as previous deep learning efforts on FA.

2.1 Background

2.1.1 Vision Transformers (ViTs)

The emergence of Vision Transformers (ViTs) has marked a new era in computer vision, demonstrating exceptional performance in a wide range of tasks by leveraging attention mechanisms and large-scale data-driven learning. Originally introduced for natural image understanding, ViTs and their increasingly powerful variants, often referred to as *visual foundation models*, have rapidly gained traction in medical imaging due to their ability to learn complex representations with minimal inductive biases.

Attention mechanism

The attention mechanism represents the heart of Transformer-based architectures. It enables models to dynamically focus on the most informative parts of the input, allowing for more efficient and context-aware processing. At its core, an attention function maps a query and a set of key-value pairs to an output, where all elements are vector representations. The output is computed as a weighted aggregation of the values, with weights determined by the relevance between the query and each key [13]. This paradigm not only improves the model's ability to handle long-range dependencies, but also reduces the reliance on strictly sequential processing.

The basis of this technique lies in the *Scaled dot-product attention*, left-side of Fig. 2.1 Given matrices $Q \in \mathbb{R}^{n \times d_k}$ (queries), $K \in \mathbb{R}^{m \times d_k}$ (keys), and $V \in \mathbb{R}^{m \times d_v}$ (values), the

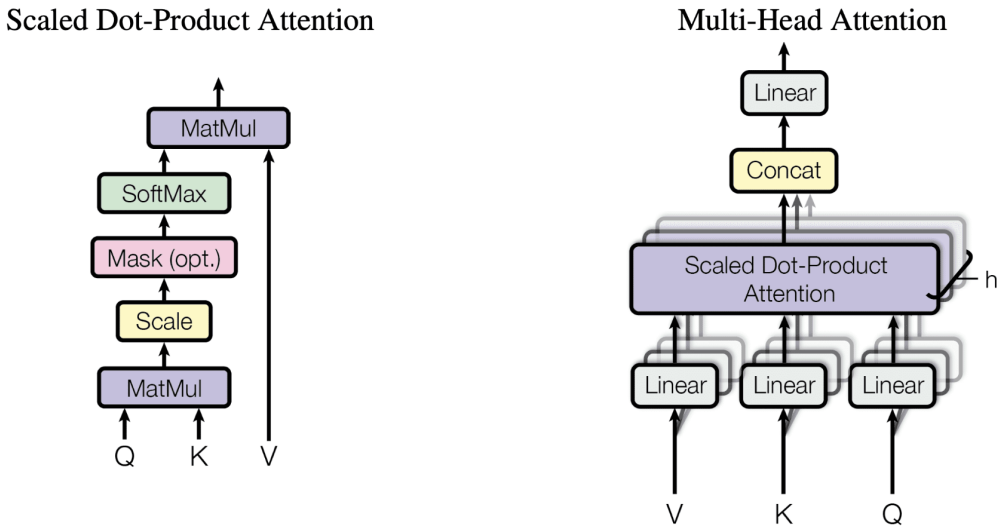

Figure 2.1: Scaled Dot-Product Attention (on the left) and Multi-Head Attention (on the right). Image reproduced from [13]

attention output is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.1}$$

where the scaling factor at the denominator in (2.1) mitigates the risk of excessively large dot products when the dimensionality d_k is high. Without scaling, large values could push the softmax into regions with vanishing gradients, thereby hindering learning.

Rather than computing attention once over d_{model} feature representations, *Multi-Head Attention* enables the model to attend to information from multiple subspaces in parallel, right-side of Fig. 2.1. Specifically, queries, keys, and values are linearly projected into h smaller subspaces using learned weight matrices. Each head i performs its own scaled dot-product attention:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.2}$$

where $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, and $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ are the learned projection matrices for the i-th head. The outputs from all heads are then concatenated and transformed through another learned linear projection:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.3}$$

where $W^O \in \mathbb{R}^{h \cdot d_v \times d_{\text{model}}}$. This multi-headed approach allows the model to jointly capture information across different representation subspaces and positions, which is especially beneficial for tasks involving complex dependencies.

Model architecture

Transformers had become the state-of-the-art method in natural language processing since they were first proposed. Inspired by this success, ViTs adapt the same core principles to image data by representing an image as a sequence of visual tokens [14].

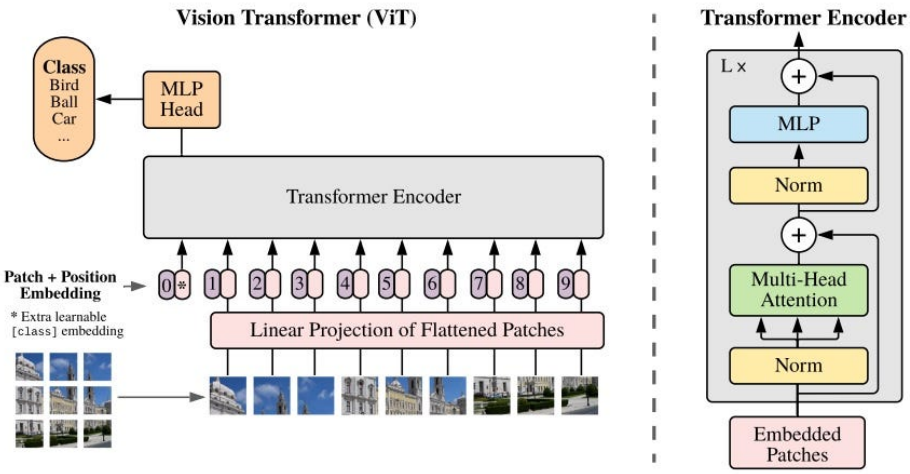

Figure 2.2: Overview of the Vision Transformer architecture. Image reproduced from [14]

The key idea is to convert images into patch-based embeddings, allowing the standard Transformer encoder to process them similarly to how it handles words in text. The step-by-step procedure to obtain the visual tokens is as follows:

1. *Patch Embedding.* Raw images are first divided into non-overlapping patches of fixed size, typically $P \times P$ pixels. For an image of resolution $H \times W$ with C channels, this results in $N = \frac{HW}{P^2}$ patches, each with dimensionality $P^2 \cdot C$. This creates a sequence of patches that can be treated similarly to tokens in natural language processing.

2. *Flattening Patches.* Each patch is flattened into a 1D vector.

3. *Linear Transformation.* The resulting vectors are passed through a linear projection layer to obtain a D-dimensional embedding:

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \ldots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{\text{pos}}, \tag{2.4}$$

   where $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ is the learnable projection matrix.

4. *CLS Token Addition.* The prepended token $\mathbf{x}_{\text{class}}$ in (2.4) is a learnable parameter that serves as a global image descriptor, later used for classification.

5. *Positional Encoding.* $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$ in (2.4) represents positional embeddings. To preserve spatial relationships between patches, these are added to each token

to help the model retain a sense of structure within the image, even though it is processed as a sequence.

The resulting embedding vectors are then fed into a standard Transformer encoder. As shown in Fig. 2.2, this is composed of two alternating layers: a *Multi-Head Self-Attention* (MSA), where the queries, keys and values are all derived from the same input sequence, and a *MLP* block. Each layer includes residual connections after every block and layer normalization (LN) before each of them:

$$\mathbf{z}'_\ell = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \tag{2.5}$$

$$\mathbf{z}_\ell = \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \tag{2.6}$$

where $\ell = 1, \ldots, L$ indicates the layer index. The final image representation is extracted from the transformed class token:

$$\mathbf{y} = \mathrm{LN}(\mathbf{z}_L^0). \tag{2.7}$$

In practice, this depends on the type of pooling applied to the output vectors of the final encoder block $\mathbf{z}_L$, but the equation in (2.7) represents the default approach.

Finally, for image classification tasks, the final state of the class token is passed through a lightweight classification head, typically a linear layer or a shallow MLP, to produce label predictions. During pre-training, a two-layer MLP may be used, while fine-tuning often uses a single linear layer initialized from scratch.

2.1.2 Adaptation of ViTs

While ViTs provide powerful general-purpose representations, effectively transferring their capabilities often requires task-specific customization. This section outlines the adaptation paradigms analyzed throughout the study.

Full Fine-Tuning

FFT is one of the most straightforward and widely used strategies for adapting pre-trained models to downstream tasks. It involves updating all the parameters of a pre-trained model using task-specific data, allowing the model to fully specialize in the target domain. This approach is significantly more data and compute-efficient than training models from scratch, particularly when leveraging large-scale supervised pre-training, such as models trained on ImageNet. Although highly effective compared to training from scratch, FFT requires meticulous hyperparameter tuning and becomes increasingly computationally expensive as model size scales, posing practical challenges for adaptation.

In the medical domain, where labeled data is often scarce due to privacy restrictions, costly annotations, and the long-tailed distribution of conditions, FFT offers a viable pathway to transfer rich visual representations to specific diagnostic tasks. However, this process must be handled with care to ensure proper adaptation without overfitting, particularly as the size and generality of foundation models continue to grow.

Parameter-Efficient Fine-Tuning through LoRA

In recent years, Parameter-Efficient Fine-Tuning (PEFT) has emerged as a promising alternative to full fine-tuning, addressing the high computational demands associated with updating all model parameters [15]. PEFT strategies introduce lightweight adapter modules that are used to update only a small subset of the pre-trained parameters, significantly reducing the number of trainable weights. This not only preserves the general knowledge encoded in the original model, but also mitigates the risk of catastrophic forgetting during adaptation. Importantly, by limiting the extent of parameter updates, PEFT helps prevent overfitting, especially when fine-tuning large models on relatively small datasets, while still achieving performance comparable to FFT.

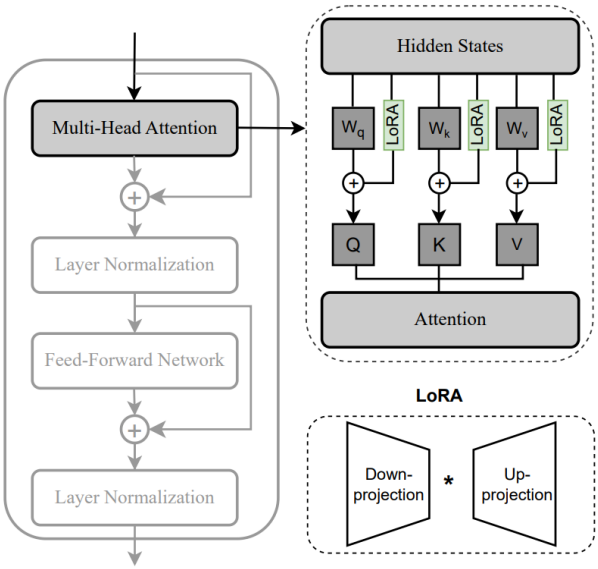

Figure 2.3: Illustration of LoRA applied to a Transformer block. The trainable components are shown in green, representing the low-rank adaptation matrices. The rest of the model components are shown in grey, indicating that they remain frozen during training. Image reproduced from [16]

One of the most popular PEFT methods is LoRA: Low-Rank Adaptation [17]. It introduces trainable low-rank matrices to adapt large pre-trained models while keeping the original model weights frozen. Rather than updating all parameters during fine-tuning, LoRA restricts weight updates to a low-dimensional subspace by representing them as a product of two smaller matrices A and B. Specifically, for a given pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA models the adaptation as $W = W_0 + \Delta W$, where $\Delta W = BA$, with $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d, k)$. Only A and B are optimized during training, while W_0 remains unchanged. This formulation significantly reduces the number of trainable parameters and the memory footprint. In the case of Transformer blocks, LoRA modules are applied in parallel to the query, key, and value projection matrices within the attention layer, as illustrated in Fig. 2.3

Self Supervised Learning

Self-supervised learning is an emerging paradigm that enables models to learn meaningful and transferable representations from unlabeled data by solving proxy or "pretext" tasks, where supervision signals are derived automatically from the data itself. This approach alleviates the dependency on large-scale annotated datasets, which, as already highlighted in previous sections, are often absent in the medical domain. SSL has shown remarkable success in natural language processing and is increasingly proving its value in computer vision by outperforming traditional supervised pre-training strategies in several benchmarks, particularly under label-scarce conditions [18]. At the same time, this generality comes with a trade-off: SSL generally requires access to a substantial volume of unlabeled data. For instance, reducing the number of pre-training images from 250k to 50k has been shown to result in over a 10% drop in accuracy on downstream tasks, while a decrease from 1 million to 250k images leads to a 2–4% performance drop [19].

The ability of SSL models to generalize across tasks and domains makes them particularly appealing for medical imaging applications, offering a promising approach for building robust and scalable solutions.

The following subsections provide a detailed explanation of the two self-supervised learning techniques employed in this study: Masked AutoEncoder (MAE) and Token Reconstruction.

Masked Autoencoder

The Masked Autoencoder [20] is a self-supervised learning framework that reconstructs images from incomplete observations, encouraging models to learn rich and generalizable visual representations.

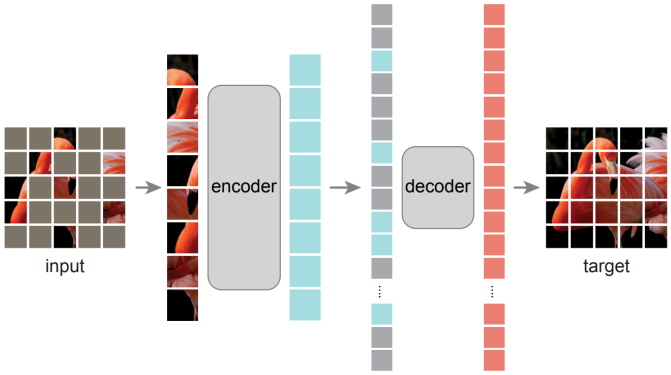

Figure 2.4: Illustration of MAE SSL technique. Image reproduced from [20]

The core idea is to randomly mask a large portion of an input image, typically around 75%, by dividing it into non-overlapping patches and selecting a subset through uniform sampling. This high masking ratio substantially reduces input redundancy and compels the model to infer global structure rather than relying on local continuity. In contrast to classical autoencoders, the architecture adopts an asymmetric encoder-decoder design:

the encoder, a ViT, processes only the visible patches and completely ignores masked regions, thus reducing computational overhead. The decoder receives a full sequence of tokens, composed of the encoded visible patches and learnable mask tokens, each representing a missing patch location. Positional embeddings are added to all tokens to preserve spatial relationships. The decoder, designed to be significantly lighter than the encoder, reconstructs the image by predicting the pixel values for each masked patch. Its output is passed through a linear projection layer to generate per-patch reconstructions, which are then reshaped to form the complete image. Fig. 2.4 illustrates the idea. The loss is calculated using the mean squared error (MSE) between the reconstructed and original pixel values, but only over the masked regions. This approach enables efficient pre-training, often reducing training time and memory usage by a factor of three or more, while maintaining or improving performance, making it an attractive choice for scaling large models.

Token Reconstruction

This technique was originally developed with the goal of creating a self-supervised learning approach specifically tailored for building domain-specific medical foundation models, rather than training models from scratch for general-purpose computer vision. Traditional methods like MAE were designed for training from randomly initialized weights and thus typically require large amounts of data and computational resources. In contrast, when working in the medical imaging domain, it is often more practical to begin with models that have already been pre-trained on large natural image datasets (e.g. ImageNet), as these models have already learned broadly transferable features. Acknowledging this, the Token Reconstruction strategy is guided by two key objectives: first, to further enhance the model's understanding of the general structure and appearance of medical images, leveraging the foundational knowledge inherited from pre-training, and second, to refine the model's representations for the medical domain without erasing or overriding the useful features acquired during the earlier stages of training.

The core idea is presented in Fig. 2.5 It involves using a frozen copy of a pre-trained model to generate feature tokens from clean, uncorrupted images. A second trainable version of the same model receives noisy versions of the same images and learns to produce output tokens that closely match those generated by the frozen model. This approach encourages the model to develop an understanding of the structural dependencies and spatial regularities inherent in medical images.

To date, retinal imaging represents the only documented application domain of this self-supervised learning technique, as it was originally developed with the specific goal of building a retinal foundation model (RETFound Green). In this context, reconstructing high-level representations of image patches containing lesions requires the model to capture not only the visual characteristics of such features but also their spatial dependencies. Lesions frequently co-occur with other anatomical structures, and their presence in one region may increase the likelihood of similar or related features appearing elsewhere in the image. Unlike pixel-level reconstruction methods such as MAE, which often produce blurred outputs due to averaging across possible pixel values, Token Reconstruction operates in an abstract, perceptual feature space. This enables the model to retain

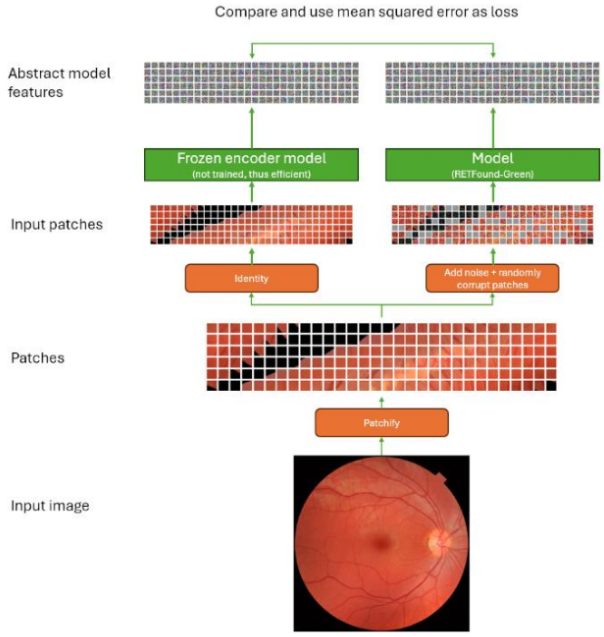

Figure 2.5: Illustration of the Token Reconstruction technique. Image reproduced from [21]

semantically meaningful signals, even when the exact pixel arrangements are ambiguous. It is hypothesized that this approach preserves critical cues, such as the presence of subtle pathological features, across multiple plausible variations. As such, reconstruction in abstract space offers a significant advantage in medical domains, where detecting small, abnormal and localized patterns is essential for the many downstream tasks of classification and diagnosis.

2.2 Related work

This section reviews prior work that informs the present study. First, it highlights recent advances in visual foundation models for retinal imaging applications. Second, it surveys previous deep learning approaches applied to FA, outlining key contributions, and identifying existing gaps in the literature. To conclude, the chapter defines the central scientific hypothesis that guides the experimental investigation presented in the subsequent chapter.

2.2.1 RETFound

RETFound is a self-supervised learning-based foundation model, based on the ViT-Large architecture, specifically developed for retinal imaging, designed to enable accurate and label-efficient adaptation to a wide range of disease detection tasks [22]. It is trained on 904,170 unlabeled CFP images leveraging the MAE strategy, which allows the model

to learn rich visual representations without requiring manual annotations. Prior to this stage, RETFound benefits from pre-training on natural images (ImageNet-1k), creating a robust initialization that enhances its learning on domain-specific retinal data. RETFound demonstrates superior generalizability and label efficiency compared to conventional supervised pre-training methods such as those based on ImageNet-21k, achieving consistently better results in the detection and prognosis of various sight-threatening eye conditions. The model excels in identifying characteristic pathological features, such as hemorrhages or hard exudates in diabetic retinopathy, by capturing structural and color-based irregularities within the retina, as clearly shown in Fig. 2.6

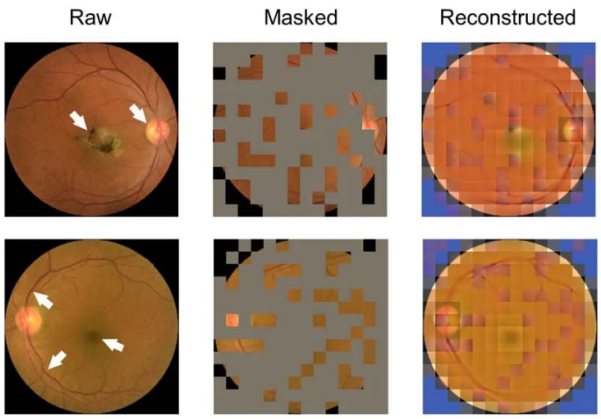

Figure 2.6: RETFound capability in reconstructing colour fundus photographs from highly masked images in pretext task. Image reproduced from [22]

Notably, RETFound has been shown to outperform both contrastive learning methods and supervised baselines on several ocular diasgnostic tasks, highlighting the strength of MAE as a self-supervised objective in the medical imaging domain.

2.2.2 RETFound Green

RETFound Green is a compact and computationally efficient retinal foundation model designed using the ViT-Small architecture and trained using the novel Token Reconstruction objective [21]. This approach offers key advantages over models like RETFound, which adopt the original MAE strategy and hyperparameters developed for general-purpose computer vision tasks, an objective that, as discussed in Subsec. 2.1.2, differs significantly from the requirements of retinal image modeling. Furthermore, while RETFound is initialized from a model pre-trained on ImageNet-1k using the standard MAE framework, RETFound-Green benefits from a more robust initialization through prior pre-training with DinoV2 on the large-scale LVD-142M natural image dataset.

This setup enabled RETFound Green to be trained on a comparatively smaller dataset of only 75,000 retinal images, and achieving 68 statistically significant wins across a wide range of downstream tasks for retinal disease diagnosis, far outperforming RETFound, which recorded only 13. Additionally, RETFound Green requires approximately 400 times less computational resources for training and results in a model that is 14 times smaller

in size. This compact design facilitates faster inference, denser embeddings, and easier storage and deployment, making it especially suitable for low-resource environments.

2.2.3 Prior DL works on FA

This subsection aims to provide a general overview of research efforts focused on automated inflammation detection in uveitis, highlighting the limited progress in this area, primarily due to the challenges associated with collecting large annotated datasets for such a rare condition.

Quantitative analysis of retinal vascular leakage

Keino et al. [23] developed a quantitative analysis tool for retinal vascular leakage, a key biomarker for monitoring disease activity in posterior uveitis (Cite). Its main contribution is the design of a deep learning system capable of segmenting vascular leakage in ultrawide field fluorescein angiography (UWF FA) by comparing early and late-phase images. The model is based on a U-Net architecture, trained for vessel segmentation using the publicly available DRIVE and STARE datasets, which consist of standard 30° CFP images. Since these datasets differ significantly in modality and field of view from UWF FA, the authors applied pre-processing techniques to convert the color fundus images into pseudo-FA grayscale images that emphasize vascular structures and simulate the contrast characteristics of FA scans. This step was crucial for improving the model's generalization to FA data, despite not having any UWF FA examples during training. In contrast, the test set comprised real UWF FA images from 82 patients with uveitis. Nevertheless, the model produced promising results, showing a strong positive correlation between the automatically detected leakage areas and clinical FA scores. Additionally, the segmented leakage areas significantly decreased after treatment, closely mirroring clinical improvements and reinforcing the model's potential as a tool for objectively monitoring treatment response in retinal vasculitis.

Another study of this kind was conducted by Young et al. [24], who trained a modified U-Net architecture to segment regions of vascular leakage in FA images. Unlike the previously analyzed work, their approach relied on a real-world dataset of 200 FA images collected from a uveitis biobank, with expert-annotated leakage areas used as ground truth. Given the limited size of the dataset, the authors employed extensive data augmentation techniques, including fixed-angle image rotations, window-based cropping, and contrast enhancement through CLAHE. The model was evaluated on an independent test set of 50 FA images, achieving a Dice Similarity Coefficient (DSC) of 0.563. Although moderate, this performance exceeded the inter-rater agreement between clinicians, which had a DSC of 0.483. These findings further support the potential of automated tools to mitigate the variability and subjectivity inherent in clinical interpretation of FA.

Automatic Transformer-based Grading of Multiple Retinal Inflammatory Signs on Fluorescein Angiography

To date, the only study that has focused on the automatic grading of retinal inflammation severity in fluorescein angiography, while accounting for multiple inflammatory signs, is

the work by Jimenez et al. [25]. A deep learning system was developed using the Swin Transformer architecture, capable of automatically grading four key inflammatory signs in the posterior pole: *vascular leakage*, *capillary leakage*, *macular edema*, and *optic disc hyperfluorescence*. The model was trained on the largest FA dataset in uveitis at present, comprising more than 40,000 images from 752 eyes of 543 patients. A Swin Transformer backbone, pre-trained on ImageNet-21k, was fine-tuned for each inflammatory sign using resized 384×384 FA images. The model, tested on 112 eyes, achieved high performance across all tasks, with an average score across all signs of $1 - OCI = 0.87 \pm 0.02$. Notably, the automatic system outperformed inter-grader agreement in detecting vascular and capillary leakage, and Grad-CAM visualizations confirmed the model's focus on clinically relevant regions.

The added value of this work lies in its demonstration of how transformer-based models have the potential to grade retinal inflammation in real-world clinical settings. By automating this task, the proposed system offers improved consistency, reliability, and scalability in diagnosing uveitis-related inflammatory signs, thanks to the Vision Transformer architecture's ability to capture both global and subtle image context.

2.3 Defining the scientific hypotheses

This thesis, based on the work of Jimenez et al. and conducted within the same research group, further develops the investigation of Transformer-based models in greater depth and extends the analysis to multiple tasks and FA datasets, including the one used for grading retinal inflammation.

Within this broader scope, the study is guided by three central scientific hypotheses:

- **Hypothesis 1: On the Effectiveness of Retinal Foundation Models**. If pre-training on a related domain such as color fundus photography provides relevant inductive biases for downstream retinal tasks, then retinal foundation models such as RETFound and RETFound Green should outperform generic vision models pretrained on natural images when fine-tuned on FA classification tasks.

- **Hypothesis 2: On the Use of LoRA for Efficient Adaptation**. If parameter-efficient fine-tuning methods such as LoRA can preserve model performance while updating only a small subset of parameters, then applying LoRA to vision models will yield performance comparable to full fine-tuning, with significantly reduced computational cost.

- **Hypothesis 3: On the Role of Self-Supervised Learning for Cross-Domain Adaptation**. If SSL strategies such as Masked Autoencoder and Token Reconstruction enhance a model's ability to learn transferable, domain-specific features, then continual pre-training using these techniques should lead to improved generalization on downstream FA classification tasks.

These hypotheses collectively frame the experimental investigation carried out in the following chapters and serve to evaluate how foundation models can be optimally adapted for clinically relevant applications in FA.

Chapter 3

Methods

This chapter describes the experimental design used to investigate the scientific hypotheses outlined previously. It first details the datasets used, including their clinical context, labeling and acquisition protocols, and associated tasks. Next, it presents the evaluation strategies used to comprehensively assess model performance, including the metrics and statistical analysis tools applied. Finally, the chapter outlines the training procedures, providing all necessary implementation details to ensure transparency and reproducibility of the experiments.

3.1 Datasets

In the following section, detailed information will be provided regarding the datasets used in this study and their corresponding tasks.

3.1.1 3rd Aptos Competition dataset

The 3rd Aptos Competition Dataset is the first publicly available angiographic dataset, comprising over 50,000 images of both fluorescein angiography and indocyanine green angiography (ICGA), all labeled by retinal specialists [26]. Originally published to accelerate the application of artificial intelligence in automatic report generation for angiography images, the dataset covers 24 medical conditions and provides detailed descriptions of the type, location, shape, size, and pattern of abnormal fluorescence. The dataset was initially composed of 58,520 angiographic images from 1,711 patients (3,405 eyes) collected at the Department of Ophthalmology, Rajavithi Hospital, Bangkok, Thailand. After excluding 226 eyes with media opacity that prevented proper fundus visualization for diagnosis, the final dataset consists of 55,361 images from 1,691 patients (3,179 eyes). The dataset maintains an even distribution between left (50%) and right (50%) eyes. Among the 3,179 eyes, 81.8% were examined in both FA and ICGA modes, 10.3% underwent FA only, and the remaining 7.9% had ICGA only. Since angiographic imaging methods are non-standardized and vary based on the specialist and patient, the number of images per eye in this dataset ranges from a few to several hundred. Specifically, each

eye has a median of 28 images with a minimum of 1 and a maximum of 286. A single ophthalmologist with more than five years of clinical experience reviewed the images for each patient. The images were labeled based on disease impressions and abnormal fluorescence patterns, including:

- Fluorescence type (hyperfluorescence or hypofluorescence)
- Fluorescence location
  - Depth direction (retinal/subretinal)
  - X-direction (macula, disc, or other)
- Size of abnormal fluorescence
- Special fluorescence patterns (e.g., polyps, ink dots)
- Vascular abnormalities due to diabetic retinopathy

Since the dataset was originally designed for a competition, the angiographic images were pre-divided into training (labeled), validation (unlabeled), and test (private for final evaluation) sets at the subject level with a 6:2:2 split. There was no patient overlap among the different splits. As such, this work only had access to only the training and validation sets.

The training set includes 33,559 images from 1,921 eyes. The most common image resolution was 768×434 (90.9%), followed by 384×409 (5.1%), 384×434 (3.2%), and 512×306 (0.8%). Notably, images where the width is twice the height correspond to cases captured in both FA and ICGA modes (91.7%). For all remaining images, the absence of labels distinguishing between FA and ICGA modes made it impossible to identify their origin. As a result, these images were excluded from further experiments, reducing the dataset to 30,768 images from 1,877 eyes. To ensure that only FA images were retained, a cropping operation was performed, Fig. 3.1, standardizing the final resolution to 384×384 (30,521 images) and 256×256 (247 images).

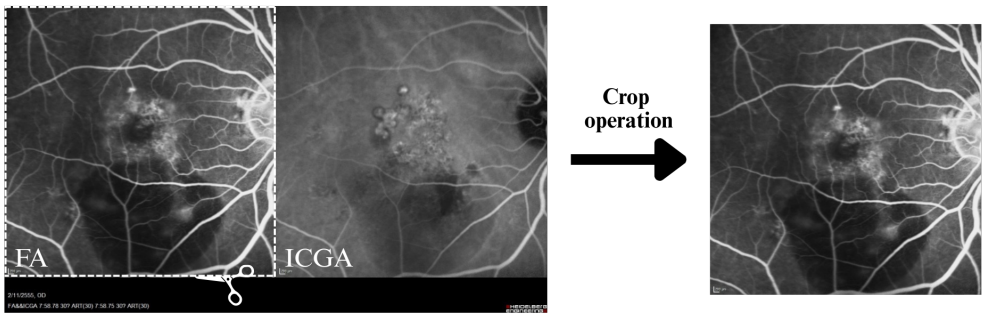

Figure 3.1: Illustration of the crop operation applied to retain only the FA portion of the dual-modality image. The white dashed box indicates the region extracted for subsequent experiments, excluding the ICGA section.

The validation set consists of 11,282 images from 630 eyes, with the following distribution of image resolutions: 768×434 (9,778 images, 86.7%), 384×409 (1,266 images, 11.2%), 384×434 (218 images, 1.9%), and 512×306 (20 images, 0.2%). The same considerations applied to the training set were also applied here, resulting in a final set of 9,608 usable validation images.

Two tasks were selected from this dataset and are described next.

Classification of hyperfluorescence type

It is a multiclass classification problem. Hyperfluorescence is an increase in fluorescence resulting from the increased transmission of normal fluorescence or an abnormal presence of fluorescein at a given time in the angiogram [27]. There are four types [28]:

- *Leakage* hyperfluorescence refers to the extravasation of fluorescein dye from leaky or incompetent blood vessels (Fig. 3.2), commonly seen in conditions such as neovascularization, retinal vasculitis, vascular malformations, tumors, and disc edema. This leakage occurs due to the rupture of the tight junctions between the vascular endothelial cells of the retina (inner blood-retinal barrier) or the breakdown of those between the epithelial cells of the retinal pigment (outer blood-retinal barrier). As the test progresses, the leakage becomes increasingly diffuse, spreading from its point of origin leading, in some cases, to late staining or pooling of dye;

Figure 3.2: Temporal progression of a leakage-labeled FA sample, shown from early to late phases (left to right). The arrow highlights the initial point of leakage, which progressively spreads, resulting in increased brightness and diffuse hyperfluorescence in the later frame.

- *Pooling* hyperfluorescence refers to the accumulation of fluorescein dye within a distinct anatomic space, such as the subretinal space or sub-RPE space, commonly seen in pigment epithelial detachments due to a breakdown of the blood-retinal barrier. As the dye fills the cavity, the hyperfluorescence gradually expands until it reaches a fixed size (Fig. 3.3);

- *Staining* hyperfluorescence is characterized by the late accumulation of fluorescein dye within certain tissues, causing a gradual increase in brightness while maintaining a constant size (Fig. 3.4). Unlike other forms of hyperfluorescence, staining remains

Figure 3.3: Temporal progression of a pooling-labeled FA sample, shown from early to late phases (left to right). The arrow highlights the accumulation of fluorescein dye within a specific area.

relatively mild and never appears intensely bright. It is a normal finding in the optic disc and is also commonly seen in drusen and fibrosis;

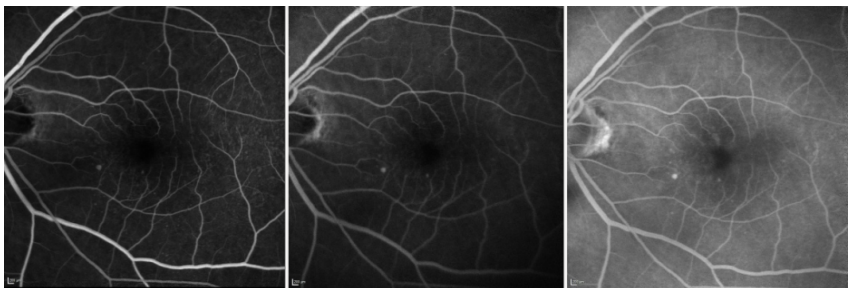

Figure 3.4: Temporal progression of a staining-labeled FA sample, showing a gradual growing in hyperfluorescence from early to late phases (left to right).

- *Window defect* hyperfluorescence occurs when there is an absence or reduction of pigmentation due to damage of the RPE, allowing increased visibility of the underlying choroidal fluorescence. This transillumination effect appears early in the angiogram, even before arterial filling, and remains static in both size and brightness throughout the test (Fig. 3.5). The degree of hyperfluorescence can vary depending on retinal pigmentation density, with greater transparency revealing more background fluorescence from the choriocapillaris.

Due to the complexity of this task, five additional binary classification tasks were derived, where each individual label serves as the positive class, while all the others are treated as negative, namely *no vs rest* (indicating the absence of any hyperfluorescence, i.e., a healthy eye), *leakage vs rest*, *pooling vs rest*, *staining vs rest*, *window defect vs rest*. This approach highlights the difficulty of each specific label, providing deeper insights into the model's challenges in understanding and accurately classifying the different types of hyperfluorescence.

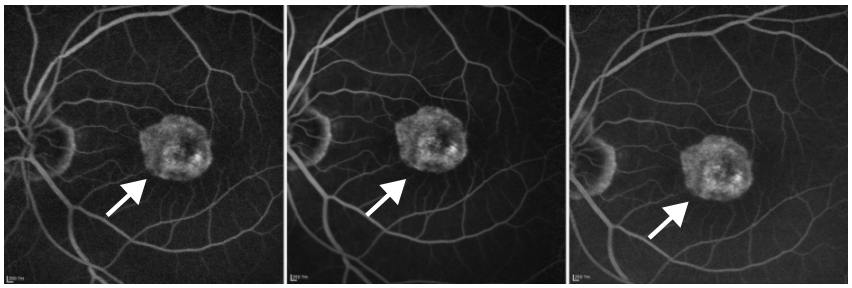

Figure 3.5: Temporal progression of a sample labeled with window defect hyperfluorescence. Its visibility remains constant from early to late phases (left to right).

Diagnosis of choroidal neovascularization

Choroidal neovascularization (CNV) is the growth of new blood vessels that originate from the choroid through a break in the Bruch membrane into the sub-RPE or subretinal space [29]. The choroid is the blood vessel-filled middle layer of the eye, which lies between the sclera and the retina. These fragile vessels can leak fluid and blood, creating a blister-like elevation in the retina, which disrupts its normally flat structure and immediately distorts vision. Over days to months, the accumulated fluid can damage the retina, leading to the loss of photoreceptor cells and resulting in progressive vision loss.

FA is a key diagnostic tool for CNV, allowing visualization of abnormal vessels in the choroid. As the fluorescein dye moves through the blood vessels of the eye, it highlights areas of CNV, helping to identify the location of leaking vessels in the retina.

This is a binary classification problem.

Label distributions

Table 3.1 presents the label distribution for the binary classification tasks derived from the 3rd Aptos competition dataset. The distribution for the *hyperfluorescence type* multiclass classification task is not shown in a separate table, as it directly corresponds to the positive class counts from each of the binary tasks (excluding CNV). This is because, as already explained, the multiclass hyperfluorescence task serves as the source from which these binary tasks were derived, with each class in the multiclass setup representing the positive class in its respective one-vs-rest binary formulation.

	no eyes (%)	leakage eyes (%)	staining eyes (%)	pooling eyes (%)	window defect eyes (%)	CNV eyes (%)
Positives	463 (24.7)	738 (39.3)	357 (19)	195 (10.4)	124 (6.6)	659 (35.1)
Negatives	1414 (75.3)	1139 (60.7)	1520 (81)	1682 (89.6)	1753 (93.4)	1218 (64.9)

Table 3.1: Label distribution of the binary classification tasks derived from the 3rd Aptos competition dataset. Values are reported as number of eyes (percentages).

3.1.2 HOJG dataset

The HOJG dataset includes 543 patients (248 male, 295 female), both with active uveitis and those with resolved or no inflammation, with an average age of 48 ± 19 years, all recruited from the Ocular Immune-Infectiology Department of Jules-Gonin Eye Hospital, Lausanne, Switzerland. These patients were treated between 2018 and 2021 and underwent FA using the Spectralis-Heidelberg device (Heidelberg Engineering, Heidelberg, Germany).

Medical experts evaluated the image quality, excluding those deemed uninterpretable or of low quality. Specifically, out of the 1,042 eyes examined, 88 were excluded due to the unavailability of late-frame images, 53 due to poor image quality, 149 due to alternative diagnoses that could interfere with angiography interpretation, such as diabetic retinopathy or vascular occlusions. The final population is composed by 752 eyes.

The FA exam followed a standardized protocol: images of the posterior pole were captured using a 55-degree field of view at approximately 1, 2, 5, and 10 minutes after the intravenous injection of fluorescein (10%, 5 ml) in both eyes. The final image resolution is 768×768 pixels. To optimize image quality, the photographer manually adjusted exposure and gain settings, as these parameters play a crucial role in enhancing the visibility of retinal structures. This adjustment was necessary due to variations in fluorescence levels across patients and different stages of angiography.

All data were extracted in DICOM format from the Spectralis-Heidelberg angiograph database. Data annotation was made through the Discovery® software by RetinAI, an Image Management System (IMS) and Image Viewer for ophthalmology, which allows the attachment of Electronic Case Report Forms (eCRF) to patient data. Finally, the manual grading of uveitis in FA was performed using the ASUWOG-validated scoring system [30] by a senior uveitis expert. Additionally, 112 eyes (the test set of split 1) were independently annotated by three additional experts from different hospitals, each with varying training backgrounds.

Classification of inflammatory signs

According to the ASUWOG scoring system, the fundus was divided into four quadrants using a horizontal and a vertical line passing through the optic disc. The posterior pole was defined as the area between the temporal vascular arcades. This dataset focuses on four key inflammatory signs:

- *Macular edema* (5 classes). It refers to the accumulation of fluid and protein deposits on or beneath the macula, the central portion of the eye responsible for sharp, detailed central vision. This buildup causes the macula to thicken and swell, a condition known as edema, which can distort vision [31]. An example of this condition is presented in Fig. 3.6

  The labels in the dataset are as follow:

  - None
  - Faint fluorescence

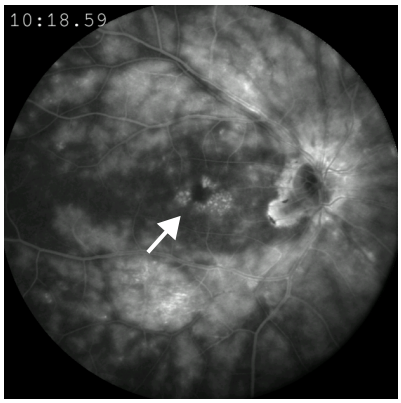

Figure 3.6: Late-phase FA showing macular edema, indicated by pooling of dye at the site marked by the arrow.

  - Incomplete ring
  - Complete ring
  - Pooling of dye

- *Optic disc hyperfluorescence* (4 classes). It denotes increased fluorescence at the optic disc during FA, suggesting inflammation or leakage at the optic nerve head. An example of this condition is presented in Fig. 3.7

Figure 3.7: Late-phase FA showing optic disc hyperfluorescence, with diffuse staining highlighted by the arrow.

The labels in the dataset are as follow:

  - None
  - Partial staining
  - Diffuse staining
  - Margin blurring

- *Retinal vascular staining and/or leakage* (4 classes). It refers to the abnormal leakage of fluorescein dye from retinal blood vessels, indicating compromised vascular integrity due to inflammation. An example of this condition is presented in Fig. 3.8

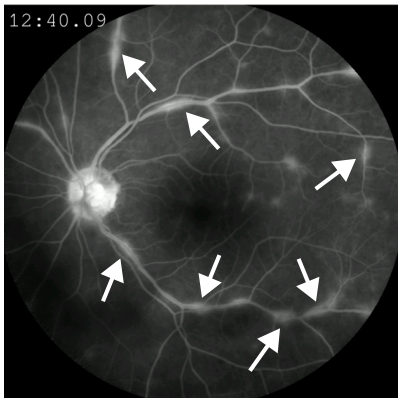

Figure 3.8: Late-phase FA showing showing multiple discrete areas of hyperfluorescence scattered along the retinal vasculature, as indicated by the arrows, suggesting a multifocal vascular leakage.

The labels in the dataset are as follow:

  - None
  - Focal
  - Multifocal
  - Diffuse

- *Capillary leakage* (3 classes). It involves the extravasation of fluorescein dye from the retinal capillaries, indicative of microvascular damage. An example of this condition is presented in Fig. 3.9

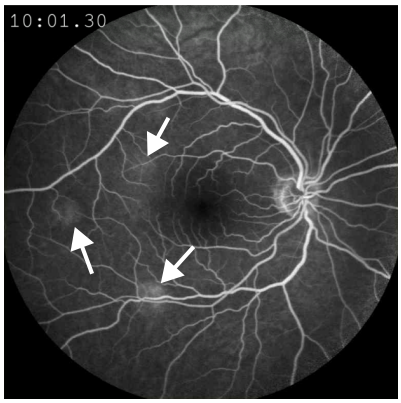

Figure 3.9: Late-phase FA showing showing localized areas of capillary leakage, indicated by the arrows.

The labels in the dataset are as follow:

- None
- Limited
- Diffuse

In the study population, the most common inflammatory sign was optic disc hyperfluorescence, observed in 305 eyes (40.5%), followed by macular edema (223 eyes, 29.6%), vascular leakage (174 eyes, 23.1%), and capillary leakage (141 eyes, 18.7%). Table 3.2 shows the exact distribution of each label. Manual grading revealed that at least one inflammatory sign was present in 49.2% of cases.

	Total distribution eyes (%)
Optic disc hyperfluorescence	
None	447 (59.5)
Partial staining	99 (13.1)
Diffuse staining	155 (20.6)
Blurring of margins	51 (6.8)
Macular edema	
None	529 (70.4)
Faint hyperfluorescence	67 (8.9)
Incomplete ring of leakage	47 (6.3)
Complete ring of hyperfluorescence	41 (5.4)
Pooling of dye into cystic spaces	68 (9.0)
Vascular leakage	
None	578 (76.8)
Focal	72 (9.6)
Multifocal	54 (7.2)
Diffuse	48 (6.4)
Capillary leakage	
None	611 (81.3)
Limited	93 (12.4)
Diffuse	48 (6.4)

Table 3.2: Distribution of inflammatory signs in the HOJG dataset. Values are reported as number of eyes (percentages).

3.2 Evaluation strategies

To comprehensively evaluate the performance of the pre-trained models used in this study, two evaluation techniques have been employed:

1. Direct probing to assess model's effectiveness as a feature extractor;

2. End-to-end fine-tuning to evaluate model's full potential and best possible performance.

3.2.1 Probing

Linear probing

Linear probing involves training a simple linear classifier, such as logistic regressor or a single linear layer, on top of the frozen feature representations extracted by the pre-trained model. The linear classifier's parameters are randomly initialized, while the backbone remains frozen, meaning its weights are not updated during training.

This approach is commonly used to assess the quality of deep feature representations. Since the classifier is relatively simple, it cannot provide additional discrimination beyond what is already present in the feature space. As a result, its performance is heavily dependent on the quality of the extracted features, specifically measuring how linearly separable they are for a given downstream task. The underlying assumption is that strong representations should be linearly separable.

MLP probing

While linear probing has been widely used in recent years, it does not capture non-linear structures in the feature space; that is one of the key strengths of deep learning. To address this limitation, MLP probing extends linear probing by replacing the linear classifier with a Multi-Layer Perceptron, which has greater complexity and can learn the task more effectively.

This method helps determine whether the information required for the downstream task is stored in a non-linear form within the model's embeddings [32]. Analyzing the performance difference between linear and MLP probing provides insights into the complexity of the learned feature space. A high performance gap between the two indicates that the model encodes information in a non-linear manner, which is not a weakness but rather a different way of structuring knowledge.

3.2.2 End-to-end fine-tuning

In end-to-end fine-tuning, all layers of the model are updated using supervised learning during training. According to previous studies [33], this approach typically yields the best performance for medical imaging tasks and allows the measurement of the maximum achievable performance for each pre-trained model.

In this study, two techniques to evaluate model fine-tuning are employed.

Full Fine-Tuning (FFT)

It is the standard approach, where all parameters of a pre-trained model are updated during training. This method allows the model to fully adapt to the target task but

requires significant computational resources and may risk overfitting, especially when working with limited labeled data.

Parameter-Efficient Fine-Tuning (PEFT)

PEFT aims to preserve most of the pre-trained model's structure by freezing the original backbone and fine-tuning only a small subset of existing parameters or a newly introduced set of layers, known as adapters. In this study, the PEFT method analyzed was LoRA, which represents weight updates as the product of two low-rank matrices, significantly reducing computational and memory requirements compared to FFT.

Fine-tuning foundation models for medical tasks presents a key challenge, as it requires striking a delicate balance between specializing the model for the downstream task and avoiding overfitting, given the limited availability of labeled samples. LoRA has the potential to address both of these challenges. Additionally, the effectiveness of PEFT in the medical domain remains underexplored, making its analysis a crucial aspect of this study's evaluation.

3.3 Evaluation metrics

In the medical field, evaluating diagnostic systems is crucial for ensuring the deployment of high-quality solutions by assessing their reliability and effectiveness, particularly given their context-sensitive nature.

In this study, the selected metrics provide a quantitative framework to assess model performance from different perspectives and compare the various strategies used to adapt foundation models.

This section presents a detailed discussion of the evaluation metrics, including their mathematical formulation and interpretation.

3.3.1 ROC and AU-ROC

The Receiver Operating Characteristic (ROC) curve is a graphical representation of a classifier's performance across all possible decision thresholds [34].

Historically, ROC analysis originated in signal detection theory [35] and was later extended to the visualization and analysis of diagnostic systems' behaviour [36]. After Spackman [37] first introduced the use of ROC curves in 1989 to evaluate and compare different algorithms, their adoption in machine learning has grown significantly. This increased use stems from the realization that simple accuracy can be misleading, particularly in imbalanced datasets, where a classifier predicting only the majority class can still achieve high accuracy.

Mathematical formulation

In a binary classification problem each sample s has to be mapped to one of two class labels: positive (p) or negative (n). A classification model (or classifier) is a function $f(s)$ that assigns a predicted class to each sample. Most classifiers produce a continuous

output, such as a probability estimate, to which different decision thresholds can be applied to determine class membership. Based on this threshold, four possible outcomes arise:

- True Positive (TP): The instance is positive and correctly classified as positive.
- False Negative (FN): The instance is positive but classified as negative.
- True Negative (TN): The instance is negative and correctly classified as negative.
- False Positive (FP): The instance is negative but classified as positive.

Given a classifier and a test set, these outcomes can be organized into a confusion matrix, which serves as the basis for many common performance metrics.

The ROC curve plots the *True Positive Rate* (TPR) (3.1), also known as recall, on the y-axis, and the *False Positive Rate* (FPR) (3.2) on the x-axis.

$$TPR = \frac{TP}{TP + FN} \tag{3.1}$$

$$FPR = \frac{FP}{FP + TN} \tag{3.2}$$

Each point on the ROC curve corresponds to a different classification threshold.

Interpretation

A well-performing model will have a ROC curve that approaches the top-left corner of the plot, indicating high TPR and low FPR. The Area Under the ROC Curve (AU-ROC or AUC for simplicity) quantifies the overall performance of the model by measuring the area beneath the ROC curve. Fig. 3.10 illustrates this concept. This metric ranges from 0 to 1, where:

- $AUC = 1.0$: Perfect classification.
- $AUC = 0.5$: No discriminative power (equivalent to random guessing).
- $AUC < 0.5$: Worse than random (indicating systematic misclassification).

Since the ROC curve evaluates performance across all possible thresholds, the AUC provides a threshold-independent measure of classification effectiveness, making it particularly useful for comparing different models without being restricted to a single decision threshold.

Multiclass ROC

ROC curves are commonly used in binary classification, where the TPR and FPR are clearly defined, as showed previously. In multiclass classification, however, these metrics can only be determined after binarizing the output. This can be achieved in two different ways:

- *One-vs-Rest* scheme (OvR) compares each class against all the others;

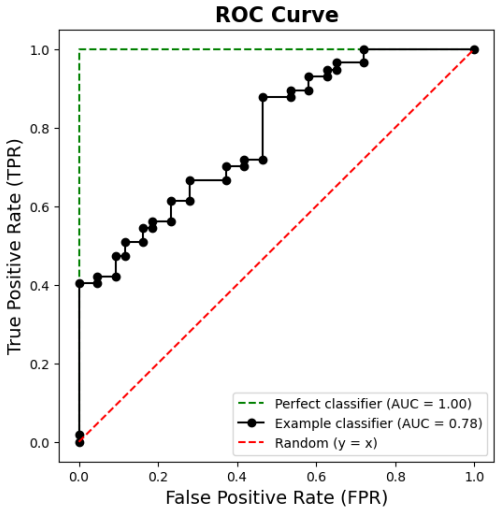

Figure 3.10: Illustrative ROC curve highlighting the behavior of a well-performing model (black) compared to an ideal model (green) and random guessing (red).

- *One-vs-One* scheme (OvO) compares every unique pairwise combination of classes.

After computing the individual AUCs with one of these strategies, there are multiple ways to derive a single overall score. The macro-average AUC is obtained by averaging the AUC scores of all classes, providing insight into the model's performance across classes regardless of their distribution. In contrast, the micro-average AUC accounts for class imbalance by weighting the contribution of each class based on its proportion in the dataset. As a result, the performance of larger classes has a greater impact on the final score compared to smaller classes.
For all experiments, evaluation suing AUC with a OvR scheme and macro-averaging has been chosen for multiclass problems [38]. This decision is based on the fact that class imbalance is a significant challenge across all selected tasks in this study. By using the OvR scheme and macro-averaging, the classifier is optimized to correctly recognize each class rather than favoring the majority class. Since smaller classes are typically harder to identify, this approach may lead to a slight trade-off in performance on larger classes, but it ensures a more balanced evaluation across all categories.

3.3.2 IMCP and AU-IMCP

Although the ROC curve presents many interesting properties, it has two significant shortcomings worth mentioning.

Firstly, it is designed for binary classification and does not inherently support multi-class datasets. As seen in the previous susbection, various aggregation strategies, such as OvR or OvO with macro or micro averaging, can be used to summarize classifier performance across multiple classes. However, this approach does not fully capture the

global performance of the classifier, and some information is inevitably lost when reducing a multi-class problem to a dichotomy.

Secondly, the ROC curve provides only a relative measure of a model's performance. It evaluates how well a classifier distinguishes between positive and negative instances but does not assess the accuracy or calibration of the predicted probabilities. In other words, a classifier that appears strong based on its AUC metric only needs to generate scores that effectively differentiate between classes, without considering the actual confidence levels of its predictions.

To address these limitations, the Multiclass Classification Performance (MCP) curve was recently introduced [39]. Unlike the ROC curve, MCP is not directly based on the confusion matrix but instead leverages the probabilities of class assignment. A key advantage of MCP is that it is independent of the number of classes in the dataset and provides richer information by extending over the entire unit square. The MCP curve distributes all instances within the range [0,1] on the X-axis. The Y-axis represents the probability-based distance between the predicted class and the true class, offering a more nuanced evaluation of classification performance. To further address the impact of class imbalance, MCP has been extended into the Imbalanced Multiclass Classification Performance (IMCP) curve [40], which adjusts for unequal class distributions, providing a more balanced assessment of model performance across all classes.

Mathematical formulation

Let $D = \{s_i \mid s = (\bar{x}, y), \bar{x} \in X^m, y \in \mathcal{Y}, i = 1, 2, \ldots, n\}$ be a dataset with n samples, where each sample belongs to an m-dimensional feature space and has a corresponding outcome y in the space $\mathcal{Y}$. In a multi-class classification problem $\mathcal{Y} = \{1, 2, \ldots, K\}$. The construction of the IMCP curve goes through five steps:

Step 1. Evaluation of the classifier on the test set.

Step 2. Computation of the *prediction probability* and *true probability* matrices. The true probability of a sample, denoted as $t(s_i)$, can be encoded as a K-dimensional one-hot vector, where all values are 0 except for a 1 at position k. Basically the true probability matrix is the binarization of the target label.
The prediction probability of a sample, denoted by $p(s_i)$, is also a K-dimensional vector, where each value represents the classifier's predicted probability of belonging to a class $k \in \mathcal{Y}$. Since it is a probability distribution, the sum of all values in the vector is 1.

Step 3. Computation of matrix $\mathcal{M}$. This matrix contains two values for each sample:

- δ_i represents the width associated with each test sample s_i depending on its class distribution (3.3)

$$\delta_i = \frac{1}{K \cdot |D_{class(s_i)}|} \tag{3.3}$$

All these values are arranged in a column vector Δ, which corresponds to the X-axis of the plot and ensures that the instances are evenly distributed along it according to their respective classes (examples from the overrepresented class will have lower values).

- φ_i quantifies the distance between the prediction and the true probability for each test sample. The authors of IMCP selected *Hellinger distance*, a metric specifically designed for probability distributions, due to its favorable properties [41]. Let $P = [p_1, \ldots, p_K]$ and $Q = [q_1, \ldots, q_K]$ be two discrete probability distributions, the Hellinger distance is defined as (3.4)

$$H(P,Q) = \frac{1}{\sqrt{2}} \left(\sum_{j=1}^{K} \left(\sqrt{p_j} - \sqrt{q_j} \right)^2 \right)^{\frac{1}{2}} \tag{3.4}$$

In our case, the computed distance is $H(p(s_i), t(s_i))$, where a value close to 0 indicates that the classifier is making a confident prediction, while a value close to 1 suggests the contrary. This value provides interesting insights in assessing classifier uncertainty, as correct predictions may still have similar probability distributions.
Column Φ, associated with the Y-axis of the plot, contains all the values φ_i:

$$\varphi_i = 1 - H(p(s_i), t(s_i)) \tag{3.5}$$

Equation (3.5) can be interpreted as the probability that the classifier correctly assigns the observed class label to the test sample s_i.

Step 4. Calculation of X-axis and Y-axis values from the Δ and Φ columns, respectively. First, the $\mathcal{M}$ matrix is sorted in ascending order based on the Φ column. Then, the X-axis values are computed using (3.6), while the Y-axis values are extracted from the newly ordered Φ' (3.7).

$$X \equiv x_i = \frac{\delta'_i}{2} + \sum_{j=1}^{i-1} \delta'_j \tag{3.6}$$

$$Y \equiv y_i = \varphi'_i \tag{3.7}$$

Step 5. Plot of the curve on the unit square and computation of Area Under IMCP (AUIMCP). The first $(0, \varphi_1)$ and last point $(0, \varphi_n)$ retain the first and last values of the Φ' column.

Interpretation

The IMCP curve offers valuable insight regarding the interpretability and explainability of predictive models:

- Correct classification threshold (η): The Hellinger distance can be expressed in terms of the symmetric Bhattacharyya coefficient (BC), where $BC(P,Q) = \sum_{j=1}^{K} \sqrt{p_j q_j}$

and $H(P,Q) = \sqrt{1 - BC(P,Q)}$. If $BC > \sqrt{0.5}$, the predicted class corresponds to the true class. As a consequence, if $\varphi > 1 - \eta = 1 - \sqrt{1 - \sqrt{0.5}} = 0.459$, the test sample will be correctly classified, although the reverse is not necessarily true.

- Incorrect classification threshold (θ): The threshold for a correct prediction in a K-class classification problem is $\theta = \sqrt{1 - \frac{1}{\sqrt{K}}}$. Therefore, $\varphi < 1 - \theta$ means that the corresponding test sample will not be correctly classified.
- Uncertainty: The classification results, whether correct or incorrect, depend on the values of φ. However, there is an uncertainty interval $[1 - \theta, 1 - \eta]$ where all class prediction probabilities remain below 0.5. In this range, the difference between the highest and second highest probabilities may be very small.

An example plot from the original paper is shown in Fig. 3.11

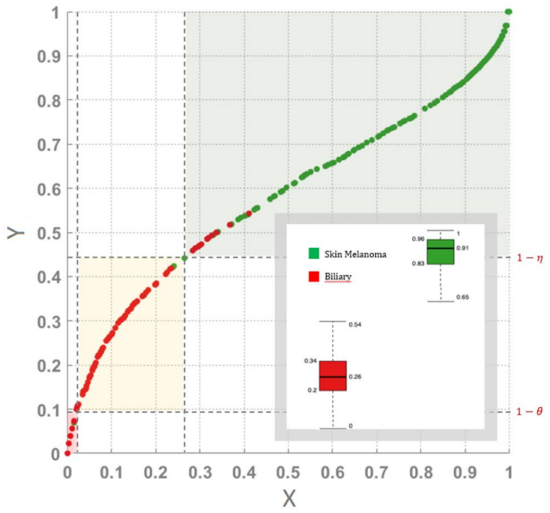

Figure 3.11: IMCP curve highlighting the different regions of confidence for two specific classes in the tumor dataset. Image reproduce from [40]

These thresholds retain a significant importance, especially in clinical settings, as they define the model's confidence landscape and provide valuable information into the certainty of its predictions for the analyzed dataset.

3.3.3 F1-score

The F1-score is a widely used performance metric in classification tasks that combines precision and recall through their harmonic mean, ensuring that both contribute equally.

Mathematical formulation

Mathematically, it is defined as (3.8)

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3.8}$$

where precision (3.9) measures the proportion of correctly predicted positive instances among all positive predictions, and recall (3.1) quantifies the proportion of actual positive instances correctly identified by the model.

$$Precision = \frac{TP}{TP + FP} \tag{3.9}$$

Interpretation

The F1-score provides a more informative evaluation of model performance by emphasizing the trade-off between these two metrics. A classifier that prioritizes precision is more strict in labeling instances as positive, which reduces recall by increasing false negatives. Conversely, a classifier that maximizes recall aims to capture as many positive cases as possible, often at the cost of misclassifying some borderline negative samples as positive, thereby reducing precision.

This trade-off makes the F1-score particularly valuable in scenarios where rare classes are of particular interest, like the one analyzed in this study.

Multiclass F1-score

In multi-class classification, the F1-score is computed using the OvR approach, where the harmonic mean between precision and recall is calculated separately for each class, treating it as positive while considering all other classes as negative. The macro-averaging technique is then used to obtain a single overall score. In this way all classes contribute equally, regardless of their frequency in the dataset, giving an idea of the general performance of the model.

3.4 Statistical analysis

In machine learning, statistical tests are often used to compare models and determine whether observed differences in their performance are statistically significant or simply due to random variation [42].

To do this, a null hypothesis H_0 is formulated, stating that there is no real difference between the models' performances. A test statistic Z, which follows a known probability distribution under H_0, is then defined, and the p-value is computed. The p-value quantifies the probability of obtaining a difference at least as extreme as the observed one under the assumption that H_0 is true. If the p-value is lower than a predefined significance level α (e.g., 0.05), the null hypothesis can be rejected.

In this context, a Type I error refers to the incorrect conclusion that one model is superior when no true difference exists, while a Type II error refers to the failure to detect an actual difference between models. By setting α, the probability of making a Type I error is controlled, ensuring that conclusions are based on robust statistical evidence rather than random variation.

3.4.1 Wilcoxon signed-rank test

Given the nature and size of the datasets analyzed, the *Wilcoxon signed-rank test* has been chosen as the appropriate statistical method. This test serves as a non-parametric alternative to the paired Student's t-test and is particularly useful when the sample size is small or when the assumption of normality in the differences between the paired observations does not hold, as is the case in this study, because it only requires it to be symmetrically distributed around a central value.

To compare Model A and Model B, one performs k-fold cross-validation, generating k performance scores for each model. The test then assesses whether the differences in these scores ($A_i - B_i$) are statistically significant, helping to determine whether one model consistently outperforms the other. The hypotheses are defined as follows.

- **Null hypothesis (H_0)**: The observations $A_i - B_i$ are symmetric about $\mu = 0$. This means no difference between the two models.

- **Alternative hypothesis (H_1)**: the distribution underlying $A_i - B_i$ is stochastically greater than a distribution symmetric about zero. This means that Model A performs significantly better than Model B.

The chosen significance level for α in the following experiments is set to 0.05, ensuring a 95% confidence level in the statistical significance of the observed results.

3.5 Training details

This section outlines the essential details and information necessary to understand the training process.

3.5.1 K-fold cross-validation

Cross-validation is a widely used validation technique in machine learning for assessing how well a model will generalize to an independent dataset [43]. It involves dividing the available data into K multiple folds, where, in each iteration, one fold is used for testing while the remaining ones are used for training. The results from each iteration are then averaged to provide a more reliable estimate of the model's performance. This approach not only helps evaluating the quality of the fitted model and the stability of its hyperparameters but also flags potential issues like overfitting.

The data were partitioned for each split into training, validation, and test sets using a 75%-15%-15% ratio. Given the class imbalance present in most tasks, a stratified random splitting approach was applied across 5 different seeds to balance the representation of each label. To prevent potential biases, both eyes from the same patient were always assigned to the same subset. Consequently, the reported results int this study represent the mean and standard deviation of the evaluation metric computed across the 5 test sets.

3.5.2 Data selection and preprocessing

Data selection

Since classification is performed per eye, regardless of the number of available images, and the analyzed models can process only one image at a time, only the *last frame* is used for analysis. This decision is based on the fact that all selected tasks contain relevant information in the late frame across both the HOJG and Aptos datasets, as already outlined in Sec. 3.1. Specifically, for the HOJG dataset, the last 55-degree angle frame captured between 5 and 15 minutes is selected to ensure temporal relevance [25]. In contrast, for the Aptos dataset, where temporality is not explicitly encoded beyond the numerical order of saved images, the last available image for each eye is selected.

Preprocessing

Only two preprocessing strategies were applied to the data. First, images were resized to match the model's required input size using the `Resize()` method with BICUBIC interpolation, following the approach recommended by the authors of the original retinal foundation models. Secondly, the grayscale images (FA) were converted into an RGB format by replicating the single channel on the other ones using the `RGB()` method [44].

3.5.3 Model's checkpoint

It is often argued in the literature that the Area Under the Precision-Recall Curve (AU-PR) is preferable to the AU-ROC for evaluating models on imbalanced datasets. However, recent research by McDermott et al. [45] challenges this claim, providing both theoretical and empirical evidence that AU-ROC remains robust to class imbalance. Their findings also highlight that selecting models based on AU-PR can introduce disparities across different subpopulations, suggesting that AU-ROC may be more suitable for domains like healthcare, where fairness and equity are critical.

Following this interpretation, the best model checkpoint used one the test set for each task was selected based on the highest AUC score achieved on the validation set. Additionally, a minor hyperparameter tuning was performed in each fine-tuning experiment to optimize performance. This tuning was carried out heuristically by manually inspecting the loss curves, without the use of any automated tools or libraries. As such, the reported evaluation is intended to reflect the best achievable performance in terms of AUC based on empirical observation.

3.5.4 The problem of class imbalance

As discussed in Sec. 3.1, class imbalance is a significant challenge affecting most tasks across both datasets. In this study, depending on the task, one of the following strategies was employed to address it.

- **Weighted Loss**: higher weights were assigned to minority classes, increasing their contribution to the loss function. There are several methods for determining class weights; in this work, they were computed using the `compute_class_weight()`

function from the scikit-learn library [46]. This method adjusts the weights proportionally to the total number of samples in each class, ensuring consistency across datasets of different sizes.

- **Data Augmentation**: This technique artificially increases the size and diversity of the training dataset by applying random transformations to existing data. It tipically offers several benefits, including improved model generalization, increased robustness to variations in the data, and a reduction in overfitting. Additionally, when combined with an oversampling technique, data augmentation can help mitigate class imbalance by reducing distribution skew.

  In recent years, powerful augmentation techniques like CutMix [47] and AugMix [48] have been developed to enhance model performance. However, in the medical imaging domain, these transformations can introduce strong variations that may alter the semantic meaning of images [49]. Unlike in general computer vision, where the focus is often on objects, medical imaging primarily relies on detecting the presence or absence of pathological abnormalities within organs or tissues. Many of these abnormalities are characterized by subtle and localized visual cues, which can become ambiguous or even obscured by aggressive augmentations, potentially compromising diagnostic accuracy. In light of these challenges, the transformations used in this study were carefully selected from a subset identified in a previous study that analyzed the most suitable augmentations for color fundus images [50]. To ensure compatibility with grayscale images, only transformations that did not rely on colored inputs were considered. As a result, the chosen augmentations include `RandomHorizontalFlip()`, `RandomVerticalFlip`, scaling (`RandomAffine(degrees=0, scale=(0.8, 1.2))`), and `RandomRotation()` between -25 and 25 degrees [51].

3.5.5 Probing architectures and hyperparameters

	Linear Probing	**MLP Probing**
training epochs	500	500
learning rate (lr)	6e-5	5e-6
optimizer	SGD (momentum=0.9)	Adam
weight decay	5e-2	5e-3
loss	weighted version	weighted version
batch size	64	32
hidden size	/	728
dropout probability	/	0.3
layer initialization	`trunc_normal()`	`kaiming_uniform()`

Table 3.3: Hyperparameters for linear and MLP probing

Linear probing was performed using a simple `torch.nn.Linear()` layer, where the input size corresponds to the embedding dimension of the model being probed, and the

output size matches the number of classes in the downstream task. MLP probing, on the other hand, was implemented using a two-layer architecture with a ReLU activation function in between and a dropout layer for regularization.

No data augmentation was applied. As is common practice when training a standard linear classifier, to normalize the input, similarly, normalizing the pre-trained features (`StandardScaler()` [52]) proved beneficial when probing.

Since the goal of probing is to analyze how models encode information, the same set of hyperparameters, showed in Table 3.3, was applied across all tasks. These hyperparameters were selected based on a series of tuning experiments conducted using the Optuna library [53] across various tasks.

Specifically, Binary Cross-Entropy was used for binary classification tasks, while Weighted Cross-Entropy was applied for multiclass problems.

Chapter 4

Experiments and results

This chapter presents and analyzes the results obtained throughout the development of the thesis. First, the initial results from the baseline models are reported to establish a performance reference. The evaluation then focuses the two retinal foundation models, *RETFound* and *RETFoundGreen*, to determine whether they provide improvements over their respective baselines. Additionally, an extra pretraining step on FA images is introduced for both models to investigate the impact of *self-supervised learning* techniques on cross-domain adaptation from color fundus to fluorescein angiography. Finally, a broader comparison of the models, incorporating these adaptation techniques, is conducted to highlight their differences and overall effectiveness.

4.1 Vision Transformer Large

The first model analyzed in this study is the ViT-Large, a Vision Transformer with 303 million parameters. It is a BERT-like transformer encoder pre-trained on the ImageNet-21K dataset in a supervised manner at a resolution of 224×224 pixels. ImageNet-21K is a large-scale dataset designed to support advanced research in image classification, containing 14,197,122 images across 21,841 classes, making it one of the most extensive datasets available for training deep learning models [54]. The dataset encompasses a diverse range of natural images, covering various subjects from animals to everyday objects, enhancing the model's ability to generalize across different classes.

4.1.1 Probing results

Table 4.1 presents the performance of representations obtained from ViT-Large under a linear probing evaluation on the 3rd Aptos Competition dataset. As already explained in Sec. 3.2, this evaluation involves training a linear classifier on top of the frozen backbone's output. The classifier consists of a fully connected layer with an output size that matches the number of classes in the downstream task. Since the backbone remains frozen, the classifier's performance primarily reflects the quality of the representations learned during the pre-training.

Table 4.2 shows the results of the linear probing for the HOJG dataset.

	CNV	no vs rest	leakage vs rest	pooling vs rest	staining vs rest	window defect vs rest	HyperF Type
AUC	0.88 ± 0.02	0.93 ± 0.02	0.84 ± 0.02	0.71 ± 0.04	0.70 ± 0.03	0.68 ± 0.05	0.78 ± 0.02
F1	0.74 ± 0.03	0.72 ± 0.02	0.72 ± 0.04	0.28 ± 0.05	0.41 ± 0.03	0.18 ± 0.02	0.43 ± 0.01
AU(IMCP)	0.65 ± 0.01	0.68 ± 0.02	0.60 ± 0.01	0.52 ± 0.02	0.51 ± 0.01	0.51 ± 0.01	0.36 ± 0.01

Table 4.1: ViT-Large linear probing evaluation on the 3rd Aptos competition dataset.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.89 ± 0.03	0.86 ± 0.04	0.85 ± 0.03	0.83 ± 0.03
F1	0.59 ± 0.03	0.48 ± 0.06	0.52 ± 0.04	0.46 ± 0.03
AU(IMCP)	0.54 ± 0.03	0.43 ± 0.03	0.45 ± 0.02	0.39 ± 0.03

Table 4.2: ViT-Large linear probing evaluation on the HOJG dataset.

To draw clear and meaningful conclusions about the quality of ViT-Large's representations, it is essential to compare the results obtained from linear probing with those from MLP probing, shown in the Tables 4.3 and 4.4.

	CNV	no vs rest	leakage vs rest	pooling vs rest	staining vs rest	window defect vs rest	HyperF Type
AUC	0.88 ± 0.02	0.93 ± 0.02	0.84 ± 0.03	0.72 ± 0.05	0.70 ± 0.03	0.64 ± 0.03	0.78 ± 0.02
F1	0.75 ± 0.03	0.78 ± 0.03	0.70 ± 0.05	0.31 ± 0.03	0.39 ± 0.04	0.13 ± 0.07	0.45 ± 0.03
AU(IMCP)	0.67 ± 0.01	0.74 ± 0.02	0.62 ± 0.01	0.52 ± 0.03	0.52 ± 0.02	0.51 ± 0.01	0.37 ± 0.01

Table 4.3: ViT-Large MLP probing evaluation on the 3rd Aptos competition dataset.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.92 ± 0.03	0.90 ± 0.03	0.88 ± 0.02	0.87 ± 0.03
F1	0.72 ± 0.04	0.60 ± 0.09	0.56 ± 0.07	0.52 ± 0.05
AU(IMCP)	0.63 ± 0.06	0.54 ± 0.01	0.48 ± 0.03	0.44 ± 0.01

Table 4.4: ViT-Large MLP probing evaluation on the HOJG dataset.

The first noticeable trend in the probing results is that some tasks are more challenging than others. In particular, the classification of *pooling*, *staining*, and *window defect* appears to be the most difficult, as indicated by lower AUC and F1 scores. This is likely due to the low representation of these classes in the dataset, with their respective proportions being 10.4%, 18.9%, and 6.7%.

A comparison between linear and MLP probing results for the Aptos dataset reveals that AUC remains unchanged. The fact that an MLP, despite being more expressive than a simple linear layer, does not significantly improve AUC suggests that the features extracted by the ViT-Large backbone are already highly informative for distinguishing between classes. The slight improvement in F1 scores across all tasks in MLP probing,

except for *window defect vs rest*, is due to refined predictions, leading to a better balance between precision and recall. This is further supported by the higher AU(IMCP) values, indicating improved separability of the decision boundaries. The different behavior observed in the *window defect vs rest* task is likely due to the extremely small number of positive samples in the dataset (only 128 out of 1877 eyes). With such limited data, the model struggles to learn a generalizable decision boundary, leading to overfitting, even in a relatively simple classifier like the MLP.

Unlike the observations on the Aptos dataset, MLP probing shows a consistent improvement across all metrics and tasks in the HOJG dataset. This suggests that the representation on these data contains nonlinear structures that require a more complex classifier to be fully leveraged.

4.1.2 FFT results

This training technique fully adapts the model to the downstream task. The goal of full fine-tuning is to determine whether the pre-trained weights of ViT-Large, obtained through supervised training on ImageNet-21K, serve as a good initializer for FA-related tasks.

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**
training epochs	50	50	50	50
learning rate (lr)	5e-6	2e-6	5e-6	3e-6
scheduler (min lr)	None	None	None	None
loss	weighted BinaryCE	weighted BinaryCE	weighted BinaryCE	weighted BinaryCE
augmentations	Yes	Yes	Yes	Yes
oversampling	No	No	No	No

	staining vs rest	**window defect vs rest**	**HyperF Type**
training epochs	50	30	30
learning rate (lr)	2e-6	8e-6	1e-5
scheduler (min lr)	None	WarmupCosine (5e-8)	WarmupCosine (5e-8)
loss	weighted BinaryCE	BinaryCE	weighted CrossEntropy
augmentations	Yes	Yes	Yes
oversampling	No	Yes	No

Table 4.5: FFT ViT-Large hyperparameters for the 3rd Aptos competition's tasks.

Given the diverse nature and varying difficulty levels of each task, a task-specific set

of hyperparameters was chosen, each heuristically tuned to achieve the best possible performance, as anticipated in Subsec. 3.5.3. Instead of using a single fixed configuration across all tasks, this approach allows for better adaptation to the unique challenges presented by each classification problem. The Tables 4.5 and 4.6 provide an overview of the selected hyperparameter configurations for each task, corresponding to the HOJG and aptos datasets, respectively. The *AdamW* optimizer, with a *weight decay* of 0.05, a *batch size* of 32, and a *drop path rate* of 0.1, were kept the same for all tasks and are therefore omitted from the tables for brevity.

	capillary leakage	**vascular leakage**	**optic disk hyperfluorescence**	**macular edema**
training epochs	30	30	30	30
learning rate (lr)	9e-7	1e-6	3e-6	1.5e-6
scheduler (min lr)	WarmupCosine (5e-9)	None	WarmupCosine (5e-8)	None
loss	weighted CrossEntropy	CrossEntropy	CrossEntropy	CrossEntropy
augmentations	Yes	Yes	Yes	Yes
oversampling	No	Yes	Yes	Yes

Table 4.6: FFT ViT-Large hyperparameters for the HOJG's tasks.

The results presented in Tables 4.7 and 4.8 confirm the varying levels of difficulty among the analyzed tasks. Specifically, when refining the entire model's weights, the performance improvements over MLP probing appear to follow a consistent trend: the more complex tasks benefit the most from full fine-tuning. For instance, in the Aptos dataset, tasks such as *pooling* show a notable increase (+0.05 in AUC and +0.07 in F1), along with *staining* (+0.05 in AUC and +0.07 in F1) and *window defect* (+0.13 in AUC and +0.12 in F1).

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**	**staining vs rest**	**window defect vs rest**	**HyperF Type**
AUC	0.90 ± 0.01	0.95 ± 0.01	0.87 ± 0.03	0.77 ± 0.05	0.74 ± 0.03	0.77 ± 0.05	0.81 ± 0.02
F1	0.77 ± 0.02	0.82 ± 0.03	0.74 ± 0.04	0.38 ± 0.05	0.43 ± 0.03	0.25 ± 0.04	0.46 ± 0.04
AU(IMCP)	0.67 ± 0.01	0.65 ± 0.03	0.60 ± 0.04	0.54 ± 0.05	0.51 ± 0.02	0.54 ± 0.02	0.38 ± 0.02

Table 4.7: Full fine-tuning of ViT-Large on the 3rd Aptos competition dataset.

	capillary leakage	**vascular leakage**	**optic disk hyperfluorescence**	**macular edema**
AUC	0.92 ± 0.02	0.91 ± 0.02	0.91 ± 0.01	0.94 ± 0.01
F1	0.69 ± 0.08	0.63 ± 0.03	0.59 ± 0.06	0.59 ± 0.04
AU(IMCP)	0.59 ± 0.02	0.56 ± 0.01	0.54 ± 0.02	0.52 ± 0.03

Table 4.8: Full fine-tuning of ViT-Large on the HOJG dataset.

Similarly, in the HOJG dataset, the most substantial improvements are observed

in the scoring of *optic disk hyperfluorescence* (+0.03 in AUC and +0.03 in F1) and *macular edema* (+0.07 in AUC and +0.07 in F1), which were the tasks reaching the lower performances between the four.

4.1.3 PEFT results with LoRA

Although full fine-tuning provides a significant performance boost, it comes with several challenges that must be carefully considered. First, its efficiency and computational demand are heavily influenced by the model size: larger models require more computational resources and data, making the process potentially costly and inefficient. Additionally, a meticulous hyperparameter tuning is essential, as suboptimal settings can lead to poor generalization or unstable training dynamics. In fact, these models are inherently designed to leverage large datasets, and when applied to low-data regimes, such as the one analyzed in this study, they become highly prone to overfitting.

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**
lr	1e-4	1e-5	1e-5	8e-5
weight decay	1e-4	5e-2	5e-2	2e-4
scheduler (min lr)	WarmupCosine (5e-9)	None	None	WarmupCosine (5e-9)
batch size	64	32	32	16
r (train. params)	32 (1%)	32 (1%)	64 (2%)	64 (2%)
dropout	0.1	0.1	0.1	0.1

	staining vs rest	**window defect vs rest**	**HyperF Type**
lr	6e-5	2e-6	3e-5
weight decay	5e-1	5e-2	1e-4
scheduler (min lr)	WarmupCosine (5e-6)	None	None
batch size	64	32	32
r (train. params)	64 (2%)	64 (2%)	64 (2%)
dropout	0.1	None	0.15

Table 4.9: LoRA hyperparameters for the ViT-Large on the 3rd Aptos competition's tasks. The percentage in brackets next to the rank indicates the proportion of trainable parameters relative to the full model.

To address these challenges, adapters offer an effective solution through parameter-efficient fine-tuning. Adapters are lightweight architectures inserted at specific points within the large pre-trained model, allowing for task-specific adaptation without modifying the original model's weights. During fine-tuning, the pre-trained model remains frozen, and only the adapter parameters are updated. This approach has several key advantages: it significantly reduces computational and storage costs, as only a small

subset of parameters is trained, and it makes the process faster and less keen to overfit. Additionally, since the core model weights remain unchanged, adapters help preserve previously learned knowledge, mitigating the risk of catastrophic forgetting while enabling efficient adaptation to new tasks.

The adapter employed in this study is LoRA, Low-Rank Adaptation, which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each attention layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks.

In this study, different configurations were explored to optimize performance across tasks. Tables 4.9 and 4.10 provide a detailed overview of the LoRA hyperparameters used, including rank (r), scaling factor (α), and dropout rate. For each task, the model was trained for 20 epochs with the $AdamW$ optimizer.

	capillary leakage	**vascular leakage**	**optic disk hyperfluorescence**	**macular edema**
lr	3e-6	3e-6	2e-6	2e-6
weight decay	5e-2	5e-4	5e-2	5e-2
scheduler (min lr)	WarmupCosine (5e-8)	None	None	None
batch size	32	32	32	32
r (train. params)	128 (4%)	128 (4%)	128 (4%)	128 (4%)
dropout	0.1	0.1	0.1	0.1

Table 4.10: LoRA hyperparameters for the ViT-Large on the HOJG's tasks. The percentage in brackets next to the rank indicates the proportion of trainable parameters relative to the full model.

The rank (r) determines the dimensionality of the low-rank decomposition used to approximate weight updates during fine-tuning. It controls the adapter's flexibility in modifying the pre-trained model's parameters while keeping the number of trainable parameters low. Typically, low-rank values (from 4 to 32) are preferred, not only because they significantly reduce the number of trainable parameters but also because higher ranks tend to saturate performance improvements. This limitation was later addressed through rank-stabilized initialization ($rslora$) [55], which was consistently applied in all the following experiments.

The scaling factor (α) regulates the influence of fine-tuning updates on the model. Empirical studies have shown that setting $\alpha = 2r$ improves results, particularly for higher rank values, and is also theoretically well-motivated [56]. For this reason, this parameterization is adopted across all tasks.

Finally, dropout is applied within the LoRA adapter layer to prevent overfitting, especially in low-data scenarios.

The results presented in the tables, where the relative change in performance compared to FFT is shown in brackets, suggest that the discriminative power of the model remains unchanged despite the significantly lower number of trainable parameters. Across tasks, AUC and F1 remain largely stable, with slight variations. Notably, *capillary* and *vascular*

leakage show statistically significant improvements. Conversely, the scoring of *macular edema* experiences a decline in AUC and AU(IMCP), making the model trained through LoRA statistically worse for that condition. However, when analyzing AU(IMCP), LoRA appears to provide a significant boost, resulting in a model that is overall more confident and reliable across almost all tasks. In particular, this improvement is statistically significant in three (*no vs rest*, *staining vs rest*, *HyperF Type*) out of the eleven tasks analyzed.

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**	**staining vs rest**	**window defect vs rest**	**HyperF Type**
AUC	0.90 ± 0.01 (−)	0.96 ± 0.01 (+0.01)	0.87 ± 0.03 (−)	0.75 ± 0.07 (-0.02)	0.73 ± 0.03 (-0.01)	0.78 ± 0.02 (+0.01)	0.82 ± 0.02 (+0.01)
F1	0.77 ± 0.02 (−)	0.82 ± 0.02 (−)	0.74 ± 0.05 (−)	0.36 ± 0.06 (-0.02)	0.43 ± 0.06 (−)	0.21 ± 0.05 (-0.04)	0.46 ± 0.05 (−)
AU(IMCP)	0.69 ± 0.04 (+0.02)	0.73 ± 0.02 (+0.08)	0.64 ± 0.01 (+0.04)	0.55 ± 0.05 (+0.01)	0.55 ± 0.03 (+0.04)	0.53 ± 0.02 (-0.01)	0.40 ± 0.02 (+0.02)

Table 4.11: PEFT (LoRA) of ViT-Large on the 3rd Aptos competition dataset. Between brackets, the relative change in performance compared to FFT. If the number in brackets is underlined, the change is statistically significant.

	capillary leakage	**vascular leakage**	**optic disk hyperfluorescence**	**macular edema**
AUC	0.93 ± 0.01 (+0.01)	0.93 ± 0.03 (+0.02)	0.91 ± 0.01 (−)	0.92 ± 0.01 (-0.02)
F1	0.70 ± 0.08 (+0.01)	0.66 ± 0.08 (+0.03)	0.57 ± 0.03 (-0.02)	0.57 ± 0.05 (-0.01)
AU(IMCP)	0.64 ± 0.03 (+0.05)	0.59 ± 0.04 (+0.03)	0.53 ± 0.02 (-0.01)	0.48 ± 0.02 (-0.04)

Table 4.12: PEFT (LoRA) of ViT-Large on the HOJG dataset. Between brackets, the relative change in performance compared to FFT. If the number in brackets is underlined, the change is statistically significant.

4.2 Vision Transformer Small

The Vision Transformer (ViT)-Small is a compact variant (22.1 milion parameters) of the Transformer architecture analyzed in the previous chapter. In this study, ViT-Small serves as the baseline model for RETFound Green. It has been pre-trained on the LVD-142M dataset, which consists of 142 million natural images curated from various datasets. The pretraining was conducted using the self-supervised learning method DINOv2 [57]. This technique has been shown to produce high-performance visual features that can be directly utilized with simple classifiers across various computer vision tasks, even without fine-tuning. Its success stems from the refinement of existing self-supervised learning

approaches, combining multiple strategies to scale pre-training efficiently in terms of data volume and model size. In particular, the ViT-Small was obtained through a self-distillation approach to transfer knowledge from the larger model ViT-Giant (ViT-G), enabling the smaller model to inherit the performance advantages of the larger one, resulting in efficient training and deployment without compromising accuracy.

To ensure a fair comparison between models, the same input size as ViT-Large, and consequently RETFound, was used, specifically 224×224 pixels.

4.2.1 Probing results

Despite the proven effectiveness of features extracted from DINOv2-pretrained models in linear probing evaluations, the results in Tables 4.13 and 4.14 indicate an overall lower performance compared to ViT-Large. For instance, in simpler tasks such as *no vs other* and *CNV* diagnosis, performance remains comparable between the two models. However, for most other tasks, ViT-Small consistently lags behind, with performance differences of approximately 0.02 to 0.03 in AUC across nearly all cases.

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**	**staining vs rest**	**window defect vs rest**	**HyperF Type**
AUC	0.87 ± 0.01	0.93 ± 0.02	0.82 ± 0.02	0.67 ± 0.03	0.67 ± 0.03	0.70 ± 0.02	0.76 ± 0.02
F1	0.73 ± 0.02	0.73 ± 0.03	0.72 ± 0.01	0.25 ± 0.04	0.37 ± 0.02	0.21 ± 0.02	0.41 ± 0.02
AU(IMCP)	0.63 ± 0.00	0.67 ± 0.01	0.58 ± 0.01	0.50 ± 0.02	0.50 ± 0.02	0.51 ± 0.02	0.34 ± 0.01

Table 4.13: ViT-Small linear probing evaluation on the 3rd Aptos competition dataset.

	capillary leakage	**vascular leakage**	**optic disk hyperfluorescence**	**macular edema**
AUC	0.91 ± 0.02	0.84 ± 0.04	0.84 ± 0.01	0.80 ± 0.03
F1	0.57 ± 0.08	0.42 ± 0.02	0.48 ± 0.01	0.42 ± 0.03
AU(IMCP)	0.52 ± 0.09	0.41 ± 0.04	0.42 ± 0.03	0.37 ± 0.01

Table 4.14: ViT-Small linear probing evaluation on the HOJG dataset.

Once again, by comparing the results of linear probing and MLP probing, as shown in Tables 4.15 and 4.16, it is possible to confirm the same trend observed with ViT-Large. Specifically, the only tasks that show improvement are those from the HOJG dataset, suggesting that these tasks exhibit more complex patterns in the latent space. This complexity likely requires the feature interactions enabled by the greater depth of the MLP, even though it consists of only a single layer.

4.2.2 FFT results

Before analyzing the results, the Tables 4.17 and 4.18 provide an overview of the hyper-parameters used for each task. As in previous experiments, the *AdamW* optimizer with a *weight decay* of 5e-2 and a *drop path rate* of 0.1 (to mitigate overfitting) was used consistently across all tasks, except for *capillary leakage*, where a lower weight decay (5e-3)

	CNV	no vs rest	leakage vs rest	pooling vs rest	staining vs rest	window defect vs rest	HyperF Type
AUC	0.87 ± 0.02	0.93 ± 0.02	0.83 ± 0.02	0.67 ± 0.04	0.69 ± 0.03	0.69 ± 0.04	0.77 ± 0.02
F1	0.74 ± 0.03	0.77 ± 0.03	0.72 ± 0.03	0.26 ± 0.04	0.40 ± 0.04	0.19 ± 0.05	0.44 ± 0.03
AU(IMCP)	0.65 ± 0.02	0.74 ± 0.02	0.60 ± 0.02	0.51 ± 0.03	0.52 ± 0.01	0.52 ± 0.02	0.36 ± 0.01

Table 4.15: ViT-Small MLP probing evaluation on the 3rd Aptos competition dataset.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.94 ± 0.01	0.87 ± 0.03	0.87 ± 0.02	0.85 ± 0.03
F1	0.68 ± 0.03	0.54 ± 0.07	0.53 ± 0.05	0.46 ± 0.05
AU(IMCP)	0.62 ± 0.04	0.48 ± 0.03	0.48 ± 0.03	0.41 ± 0.02

Table 4.16: ViT-Small MLP probing evaluation on the HOJG dataset.

proved to be more effective.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
training epochs	30	30	30	30
lr	2e-6	7e-6	5e-6	5e-6
scheduler (min lr)	None	None	warmupCosine (5e-8)	None
batch size	16	64	32	64
loss	weighted CrossEntropy	CrossEntropy	CrossEntropy	CrossEntropy
augmentations	Yes	Yes	Yes	Yes
oversampling	No	Yes	Yes	Yes

Table 4.17: FFT ViT-Small hyperparameters for the HOJG's tasks.

A noteworthy aspect of ViT-Small is that, thanks to its smaller size, it was significantly faster to train, making its study more manageable. In contrast, training ViT-Large took approximately 1 minute 15 seconds to 1 minute 45 seconds per epoch, whereas ViT-Small completed an entire epoch in just 10-30 seconds. These estimates, while purely qualitative and specific to the dataset sizes used in this study, are important to highlight, especially when considering computational resources, which are often limited in clinical settings.

The results of FFT on ViT-Small, as shown in Tables 4.19 and 4.20, reveal a substantial improvement compared to the less promising performance observed in the probing evaluations, bringing it on par with ViT-Large. In some cases, it even demonstrates a slightly higher discriminative power, as reflected in the AUC scores, particularly in tasks related to the classification of inflammatory signs (+0.03 for *capillary leakage*, +0.02 for *vascular leakage* and +0.01 for *macular edema*). Interestingly, the only task in this category (*optic disk hyperfluorescence*) that does not show improvement in AUC still exhibits a lower standard deviation and a higher F1 score with respect to the results obtained on the ViT-Large, suggesting greater precision in the model's predictions.

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**
training epochs	50	50	50	50
lr	2e-6	1e-5	5e-6	4e-6
scheduler (min lr)	None	WarmupCosine (5e-8)	WarmupCosine (5e-8)	None
batch size	128	128	128	128
loss	weighted BinaryCE	weighted BinaryCE	weighted BinaryCE	weighted BinaryCE
augmentations	Yes	Yes	Yes	Yes
oversampling	No	No	No	No

	staining vs rest	**window defect vs rest**	**Hyperf Type**
training epochs	50	30	50
lr	5e-6	5e-6	5e-6
scheduler (min lr)	WarmupCosine (5e-8)	None	None
batch size	64	32	128
loss	weighted BinaryCE	BinaryCE	weighted CrossEntropy
augmentations	Yes	Yes	Yes
oversampling	No	Yes	No

Table 4.18: FFT ViT-Small hyperparameters for the 3rd Aptos competition's tasks.

An even more remarkable finding is the consistent advantage of ViT-Small in AU(IMCP) across all tasks. This indicates that ViT-Small is more reliable and confident in its predictions compared to its larger counterpart, further reinforcing its effectiveness despite its smaller size.

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**	**staining vs rest**	**window defect vs rest**	**HyperF Type**
AUC	0.89 ± 0.02	0.95 ± 0.02	0.87 ± 0.02	0.76 ± 0.04	0.74 ± 0.03	0.80 ± 0.03	0.83 ± 0.02
F1	0.75 ± 0.02	0.83 ± 0.02	0.75 ± 0.04	0.37 ± 0.05	0.41 ± 0.04	0.31 ± 0.08	0.50 ± 0.02
AU(IMCP)	0.71 ± 0.01	0.79 ± 0.01	0.66 ± 0.02	0.57 ± 0.02	0.56 ± 0.02	0.61 ± 0.02	0.42 ± 0.01

Table 4.19: Full fine-tuning of ViT-Small on the 3rd Aptos competition dataset.

4.2.3 PEFT results with LoRA

Although LoRA was originally designed for large models, its application to smaller models provides the same key advantages for which it was developed. This subsection examines the response of the ViT-Small to PEFT using LoRA. The hyperparameters used for these experiments are presented in Tables 4.21 and 4.22. The considerations and decisions

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.95 ± 0.01	0.93 ± 0.02	0.91 ± 0.00	0.95 ± 0.01
F1	0.70 ± 0.06	0.63 ± 0.01	0.61 ± 0.03	0.64 ± 0.05
AU(IMCP)	0.65 ± 0.03	0.59 ± 0.04	0.53 ± 0.01	0.54 ± 0.05

Table 4.20: Full fine-tuning of ViT-Small on the HOJG dataset.

regarding rank stabilization and the choice of α for ViT-Large also apply to its smaller variant.

	CNV	no vs rest	leakage vs rest	pooling vs rest
lr	9e-6	2e-5	4e-5	1e-5
weight decay	5e-2	1e-2	5e-2	1e-2
scheduler (min lr)	None	WarmupCosine (5e-8)	None	None
batch size	128	64	128	64
r (train. params)	32 (2.67%)	32 (2.67%)	32 (2.67%)	64 (5.18%)
dropout	0.1	0.1	0.1	0.1
drop path	-	-	0.1	0.1

	staining vs rest	window defect vs rest	HyperF Type
lr	5e-5	1e-5	8e-6
weight decay	5e-2	5e-2	1e-2
scheduler (min lr)	None	None	None
batch size	128	64	128
r (train. params)	64 (5.18%)	64 (5.18%)	64 (5.18%)
dropout	0.1	0.1	0.1
drop path	0.1	-	-

Table 4.21: LoRA hyperparameters for the ViT-Small on the 3rd Aptos competition's tasks. The percentage in brackets next to the rank indicates the proportion of trainable parameters relative to the full model.

Once again, LoRA proves its effectiveness, as the results presented in the Tables 4.23 and 4.24 show that the model's performance remains comparable to FFT.

In fact, there is only one statistically significant decrease in AUC, which affects the *optic disk hyperfluorescence* task. However, unlike what was observed for ViT-Large, there is a consistent decline in model confidence, as indicated by the lower AU(IMCP) values across almost all tasks, with four of them being statistically significant. This could suggest that smaller models are more vulnerable to LoRA, leading to less certain predictions compared to their larger counterparts.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
lr	9e-7	1e-5	2e-6	1e-5
weight decay	5e-3	5e-2	5e-2	5e-2
scheduler (min lr)	None	WarmupCosine (5e-8)	None	None
batch size	16	64	64	64
r (train. params)	128 (9.8%)	128 (9.8%)	128 (9.8%)	128 (9.8%)
dropout	0.1	0.1	0.1	0.1
drop path	-	0.1	-	-

Table 4.22: LoRA hyperparameters for the ViT-Small on the HOJG's tasks. The percentage in brackets next to the rank indicates the proportion of trainable parameters relative to the full model.

	CNV	no vs rest	leakage vs rest	pooling vs rest	staining vs rest	window defect vs rest	HyperF Type
AUC	0.89 ± 0.02 (−)	0.95 ± 0.01 (−)	0.87 ± 0.02 (−)	0.76 ± 0.03 (−)	0.75 ± 0.02 (+0.01)	0.80 ± 0.04 (−)	0.82 ± 0.02 (-0.01)
F1	0.76 ± 0.02 (+0.01)	0.82 ± 0.03 (-0.01)	0.75 ± 0.03 (−)	0.33 ± 0.03 (-0.04)	0.47 ± 0.03 (+0.06)	0.23 ± 0.08 (-0.08)	0.49 ± 0.02 (-0.01)
AU(IMCP)	0.67 ± 0.01 (-0.04)	0.76 ± 0.01 (-0.03)	0.66 ± 0.02 (−)	0.56 ± 0.01 (-0.01)	0.57 ± 0.02 (+0.01)	0.57 ± 0.03 (-0.04)	0.40 ± 0.01 (-0.02)

Table 4.23: PEFT (LoRA) of ViT-Small on the 3rd Aptos competition dataset. Between brackets, the relative change in performance compared to FFT. If the number in brackets is underlined, the change is statistically significant.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.94 ± 0.02 (-0.01)	0.93 ± 0.02 (−)	0.90 ± 0.00 (-0.01)	0.94 ± 0.02 (-0.01)
F1	0.69 ± 0.05 (-0.01)	0.63 ± 0.08 (−)	0.60 ± 0.02 (-0.01)	0.62 ± 0.06 (-0.02)
AU(IMCP)	0.62 ± 0.03 (-0.03)	0.53 ± 0.04 (-0.06)	0.49 ± 0.02 (-0.04)	0.54 ± 0.05 (−)

Table 4.24: PEFT (LoRA) of ViT-Small on the HOJG dataset. Between brackets, the relative change in performance compared to FFT. If the number in brackets is underlined, the change is statistically significant.

4.3 RETFound

RETFound is the first foundation model for retinal images analyzed in this study. It is trained in a self-supervised manner on approximately 1 million unlabeled retinal images, specifically color fundus photography (CFP). The choice of CFP is driven by the nature

of self-supervised learning, which relies on large amounts of unlabeled data to develop general-purpose feature representations. Since CFP is one of the most widely used imaging techniques in ophthalmology, these retinal images are routinely collected in clinical practice, making them an ideal dataset for large-scale self-supervised pretraining. RETFound was developed using an advanced SSL approach, specifically a masked autoencoder (MAE), applied in two sequential stages:

1. SSL pre-training on natural images from ImageNet-1k, ensuring broad feature extraction capabilities.

2. Domain-specific pre-training on retinal images, using data from the private Moorfields Diabetic Image Dataset (MEH-MIDAS) and publicly available datasets, totaling 904,170 CFPs.

The schematics illustrating this process are shown in Fig. 4.1, where the previously analyzed baseline of the model is included for clarity.

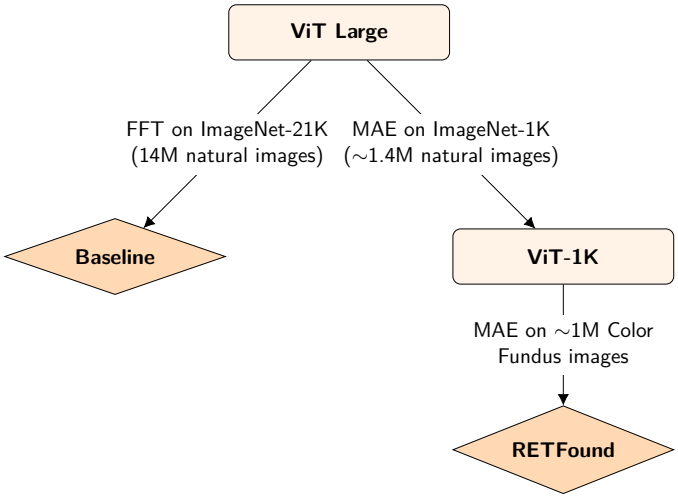

Figure 4.1: Schematic tree illustrating the steps leading to RETFound and its baseline. Each step includes the training method used, the dataset on which it was performed, and its corresponding size.

RETFound has demonstrated consistently superior performance and label efficiency in the diagnosis and prognosis of sight-threatening eye diseases on CFP images, outperforming state-of-the-art competing models, including its baseline model (named ViT-Large in this study), which was pre-trained on ImageNet-21k using traditional transfer learning. This advantage arises from the MAE pretext task, which enabled the model to learn retina-specific contextual features, including anatomical structures and disease lesions. As visual evidence of this capability, Fig. 4.2 illustrates how RETFound successfully reconstructs major anatomical structures, such as the optic nerve and large vessels on FA, even though it was pretrained on CFP. Remarkably, it achieves this despite 75% of the retinal image being masked.

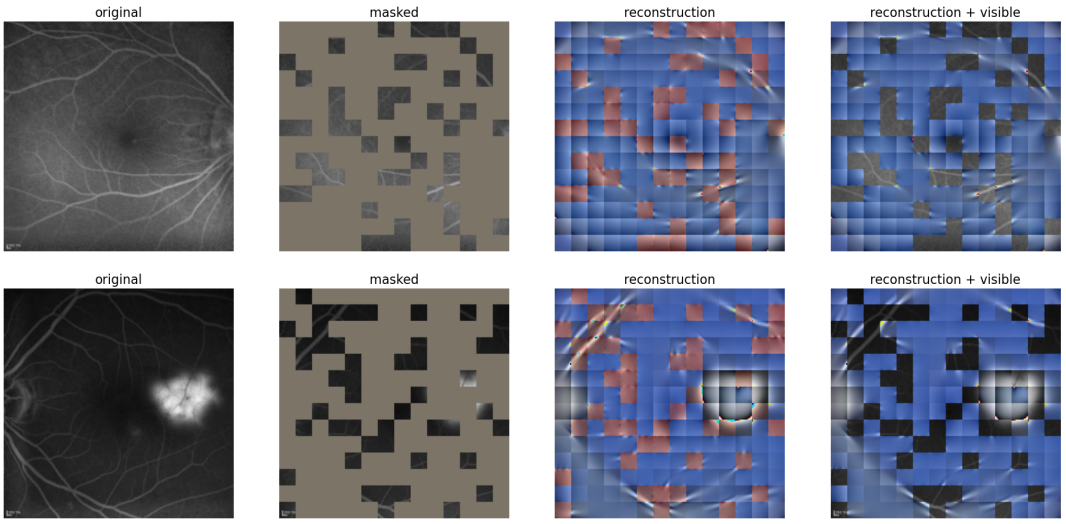

Figure 4.2: Reconstructed fluorescein angiography from highly masked images in pretext task using original RETFound weights.

4.3.1 Probing results

The probing results of RETFound's extracted features are shown in Tables 4.25 and 4.26 for the linear evaluation and in Tables 4.27 and 4.28 for the MLP.

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**	**staining vs rest**	**window defect vs rest**	**HyperF Type**
AUC	0.86 ± 0.02	0.91 ± 0.02	0.81 ± 0.02	0.68 ± 0.04	0.63 ± 0.03	0.63 ± 0.07	0.73 ± 0.02
F1	0.72 ± 0.02	0.71 ± 0.03	0.69 ± 0.03	0.26 ± 0.03	0.35 ± 0.02	0.15 ± 0.02	0.38 ± 0.03
AU(IMCP)	0.62 ± 0.01	0.66 ± 0.01	0.56 ± 0.02	0.51 ± 0.01	0.49 ± 0.01	0.49 ± 0.02	0.33 ± 0.01

Table 4.25: RETFound linear probing evaluation on the 3rd Aptos competition dataset.

If pre-training on a more related domain, such as CFP, provided a real advantage over natural images, the extracted features in RETFound would be expected to demonstrate greater discriminative power already at the linear probing stage. However, the results do not seem to support this hypothesis. The linear evaluation suggests that the features extracted, presumably capturing retinal structural information, do not appear to be particularly informative for FA-related tasks. This is an unexpected finding, and a more detailed comparison of all models will be conducted later in the chapter to further investigate these observations.

As observed in ViT-Large, only a small subset of tasks benefit from the non-linear interactions introduced by the MLP. Specifically, improvements are seen in *window defect* recognition and *optic disk hyperfluorescence* (+0.03 AUC for both), as well as in *macular edema* classification (+0.04 AUC). However, the overall improvement is limited and less significant compared to what was observed in ViT-Large. This is likely due to the lower quality of the extracted features, which reduces the effectiveness of the MLP.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.90 ± 0.01	0.81 ± 0.04	0.80 ± 0.03	0.82 ± 0.03
F1	0.57 ± 0.03	0.45 ± 0.06	0.46 ± 0.02	0.46 ± 0.03
AU(IMCP)	0.54 ± 0.01	0.42 ± 0.04	0.42 ± 0.03	0.39 ± 0.03

Table 4.26: RETFound linear probing evaluation on the HOJG dataset.

	CNV	no vs rest	leakage vs rest	pooling vs rest	staining vs rest	window defect vs rest	HyperF Type
AUC	0.86 ± 0.01	0.91 ± 0.02	0.82 ± 0.02	0.69 ± 0.03	0.63 ± 0.03	0.66 ± 0.08	0.74 ± 0.02
F1	0.72 ± 0.02	0.73 ± 0.02	0.68 ± 0.04	0.23 ± 0.09	0.34 ± 0.04	0.11 ± 0.04	0.40 ± 0.02
AU(IMCP)	0.64 ± 0.01	0.70 ± 0.01	0.58 ± 0.02	0.52 ± 0.02	0.49 ± 0.02	0.52 ± 0.03	0.34 ± 0.01

Table 4.27: RETFound MLP probing evaluation on the 3rd Aptos competition dataset.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.91 ± 0.01	0.82 ± 0.04	0.83 ± 0.02	0.86 ± 0.02
F1	0.65 ± 0.03	0.49 ± 0.07	0.46 ± 0.04	0.44 ± 0.06
AU(IMCP)	0.60 ± 0.02	0.46 ± 0.05	0.44 ± 0.03	0.42 ± 0.03

Table 4.28: RETFound MLP probing evaluation on the HOJG dataset.

4.3.2 FFT results

For the full fine-tuning of RETFound, the hyperparameter selection was initially based on the settings suggested by the original authors of the model. As part of this process, the *layer decay* parameter was introduced. Layer-Wise Learning Rate Decay (LLRD) is a technique in which different layers of a neural network are assigned different learning rates during training based on their depth. A common formula is the one displayed in Eq. (4.1)

$$\eta_i = \eta_0 \times \alpha^{(max\_depth - i)} \tag{4.1}$$

where α is a number between 0 and 1, η_0 is the base learning rate and η_i indicates the scaled lr assigned to the i-th layer. This means that shallower layers (closer to the input) receive smaller learning rates, while deeper layers (closer to the output) are updated with larger ones. The reasoning behind this approach is that earlier layers tend to capture generic and transferable features from the original pre-training, while deeper layers are more task-specific. Applying large updates to the earlier layers could disrupt these useful features, so they are updated more conservatively. Although the full impact of LLRD was not extensively explored in this study, it showed some effectiveness in improving performance for the HyperF Type classification, where the *layer decay* was set to 0.75. All remaining hyperparameters are presented in Tables 4.29 and 4.30.

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**
training epochs	50	50	50	50
lr	8e-6	5e-6	5e-6	5e-5
scheduler (min lr)	None	None	None	None
batch size	32	64	64	64
loss	weighted BinaryCE	weighted BinaryCE	weighted BinaryCE	weighted BinaryCE
augmentations	Yes	Yes	Yes	Yes
oversampling	No	No	No	No

	staining vs rest	**window defect vs rest**	**Hyperf Type**
training epochs	50	30	30
lr	1e-5	1e-5	3e-5
scheduler (min lr)	WarmupCosine (5e-8)	WarmupCosine (5e-8)	WarmupCosine (1e-7)
batch size	32	32	32
loss	weighted BinaryCE	BinaryCE	weighted CrossEntropy
augmentations	Yes	Yes	Yes
oversampling	No	Yes	No

Table 4.29: FFT RETFound hyperparameters for the 3rd Aptos competition's tasks.

	capillary leakage	**vascular leakage**	**optic disk hyperfluorescence**	**macular edema**
training epochs	30	30	30	30
lr	3e-6	1e-5	3e-6	8e-6
scheduler (min lr)	None	warmupCosine (5e-8)	None	None
batch size	32	32	16	32
loss	weighted CrossEntropy	CrossEntropy	weighted CrossEntropy	CrossEntropy
augmentations	Yes	Yes	Yes	Yes
oversampling	No	Yes	No	Yes

Table 4.30: FFT RETFound hyperparameters for the HOJG's tasks.

The results of FFT on the Aptos dataset presented in the Table 4.31 show a general improvement in performance compared to the MLP probing, with the most notable gains observed in the more complex tasks, such as *pooling*, *staining*, and *HyperF Type*. These tasks show an AUC increase of +0.04, +0.08, and +0.05, respectively, along with improvements in F1-score, suggesting a better balance between precision and recall.

	CNV	no vs rest	leakage vs rest	pooling vs rest	staining vs rest	window defect vs rest	HyperF Type
AUC	0.88 ± 0.01	0.93 ± 0.02	0.84 ± 0.02	0.73 ± 0.02	0.71 ± 0.02	0.68 ± 0.05	0.79 ± 0.01
F1	0.75 ± 0.02	0.78 ± 0.04	0.72 ± 0.02	0.28 ± 0.04	0.42 ± 0.02	0.19 ± 0.04	0.44 ± 0.03
AU(IMCP)	0.64 ± 0.02	0.71 ± 0.06	0.62 ± 0.01	0.58 ± 0.03	0.53 ± 0.02	0.51 ± 0.02	0.39 ± 0.02

Table 4.31: Full fine-tuning of RETFound on the 3rd Aptos competition dataset.

A significant performance boost is also observed in the Table 4.32 for the HOJG dataset, where the *vascular leakage*, *optic disk hyperfluorescence* and *macular edema* classification tasks achieve an AUC improvement of at least +0.05. However, the *capillary leakage* task performs slightly worse, with a decrease of 0.01 in both AUC and F1-score. This decline is likely due to training instability, possibly caused by suboptimal hyperparameter selection, as indicated by the higher standard deviation values.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.90 ± 0.04	0.87 ± 0.04	0.89 ± 0.02	0.93 ± 0.02
F1	0.64 ± 0.05	0.52 ± 0.05	0.54 ± 0.03	0.58 ± 0.06
AU(IMCP)	0.56 ± 0.08	0.49 ± 0.04	0.49 ± 0.02	0.52 ± 0.04

Table 4.32: Full fine-tuning of RETFound on the HOJG dataset.

Finally, the higher AU(IMCP) values suggest that the model's predictions are more reliable and confident.

4.3.3 PEFT results with LoRA

The hyperparameters for this set of experiments are shown in Tables 4.33 and 4.34.

The results shown in Tables 4.35 and 4.36 indicate that LoRA, once again, does not lead to significant changes in the model's discriminative power, with AUC and F1 score fluctuations mostly around ± 0.01. Even in cases where a larger difference is observed, such as in *staining vs rest*, the change is not statistically significant. The most noticeable impact of LoRA is on model confidence. In fact, AU(IMCP) decreases in a majority of tasks, with the decline being statistically relevant in 4 out of 11 cases.

4.3.4 SSL results

The core reason driving this experiment is to investigate whether an additional SSL pre-training step can adapt a foundation model to a more specific medical imaging modality. Specifically, the experiment aims to verify whether applying an extra round of Masked Autoencoding on fluorescein angiography images can enhance the ability of RETFound, originally pre-trained on color fundus photography, to learn FA-specific features.

This idea is rooted in the concept of Hierarchical Pretraining (HPT), which suggests that progressively pre-training on datasets increasingly similar to the target domain leads

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**
lr	1e-4	8e-5	8e-5	8e-5
weight decay	1e-4	1e-4	1e-4	2e-4
scheduler (min lr)	WarmupCosine (5e-9)	None	None	WarmupCosine (5e-9)
batch size	64	64	64	16
r (train. params)	32 (1%)	32 (1%)	32 (1%)	64 (2%)
dropout	0.1	0.1	0.1	0.1

	staining vs rest	**window defect vs rest**	**HyperF Type**
lr	6e-5	5e-4	5e-5
weight decay	5e-1	5e-4	5e-2
scheduler (min lr)	WarmupCosine (5e-6)	WarmupCosine (5e-6)	None
batch size	64	64	32
r (train. params)	64 (2%)	64 (2%)	64 (2%)
dropout	0.1	0.1	0.1

Table 4.33: LoRA hyperparameters for RETFound on the 3rd Aptos competition's tasks. The percentage in brackets next to the rank indicates the proportion of trainable parameters relative to the full model.

	capillary leakage	**vascular leakage**	**optic disk hyperfluorescence**	**macular edema**
lr	4e-5	4e-5	4e-5	2e-5
weight decay	5e-1	5e-1	5e-1	5e-2
scheduler (min lr)	WarmupCosine (1e-6)	WarmupCosine (1e-6)	WarmupCosine (1e-6)	None
batch size	64	64	64	32
r (train. params)	128 (4%)	128 (4%)	128 (4%)	128 (4%)
dropout	0.1	0.1	0.1	0.1

Table 4.34: LoRA hyperparameters for RETFound on the HOJG's tasks. The percentage in brackets next to the rank indicates the proportion of trainable parameters relative to the full model.

to improved downstream performance [58]. HPT typically involves an initial domain-general pre-training (e.g., ImageNet-1K), followed by domain-specific (specialist) pre-training, and finally, an optional target-specific pre-training.

A major challenge in applying this strategy to FA imaging is the limited availability of large-scale datasets and computational resources in clinical settings. This experiment therefore explores whether pretraining on a closely related domain (CFP) can reduce the need for extensive in-domain data. The aim is to determine if a lightweight SSL step

	CNV	no vs rest	leakage vs rest	pooling vs rest	staining vs rest	window defect vs rest	HyperF Type
AUC	0.87 ± 0.02 (-0.01)	0.94 ± 0.01 (+0.01)	0.84 ± 0.02 (−)	0.73 ± 0.03 (−)	0.68 ± 0.02 (-0.03)	0.70 ± 0.06 (+0.02)	0.78 ± 0.02 (-0.01)
F1	0.73 ± 0.03 (-0.02)	0.79 ± 0.02 (+0.01)	0.72 ± 0.04 (−)	0.31 ± 0.04 (+0.03)	0.39 ± 0.03 (-0.03)	0.15 ± 0.10 (-0.04)	0.42 ± 0.03 (-0.02)
AU(IMCP)	0.64 ± 0.02 (−)	0.74 ± 0.03 (+0.03)	0.62 ± 0.02 (−)	0.54 ± 0.02 (-0.04)	0.52 ± 0.02 (-0.01)	0.54 ± 0.05 (+0.03)	0.36 ± 0.01 (-0.03)

Table 4.35: PEFT (LoRA) of RETFound on the 3rd Aptos competition dataset. Between brackets, the relative change in performance compared to FFT. If the number in brackets is underlined, the change is statistically significant.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.91 ± 0.02 (+0.01)	0.86 ± 0.04 (-0.01)	0.88 ± 0.01 (−)	0.93 ± 0.02 (-0.01)
F1	0.63 ± 0.06 (-0.01)	0.52 ± 0.04 (−)	0.53 ± 0.04 (-0.01)	0.56 ± 0.04 (-0.02)
AU(IMCP)	0.51 ± 0.04 (-0.05)	0.42 ± 0.03 (-0.07)	0.43 ± 0.02 (-0.06)	0.51 ± 0.02 (-0.01)

Table 4.36: PEFT (LoRA) of RETFound on the HOJG dataset. Between brackets, the relative change in performance compared to FFT. If the number in brackets is underlined, the change is statistically significant.

on a relatively small set of FA images can still yield effective adaptation and improved downstream task performance, leveraging the prior knowledge already captured from the retinal domain.

For this experiment, the set of unlabeled images from the 3rd Aptos competition dataset was used, comprising a total of 9,608 FA images. To better illustrate the full pretraining process, the diagram shown in the previous section was expanded, Fig. 4.3, to include the additional SSL pretraining step on FA images.

The decoder architecture and weights used are those provided in the official GitHub repository of the original paper. The same hyperparameters suggested in the paper were adopted, including a mask ratio of 75%. The model was trained for a total of 150 epochs and the checkpoint corresponding to the lowest validation loss was saved. The loss curves are shown in Fig. 4.4.

Fig. 4.5 shows the same images previously used to visualize reconstruction quality, now used to highlight improvements after the additional pre-training step.

Linear probing results

As a first step, the effectiveness of the technique is evaluated through a linear probing analysis, with the goal of assessing whether the feature representations in the latent space

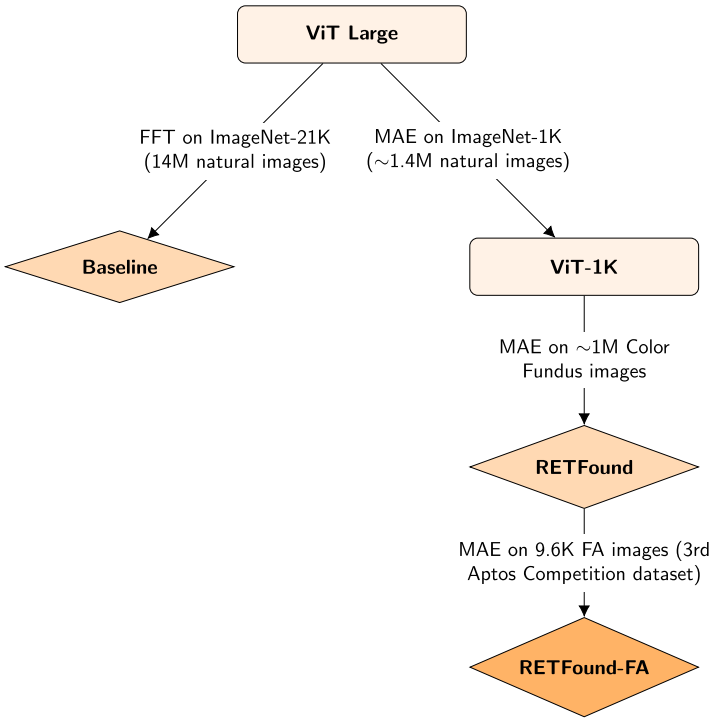

Figure 4.3: Schematic tree illustrating the extra-step leading to RETFound-FA.

have become more linearly separable after the additional pre-training step. The results shown in Tables 4.37 and 4.38 suggest that the model has learned more effective and discriminative features for the Aptos tasks. However, at this stage, there is no clear evidence that the new representations are of overall higher quality, as the performance boost is not consistently observed across the HOJG dataset.

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**	**staining vs rest**	**window defect vs rest**	**HyperF Type**
AUC	0.87 ± 0.02 (+0.01)	0.92 ± 0.02 (+0.01)	0.82 ± 0.02 (+0.01)	0.67 ± 0.02 (+0.01)	0.66 ± 0.03 (+0.03)	0.67 ± 0.01 (+0.04)	0.76 ± 0.01 (+0.03)
F1	0.73 ± 0.02 (+0.01)	0.72 ± 0.02 (+0.01)	0.71 ± 0.03 (+0.02)	0.25 ± 0.01 (-0.01)	0.38 ± 0.02 (+0.03)	0.18 ± 0.01 (+0.03)	0.42 ± 0.02 (+0.04)
AU(IMCP)	0.63 ± 0.01 (+0.01)	0.67 ± 0.01 (+0.01)	0.58 ± 0.02 (+0.02)	0.50 ± 0.02 (+0.01)	0.49 ± 0.02 (−)	0.49 ± 0.01 (−)	0.34 ± 0.01 (+0.01)

Table 4.37: Linear probing analysis of RETFound-FA on the 3rd Aptos competition dataset. Between brackets, the relative change in performance compared to original RETFound linear evaluation.

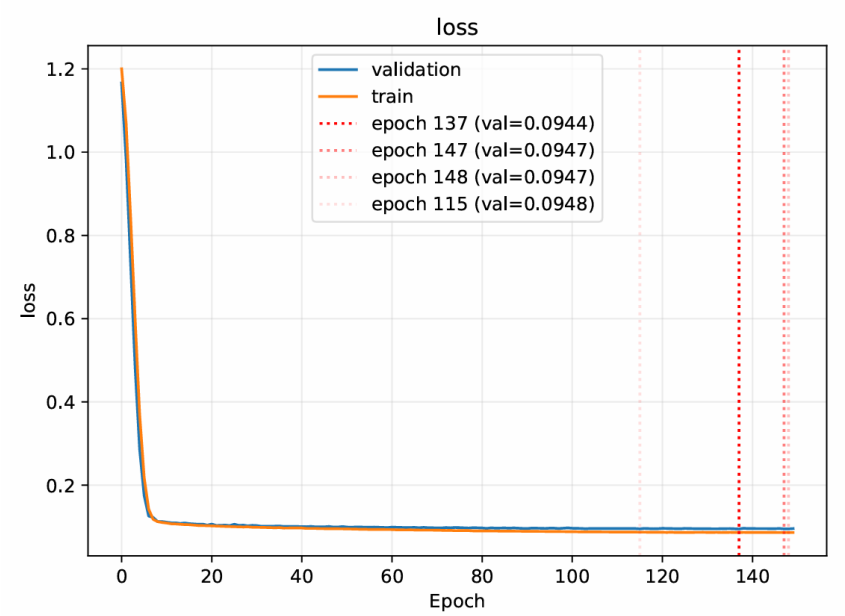

Figure 4.4: Loss curve of the MAE training on the unlabeled subset of 9,608 FA images from the 3rd Aptos competition dataset. Red lines indicate the top 5 checkpoints with the lowest validation loss during training

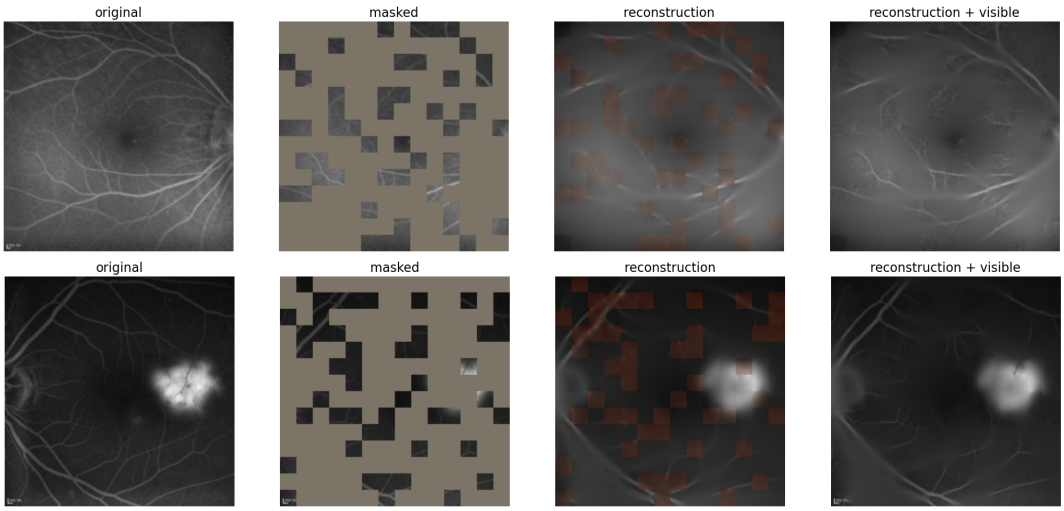

Figure 4.5: Reconstructed fluorescein angiography from highly masked images in pretext task after an extra round of Masked Autoencoding on FA images.

FFT results

FFT serves as a critical benchmark to determine if the adaptation introduced through SSL effectively improves downstream performance on FA-related tasks. In order to maintain a

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.88 ± 0.02 (-0.02)	0.84 ± 0.04 (+0.03)	0.81 ± 0.02 (+0.01)	0.82 ± 0.03 (−)
F1	0.54 ± 0.04 (-0.03)	0.45 ± 0.05 (−)	0.42 ± 0.04 (-0.04)	0.43 ± 0.05 (-0.03)
AU(IMCP)	0.52 ± 0.01 (-0.02)	0.41 ± 0.03 (-0.01)	0.41 ± 0.02 (-0.01)	0.38 ± 0.03 (-0.01)

Table 4.38: Linear probing analysis of RETFound-FA on the HOJG dataset. Between brackets, the relative change in performance compared to original RETFound linear evaluation.

fair comparison with the original RETFound, the same set of fine-tuning hyperparameters was initially used. However, given that the additional SSL pre-training may alter the model's learning dynamics, a light hyperparameter tuning was applied where needed. No changes were applied to the hyperparameters for the HOJG dataset; therefore, reference can be made to the table presented in Subsec. 4.3.2. Conversely, for the Aptos tasks, some minor modifications were necessary, Table 4.39 report the final set of hyperparameters used for each of them.

The final results, presented in Tables 4.40 and 4.41, suggest that the additional self-supervised pre-training step enabled the model to learn generalizable features specific to fluorescein angiography. While the improvements on the Aptos dataset are notable, they should be interpreted with caution, as the pre-training and evaluation data share the same source and distribution from the same dataset, acquired with the same imaging device, and reflecting the same population. However, the consistent performance gains observed in the classification of inflammatory signs, which come from a distinct dataset, provide stronger evidence of the effectiveness of SSL in adapting RETFound to FA-specific tasks.

4.4 RETFound Green

Following the success of RETFound, researchers in the ophthalmology and AI communities began to further explore the development of scalable, efficient models for this domain. One of the first extensions was DERETFound, which leveraged generative AI and expert clinical insights to produce approximately one million synthetic retinal images from only 16.7% of the real-world color fundus photography data originally used in RETFound [59]. While these efforts marked an important advancement in data efficiency, both these models still required considerable computational resources to train, limiting their accessibility and practical deployment in resource-constrained environments.

To address this, RETFound-Green was introduced as a lightweight, eco-friendly alternative. Trained with a novel SSL technique, called Token Reconstruction, on just 75,000 publicly available CFP images, RETFound-Green reduced the compute demand by a factor of 400, significantly lowering the barrier to entry for research and clinical use.

	CNV*	**no vs rest**	**leakage vs rest**	**pooling vs rest**
training epochs	50	50	50	50
lr	1e-5	5e-6	5e-6	5e-5
scheduler (min lr)	WarmupCosine (1e-7)	None	None	None
batch size	32	64	64	64
loss	weighted BinaryCE	weighted BinaryCE	weighted BinaryCE	weighted BinaryCE
augmentations	Yes	Yes	Yes	Yes
oversampling	No	No	No	No

	staining vs rest*	**window defect vs rest**	**Hyperf Type***
training epochs	50	30	30
lr	1e-5	1e-5	3e-5
scheduler (min lr)	WarmupCosine (1e-7)	WarmupCosine (5e-8)	WarmupCosine (5e-8)
batch size	32	32	32
loss	weighted BinaryCE	BinaryCE	weighted CrossEntropy
augmentations	Yes	Yes	Yes
oversampling	No	Yes	No

Table 4.39: FFT RETFound-FA hyperparameters for the 3rd Aptos competition's tasks. An asterisk (*) indicates tasks for which the hyperparameters differ from those used for the original RETFound.

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**	**staining vs rest**	**window defect vs rest**	**HyperF Type**
AUC	0.88 ± 0.02 (−)	0.94 ± 0.01 (+0.01)	0.85 ± 0.02 (+0.01)	0.76 ± 0.03 (+0.03)	0.73 ± 0.02 (+0.02)	0.71 ± 0.06 (+0.01)	0.80 ± 0.02 (+0.01)
F1	0.75 ± 0.03 (−)	0.80 ± 0.02 (+0.02)	0.73 ± 0.04 (+0.01)	0.30 ± 0.04 (+0.02)	0.44 ± 0.03 (+0.02)	0.18 ± 0.03 (+0.03)	0.45 ± 0.06 (+0.01)
AU(IMCP)	0.65 ± 0.02 (+0.01)	0.72 ± 0.03 (+0.01)	0.64 ± 0.01 (+0.02)	0.62 ± 0.04 (+0.04)	0.54 ± 0.02 (+0.01)	0.51 ± 0.03 (-0.03)	0.40 ± 0.02 (+0.01)

Table 4.40: FFT of RETFound-FA on the 3rd Aptos competition dataset. Between brackets, the relative change in performance compared to original FFT of RETFound.

What sets this model apart is that, despite its efficiency, it does not systematically underperform when compared to its larger counterparts. In fact, RETFound-Green achieves strong downstream performance while also being faster and lightweight.

The model is built on the ViT-Small architecture, which was previously analyzed in Sec. 4.2. As with RETFound, a schematic diagram summarizing the pre-training steps and design choices leading to RETFound Green is presented in Fig. 4.6, offering a clear

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.92 ± 0.02 (+0.02)	0.88 ± 0.03 (+0.01)	0.90 ± 0.02 (+0.01)	0.93 ± 0.02 (−)
F1	0.65 ± 0.05 (+0.01)	0.56 ± 0.04 (+0.04)	0.57 ± 0.04 (+0.03)	0.58 ± 0.05 (−)
AU(IMCP)	0.62 ± 0.02 (+0.06)	0.52 ± 0.04 (+0.03)	0.51 ± 0.01 (+0.02)	0.55 ± 0.05 (+0.03)

Table 4.41: FFT of RETFound-FA on the HOJG dataset. Between brackets, the relative change in performance compared to original FFT of RETFound.

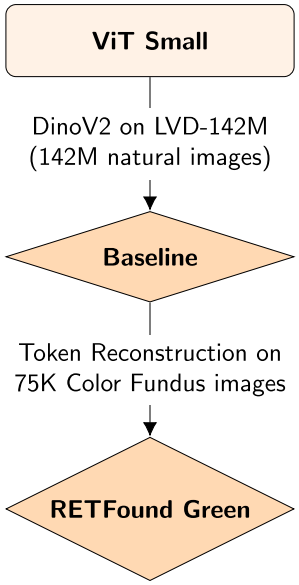

Figure 4.6: Schematic tree illustrating the steps leading to RETFound Green and its baseline.

overview of the pipeline behind this compact yet capable foundation model.

4.4.1 Probing results

The linear probing results shown in Table 4.42 and 4.43 exhibit a different trend compared to what was observed between RETFound and its baseline.

In this case, there is no clear winner between the two models, as most tasks show comparable performance. RETFound Green achieves slightly better results on tasks such as *leakage vs rest*, *staining vs rest*, and *HyperF Type*, while the baseline performs marginally better on *capillary* and *vascular leakage* classification. This is further supported by the MLP probing results presented in Tables 4.44 and 4.45, which show overall performance comparable to that of the ViT-Small.

	CNV	no vs rest	leakage vs rest	pooling vs rest	staining vs rest	window defect vs rest	HyperF Type
AUC	0.87 ± 0.02	0.93 ± 0.02	0.84 ± 0.01	0.67 ± 0.04	0.68 ± 0.03	0.70 ± 0.02	0.77 ± 0.02
F1	0.73 ± 0.03	0.73 ± 0.03	0.74 ± 0.02	0.28 ± 0.04	0.40 ± 0.03	0.17 ± 0.02	0.43 ± 0.01
AU(IMCP)	0.63 ± 0.01	0.66 ± 0.03	0.60 ± 0.01	0.51 ± 0.02	0.51 ± 0.01	0.51 ± 0.02	0.35 ± 0.01

Table 4.42: RETFound Green linear probing evaluation on the 3rd Aptos competition dataset.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.90 ± 0.02	0.81 ± 0.01	0.84 ± 0.02	0.80 ± 0.03
F1	0.59 ± 0.07	0.41 ± 0.02	0.49 ± 0.04	0.39 ± 0.05
AU(IMCP)	0.53 ± 0.05	0.37 ± 0.04	0.42 ± 0.03	0.37 ± 0.02

Table 4.43: RETFound Green linear probing evaluation on the HOJG dataset.

	CNV	no vs rest	leakage vs rest	pooling vs rest	staining vs rest	window defect vs rest	HyperF Type
AUC	0.87 ± 0.02	0.93 ± 0.02	0.84 ± 0.01	0.69 ± 0.04	0.68 ± 0.04	0.71 ± 0.02	0.77 ± 0.02
F1	0.73 ± 0.03	0.79 ± 0.02	0.73 ± 0.02	0.29 ± 0.04	0.37 ± 0.05	0.20 ± 0.05	0.44 ± 0.01
AU(IMCP)	0.63 ± 0.01	0.69 ± 0.00	0.59 ± 0.01	0.51 ± 0.01	0.50 ± 0.01	0.53 ± 0.02	0.37 ± 0.01

Table 4.44: RETFound Green MLP probing evaluation on the 3rd Aptos competition dataset.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.93 ± 0.01	0.87 ± 0.03	0.87 ± 0.01	0.86 ± 0.02
F1	0.70 ± 0.03	0.49 ± 0.04	0.53 ± 0.03	0.45 ± 0.03
AU(IMCP)	0.60 ± 0.04	0.46 ± 0.03	0.46 ± 0.03	0.40 ± 0.02

Table 4.45: RETFound Green MLP probing evaluation on the HOJG dataset.

4.4.2 FFT results

The complete list of hyperparameters used for training RETFound Green is depicted in Table 4.46 and 4.47.

Once again, Table 4.48 and 4.49 show that the improvement achieved through full fine-tuning is substantial, both in terms of discriminative power (as reflected by higher AUC and F1 scores) and in terms of prediction confidence, as indicated by the consistently higher AU(IMCP) values across all tasks. As previously emphasized, confidence is a particularly important factor in the medical domain, where reliable predictions are critical. These results confirm the effectiveness of full FFT in producing high-performing, task-specialized models. Moreover, thanks to the compact size of RETFound Green, such improvements can be achieved without the need for extensive computational resources. In fact, the model was trained in just a few minutes on a single RTX 3090 GPU, making

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**
training epochs	50	50	50	50
lr	1e-5	1e-5	1e-6	2e-6
weight decay	5e-2	5e-2	1e-4	5e-2
scheduler (min lr)	WarmupCosine (5e-9)	WarmupCosine (5e-9)	None	None
batch size	256	256	64	64
loss	weighted BinaryCE	weighted BinaryCE	BinaryCE	BinaryCE
augmentations	Yes	Yes	Yes	Yes
oversampling	No	No	Yes	Yes

	staining vs rest	**window defect vs rest**	**Hyperf Type**
training epochs	50	30	50
lr	1e-6	5e-6	3e-6
weight decay	1e-2	5e-2	5e-2
scheduler (min lr)	None	None	None
batch size	64	32	128
loss	BinaryCE BinaryCE	BinaryCE	weighted CrossEntropy
augmentations	Yes	Yes	Yes
oversampling	Yes	Yes	No

Table 4.46: FFT RETFound Green hyperparameters for the 3rd Aptos competition's tasks.

	capillary leakage	**vascular leakage**	**optic disk hyperfluorescence**	**macular edema**
training epochs	30	30	30	30
lr	4e-6	8e-6	1e-5	8e-6
weight decay	5e-3	5e-2	5e-2	5e-3
scheduler (min lr)	WarmupCosine (5e-8)	None	WarmupCosine (5e-8)	WarmupCosine (5e-8)
batch size	16	32	32	32
loss	weighted CrossEntropy	CrossEntropy	CrossEntropy	CrossEntropy
augmentations	Yes	Yes	Yes	Yes
oversampling	No	Yes	Yes	Yes

Table 4.47: FFT RETFound Green hyperparameters for the HOJG's tasks.

it highly accessible for real-world applications.

	CNV	no vs rest	leakage vs rest	pooling vs rest	staining vs rest	window defect vs rest	HyperF Type
AUC	0.90 ± 0.02	0.95 ± 0.01	0.87 ± 0.02	0.76 ± 0.02	0.75 ± 0.03	0.78 ± 0.04	0.82 ± 0.02
F1	0.75 ± 0.02	0.81 ± 0.02	0.75 ± 0.02	0.33 ± 0.02	0.44 ± 0.04	0.29 ± 0.05	0.48 ± 0.02
AU(IMCP)	0.71 ± 0.01	0.78 ± 0.00	0.66 ± 0.01	0.58 ± 0.03	0.55 ± 0.01	0.61 ± 0.02	0.41 ± 0.02

Table 4.48: Full fine-tuning of RETFound Green on the 3rd Aptos competition dataset.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.94 ± 0.02	0.95 ± 0.02	0.92 ± 0.01	0.95 ± 0.01
F1	0.71 ± 0.06	0.69 ± 0.06	0.62 ± 0.04	0.62 ± 0.04
AU(IMCP)	0.64 ± 0.03	0.62 ± 0.03	0.55 ± 0.03	0.54 ± 0.03

Table 4.49: Full fine-tuning of RETFound Green on the HOJG dataset.

4.4.3 PEFT results with LoRA

The results in Table 4.50 and 4.51 show that, also with RETFound Green, LoRA demonstrates performance comparable to FFT, despite updating only a small fraction of the model's parameters. However, its application to smaller models appears to reveal certain limitations. In fact, as observed with ViT-Small, LoRA tends to degrade the model's confidence, as indicated by lower AU(IMCP) scores across multiple tasks (with 5 of them being statistically worse).

	CNV	no vs rest	leakage vs rest	pooling vs rest	staining vs rest	window defect vs rest	HyperF Type
AUC	0.89 ± 0.01 (-0.01)	0.95 ± 0.01 (−)	0.86 ± 0.02 (-0.01)	0.77 ± 0.04 (+0.01)	0.74 ± 0.04 (-0.01)	0.79 ± 0.04 (+0.01)	0.83 ± 0.02 (+0.01)
F1	0.77 ± 0.02 (+0.02)	0.81 ± 0.04 (−)	0.75 ± 0.03 (−)	0.39 ± 0.03 (+0.06)	0.46 ± 0.05 (+0.02)	0.24 ± 0.04 (-0.05)	0.49 ± 0.04 +0.01
AU(IMCP)	0.67 ± 0.02 (-0.04)	0.75 ± 0.01 (-0.03)	0.65 ± 0.02 (-0.01)	0.57 ± 0.02 (-0.01)	0.55 ± 0.03 (−)	0.56 ± 0.01 (-0.05)	0.41 ± 0.02 (−)

Table 4.50: PEFT (LoRA) of RETFound Green on the 3rd Aptos competition dataset. Between brackets, the relative change in performance compared to FFT. If the number in brackets is underlined, the change is statistically significant.

One possible explanation is that partial updates introduced by LoRA may not significantly alter the discriminative power of the model, but still impact the stability and certainty of predictions. This effect may be more pronounced in smaller models, which could be less robust to partial adaptation techniques and instead benefit more from full fine-tuning. Prior work [60] has shown that LoRA tends to learn less expressive adaptations than FFT, and in the case of compact architectures, this limitation may translate into reduced model's reliability. Further investigation, such as analyzing the influence of different rank values, would be necessary to better understand this phenomenon, though it lies beyond the scope of this study.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.95 ± 0.01 (+0.01)	0.94 ± 0.01 (-0.01)	0.91 ± 0.02 (−)	0.95 ± 0.01 (-0.01)
F1	0.71 ± 0.03 (−)	0.68 ± 0.05 (-0.01)	0.63 ± 0.04 (+0.01)	0.63 ± 0.05 (+0.01)
AU(IMCP)	0.63 ± 0.03 (-0.01)	0.58 ± 0.01 (-0.04)	0.54 ± 0.02 (-0.01)	0.56 ± 0.04 (+0.02)

Table 4.51: PEFT (LoRA) of RETFound Green on the HOJG dataset. Between brackets, the relative change in performance compared to FFT. If the number in brackets is underlined, the change is statistically significant.

	CNV	no vs rest	leakage vs rest	pooling vs rest
lr	8e-6	9e-5	9e-5	1e-5
weight decay	5e-1	5e-2	5e-1	5e-2
scheduler (min lr)	None	WarmupCosine (5e-6)	WarmupCosine (5e-6)	None
batch size	128	256	256	64
r (train. params)	32 (2.67%)	32 (2.67%)	32 (2.67%)	64 (5.18%)
dropout	0.1	0.1	0.1	None

	staining vs rest	window defect vs rest	HyperF Type
lr	5e-5	1e-5	1e-5
weight decay	5e-1	5e-2	1e-2
scheduler (min lr)	WarmupCosine (5e-6)	None	None
batch size	256	64	8
r (train. params)	64 (5.18%)	64 (5.18%)	64 (5.18%)
dropout	0.1	0.1	0.1

Table 4.52: LoRA hyperparameters for RETFound Green on the 3rd Aptos competition's tasks. The percentage in brackets next to the rank indicates the proportion of trainable parameters relative to the full model.

Hyperparameters used for the PEFT on RETFound Green are shown in Table 4.52 and 4.53

4.4.4 SSL results

The same motivation previously outlined for RETFound also guides this experiment, in which an additional self-supervised pretraining step is applied to RETFound Green, this time based on Token Reconstruction objective. The goal is to assess the effectiveness of

	capillary leakage	**vascular leakage**	**optic disk hyperfluorescence**	**macular edema**
lr	8e-6	1e-5	5e-6	1e-5
weight decay	5e-3	5e-3	5e-1	5e-2
scheduler (min lr)	WarmupCosine (5e-8)	WarmupCosine (5e-8)	WarmupCosine (5e-8)	WarmupCosine (5e-7)
batch size	16	16	32	32
r (train. params)	128 (9.8%)	128 (9.8%)	128 (9.8%)	128 (9.8%)
dropout	0.1	0.2	0.1	0.1

Table 4.53: LoRA hyperparameters for RETFound Green on the HOJG's tasks. The percentage in brackets next to the rank indicates the proportion of trainable parameters relative to the full model.

this alternative SSL strategy in enabling cross-domain adaptation.

Given the novelty of Token Reconstruction, there is no guidance on how to effectively implement an additional SSL step using this method. This study represents the first empirical attempt to explore this direction. The adopted strategy was to treat RETFound Green as the frozen model, while using the ViT-Small initialized with DINOv2 weights (pre-trained on LVD-142M) as the model to be adapted. This design choice was driven by training observations: several attempts to directly adapt RETFound Green led to severe overfitting, which could not be mitigated through standard regularization techniques.

Since Token Reconstruction aims to instill knowledge at the abstract level, using RETFound Green as the fixed source of information appeared to be a plausible and effective choice, both in terms of training stability and learning outcomes. Both models were exposed to the same unlabeled set of 9,608 FA images from the 3rd Aptos competition dataset, previously used in the MAE-based adaptation of RETFound. In this setup, the unfrozen model is trained to reconstruct the latent output tokens generated by the frozen model, using noisy versions of the same inputs.

The diagram of RETFound Green is therefore expanded to include the additional Token Reconstruction step, as illustrated in Fig. 4.7

No hyperparameter tuning was performed; instead, the same configuration used by the original paper was adopted. RETFound Green was trained for 120 epochs with a batch size of 128, using the AdamW optimizer with a maximum learning rate of 5×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.99$, and a weight decay of 5×10^{-4}, applied to all parameters except the biases. A cosine learning rate scheduler was used, with a warmup phase of 10 epochs, during which the learning rate increased linearly from 5×10^{-9} to its maximum (WarmupCosine), followed by a 20-epoch cooldown where the learning rate remained at its minimum. The maximum gradient norm was clipped to 0.1.

In terms of data corruption strategies, a random subset of input tokens was replaced with a trainable corruption token, effectively masking parts of the input image. Additionally, pixel-wise Gaussian noise, sampled from $\mathcal{N}(0, 0.2)$, was applied to the images. For the Token Reconstruction objective, the corruption ratio was sampled from $\mathcal{U}(0, \frac{1}{3})$. The loss curves are shown in Fig. 4.8.

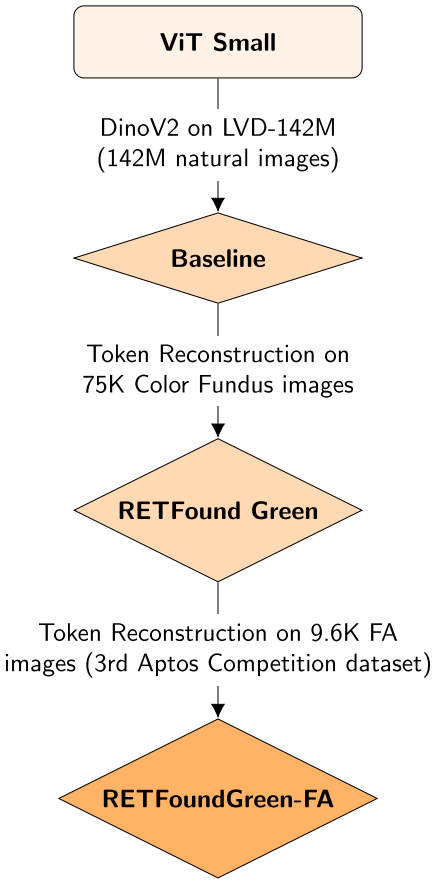

Figure 4.7: Schematic tree illustrating the extra-step leading to RETFoundGreen-FA.

Linear probing results

Table 4.54 and 4.55 present the promising results of the linear probing evaluation, showing a general improvement in both AUC and F1 score not only on the 3rd Aptos competition dataset but also on the HOJG dataset, in contrast to what was observed with RETFound-MAE. This suggests that the model has learned more globally effective and transferable feature for FA.

This improvement is further supported by the higher AU(IMCP) values. The fact that a simple linear layer can separate the features more effectively, while also doing so with higher confidence, implies that the learned representations are not only more discriminative but also more informative and reliable. In other words, the features encoded in the latent space are better structured, allowing for more certain predictions even with minimal task-specific adaptation.

FFT results

Finally, the effectiveness of the Token Reconstruction technique is assessed through FFT. Following the same reasoning applied in the evaluation of RETFound, and to ensure the

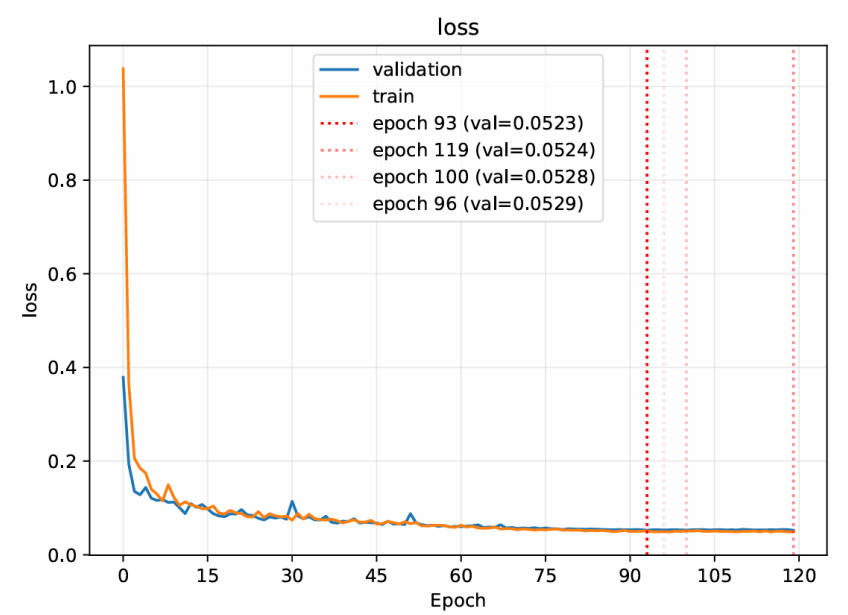

Figure 4.8: Loss curve of the Token Reconstruction training on the unlabeled subset of 9,608 FA images from the 3rd Aptos competition dataset. Red lines indicate the top 5 checkpoints with the lowest validation loss during training

	CNV	**no vs rest**	**leakage vs rest**	**pooling vs rest**	**staining vs rest**	**window defect vs rest**	**HyperF Type**
AUC	0.88 ± 0.01 (+0.01)	0.94 ± 0.02 (+0.01)	0.84 ± 0.02 (−)	0.69 ± 0.05 (+0.02)	0.68 ± 0.04 (−)	0.73 ± 0.01 (+0.03)	0.78 ± 0.02 (+0.01)
F1	0.75 ± 0.02 (+0.02)	0.76 ± 0.02 (+0.03)	0.73 ± 0.01 (-0.01)	0.28 ± 0.05 (−)	0.39 ± 0.04 (-0.01)	0.20 ± 0.02 (+0.03)	0.45 ± 0.02 (+0.02)
AU(IMCP)	0.64 ± 0.01 (+0.01)	0.69 ± 0.01 (+0.03)	0.60 ± 0.01 (−)	0.51 ± 0.02 (−)	0.51 ± 0.01 (−)	0.52 ± 0.01 (+0.01)	0.36 ± 0.01 (+0.01)

Table 4.54: Linear probing analysis of RETFoundGreen-FA on the 3rd Aptos competition dataset. Between brackets, the relative change in performance compared to original RETFound Green linear evaluation.

fairest possible comparison with RETFound Green, the same set of fine-tuning hyperparameters was used. No additional hyperparameter tuning was performed, as the training dynamics proved to be satisfactory. For reference, the hyperparameters are reported in the table presented in Subsection 4.4.2.

The results are presented in Tables 4.56 and 4.57. Interestingly, although the linear probing evaluation yielded significantly better performance than that observed for RETFound Green, the gains achieved through FFT do not reflect the same level of improvement. Overall, the model shows decent gains in F1 scores, indicating a better balance between precision and recall, with statistically significant improvements observed in the *no vs rest* and *HyperF Type* classification tasks. However, the remaining tasks show little

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.91 ± 0.02 (+0.01)	0.84 ± 0.02 (+0.03)	0.85 ± 0.01 (+0.01)	0.82 ± 0.02 (+0.02)
F1	0.62 ± 0.05 (+0.03)	0.43 ± 0.03 (+0.02)	0.50 ± 0.05 (+0.01)	0.41 ± 0.04 (+0.02)
AU(IMCP)	0.54 ± 0.05 (+0.01)	0.41 ± 0.04 (+0.04)	0.43 ± 0.03 (+0.01)	0.38 ± 0.02 (+0.01)

Table 4.55: Linear probing analysis of RETFoundGreen-FA on the HOJG dataset. Between brackets, the relative change in performance compared to original RETFound Green linear evaluation.

	CNV	no vs rest	leakage vs rest	pooling vs rest	staining vs rest	window defect vs rest	HyperF Type
AUC	0.90 ± 0.02 (–)	0.95 ± 0.02 (–)	0.87 ± 0.02 (–)	0.76 ± 0.04 (–)	0.75 ± 0.04 (–)	0.78 ± 0.06 (–)	0.83 ± 0.02 (+0.01)
F1	0.76 ± 0.03 (+0.01)	0.83 ± 0.03 (+0.02)	0.76 ± 0.03 (+0.01)	0.32 ± 0.06 (-0.01)	0.44 ± 0.03 (–)	0.30 ± 0.06 (+0.01)	0.51 ± 0.02 (+0.03)
AU(IMCP)	0.71 ± 0.02 (–)	0.78 ± 0.01 (–)	0.65 ± 0.01 (-0.01)	0.56 ± 0.02 (-0.02)	0.55 ± 0.01 (–)	0.61 ± 0.04 (–)	0.43 ± 0.01 (+0.02)

Table 4.56: FFT of RETFoundGreen-FA on the 3rd Aptos competition dataset. Between brackets, the relative change in performance compared to original FFT of RETFound Green.

	capillary leakage	vascular leakage	optic disk hyperfluorescence	macular edema
AUC	0.95 ± 0.01 (+0.01)	0.95 ± 0.02 (–)	0.92 ± 0.01 (–)	0.94 ± 0.01 (-0.01)
F1	0.70 ± 0.02 (-0.01)	0.69 ± 0.09 (–)	0.63 ± 0.03 (+0.01)	0.59 ± 0.06 (-0.03)
AU(IMCP)	0.66 ± 0.02 (+0.02)	0.65 ± 0.05 (+0.02)	0.55 ± 0.01 (–)	0.54 ± 0.07 (–)

Table 4.57: FFT of RETFoundGreen-FA on the HOJG dataset. Between brackets, the relative change in performance compared to original FFT of RETFound Green.

to no improvement and in these cases the model exhibits higher standard deviation, suggesting increased variability in performance. Moreover, in tasks such as *pooling vs rest* and *macular edema* classification, the model performs slightly worse than the original RETFound Green.

While a definitive explanation for this behavior cannot be provided at this stage, the

results highlight the need for further investigation into the Token Reconstruction technique. In particular, additional research is needed to develop a more effective training strategy that allows the method to scale better and fully realize its potential for adaptation. This need is underscored by the improvements observed in the linear probing evaluation, which suggest that Token Reconstruction is capable of producing strong and meaningful representations, even if the current fine-tuning setup did not fully capitalize on them.

4.5 Model comparison

In this final section, a comprehensive comparison is carried out across all the models analyzed in the study, with a particular focus on evaluating whether retinal foundation models offer a significant advantage in cross-domain adaptation over their respective baselines. While previous sections examined each model individually, by analyzing their performance under probing, FFT, and PEFT, and included some preliminary discussion on the different quality of the learned representations, a direct comparison was intentionally deferred. This section now addresses that gap by systematically comparing the results obtained in the FFT and PEFT settings, in order to draw more conclusive insights about the strengths and weaknesses of retinal foundation models.

4.5.1 RETFound vs ViT-Large

The first comparison focuses on RETFound and its baseline, the ViT-Large model pretrained in a supervised manner on ImageNet-21K. Figure 4.9 and 4.10 provide a side-by-side comparison of the two models under both FFT and LoRA-based PEFT. An asterisk in the figure indicates that the difference in performance between the models is statistically significant, as determined by the Wilcoxon signed-rank test.

From the plots, it can be observed that, in the *FFT* scenario, the baseline ViT-Large demonstrates stronger discriminative performance, achieving five statistically significant wins out of eleven tasks on the AUC score. However, notable differences emerge in terms of prediction reliability, where RETFound consistently outperforms the baseline in four out of six binary classification tasks, with one statistical win in the *pooling vs rest*. This increased reliability may also contribute to higher confidence in the multiclass *HyperF Type* classification. Conversely, ViT-Large appears to be more confident in certain multiclass scenarios, with a statistically significant improvement observed for the *optic disk hyperfluorescence* classification.

Another particularly interesting outcome, clearly visible in the plots, is that the use of LoRA, as previously discussed in Sec. 4.1, leads to a notable improvement in the confidence of the ViT-Large model, resulting in six statistically significant wins in terms of AU(IMCP). To further demonstrate the importance of this behaviour, Figure 4.11 presents the IMCP curves for two representative tasks in which ViT-Large outperformed RETFound. The selected tasks, *HyperF Type* classification from the Aptos dataset and *vascular leakage* classification from the HOJG dataset, correspond to cases where the difference was found to be statistically significant. The plots, based on split 1 of the evaluation, clearly show a notable reduction in the area of uncertainty, as measured by

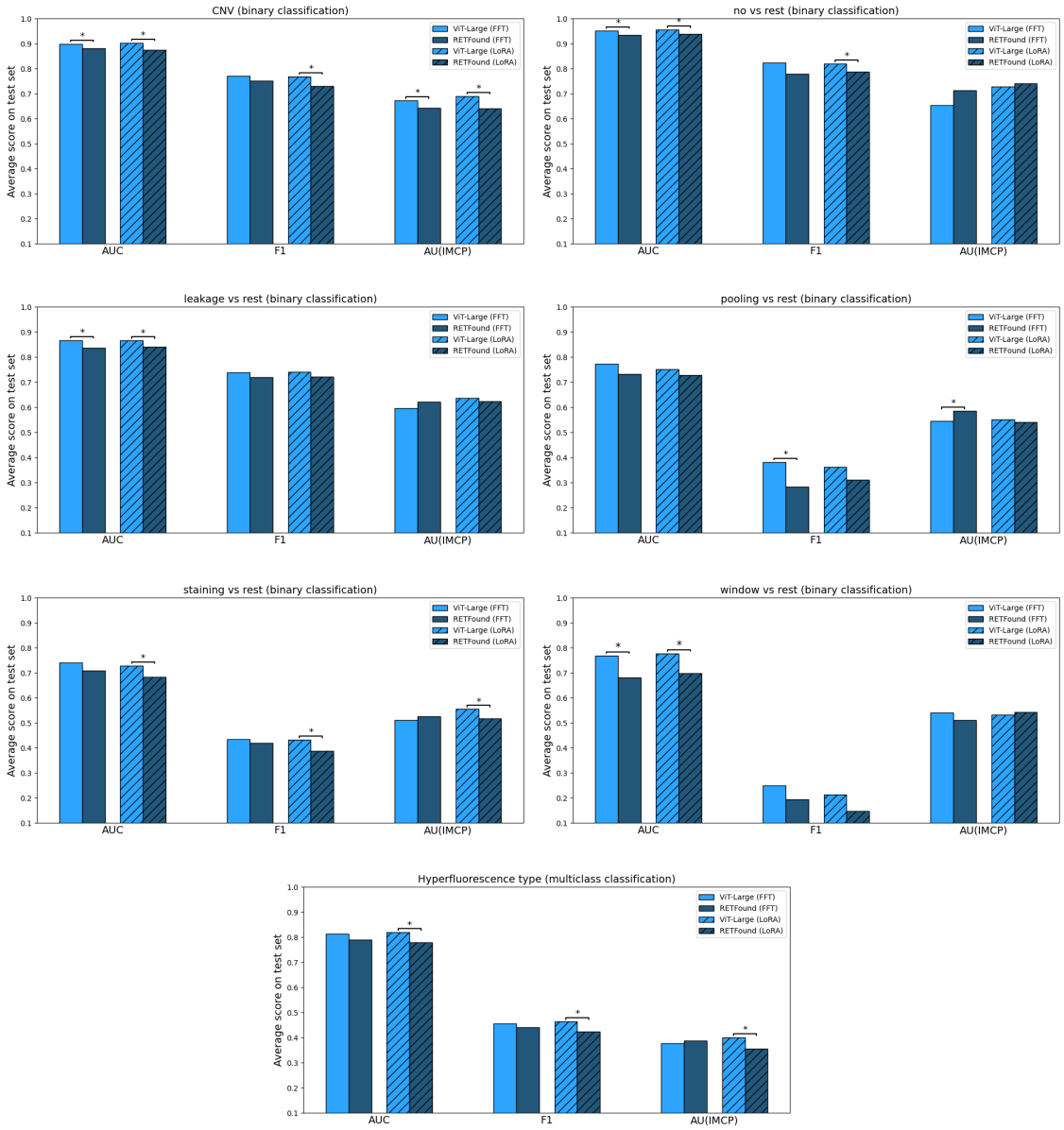

Figure 4.9: Overview of the FFT and PEFT results for RETFound and ViT-Large obtained on the 3rd Aptos competition dataset. (*) the asterisk indicates a p-value < 0.05.

the IMCP curves. It is important to note that this area does not indicate whether a sample is correctly classified or not, but rather it captures the proportion of predictions for which the output probability was not sufficiently high to be considered reliable, regardless of whether the predicted label was correct. In the case of the Aptos dataset, this reduction in uncertainty leads to a reclassification of nearly 30% of the samples, while for vascular leakage classification in the HOJG dataset, the improvement affected 11.7% of the total test samples.

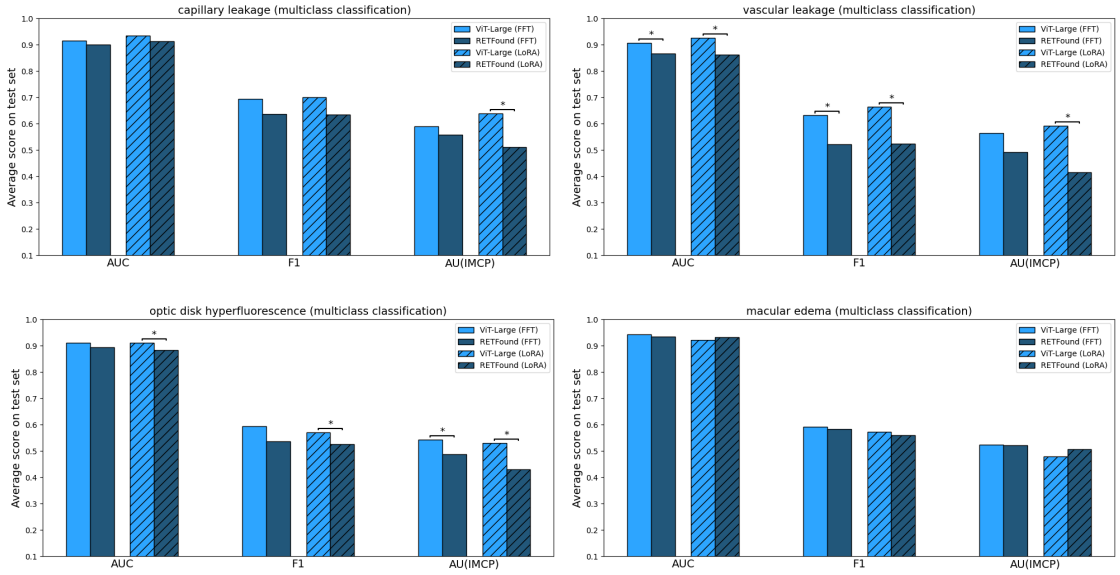

Figure 4.10: Overview of the FFT and PEFT results for RETFound and ViT-Large obtained on the HOJG dataset. (*) the asterisk indicates a p-value < 0.05.

4.5.2 RETFound Green vs ViT-Small

The second pair of models analyzed consists of RETFound Green and the ViT-Small model pre-trained on LVD-142M using the DINOv2 self-supervised learning technique. Figures 4.12 and 4.13 provide an overview of their performance under both FFT and PEFT using LoRA.

The situation differs significantly from what was observed with the larger model counterparts. In this case, RETFound Green and the ViT-Small model exhibit comparable performance across nearly all tasks, with only minor differences, most of which are not statistically significant, except in a few isolated cases. As a result, it is not possible to determine a clear winner between the two models, either in terms of discriminative power, as measured by AUC and F1 scores, or in terms of prediction reliability.

As previously noted, LoRA maintains the discriminative capabilities of both models, but introduces a slight decline in the confidence of final predictions. This is evidenced by the lower AU(IMCP) values observed in the LoRA-based PEFT setting compared to full fine-tuning.

4.5.3 Visual and quantitative aggregation of model performance

In order to identify the most effective model among those evaluated, a heuristic strategy was adopted based on the use of a radar chart and the computation of the area it encloses, based on the three evaluation metrics consistently used throughout this study: AUC, F1 score, and AU(IMCP). Given the wide range of tasks involved, each potentially highlighting different aspects of model behavior, establishing a theoretically optimal aggregation strategy is non-trivial. Consequently, a heuristic approach was chosen to synthesize model

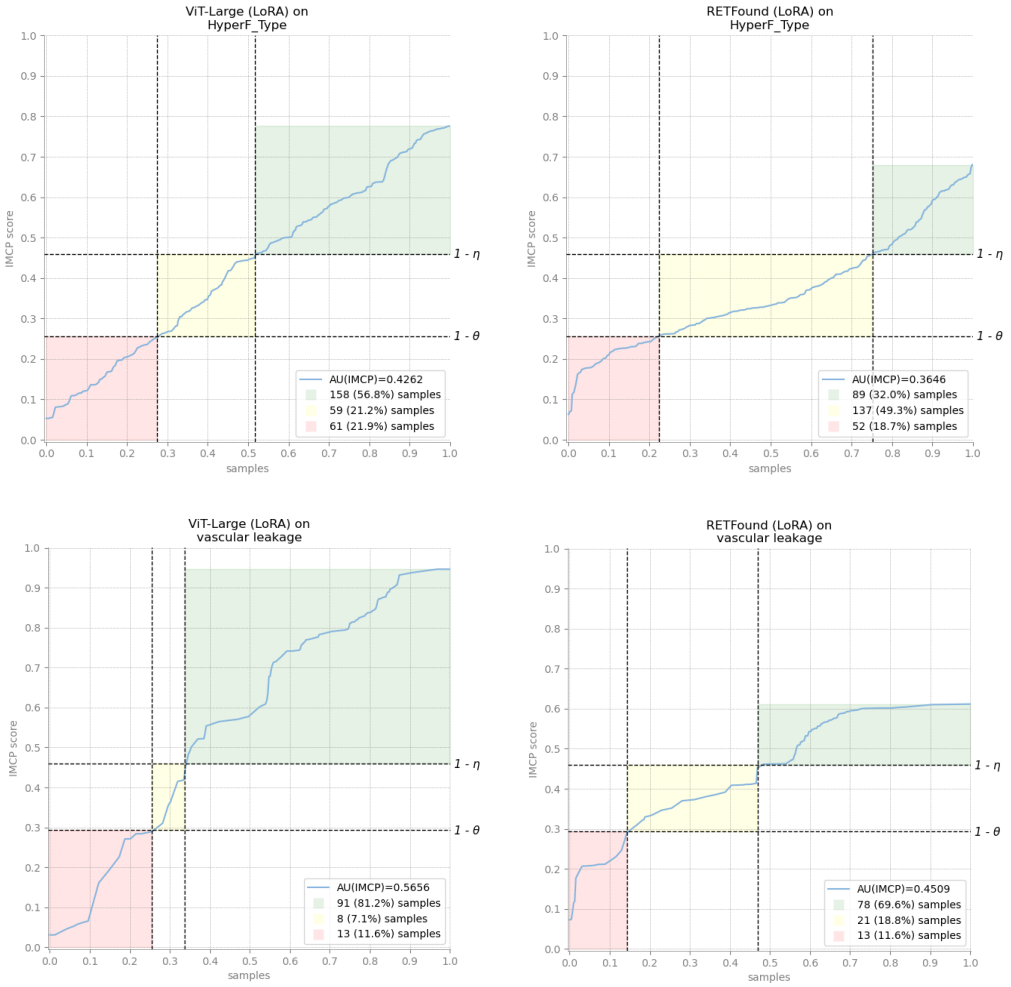

Figure 4.11: IMCP curves of ViT-Large (on the left) and RETFound (on the right) on two multiclass classification problem, showing the difference in model's confidence caused by LoRA training.

performance across tasks.

For each model and each metric, the mean value across all tasks was computed using the *inverse-variance weighting* [61], which allows higher influence from tasks where the metric estimates are more stable. Concretely, let m_i denote the average value of a given metric on task i, computed over k=5 splits, and let σ_i denote the corresponding standard deviation. Then, the aggregated score $\bar{M}_{\text{weighted}}$ for each metric is given by:

$$\bar{M}_{\text{weighted}} = \frac{\sum_{i=1}^{n} w_i \cdot m_i}{\sum_{i=1}^{n} w_i}, \quad \text{with} \quad w_i = \frac{1}{\sigma_i^2 + \varepsilon} \tag{4.2}$$

where n is the number of tasks and ε is a small positive constant introduced to avoid division by zero. This formulation ensures that metric values with lower uncertainty (i.e., smaller variance across splits) contribute more significantly to the final aggregated score.

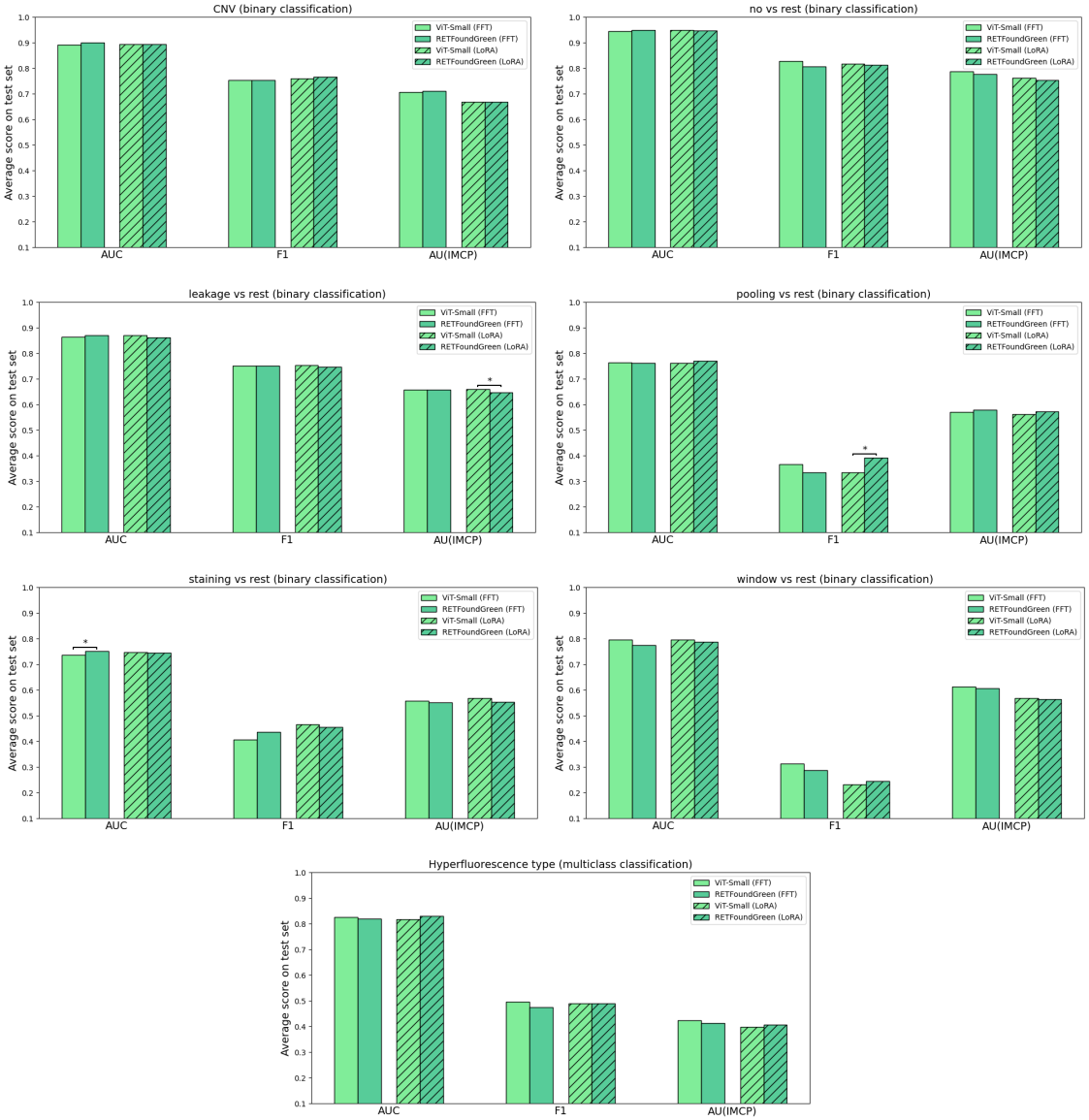

Figure 4.12: Overview of the FFT and PEFT results for RETFound Green and ViT-Small obtained on the 3rd Aptos competition dataset. (*) the asterisk indicates a p-value < 0.05.

Once the three metrics are aggregated in this manner, they are plotted on a radar chart to visually compare the models. The area of the polygon formed by the three points (each corresponding to one metric) is then computed and used as an overall score for model comparison.

Although this method does not possess a formal statistical foundation for model selection, it provides a compact and interpretable summary of performance in the presence of multiple, heterogeneous tasks and evaluation dimensions.

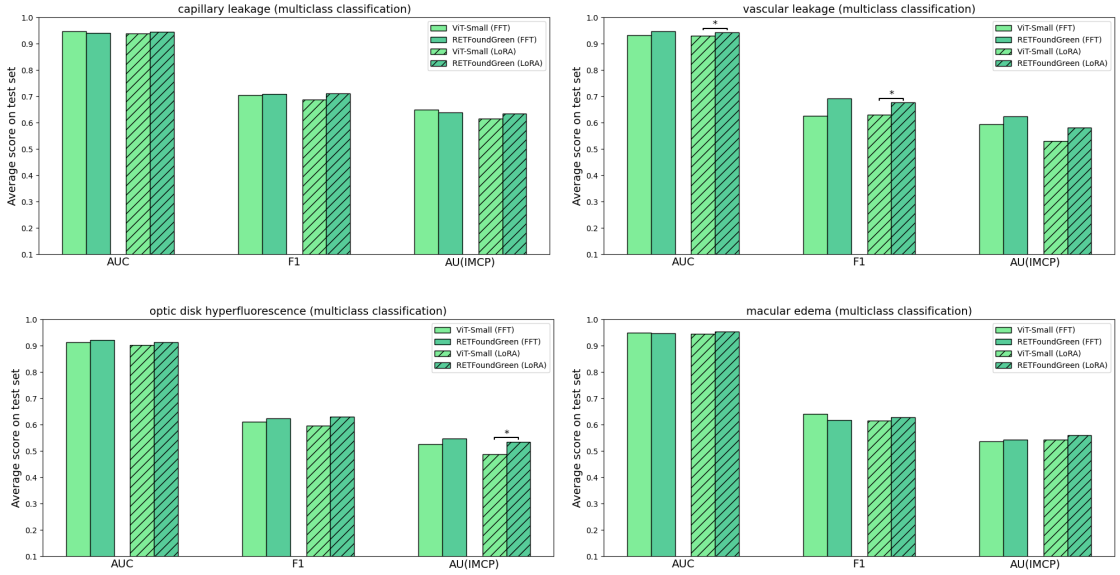

Figure 4.13: Overview of the FFT and PEFT results for RETFound Green and ViT-Small obtained on the HOJG dataset. (*) the asterisk indicates a p-value < 0.05.

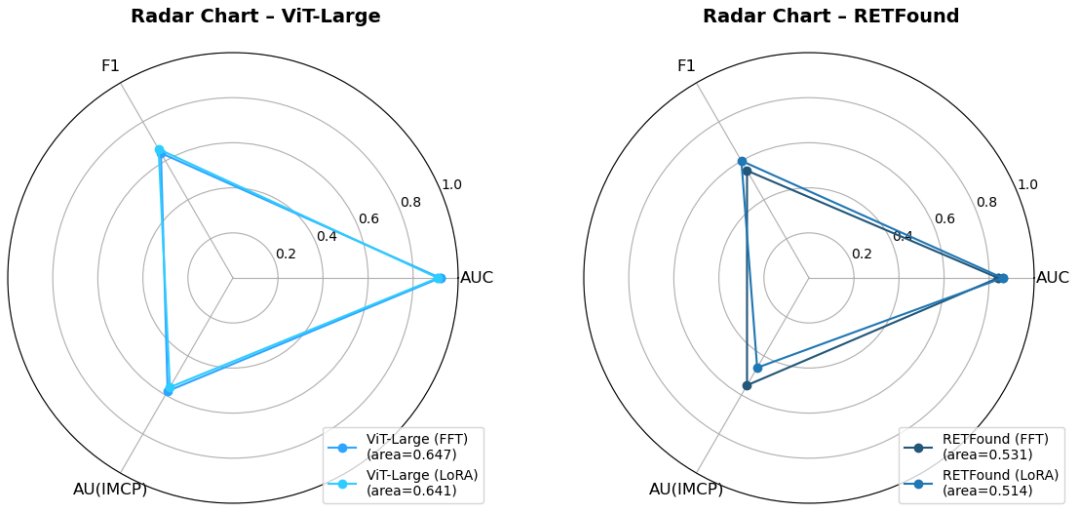

Figure 4.14: Radar charts of ViT-Large (on the left) and RETFound (on the right) showing the quantitive differences in performance between FFT and PEFT.

As shown in Fig. 4.14 the ViT-Large wins over RETFound and has a major resistence to LoRA, in fact there is almost no difference with respect to the behaviour obtained through a FFT.

The comparison between RETFound Green and its baseline, Fig. 4.15, shows the retinal foundation model emerging as the winner, achieving a final score of 0.708. However, it also appears to be less robust under LoRA-based PEFT, showing a 7.91% degradation

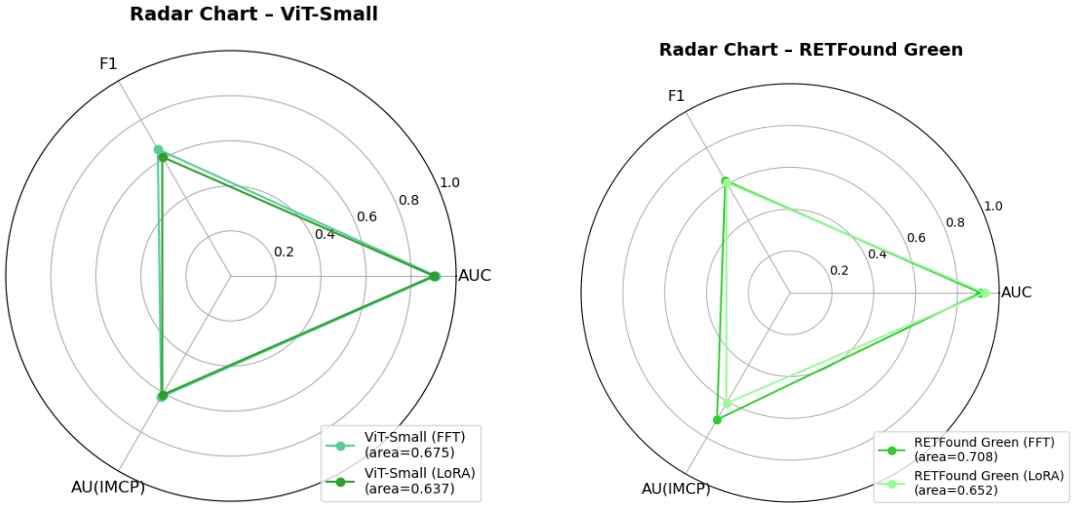

Figure 4.15: Radar charts of ViT-Small (on the left) and RETFound Green (on the right) showing the quantitive differences in performance between FFT and PEFT.

in performance compared to its own FFT results.

Effectiveness of continual pre-training through SSL

Figures 4.16 and 4.17 present the final performance outcomes of RETFound and RETFound Green, respectively, after the additional SSL step conducted on the 9,608 FA unlabeled images from the Aptos dataset.

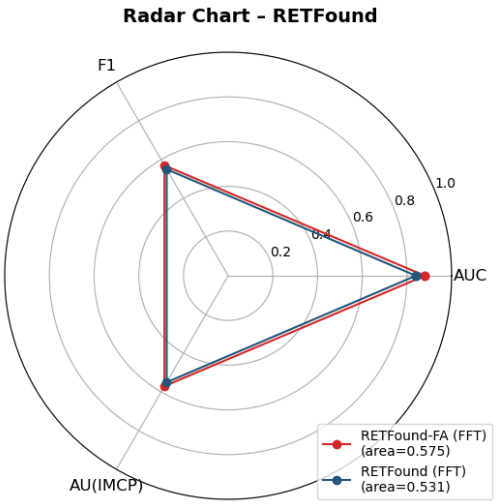

Figure 4.16: Radar chart showing the difference in performance of RETFound before (in blue) and after (in red) making an extra MAE step on FA images.

As previously discussed in Subsec. 4.3.4, The substantial 360-degree performance

gains observed for RETFound are clearly reflected in the corresponding radar chart. In particular, the model exhibits an overall performance improvement of 8.29%, increasing from an area of 0.531 to 0.575.

In contrast, RETFound Green demonstrates a 6.78% decline in performance following the same SSL step. Although this negative trend was not immediately apparent in the tabular results, it is primarily attributed to high variance across tasks, which reduces the relative impact of minor improvements.

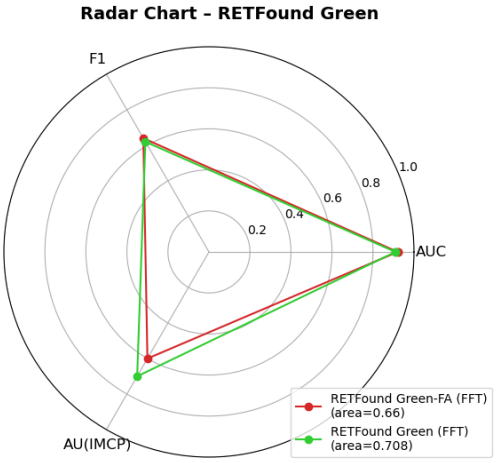

Figure 4.17: Radar chart showing the difference in performance of RETFound Green before (in green) and after (in red) making an extra Token reconstruction step on FA images.

These findings, already highlighted in Subsec. 4.4.4, suggest the need for further investigation into more robust and consistent strategies for extra pre-training through Token Reconstruction. The current approach appears to introduce considerable variability, ultimately impairing the overall effectiveness of the model.

Chapter 5

Conclusion

This study aimed to evaluate the effectiveness of retinal foundation models in the context of fluorescein angiography, with a particular focus on understanding their cross-domain adaptation capabilities. Specifically, the work investigated whether leveraging knowledge from a closely related domain, color fundus photography, through retinal-specific foundation models such as RETFound and RETFound Green could offer tangible benefits compared to transfer learning from models pre-trained on natural images.

The research was guided by three key hypotheses, each motivated by a specific rationale:

1. **Effectiveness of Retinal Foundation models**. Leveraging inductive biases from related retinal modalities may improve downstream performance on FA tasks compared to using generic models pre-trained on natural images.

2. **LoRA for efficient adaptation**. Reducing the number of trainable parameters through methods like LoRA could offer a more efficient alternative to full fine-tuning, particularly important in medical applications with limited data or compute.

3. **The role of Self-Supervised Learning for cross-domain adaptation**. Incorporating SSL techniques during continual pre-training may help models learn features that generalize better across imaging modalities, thereby enhancing cross-domain adaptation.

To investigate these hypotheses, two FA datasets were employed: the publicly available 3rd Aptos competition dataset, which offers a diverse range of tasks for the detailed description of the type and patterns of abnormal fluorescence; and the HOJG dataset, which covers some clinically relevant tasks in a real word scenario for the classification of four key inflammatory signs connected to uveitis. A total of four models were analyzed, RETFound, RETFound Green, and their respective baselines, using three different adaptation strategies: full fine-tuning, parameter-efficient fine-tuning via LoRA and continual pre-training through self-supervised learning. These techniques enabled a thorough evaluation of each model's ability to extract and adapt meaningful representations for FA tasks.

Based on the experiments and results discussed in the previous chapters, three key takeaways emerge.

5.1 On the effectiveness of Retinal Foundation models

While RETFound and RETFound Green have demonstrated strong performance on CFP tasks, there is limited evidence that this advantage transfers effectively to FA.

For instance, RETFound underperformed compared to its baseline (ViT-Large pretrained on ImageNet-21K), suggesting that the knowledge it encodes, though connected to the structure of the retina, may not generalize adequately to FA. However, the opposite conclusion cannot be firmly established either. In fact, the comparison between RETFound Green and its ViT-Small baseline tells a different story: not only does RETFound Green outperform it, but it also achieves the best overall performance across all models and metrics, consistently ranking highest on AUC, F1 score, and AU(IMCP) across the 11 tasks.

These contrasting outcomes suggest that the effectiveness of domain-specific pretraining may depend not only on the choice of source domain, but also on other factors, potentially including the SSL strategy, the scale and quality of the datasets used, and the degree of alignment between source and target domains, which should be further investigated in future research.

5.2 On the use of LoRA for efficient adaptation

LoRA, the most popular PEFT method, was analyzed to assess its applicability in the medical imaging domain, where its use remains underexplored.

Throughout this study, LoRA demonstrated strong efficiency, maintaining the discriminative power of models while updating only a very small fraction of parameters (between 1% and 4% for larger models). However, the results indicate that it is particularly effective when applied to large-scale models. In these cases, it not only provides significant computational and storage savings, but also preserves final model performance, as seen with the ViT-Large, where LoRA-based fine-tuning resulted in only a 0.93% performance degradation compared to FFT. In contrast, the smaller models appeared more sensitive to the PEFT technique, experiencing more pronounced declines in confidence and prediction stability. This suggests that LoRA is better suited for high-capacity models, where its lightweight updates suffice to retain representational strength.

5.3 On the role of SSL for cross-domain adaptation

Finally, this thesis work assessed the impact of self-supervised learning techniques, specifically Masked Autoencoder and Token Reconstruction, as strategies for continual pretraining to enhance cross-domain adaptation.

The results indicate that both SSL strategies improved model performance on FA tasks, including on the dataset not seen during the SSL phase. Although improvements were more consistent for MAE, both methods contributed to learning richer FA-specific

features that generalized beyond the training distribution, as shown by the linear probing evaluation.

These findings reinforce the value of SSL in scenarios where there is a domain shift between the pre-training and downstream tasks. By equipping models with better representations, SSL serves as a key enabler for cross-domain adaptation, particularly in underrepresented medical imaging modalities.

5.4 Final remarks

Taken together, the results of this study provide valuable insights into the strengths and limitations of foundation models and adaptation strategies in the FA setting. While the benefits of pre-training on related domains are not universally observed, techniques like SSL and PEFT play a critical role in enhancing adaptability and efficiency.

At the same time, several limitations should be acknowledged, which may have influenced the outcomes and should inform the interpretation of the results. First, despite utilizing two of the most extensive FA datasets available to date, the effective sample size was limited. This was due to the methodological decision to use only the final frame of each angiographic sequence, reducing the total data from a large pool of images to a relatively small number of unique eyes (1,877 for Aptos and 752 for HOJG). This restriction was necessary to maintain consistency across samples but inevitably constrained the amount of training data available for each task.

Secondly, differences in acquisition protocols between the datasets may have introduced additional variability. The Aptos dataset predominantly uses 35-degree field-of-view images, while HOJG focused exclusively on 55-degree images for the inflammatory signs classification task. Moreover, the latter was carefully curated through a selection process aimed at including only samples with minimal confounding factors.

Finally, the hyperparameter tuning process relied on heuristic adjustments rather than systematic optimization using fixed search spaces or automated tools. This choice was driven by the computational demands of large-scale vision models, which limited the feasibility of more exhaustive search strategies. As a result, some configurations may not reflect optimal parameter settings.

To fully understand the potential of domain-specific pre-training, future research should investigate the factors that influence its effectiveness, whether it stems from the relatedness of the source domain, the pre-training strategy employed, or inherent limitations in the ability of foundation models to learn real exam-agnostic representations. In parallel, it is essential to assess the generalization capabilities of these models on external datasets that differ from those used during fine-tuning. Such evaluations would provide insight into potential performance shifts introduced by adaptation techniques like PEFT. Finally, exploring a broader set of adapter layer designs and configurations may reveal more effective strategies for cross-domain transfer, ultimately contributing to the development of robust and scalable solutions in medical imaging.

Bibliography

[1] Wikipedia contributors. *Artificial intelligence — Wikipedia, The Free Encyclopedia*. In: 2025. URL: https://en.wikipedia.org/w/index.php?title=Artificial_intelligence&oldid=1282286553 (visited on 03/26/2025).

[2] A. M. Turing. "Computing machinery and intelligence". In: *Mind* 59 (1950), pp. 433–460. ISSN: 1460-2113. DOI: 10.1093/mind/LIX.236.433.

[3] *Artificial Intelligence: A Modern Approach | BibSonomy*. URL: https://www.bibsonomy.org/bibtex/53908a52dd4c6c8e39f93f4ffc8341be (visited on 03/24/2025).

[4] M. I. Jordan and T. M. Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349 (July 17, 2015), pp. 255–260. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaa8415.

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: http://www.deeplearningbook.org.

[6] *Difference in Data Mining Vs Machine Learning Vs Artificial Intelligence*. URL: https://www.softwaretestinghelp.com/data-mining-vs-machine-learning-vs-ai/ (visited on 03/26/2025).

[7] Pranav Rajpurkar et al. "AI in health and medicine". In: *Nature Medicine* 28 (Jan. 2022), pp. 31–38. ISSN: 1546-170X. DOI: 10.1038/s41591-021-01614-0.

[8] Emma Beede et al. "A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–12. ISBN: 9781450367080. DOI: 10.1145/3313831.3376718.

[9] Zhiyuan Gao et al. "Automatic interpretation and clinical evaluation for fundus fluorescein angiography images of diabetic retinopathy patients by deep learning". In: *The British Journal of Ophthalmology* 107 (Nov. 22, 2023). ISSN: 1468-2079. DOI: 10.1136/bjo-2022-321472.

[10] Risa M. Wolf et al. "Cost-effectiveness of Autonomous Point-of-Care Diabetic Retinopathy Screening for Pediatric Patients With Diabetes". In: *JAMA ophthalmology* 138.10 (Oct. 1, 2020), pp. 1063–1069. ISSN: 2168-6173. DOI: 10.1001/jamaophthalmol.2020.3190.

[11] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. July 12, 2022. DOI: 10.48550/arXiv.2108.07258. arXiv: 2108.07258.

[12] Theodora Tsirouki et al. "A Focus on the Epidemiology of Uveitis". In: *Ocular immunology and inflammation* 26 (July 2016), pp. 1–15. DOI: 10.1080/09273948.2016.1196713.

[13] Ashish Vaswani et al. *Attention Is All You Need*. Aug. 2, 2023. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762.

[14] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. June 3, 2021. DOI: 10.48550/arXiv.2010.11929. arXiv: 2010.11929.

[15] Neil Houlsby et al. *Parameter-Efficient Transfer Learning for NLP*. 2019. arXiv: 1902.00751 [cs.LG]. URL: https://arxiv.org/abs/1902.00751.

[16] Lingling Xu et al. *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment*. 2023. arXiv: 2312.12148 [cs.CL]. URL: https://arxiv.org/abs/2312.12148.

[17] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. Oct. 16, 2021. DOI: 10.48550/arXiv.2106.09685. arXiv: 2106.09685.

[18] Shekoofeh Azizi et al. *Robust and Efficient Medical Imaging with Self-Supervision*. 2022. arXiv: 2205.09723 [cs.CV]. URL: https://arxiv.org/abs/2205.09723.

[19] Elijah Cole et al. *When Does Contrastive Visual Representation Learning Work?* May 2021. DOI: 10.48550/arXiv.2105.05837.

[20] Kaiming He et al. *Masked Autoencoders Are Scalable Vision Learners*. 2021. arXiv: 2111.06377 [cs.CV]. URL: https://arxiv.org/abs/2111.06377.

[21] Justin Engelmann and Miguel O. Bernabeu. *Training a high-performance retinal foundation model with half-the-data and 400 times less compute*. Sept. 22, 2024. DOI: 10.48550/arXiv.2405.00117. arXiv: 2405.00117.

[22] Yukun Zhou et al. "A foundation model for generalizable disease detection from retinal images". In: *Nature* 622.7981 (Oct. 2023), pp. 156–163. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06555-x.

[23] Hiroshi Keino et al. "Quantitative Analysis of Retinal Vascular Leakage in Retinal Vasculitis Using Machine Learning". In: *Applied Sciences* 12 (2022). ISSN: 2076-3417. DOI: 10.3390/app122412751.

[24] LeAnne H. Young et al. "Automated Detection of Vascular Leakage in Fluorescein Angiography – A Proof of Concept". In: *Translational Vision Science & Technology* 11 (July 2022), p. 19. ISSN: 2164-2591. DOI: 10.1167/tvst.11.7.19.

[25] Victor Amiot et al. *Automatic Transformer-Based Grading of Multiple Retinal Inflammatory Signs on Fluorescein Angiography*. Sept. 24, 2024. DOI: 10.2139/ssrn.4960069.

[26] Weiyi Zhang et al. "Angiographic Report Generation for the 3rd APTOS's Competition: Dataset and Baseline Methods". In: *medRxiv* (2023). DOI: 10.1101/2023.11.26.23299021.

[27] *Interpretation - Ophthalmic Photographers' Society*. URL: https://www.opsweb.org/page/FAinterpretation (visited on 03/26/2025).

[28] EyeGuru. *How to interpret fluorescein angiography: 6 types of defects*. June 26, 2016. URL: https://eyeguru.org/essentials/fluorescein-angiography/ (visited on 03/26/2025).

[29] BrightFocus Foundation. *What is Choroidal Neovascularization?* URL: https://www.brightfocus.org/resource/what-is-choroidal-neovascularization/ (visited on 03/26/2025).

[30] Ilknur Tugal-Tutkun, Carl Herbort, and Moncef Khairallah. "Scoring of dual fluorescein and ICG inflammatory angiographic signs for the grading of posterior segment inflammation (dual fluorescein and ICG angiographic scoring system for uveitis)". In: *International ophthalmology* 30 (Oct. 2008), pp. 539–52. DOI: 10.1007/s10792-008-9263-x.

[31] Wikipedia contributors. *Macular edema — Wikipedia, The Free Encyclopedia*. In: 2025. URL: https://en.wikipedia.org/w/index.php?title=Macular_edema&oldid=1276182647 (visited on 03/26/2025).

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. *Representation Learning with Contrastive Predictive Coding*. 2019. arXiv: 1807.03748 [cs.LG]. URL: https://arxiv.org/abs/1807.03748.

[33] Shekoofeh Azizi et al. *Big Self-Supervised Models Advance Medical Image Classification*. Apr. 1, 2021. DOI: 10.48550/arXiv.2101.05224. arXiv: 2101.05224.

[34] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (June 1, 2006), pp. 861–874. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2005.10.010.

[35] J. P. Egan. *Signal detection theory and ROC analysis*. Series in Cognition and Perception. New York, NY: Academic Press, 1975.

[36] J. A. Swets. "Measuring the accuracy of diagnostic systems". In: *Science (New York, N.Y.)* 240 (June 3, 1988), pp. 1285–1293. ISSN: 0036-8075. DOI: 10.1126/science.3287615.

[37] Kent A. Spackman. "SIGNAL DETECTION THEORY: VALUABLE TOOLS FOR EVALUATING INDUCTIVE LEARNING". In: *Proceedings of the Sixth International Workshop on Machine Learning*. Ed. by Alberto Maria Segre. San Francisco (CA): Morgan Kaufmann, 1989, pp. 160–163. ISBN: 978-1-55860-036-2. DOI: https://doi.org/10.1016/B978-1-55860-036-2.50047-3.

[38] *Multiclass Receiver Operating Characteristic (ROC)*. scikit-learn. URL: https://scikit-learn/stable/auto_examples/model_selection/plot_roc.html (visited on 03/26/2025).

[39] Jesús S. Aguilar-Ruiz and Marcin Michalak. "Multiclass Classification Performance Curve". In: *IEEE Access* 10 (July 2022). ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3186444.

[40] Jesús Aguilar-Ruiz and Marcin Michalak. "Classification performance assessment for imbalanced multiclass data". In: *Scientific Reports* 14 (May 10, 2024). ISSN: 2045-2322. DOI: 10.1038/s41598-024-61365-z.

[41] Wikipedia contributors. *Hellinger distance — Wikipedia, The Free Encyclopedia*. In: 2025. URL: https://en.wikipedia.org/w/index.php?title=Hellinger_distance&oldid=1278761766 (visited on 03/26/2025).

[42] Oona Rainio, J. Teuho, and Riku Klén. "Evaluation metrics and statistical tests for machine learning". In: *Scientific Reports* 14 (Mar. 2024). DOI: 10.1038/s41598-024-56706-x.

[43] Wikipedia contributors. *Cross-validation (statistics) — Wikipedia, The Free Encyclopedia*. In: 2025. URL: https://en.wikipedia.org/w/index.php?title=Cross-validation_(statistics)&oldid=1276518877 (visited on 03/26/2025).

[44] *RGB — Torchvision main documentation*. URL: https://pytorch.org/vision/main/generated/torchvision.transforms.v2.RGB.html (visited on 03/26/2025).

[45] Giovanni Angelotti et al. *A Closer Look at AUROC and AUPRC under Class Imbalance*. May 2024.

[46] F. Pedregosa et al. *Scikit-learn: Machine Learning in Python*. `compute_class_weight` from sklearn.utils. 2011. URL: https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html (visited on 03/26/2025).

[47] Sangdoo Yun et al. *CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features*. 2019. arXiv: 1905.04899 [cs.CV]. URL: https://arxiv.org/abs/1905.04899.

[48] Dan Hendrycks et al. *AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty*. 2020. arXiv: 1912.02781 [stat.ML]. URL: https://arxiv.org/abs/1912.02781.

[49] Shih-Cheng Huang et al. "Self-supervised learning for medical image classification: a systematic review and implementation guidelines". In: *npj Digital Medicine* 6 (Apr. 26, 2023). ISSN: 2398-6352. DOI: 10.1038/s41746-023-00811-0.

[50] Evgin Goceri. "Medical image data augmentation: techniques, comparisons and interpretations". In: *Artificial Intelligence Review* 56 (Mar. 2023), pp. 1–45. DOI: 10.1007/s10462-023-10453-z.

[51] *Transforming and augmenting images — Torchvision main documentation*. URL: https://pytorch.org/vision/main/transforms.html (visited on 03/26/2025).

[52] F. Pedregosa et al. *Scikit-learn: Machine Learning in Python*. `StandardScaler` from sklearn.preprocessing. 2011. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (visited on 03/26/2025).

[53] Takuya Akiba et al. *Optuna: A Next-generation Hyperparameter Optimization Framework*. 2019. URL: https://optuna.org/.

[54] Tal Ridnik et al. *ImageNet-21K Pretraining for the Masses*. Apr. 2021. DOI: 10.48550/arXiv.2104.10972. arXiv: 2104.10972.

[55] Damjan Kalajdzievski. *A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA*. Nov. 28, 2023. DOI: 10.48550/arXiv.2312.03732. arXiv: 2312.03732.

[56] Dan Biderman et al. *LoRA Learns Less and Forgets Less*. Sept. 20, 2024. DOI: 10.48550/arXiv.2405.09673. arXiv: 2405.09673.

[57] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. Feb. 2, 2024. DOI: 10.48550/arXiv.2304.07193. arXiv: 2304.07193.

[58] Colorado J. Reed et al. *Self-Supervised Pretraining Improves Self-Supervised Pretraining*. Mar. 25, 2021. DOI: 10.48550/arXiv.2103.12718. arXiv: 2103.12718.

[59] Yuqi Sun et al. "A data-efficient strategy for building high-performing medical foundation models". In: *Nature Biomedical Engineering* (Mar. 5, 2025). ISSN: 2157-846X. DOI: 10.1038/s41551-025-01365-0.

[60] Reece Shuttleworth et al. *LoRA vs Full Fine-tuning: An Illusion of Equivalence*. Oct. 28, 2024. DOI: 10.48550/arXiv.2410.21228. arXiv: 2410.21228.

[61] Wikipedia contributors. *Weighted arithmetic mean — Wikipedia, The Free Encyclopedia*. In: 2025. URL: https://en.wikipedia.org/w/index.php?title=Weighted_arithmetic_mean&oldid=1271447623 (visited on 03/26/2025).