A Vision Language Correlation Framework for Screening Disabled Retina

Taimur Hassan, *IEEE Senior Member*, Hina Raja, Kais Belwafi, Samet Akcay, Mohamed Jleli, Bessem Samet, Naoufel Werghi *IEEE Senior Member*, Jawad Yousaf *IEEE Senior Member*, Mohammed Ghazal, *IEEE Senior Member*

***Abstract*—Retinopathy is a group of retinal disabilities that causes severe visual impairments or complete blindness. Due to the capability of optical coherence tomography to reveal early retinal abnormalities, many researchers have utilized it to develop autonomous retinal screening systems. However, to the best of our knowledge, most of these systems rely only on mathematical features, which might not be helpful to clinicians since they do not encompass the clinical manifestations of screening the underlying diseases. Such clinical manifestations are critically important to be considered within the autonomous screening systems to match the grading of ophthalmologists within the clinical settings. To overcome these limitations, we present a novel framework that exploits the fusion of vision language correlation between the retinal imagery and the set of clinical prompts to recognize the different types of retinal disabilities. The proposed framework is rigorously tested on six public datasets, where, across each dataset, the proposed framework outperformed state-of-the-art methods in various metrics. Moreover, the clinical significance of the proposed framework is also tested under strict blind testing experiments, where the proposed system achieved a statistically significant correlation coefficient of 0.9185 and 0.9529 with the two expert clinicians. These blind test experiments highlight the potential of the proposed framework to be deployed in the real world for accurate screening of retinal diseases.**

***Index Terms*—Vision Language Models, Optical Coherence Tomography, Retina, Ophthalmology.**

I. Introduction

Human vision is formed when the light passes through the biconvex lens and falls on the retina [1]. The retina is a complex membrane composed of several light-sensitive cells. Once an object is perceived, the retina converts its projection into a sequence of electrical impulses. Subsequently, these impulses are sent to the visual cortex within the brain, where they are processed and translated as visual information [2].

The authors extend their appreciation to the King Salman Center For Disability Research for funding this work through Research Group no KSRG-2023-540.

T. Hassan, J. Yousaf, and M. Ghazal are with the Department of Electrical, Computer, and Biomedical Engineering, Abu Dhabi University, UAE. (email: {taimur.hassan, jawad.yousaf, mohammed.ghazal}@adu.ac.ae).

H. Raja is with the Department of Ophthalmology, University of Tennessee Health Center, USA. (email: hraja@uthsc.edu).

K. Belwafi is with Department of Computer Engineering, College of Computing and Informatics, University of Sharjah, UAE. (email: kbelwafi@sharjah.ac.ae).

S. Akcay is with Intel, UK. (email: samet.akcay@intel.com).

M. Jleli, and B. Samet are with the King Salman Center For Disability Research, and the Department of Mathematics, College of Science, King Saud University, Saudi Arabia. (email: {jleli, bsamet}@ksu.edu.sa).

N. Werghi is with the Center of Cyber Physical Systems (C2PS), Khalifa University, UAE. (email: naoufel.werghi@ku.ac.ae).

Corresponding authors are Taimur Hassan and Naoufel Werghi.

The retina can be divided into two distinct regions: the macular region and the peripheral region. The macular region, also known as the macula, is responsible for generating central vision, whereas the peripheral region is responsible for generating side vision. Retinopathy is a collection of retinal microvascular disabilities linked to many clinical states, such as hypertension, sickle cell disease, lupus, and other autoimmune diseases [3]. However, it is primarily associated with Type-II Diabetes and is commonly known as Diabetic Retinopathy (DR). DR from optical coherence tomography (OCT) imagery is screened by observing the presence of Diabetic Macular Edema (DME), a condition characterized by significant thickness of the macula, leading to hazy and impaired vision [4]. DME is often characterized by observing the presence of hard exudates and the macular fluid within the intra-retina or sub-retina [5]. Similarly, macular fluid can also be observed in the presence of Age-related Macular Degeneration (AMD) [6]. AMD primarily affects older individuals as a result of the accumulation of drusen (lipid and protein deposits) and choroidal neovascularization (CNV) beneath the retina. AMD is typically categorized into two stages: dry (non-neovascular) AMD and wet (neovascular) AMD. Drusen are the deposits of extracellular material developed during the initial phases of AMD, which result in the degeneration of the retinal pigment epithelium (RPE) layer. Choroidal neovascularization (CNV) is a consequence of further disease progression, characterized by the intrusion of abnormal blood vessels from the choroid onto the retina, resulting in the formation of choroidal neovascular membranes and other abnormalities in the chorioretinal region. The assessment of these retinal disabilities is commonly conducted using color fundus photography (FP), fundus fluorescein angiography (FFA), and OCT [6]. Moreover, doctors predominantly use OCT scans for the objective diagnosis of retinal diseases [6].

A. Clinical Significance of OCT Examination

Retinal OCT examination has been extensively adopted by clinicians as a standard procedure for screening many retinal diseases, such as DME and AMD [5]. In clinical settings, OCT is currently the recommended examination method for analyzing pathologies affected by DME. This is because OCT provides precise visualization of intra-retinal fluid (IRF), sub-retinal fluid (SRF), and hard exudates (HE) [6]. Additionally, clinicians and researchers also conducted a thorough analysis related to screening AMD pathologies using OCT, FFA, and

FP, and determined that OCT scans can detect even the most subtle early changes in CNV and AMD [7]. Similarly, they also emphasized the usefulness of peripheral OCT imagery in detecting the clinical manifestations related to diseases, such as retinal holes, retinal detachment, and age-related retinoschisis [8].

B. Motivation

In the past decade, several researchers have presented computer-aided diagnostic systems for screening retinal diseases. These systems utilize fundus, OCT, and a combination of both fundus and OCT imagery [9]. Similarly, recent studies utilized attention mechanisms [10] and hybrid convolutional networks [6] to offer lesion-aware screening of retinal diseases. Nevertheless, these models still cannot accommodate clinical manifestations within the automated screening process, which hinders their ability to be deployed in clinical settings to aid clinicians in mass screening retinal diseases [11]. Furthermore, due to the lack of correlating the clinical prompts within the decision-making process, these frameworks have to be re-trained or fine-tuned exhaustively to learn prognosis tasks across different datasets and scanner specifications, which is both infeasible and impractical [9]. Therefore, to produce a scalable solution that can be deployed in clinical settings to aid clinicians in their daily screening routine, there is a dire need to develop a framework that possesses the ability to incorporate clinical manifestations within the automated screening process to generate meaningful results irrespective of the datasets, or the scanner specifications.

C. Contributions

This paper presents a novel approach to correlate clinical manifestations with the deep image features to accurately screen retinal disabilities from the OCT imagery as per the clinical standards. Unlike state-of-the-art methods, the proposed framework does not require any re-training, fine-tuning, or domain adaptation process to recognize different retinal abnormalities from the OCT scans. Furthermore, the proposed framework can robustly identify retinal diseases irrespective of the scanner specifications, the pathological variations and the vendor artifacts. Additionally, the proposed framework has been rigorously tested on six public OCT datasets, where it outperformed state-of-the-art methods for screening retinal diseases (see Section V for more details). To summarize, the main contributions of this paper are:

- A vision-language model that integrates clinical prompts with the deep image features via vision-language correlations to screen retinal disabilities from OCT imagery.
- A robust framework that can effectively screen retinal diseases from diverse datasets by analyzing their clinically significant manifestations in a self-supervised manner.
- Since the proposed scheme leverages self-supervision, it does not require extensive fine-tuning or retraining efforts for screening retinal diseases.
- An extensively tested approach under clinical settings through a series of blind test experiments in which the proposed system achieves a high statistically significant correlation with the grading of the expert clinicians (see Section V-D for more details).

D. Paper Organization

The rest of the paper is organized as follows: The literature review is presented in Section II. Section III discusses the proposed framework in detail. Section IV enlists the experimental protocols. Section V presents the results obtained by the proposed framework and its performance comparison with state-of-the-art methods. Section VI discusses the proposed framework and its applicability in the clinical settings for screening retinal disabilities. Section VII concludes the paper and highlights the prospects of the proposed approach in different applications.

II. Related Work

Retinal image analysis is a widely explored area where researchers and clinicians have effectively employed various machine learning models to detect different retinal disorders using fundus [1] and OCT scans [12]. From a clinical perspective, OCT imagery is preferred over fundus photography because it possesses the capacity to detect early pathological changes in the human retina [9]. In this section, we present a detailed discussion of some of the recent advancements related to the automated screening of retinal diseases.

A. Autonomous Retinal Analysis Frameworks

Over the past few decades, many researchers have proposed automated methods to screen retinal diseases using OCT scans effectively. The initial wave of these methods relied on conventional machine learning [13], which uses handcrafted features to detect the presence of retinal diseases [12]. However, because of the inherent subjectivity within these features, the conventional machine-learning approaches were only applicable to limited datasets and experimental conditions. The deep learning models effectively mitigated this constraint by enhancing the performance of traditional machine learning methods [6]. Fang et al. [14] were among the first ones to use CNN-based methods to segment retinal layers from OCT images of non-neovascular AMD patients. Similarly, by adhering closely to established ophthalmological examination protocols in clinical settings, advanced attention mechanisms were developed based on activation maps and rules-based schemes [6]. These mechanisms, as proposed in [10] and [6], allow deep learning models to highlight the underlying lesion pathologies, enabling lesion-aware screening and grading of retinopathy. Butola et al. [15] developed a specialized CNN model called LightOCT [15]. This model consists of two convolutional layers and a fully connected layer, and it is designed to analyze OCT scans for screening AMD, DME, and healthy subjects. Moreover, Das et al. [16] employed adversarial networks to enhance the resolution of OCT scans, aiming to improve the accuracy of AMD diagnosis.

B. Advanced Retinal Analysis Frameworks

Many researchers have utilized deep-learning models for diagnosing retinal diseases. However, most of these frameworks are fully supervised. Consequently, they necessitate extensive and well-annotated data for training to achieve promising results [6], [10]. Such a requirement makes these models infeasible and impractical for deployment in the real world. Furthermore, these models lack the ability to incorporate clinical manifestations within the decision-making process, where these clinical observations are necessary to replicate the screening standards of clinicians and ophthalmologists. To overcome the first limitation, many researchers have proposed weakly supervised schemes that do not require extensive annotation efforts to screen retinal diseases [17]. Ma et al. [18] devised a method using weak supervision and multi-scale class activation maps. They employed scaling and upsampling techniques and attentional fully linked modules to accurately identify and extract geographic atrophy (GA) lesions from OCT imagery.

Weakly supervised approaches have partially alleviated the need for extensive ground truth labeling [18]. However, these methods either exhibit limited screening capabilities [19] or necessitate exhaustive training procedures to identify retinal diseases effectively [6], [10], [15], [20]. Similarly, to detect new types of retinal abnormalities obtained from different scanning devices, these models must undergo incremental training iterations on large-scale data, which is again not a scalable option [10]. To address the challenge of re-training the model using extensive data, researchers have utilized incremental few-shot learning [17] approaches. However, striking the optimal balance of retaining the model's prior knowledge while learning new knowledge representation to avoid catastrophic forgetting is again a haggling job [17], and might not achieve convergence across multi-vendor and multi-modal setup established within the clinical settings. To overcome these limitations, we present a novel approach to incorporate clinical manifestations within the machine learning systems via vision language correlations in order to identify retinal disabilities in the clinical settings irrespective of the pathological variations, the scanner differences, or vendor artifacts. Furthermore, through rigorous evaluation of the proposed scheme across six public datasets and blind-test experiments in clinical settings, we showcased that the proposed framework can remarkably screen retinal diseases compared to different state-of-the-art methods. A detailed discussion of the contributions of the proposed framework is presented in Section I-C.

C. Large Language Models in Retinal Analysis

Recently, many clinicians highlighted the potential of using large language models (LLMs) to screen retinal disabilities within clinical settings. Desideri et al. [21] studied the effects of using LLMs, such as ChatGPT 3.5 [22], Google Bard [23], and Bing AI [24], in monitoring the progression of AMD. They concluded that ChatGPT 3.5 [22] produced the most satisfactory results against the specialized questions related to monitoring the progression of AMD as compared to the other LLM models. Apart from this, Jin et al. [25] presented an overview on the usage of LLM models in managing various diseases in ophthalmology and conclude that LLMs like ChatGPT [22] are appreciated by different clinicians around the world to screen retinal and other ophthalmic diseases.

By digging into the literature, we observed that the LLMs, especially ChatGPT 3.5 [22], and ChatGPT 4.0 [26], can produce clinically significant prompts related to retinal disabilities. They also provide excellent opportunities to the researchers to use these prompts in training the deep learning models to accurately screen the retinal diseases irrespective of the scanner specification, the datasets, and the pathological variations [21], [25], [27]–[30].

III. Proposed Framework

The detailed block diagram of the proposed framework is shown in Figure 1. During the training phase, the text encoder is fine-tuned on a set of clinically verified retinal text prompts. These text prompts highlight the presence of disease-specific retinal lesions and abnormalities that appear on the retinal OCT scans during the early progression of the disease. Once the text encoder is fine-tuned, we use it to train the vision encoder in a self-supervised manner using the proposed L_{as} loss function. The L_{as} loss function transforms the features from latent space to angular space and then tries to minimize the distance between positives and anchors while maximizing the distance between negatives and anchors in the angular space. Positives are the text prompts belonging to the positive disease class, where their features are extracted using the text encoder. Negatives are the text prompts not belonging to the positive class, where the text encoder extracts their features. Similarly, anchors are the given OCT samples belonging to the positive class, and the vision encoder extracts features from them. It should be noted here that the loss values generated by the L_{as} directly showcase the deviation between the features of the text prompts and the features of the vision encoder for the given retinal disease (positive class), and minimizing these loss values allow the vision encoder to produce discriminative features from the OCT scans, irrespective of their scanner specifications, that are self-aligned with the features generated from the clinically verified text prompts. Hence, these visual features, when aligned with the textual features generated from the prompts, allow clinically significant screening of retinal diseases at the inference stage. Moreover, once the vision encoder is trained, we use it to extract distinct features from the OCT scans belonging to any dataset or acquired from different OCT machinery. We also use the fine-tuned text encoder to generate text features from the default prompt, in which we vary the lesion and disease label iteratively. These 'lesion' and 'label' words are taken from the lesion and pathology vocabulary created by the expert clinicians, and for each combination of lesion and label word in the default text prompt, we compute the vision language correlation coefficient. The combination of text and visual features that give the maximum correlation is then used to predict the overall disease label for the given OCT scan. The detailed description of each block within the training and testing phase is discussed in the subsequent section below:

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3462653

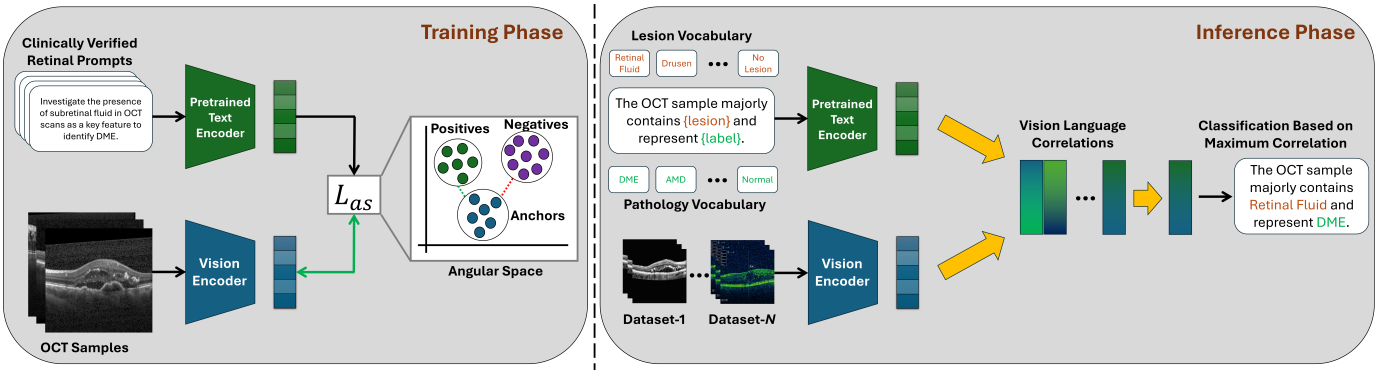

Figure 1: Block diagram of the proposed framework. During the training phase, we fine-tune the text encoder to generate discriminative features from the clinical prompts, which allows the vision encoder to learn the clinical manifestations from the OCT scans and recognize the disease patterns. The vision encoder learns the clinical manifestations of the underlying disease using the proposed L_{as} loss function, which inherently minimizes the distance between anchors and positive samples while maximizing the distance between anchors and negative samples in the angular space. During the inference phase, we use a single prompt with different combinations of words from the lesion and pathological vocabularies to generate textual features. Similarly, we extract the visual embeddings from the OCT scans belonging to any dataset using the vision encoder. Afterward, we perform vision language correlations and choose the combination of the lesion and pathological labels from the text prompt that has a maximum correlation coefficient with the visual features to identify the disease label for the given test scan.

A. *Training Phase*

In the training phase, we fine-tune the text and the vision encoders. The details about fine-tuning these models are presented below:

1) Fine-Tuning Text Encoder: During the training phase, we first fine-tune the text encoder on the set of retinal prompts that are stored in the database. We used two different methods of generating and collecting the text prompts. The first method involved requesting the clinicians to generate the prompts that represent symptomatic appearances of a healthy retina in OCT imagery. The prompts also represent the presence of symptoms for some of the widespread retinal diseases, such as DME, drusen and choroidal neovascularization (CNV), in OCT imagery. Furthermore, these prompts highlight the appearance of the clinical manifestations, such as retinal lesions and biomarkers, which develop during the mild, moderate, and severe progression of these retinal diseases. The clinicians also observe these abnormalities in the OCT scans to screen these diseases objectively in clinical practice. Following the same analogy, we used these prompts as input to fine-tune the text encoder so that it generates distant and clinically significant feature representations for each disease. The total number of prompts that the clinicians generated is 1,000 for each retinal disease. Moreover, the text encoder is tuned on these prompts for 100 epochs using ADAM [31] optimizer, having an initial learning rate of 0.01 and a categorical cross-entropy loss function with a batch size of 32. It should be noted that the text encoder is fine-tuned to predict output labels from the input prompt where the output labels are normal, CNV, drusen, and DME. Moreover, the classification head mounted on the text encoder is a neural network having these four output classes. In the second method, we generated the text prompts using ChatGPT 4.0 [26], which were clinically verified by the panel of two expert clinicians. These prompts are then used to fine-tune the text encoder, in a similar manner, using the same training configurations which were used in the first approach. The results for both approaches are discussed in the ablation experiment (reported in Section V-A3). We noticed here that by using the large pool of clinically verified prompts (10,000 in total) for training the text encoder, the proposed system achieved an overall better performance to screen retinal diseases across each dataset (see Table IV). This performance improvement is evident as more prompts highlighting the presence of retinal lesions for each type of retinal disease allow the text encoder to capture more exposure to produce generalized feature representations that can accurately train the vision encoder to generate better results. Furthermore, generating these large prompts does not incur any hectic workflows, as they are generated through ChatGPT 4.0 [26] and are cross-verified by expert ophthalmologists only once. At the same time, they allow the text encoder to produce clinically significant latent space representations to screen different types of retinal diseases irrespective of the dataset, vendor artifacts, or scanner variations. Afterward, we used this fine-tuned text encoder to train the vision encoder in a self-supervised manner using the proposed L_{as} loss function. The training details of the vision encoder are presented in a subsequent section.

2) Fine-tuning Vision Encoder: Once the text encoder is fine-tuned, we use its representations as ground truth to update the weights of the vision encoder to produce similar embeddings in a self-supervised manner using the proposed L_{as} loss function. L_{as} loss function transforms the latent space representation of the text and vision encoder in the angular space where it tries to minimize the distance between the vision encoder representations (anchors) and the text encoder representations of the positive class (positives), while simultaneously maximizing the distance between anchors and negatives (the text encoder representations for the non-positive categories). Moreover, the vision encoder is fine-tuned for 100 epochs with a batch size of 32 using ADAM [31] optimizer and the L_{as} loss function. The input size of the vision encoder is $224 \times 224 \times 3$, and no data augmentation step was applied in fine-tuning both the text and vision encoders. The details

about the L_{as} loss function are presented below.

3) L_{as} Loss Function: We propose an angular space loss function L_{as} to constrain the vision encoder toward learning the retinopathy screening tasks in alignment with the feature representations generated by the text encoder from the clinical prompts. L_{as} transforms the feature representations of both text and vision encoder from latent space to angular space, where it minimizes the distance between embeddings of anchors x_a and the positives x_p while maximizing the distance between embeddings of anchors x_a and negatives x_n. The transformation of features from latent space to angular space is based on the dot product, which allows the optimal discrimination between the features of the positive and negative samples. This is because the dot product brings the homogeneous (similar) features together and separates the heterogeneous (dissimilar) features. Moreover, L_{as} requires no supervision from the ground truth labels. Instead, it constrains the vision encoder to learn retinopathy screening tasks using the text encoder representations in a self-supervised manner. Due to this self-supervision and angular space transformation of the feature representation produced by the L_{as} loss function, the vision encoder, once trained, can optimally recognize different types of retinal diseases from the OCT scans irrespective of the dataset or scanner variations at the inference stage. Mathematically, L_{as} is expressed as:

$$L_{as} = -\frac{1}{b_s}\sum_{i=0}^{b_s-1}\log\left(\frac{\alpha_i}{\alpha_i+\beta_i}\right), \tag{1}$$

where

$$\alpha_i = e^{-d(\cos(\theta_{i,a}),\cos(\theta_p))/\tau}, \tag{2}$$

$$\beta_i = e^{-d(\cos(\theta_{i,a}),\cos(\theta_n))/\tau}. \tag{3}$$

$\cos(\theta_{i,a})$ denotes the angular space representations generated from each anchor sample x_a^i by taking its dot product with the weights of the vision encoder w_v, such that $w_v.x_a^i = |w_v||x_a^i|\cos(\theta_{i,a})$. These angular space representations are administrated by a hyperparameter $\theta_{i,a}$, which is the angle between the anchor input x_a^i and the vision encoder weights w_v, and it is computed as: $\theta_{i,a} = \arccos\left(\frac{w_v.x_a^i}{|w_v||x_a^i|}\right)$. Similarly, $\cos(\theta_p)$ denotes the angular space representations generated from the positive samples x_p, and $\cos(\theta_n)$ represents the angular space representations generated from the negative samples x_n. Apart from this, τ denotes the temperature constant, b_s denotes the batch size, and the function $d(e, f)$ computes the Euclidean distance between the vectors e and f. The proposed L_{as} loss function updates the model weights w_v in each iteration by minimizing the distance between the features extracted from anchor samples x_a and the positive samples x_p while maximizing the distance between features produced from x_a and the negative samples x_n, in the angular space. Once this fine-tuning routine is completed, we get the optimal model weights w_v for the vision encoder model that generates image features for each class, which are maximally aligned with the text features produced from the clinical prompts by the text encoder. We also highlight here that the number of features generated by the text and vision encoder backbones can differ based on their architectures. However, for the optimal text encoder, BERT [32], chosen through the ablation experiments, reported in Section V-A1, the number of text features f_t are 3072. Hence, the feature set size is 1×3072. Similarly, the number of image features f_i extracted from the optimal vision encoder, PVT [33], selected through the ablation experiments reported in Section V-A2, are 512, and the size of the respective feature set is 1×512. Since the number of text and image features does not necessarily match, we resized the text features from 3072 to 512 through sub-sampling with a sampling rate of 6. In general, when the length of text and image features do not match, we resize the larger set of features (either text or image) to length n through sub-sampling, where $n = min(len(f_t), len(f_i))$.

B. Inference Phase

Once the text and vision encoders are fine-tuned, we use them to screen retinal diseases using OCT scans from different datasets. To screen the candidate OCT scans against retinal diseases, they are passed through the vision encoder that generates the visual embeddings. Moreover, a default prompt is passed to the text encoder in which we vary the lesion and label words using the lesion and pathological vocabularies, as shown in Figure 1. The text within the default text prompt is: *'The OCT sample majorly contains* {lesion} *and represent* {label}'. For each combination of 'lesion' and 'label' words in the default prompt, the text encoder generates unique textual features that highlight the clinical manifestations related to each disease. The textual features are then correlated with the visual embeddings to generate a correlation coefficient. Afterward, the 'lesion' and 'label' word pair, which produces a maximum correlation coefficient, are used to screen the underlying disease. The details about the vision language correlations are presented below:

1) Vision Language Correlations: The classification within the proposed framework is based on vision language correlations, which measure the degree of similarities between the image features f_i and the clinically significant textual features f_t to screen the retinal diseases. Moreover, for each pair of 'lesion' and 'label' words in the default prompt, we generate the textual features using the text encoder, which are correlated with the visual features generated by the vision encoder for the given OCT scan. Let j denote the number of 'lesion' words in the lesion vocabulary, and k denote the number of disease 'labels' in the pathological vocabulary. Then, for a given OCT scan, the total number of correlations the proposed scheme performs is $j \times k$. Similarly, the total number of correlations performed by the proposed framework on a given dataset is $j \times k \times m$, where m denotes the total number of OCT scans within the respective dataset. Moreover, the correlations are computed using a Pearson correlation coefficient r [34], as expressed below:

$$r = \frac{n \times (p-q)}{\sqrt{(n \times s) \times (n \times w)}}, \tag{4}$$

where

$$p = \sum_{u=0}^{n-1} f_t^u f_i^u, \tag{5}$$

$$q = \sum_{u=0}^{n-1} f_t^u \times \sum_{u=0}^{n-1} f_i^u, \tag{6}$$

$$s = \sum_{u=0}^{n-1} (f_t^u)^2 - \left(\sum_{u=0}^{n-1} f_t^u \right)^2, \tag{7}$$

$$w = \sum_{u=0}^{n-1} (f_i^u)^2 - \left(\sum_{u=0}^{n-1} f_i^u \right)^2. \tag{8}$$

Although, the number of correlations performed within the proposed framework is directly proportional to the number of OCT scans and the number of combinations between 'lesion' and 'label' words that are taken from the lesion and pathological vocabularies. However, since it does not consume much time in computing a single correlation, the overall runtime of the proposed system is not significant (as discussed further in Section V-C), where n denotes the total number of features.

2) Classification based on Maximum Correlation: For a candidate OCT scan, once the correlation for all the combinations of textual and visual features is obtained, the proposed framework assigns the 'label' of the default text prompt to the given OCT scan that has the maximum correlation coefficient between the respective textual and visual features. The assignment of the final disease label to the given OCT scan based on the maximum correlation coefficient ensures that the proposed framework is screening the retinal diseases by considering the clinically significant manifestations (lesions), which the expert ophthalmologists also observe as a standard practice within the clinical settings.

IV. Experimental Setup

In the following section, we report a detailed description of the datasets, the training and implementation protocols, and the evaluation metrics by which we compared the proposed framework's performance with state-of-the-art works.

A. Datasets

To evaluate the classification performance of the proposed framework and to compare it with the state-of-the-art methods, we used six popular and publicly available datasets dubbed the Zhang [20], Duke-1 [35], Duke-2 [13], Duke-3 [36], Rabbani [37], and BIOMISA [38]. The description of the datasets, as well as the training and testing splits used in this paper, is presented in Table I. From Table I, we can observe that Zhang [20] dataset contains a total of 109,309 OCT scans that represent Drusen, CNV, DME, and healthy pathologies. Out of these total scans, we have used 9,500 OCT scans for training purposes, and the rest of them are used to evaluate the proposed framework at the inference stage. It should be noted that the proposed framework was trained only once on the 9,500 scans from the Zhang dataset [20], while the other datasets were used only for testing purposes. Moreover,

Table I: A detailed description of all the datasets along with the training-test splits that are used in this research. '-' indicates the samples from the respective dataset were not used in the training at all. Moreover, Drusen and CNV pathologies are considered as AMD across Duke-1 [35], Duke-2 [13], Duke-3 [36], Rabbani [37], and BIOMISA [38] datasets for fair comparison.

Dataset	Training	Testing	Pathologies
Zhang [20]	9,500	99,809	Drusen CNV DME Normal
Duke-1 [35]	-	38,300	AMD (Drusen + CNV) Controlled (Normal)
Duke-2 [13]	-	610	DME
Duke-3 [36]	-	3,231	AMD (Drusen + CNV) DME Normal
Rabbani [37]	-	4,241	AMD (Drusen + CNV) DME Normal
BIOMISA [38]	-	4,163	AMD (Drusen + CNV) DME Normal

Duke-1 [35] contains a total of 38,300 OCT scans, which depict AMD and controlled pathologies. We used all 38,300 scans from the Duke-1 dataset [35] for testing the proposed framework. The Duke-2 dataset [13] contains 610 retinal OCT scans that depict severe DME pathologies. Similarly, Duke-3 [36], Rabbani [37], and BIOMISA [38] datasets contain a total of 3,231, 4,241, and 4,163 OCT scans, respectively, which depict AMD, DME, and healthy pathologies.

B. Implementation Details

The proposed framework is implemented using TensorFlow 2.1.0 and Python 3.7.9 on the Anaconda platform with a machine having Core i9-10940@3.30 GHz processor, 128 GB RAM, and a single NVIDIA Quadro RTX 6000 GPU having CUDA toolkit v11.0.221 and cuDNN v7.5. Apart from this, the training details are explained in Section III-A.

C. Evaluation Metrics

We used the standard classification metrics, such as recall, precision, F1 score, receiver operator characteristics (ROC) curve, area under the curve (AUC), precision-recall (PR) curve, and average precision (AP) scores to evaluate the performance of the proposed framework and compare it with state-of-the-art methods.

V. Results

This section reports the detailed evaluation of the proposed framework and its comparison with state-of-the-art methods across six public retinal OCT datasets. Before diving into the comparative analysis, we first report the ablation studies through which we determined the optimal hyperparameters of the proposed framework.

A. Ablation Studies

We performed a series of ablative experiments to determine 1) the optimal text encoder, 2) the optimal vision encoder, 3) the optimal set of clinical prompts, 4) the optimal value of the

Table II: Determining the optimal text encoder across each dataset. Bold indicates the best performance, while the second-best scores are underlined. Moreover, the abbreviations are: ZH: Zhang [20], BO: BIOMISA [38], D1: Duke-1 [35], D2: Duke-2 [13], D3: Duke-3 [36], RB: Rabbani [37].

Dataset	Encoder	Recall	Precision	F1	AUC
ZH [20]	BERT [32]	**0.9128**	**0.9471**	**0.9296**	**0.9728**
	Word2Vec [39]	0.8246	0.8502	0.8372	0.9272
	XLNet [40]	0.8963	0.9346	0.9150	0.9591
BO [38]	BERT [32]	**0.9308**	**0.9564**	**0.9434**	**0.9805**
	Word2Vec [39]	0.8427	0.8751	0.8586	0.9456
	XLNet [40]	0.9196	0.9472	0.9331	0.9728
D1 [35]	BERT [32]	**0.8951**	**0.9148**	**0.9048**	**0.9246**
	Word2Vec [39]	0.7821	0.8204	0.8007	0.8859
	XLNet [40]	0.8802	0.9043	0.8920	0.9047
D2 [13]	BERT [32]	**0.8729**	**0.9042**	**0.8882**	**0.9182**
	Word2Vec [39]	0.7743	0.8263	0.7994	0.8583
	XLNet [40]	0.8618	0.8875	0.8744	0.9109
D3 [36]	BERT [32]	**0.8846**	**0.9017**	**0.8930**	**0.9208**
	Word2Vec [39]	0.7723	0.8119	0.7916	0.8813
	XLNet [40]	0.8794	0.8879	0.8836	0.8965
RB [37]	BERT [32]	**0.9217**	**0.9382**	**0.9298**	**0.9408**
	Word2Vec [39]	0.8043	0.8531	0.8279	0.8963
	XLNet [40]	0.9135	0.9261	0.9197	0.9374

Table III: Determining the optimal vision encoder across each dataset. Bold indicates the best performance, while the second-best scores are underlined.

Dataset	Encoder	Recall	Precision	F1
Zhang [20]	ResNet-50 [41]	0.8672	0.9052	0.8857
	PVT [33]	**0.9128**	**0.9471**	**0.9296**
	ViT [42]	0.8924	0.9253	0.9085
BIOMISA [38]	ResNet-50 [41]	0.8912	0.9176	0.9042
	PVT [33]	**0.9308**	**0.9564**	**0.9434**
	ViT [42]	0.9107	0.9291	0.9198
Duke-1 [35]	ResNet-50 [41]	0.8351	0.8624	0.8485
	PVT [33]	**0.8951**	**0.9148**	**0.9048**
	ViT [42]	0.8543	0.8759	0.8649
Duke-2 [13]	ResNet-50 [41]	0.8348	0.8691	0.8516
	PVT [33]	**0.8729**	**0.9042**	**0.8882**
	ViT [42]	0.8548	0.8864	0.8703
Duke-3 [36]	ResNet-50 [41]	0.8421	0.8693	0.8554
	PVT [33]	**0.8846**	**0.9017**	**0.8930**
	ViT [42]	0.8608	0.8776	0.8691
Rabbani [37]	ResNet-50 [41]	0.8863	0.9046	0.8953
	PVT [33]	**0.9217**	**0.9382**	**0.9299**
	ViT [42]	0.9054	0.9143	0.9098

temperature constant τ which gives the optimal performance of the proposed framework at the inference stage, 5) the optimal loss function to fine-tune vision encoder in a self-supervised manner, and 6) the comparison of the proposed vision language correlations with the conventional Euclidean distance and cosine similarity metric used in the vision language models [11].

1) The Optimal Text Encoder: The first ablation experiment relates to determining the text encoder, which results in the best overall performance for screening retinal diseases from OCT scans across each dataset. For this purpose, we plugged different models, such as BERT [32], Word2Vec [39], and XLNet [40], as a text encoder within the proposed framework and computed the overall performance to screen the retinal diseases. Here, we used PVT [33] as a vision encoder model, which is constrained using the proposed L_{as} loss function. The results for this experiment are reported in Table II from which we can observe that across each dataset, BERT [32] generated the best text features, which resulted in better classification performance of the proposed framework. Moreover, the second best results are generated when we used XLNet [40] across each dataset. Although the performance difference between BERT [32] and XLNet [40] is not significant, both of these models can be used as a text encoder within the proposed scheme. But since BERT [32] generated overall better results, we chose it as an optimal text encoder for the rest of the experimentation.

2) The Optimal Vision Encoder: The second ablation experiment relates to determining the vision encoder, which gives the overall best performance in screening retinal diseases from OCT scans across each dataset. For this purpose, we plugged different popular models, such as ResNet-50 [41], ViT [42], and PVT [33], as a vision encoder within the proposed framework and compared their overall performance to screen retinal diseases. Here, we used BERT [32] as a text encoder that produced the most discriminative text features that serve as a guide to the vision encoder in screening retinal diseases per the clinical standards. Moreover, we constrained the vision encoder using the proposed L_{as} loss function during training to generate distant feature representations to screen retinal diseases at the inference stage. The results for this experiment are reported in Table III from which we can observe that across each dataset, PVT [33] produces the best results in generating the most discriminative features which are aligned with the textual features in screening different types of retinal diseases across each dataset. Since PVT [33] provided the best classification performance compared to the other models, we used it as an optimal vision encoder for the rest of the experimentation.

3) Determining the Optimal Set of Clinical Prompts: In Section III-A1, we explained that the text encoder is fine-tuned using two approaches. In the first approach, we used 1,000 retinal prompts generated by expert clinicians to fine-tune the text encoder. These prompts highlighted the clinical manifestations and biomarkers in identifying retinal diseases. Moreover, in the second approach, we used 10,000 prompts generated by ChatGPT 4.0 [26] but thoroughly verified by the two expert clinicians in order to fine-tune the text encoder in discriminating DME, AMD, and healthy pathologies. The classification performance of the proposed framework using both approaches across each dataset is reported in Table IV. From Table IV, we can see that, across each dataset, the performance of the proposed framework is similar when an equal number of clinical prompts (i.e., 1,000) are used in fine-tuning the text encoder where these prompts are directly generated by both the expert clinicians and ChatGPT 4.0 [26]. However, the overall performance of the proposed framework significantly increases when we use all the 10,000 prompts generated by ChatGPT 4.0 [26] in fine-tuning the text encoder. This performance gain is evident as the text encoder, fine-tuned on 10,000 prompts, gets more exposure to the variety of biomarkers that can exist within the OCT scans in the presence of each retinal disease. On the contrary, it is difficult for clinicians to produce 10,000 prompts for training any machine learning model. So, we use the 10,000 clinical prompts generated by ChatGPT 4.0 [26] as an optimal set of prompts to fine-tune the text encoder within the proposed framework, and ChatGPT 4.0 [26] has already

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3462653

Table IV: Determining the optimal set of clinical prompts that are used in fine-tuning the text encoder to produce clinically significant screening of retinal diseases across each dataset. Bold indicates the best performance, while the second-best scores are underlined. Moreover, 'C-Prompts' stands for Clinicians' Prompts.

Dataset	Prompts	Recall	Precision	F1	AUC
Zhang [20]	GPT-4 1k	0.8775	0.8906	0.8840	0.9463
	C-Prompts	0.8813	0.9061	0.8935	0.9528
	GPT-4 10k	**0.9128**	**0.9471**	**0.9296**	**0.9728**
BIOMISA [38]	GPT-4 1k	0.9175	0.9406	0.9289	0.9524
	C-Prompts	0.9096	0.9324	0.9208	0.9614
	GPT-4 10k	**0.9308**	**0.9564**	**0.9434**	**0.9805**
Duke-1 [35]	GPT-4 1k	0.8435	0.8671	0.8551	0.8762
	C-Prompts	0.8528	0.8843	0.8682	0.8958
	GPT-4 10k	**0.8951**	**0.9148**	**0.9048**	**0.9246**
Duke-2 [13]	GPT-4 1k	0.8049	0.8237	0.8141	0.8758
	C-Prompts	0.8243	0.8332	0.8287	0.8968
	GPT-4 10k	**0.8618**	**0.8875**	**0.8744**	**0.9109**
Duke-3 [36]	GPT-4 1k	0.8371	0.8184	0.8276	0.8856
	C-Prompts	0.8417	0.8398	0.8407	0.8972
	GPT-4 10k	**0.8846**	**0.9017**	**0.8930**	**0.9208**
Rabbani [37]	GPT-4 1k	0.8786	0.8975	0.8879	0.9058
	C-Prompts	0.8952	0.9156	0.9052	0.9243
	GPT-4 10k	**0.9217**	**0.9382**	**0.9298**	**0.9408**

Table V: Determining the optimal value of temperature constant τ on the Zhang dataset. Bold indicates the best performance, while the second-best scores are underlined.

τ	Recall	Precision	F1	AUC
1	0.8961	0.9383	0.9167	0.9634
1.5	**0.9128**	**0.9471**	**0.9296**	**0.9728**
2	0.8747	0.9056	0.8898	0.9453

been recognized by many clinical researchers in generating clinically significant prompts for the retinal diseases [21], [25], [27]–[30].

4) Determining the Temperature Constant τ: In this series of experiments, we determined the optimal value of the temperature constant τ, which is a hyperparameter that softens the target probabilities produced by the model to maximize its generalization capacity. For this purpose, we varied the value of τ between 1 to 2 in the step of 0.5 within the L_{as} loss function during the training phase and then measured the overall classification performance of the proposed framework at the inference stage. From Table V, we can observe that when we varied τ from 1 to 1.5, we achieved a performance boost of 1.82%, 0.929%, 1.32%, and 0.966% in terms of recall, precision, F1, and AUC scores, respectively. However, increasing τ does not guarantee that the network's performance will always increase. For example, when $\tau = 2.0$, we observe a reduction in the overall performance of the proposed framework. Therefore, to produce the best classification results, we used $\tau = 1.5$ within all the experimentation.

5) Determining the Optimal Loss Function: The fifth ablation experiment is related to determining the optimal loss function that can be used in the training phase to update the weights of the vision encoder and align its feature representations with the textual features generated by the text encoder. For this purpose, we used the triplet loss function L_t [43], soft nearest neighbor loss function L_{sn} [44], MagFace loss function L_{mf} [45], angular loss function L_a [46], and the proposed L_{as} loss function to fine-tune the vision encoder. Afterward, we compared their performance difference across each dataset

Table VI: Determining the optimal loss function to constrain the vision encoder during the training phase. The results are reported at the inference stage across each dataset once the proposed framework is trained using the respective loss function. Bold indicates the best performance, while the second-best scores are underlined.

Dataset	Loss	Recall	Precision	F1	AUC
Zhang [20]	L_t	0.8651	0.8842	0.8745	0.9356
	L_{sn}	0.8869	0.9163	0.9013	0.9564
	L_{mf}	0.8948	0.9273	0.9107	0.9651
	L_a	0.8724	0.9158	0.8935	0.9586
	L_{as}	**0.9128**	**0.9471**	**0.9296**	**0.9728**
BIOMISA [38]	L_t	0.8752	0.9076	0.8911	0.9527
	L_{sn}	0.9051	0.9281	0.9164	0.9613
	L_{mf}	0.9143	0.9395	0.9267	0.9716
	L_a	0.9082	0.9306	0.9192	0.9684
	L_{as}	**0.9308**	**0.9564**	**0.9434**	**0.9805**
Duke-1 [35]	L_t	0.8293	0.8471	0.8381	0.8996
	L_{sn}	0.8563	0.8754	0.8657	0.9078
	L_{mf}	0.8678	0.8923	0.8798	0.9175
	L_a	0.8605	0.8846	0.8723	0.9128
	L_{as}	**0.8951**	**0.9148**	**0.9048**	**0.9246**
Duke-2 [13]	L_t	0.8042	0.8395	0.8214	0.8639
	L_{sn}	0.8253	0.8615	0.8430	0.8956
	L_{mf}	0.8436	0.8814	0.8620	0.9132
	L_a	0.8375	0.8729	0.8548	0.9028
	L_{as}	**0.8729**	**0.9042**	**0.8882**	**0.9182**
Duke-3 [36]	L_t	0.8273	0.8463	0.8366	0.8837
	L_{sn}	0.8514	0.8876	0.8691	0.9042
	L_{mf}	0.8695	0.8913	0.8802	0.9120
	L_a	0.8612	0.8895	0.8751	0.9082
	L_{as}	**0.8846**	**0.9017**	**0.8930**	**0.9208**
Rabbani [37]	L_t	0.8656	0.8769	0.8712	0.9096
	L_{sn}	0.8863	0.9089	0.8974	0.9211
	L_{mf}	0.8986	0.9143	0.9063	0.9345
	L_a	0.8914	0.9096	0.9004	0.9269
	L_{as}	**0.9217**	**0.9382**	**0.9298**	**0.9408**

to screen retinal diseases at the inference stage. The results are reported in Table VI from which we can observe that across each dataset, the proposed L_{as} generated the most aligned feature representations with the textual embeddings, which resulted in an overall performance gain of 2.03%, 1.77%, 2.76%, 2.94%, 1.43%, and 2.52% in terms of F1 score compared to its competitors across Zhang [20], BIOMISA [38], Duke-1 [35], Duke-2 [13], Duke-3 [36], and Rabbani [37] datasets, respectively. The performance improvements produced by the L_{as} loss function stems from the fact that L_{as} is specifically designed to constrain the VLM models to adapt to the new datasets in order to perform the prior learned tasks in a self-supervised manner. The self-supervision here is established when L_{as} minimizes the Euclidean distance between the image features (anchor) and the textual features (positives) in an angular space while maximizing the distances between anchors and negatives (textual features not belonging to the positive class). Moreover, the state-of-the-art loss functions, such as L_a [46] and L_{mf} [45], although work well for the vision encoders, they are not inherently built to analyze the relationship between image and text features. Due to this reason, their performance deteriorates compared to the proposed L_{as} loss function when constraining the proposed VLM scheme to learn retinopathy screening tasks across different multi-vendor datasets (see Table VI for more details).

6) The Optimal Classification Method: Once the text and vision encoders are fine-tuned during the training phase, we

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3462653

Table VII: Determining the optimal classification method to screen retinal diseases from the OCT scan of each dataset. Bold indicates the best performance, while the second-best scores are underlined.

Dataset	Method	Recall	Precision	F1	AUC
Zhang [20]	Cosine	0.8641	0.9078	0.8854	0.9536
	Euclidean	0.8453	0.8898	0.8669	0.9283
	Correlation	**0.9128**	**0.9471**	**0.9296**	**0.9728**
BIOMISA [38]	Cosine	0.8861	0.9238	0.9045	0.9536
	Euclidean	0.8617	0.8809	0.8711	0.9425
	Correlation	**0.9308**	**0.9564**	**0.9434**	**0.9805**
Duke-1 [35]	Cosine	0.8462	0.8756	0.8606	0.8954
	Euclidean	0.8236	0.8545	0.8387	0.8713
	Correlation	**0.8951**	**0.9148**	**0.9048**	**0.9246**
Duke-2 [13]	Cosine	0.8391	0.8601	0.8494	0.8753
	Euclidean	0.8064	0.8526	0.8288	0.8479
	Correlation	**0.8729**	**0.9042**	**0.8882**	**0.9182**
Duke-3 [36]	Cosine	0.8452	0.8647	0.8548	0.8891
	Euclidean	0.8291	0.8386	0.8338	0.8657
	Correlation	**0.8846**	**0.9017**	**0.8930**	**0.9208**
Rabbani [37]	Cosine	0.8705	0.8925	0.8813	0.9143
	Euclidean	0.8498	0.8643	0.8569	0.8975
	Correlation	**0.9217**	**0.9382**	**0.9298**	**0.9408**

use them at the inference stage to screen retinal diseases from the OCT scans. In order to do this, we first extract the text features from the default prompt in which we vary the 'lesion' and 'label' words using the lesion and pathological vocabularies (see Section III-B). For each set of textual features generated from the 'lesion' and 'label' words, we check its correlation with the visual features. The 'label' word generated from the textual and visual features pair that has the maximum correlation is then used as a final label for the given OCT scan. In this ablation experiment, we compared the performance of vision language correlations with the other similarity metrics, such as cosine similarity, and Euclidean distance, which can also be used as a classification method within the proposed framework. After injecting these similarity metrics with the proposed framework, we evaluated their overall performance toward screening retinal diseases and compared them with the vision language correlation method. The results for this experiment are reported in Table VII from which we can observe that although the cosine similarity and Euclidean distance method produced decent classification results. However, the best performance is obtained using the vision language correlations across each dataset. These improvements stem from the fact that vision language correlation allows the proposed framework to generate contextual-aware similarities from which the maximum correlation is obtained only when the visual features are closely aligned with the clinically significant textual features. Since the vision language correlations produced better classification results than the other similarity metrics, we chose it as an optimal classification method for the rest of the experiment.

B. Comparative Analysis

This section compares the proposed framework with state-of-the-art (SOTA) methods across all six public datasets. This section also compares the proposed framework thoroughly with the recent vision language models to screen retinal diseases from the OCT scans across each of the six datasets. Furthermore, this section reports the blind test experiments through which we compared the classification performance of the proposed framework against the grading of two expert clinicians in clinical settings using unseen OCT scans depicting different retinal abnormalities. Lastly, we present the runtime performance comparison of the proposed framework against SOTA methods and shed light on its limitations and their remedies.

1) Comparison with SOTA Methods: Table VIII reports the comparison of the proposed framework with state-of-the-art solutions across each dataset. It should be noted here that we used the same test sets for those state-of-the-art methods, whose source codes were publicly available, in order to compare their performance with the proposed scheme. However, for those methods whose source codes were not publicly available, their performance scores are taken from their corresponding papers. We have highlighted such methods with '*' within Table VIII. Apart from this, for the methods whose scores are missing in Table VIII under any metric, their performances were actually not reported for that particular metric in their paper, and their codes were also not publicly available for us to test. Also, it should be noted that the SOTA methods reported in Table VIII for each dataset are different. The reason for having different state-of-the-art methods across each dataset is because, to the best of our knowledge, there are very limited works that were thoroughly tested together on all six publicly available datasets. Moreover, from Table VIII, we can observe that the proposed framework achieved better results in most of the metrics across each dataset. For example, on Zhang dataset [20], the proposed framework achieved 3.39% better performance than the second-best method in terms of F1 score. On the BIOMISA dataset [38], the proposed framework achieved 2.36% better performance than the second-best method in terms of F1 score. Similarly, across Duke-1 [35], Duke-2 [13], Duke-3 [36] and Rabbani [37] datasets, the proposed framework also outperformed state-of-the-art methods in terms of F1 score. These performance improvements stem from the fact that the proposed framework leverages the clinical manifestations encoded within the textual features to screen retinal diseases depicted within OCT scans, and such an approach greatly eliminates the false positives and false negatives in a self-supervised manner. Apart from this, Figure 3 (A-L) reports the attention maps generated by the proposed framework from the OCT scans. From Figure 3 (A-L), we can observe that for each type of retinal disease, the proposed framework was able to focus on its clinically significant manifestations. For example, for the DME pathologies, the proposed framework was able to focus on the hard exudates, intra-retinal and sub-retinal fluid regions. For AMD pathologies, the proposed framework looks for the presence of drusen and choroidal neovascularization (CNV) to screen the scans correctly. These areas, highlighted in the attention maps, represent the clinically significant lesion pathologies that clinicians analyze within the clinical settings to screen for DME and AMD diseases.

2) Scanner-specific Screening of Retinal Diseases: We conducted another series of experiments to evaluate the performance of the proposed framework for screening retinal diseases represented within the OCT scans acquired with

Table VIII: Performance comparison of the proposed framework with state-of-the-art methods across each dataset. Bold indicates the best performance, while the second-best scores are underlined. '-' indicates that the metric is not computed by the respective method. '*' indicate those methods whose source codes are not publicly available. Hence, their results are reported from their original papers.

Dataset	Method	Recall	Precision	F1
Zhang [20]	Proposed	0.9128	0.9471	**0.9296**
	RAG-FW [6]	0.8942	0.9015	0.8980
	LightOCT [15]*	**0.9450**	-	-
	f-AnoGAN [47]	0.5021	0.6982	0.5841
	CycleGAN [48]	0.8025	0.9341	0.8633
	Ganimorph [49]	0.7897	0.9321	0.8550
	WeaklyGAN [50]*	0.8374	**0.9671**	0.8976
	LACNN [10]	0.8680	0.8620	0.8640
BIOMISA [38]	Proposed	**0.9308**	**0.9564**	**0.9434**
	RAG-FW [6]	0.9053	0.9281	0.9165
	STGS [12]	0.8929	0.9186	0.9053
	ICDA [17]	0.9085	0.9341	0.9211
Duke-1 [35]	Proposed	**0.8951**	**0.9148**	**0.9048**
	RAG-FW [6]	0.8738	0.8863	0.8794
	IR [51]*	0.8872	0.8945	0.8908
	SR-GAN [16]*	0.8543	0.8724	0.8632
	Semivariogram [52]*	0.8498	0.8516	0.8506
Duke-2 [13]	Proposed	**0.8729**	**0.9042**	**0.8882**
	RAG-FW [6]	0.8548	0.8895	0.8718
	STGS [12]	0.8276	0.8403	0.8339
Duke-3 [36]	Proposed	**0.8846**	**0.9017**	**0.8930**
	RAG-FW [6]	0.8769	0.8941	0.8854
	MS-CNN [53]*	0.8682	0.8876	0.8777
Rabbani [37]	Proposed	**0.9217**	**0.9382**	**0.9298**
	RAG-FW [6]	0.8736	0.8941	0.8836
	WeaklyGAN [50]*	0.8468	-	-
	SC-CNN [54]*	0.8294	-	-
	RAGNet [55]	0.8547	0.8606	0.8576

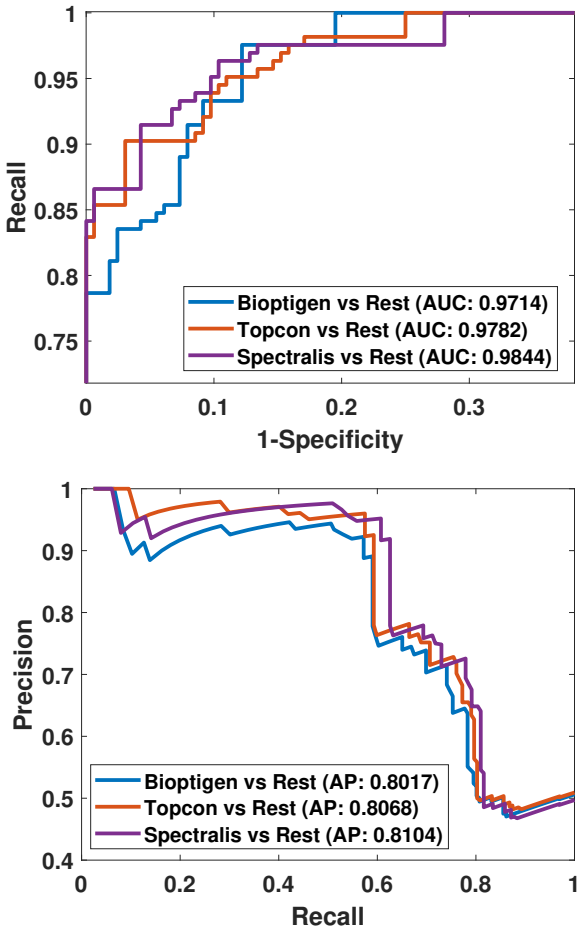

Figure 2: Performance of the proposed framework in terms of ROC (up) and PR (down) curves to screen retinal diseases from OCT scans acquired with different scanners.

different scanners, such as Bioptigen, Spectralis, and Topcon. From Table I, we can observe that Duke-2, 3, and Zhang datasets are acquired with the Spectralis machine. Moreover, Duke-1 is acquired using a Bioptigen OCT machine. Similarly, the BIOMISA [38] dataset is acquired using the Topcon OCT machine. Therefore, we combined all of these datasets together with respect to their acquisition machinery and passed them directly to the proposed framework at the inference stage. The results for this experiment are reported with ROC and PR curves within Figure 2. From Figure 2, we can observe that regardless of the scanner, the proposed framework is able to produce promising results in screening retinal diseases. These performance gains are again produced due to the ability of the proposed framework to predict the disease label based on the maximum correlation between the visual features and the textual features that represent clinically significant embeddings of the retinal biomarkers. The performance of the proposed framework in Figure 2 also demonstrates the ability of the proposed framework to be attached to these types of scanners in clinical settings to mass screen retinal OCT scans against these types of retinal diseases. Additionally, we verify this ability further in blind test experiments reported in Section V-D.

3) Comparison with VLM Models: In this section, we report a series of experimentation through which we compared the performance of the proposed framework with state-of-the-art vision-language models (VLMs), such as CLIP [11], BLIP [56], FLIP [57], CLIPS [58], and MedCLIP [59] for screening AMD, DME, and healthy pathologies from OCT scans. Since the proposed framework also belongs to the category of VLM models, we conducted this set of experiments to showcase how well the proposed framework performs against the state-of-the-art models within the VLM territory. Here, to make the comparison fair, we followed the same experimental protocol reported in Section IV for training and evaluating each of the models. The comparison is reported in Table IX from which we can observe that, across each dataset, the proposed framework outperforms the second-best model. For example, across Zhang dataset [20], the proposed model achieves 1.72% better performance in terms of F1 score. Across the BIOMISA dataset [38], the proposed model achieves a 1.47% improvement in terms of F1 score. Across the Duke-1 dataset [35], the proposed model outperforms the second-best method by 2.54% in terms of F1 score. Across the Duke-2 dataset [13], the proposed model outperforms others by 2.79% in terms of F1 score. Across the Duke-3 dataset [36], the proposed model achieves 1.01% performance improvements in terms of F1 score, and across the Rabbani dataset [37], the proposed model achieves 1.55% better performance than others in terms of F1 score. These improvements stem from the fact that the proposed framework uses vision language correlation to screen

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3462653

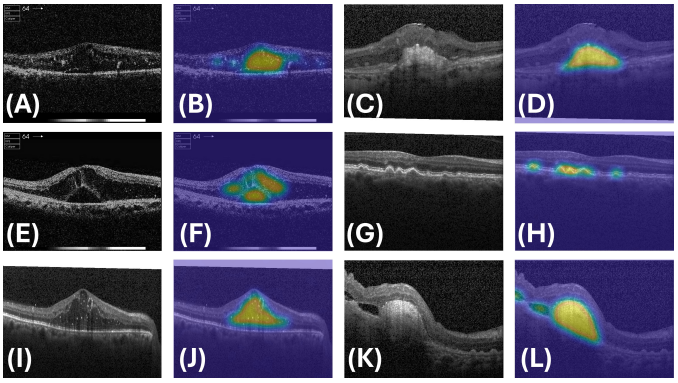

Figure 3: Visual examples showcasing the attention maps generated by the proposed framework. The first and third column shows the original scans, whereas the second and fourth column shows the attention maps overlaid on the original scans. These attention maps highlight different retinal lesions, such as intra-retinal fluid, sub-retinal fluid, hard exudates, drusen, and CNV, contained within DME and AMD affected OCT scans.

retinal diseases as compared to its competitors, which uses either cosine similarities [11], [58], [59], image-text matching mechanism [56], or image masking [57]. Contrary to these models, the vision language correlations allow the proposed framework to generate contextual-aware alignment between the visual and textual features in a robust and efficient manner, from which the maximum correlation is obtained only when the visual features are closely aligned with the clinically significant textual features.

We also want to highlight that although the CLIP [11], MedCLIP [59], and the proposed model are all VLMs. However, the way in which each of the models performs the classification tasks is different. For example, CLIP [11] and MedCLIP [59] adapt to multiple datasets in order to perform the classification tasks through zero-shot transfer [11], [59]. Similarly, their performance toward solving the classification tasks could be improved, especially for medical applications. For example, CLIP [11] and MedCLIP [59] achieved the best accuracy of 0.5055 and 0.7682, respectively, when they were evaluated for screening pulmonary diseases from the chest X-ray scans [59]. When CLIP [11] and MedCLIP [59] are adapted to retinal datasets, they lag behind the proposed model by 4.17% and 4.56% on Zhang dataset [20], 2.93% and 2.66% on BIOMISA dataset [38], 5.72% and 5.31% on Duke-1 dataset [35], 5.34% and 4.43% on Duke-2 dataset [13], 4.42% and 3.03% on Duke-3 dataset [36], 4.67% and 3.10% on Rabbani dataset [37], respectively, in terms of F1 score. Therefore, considering the classification performance, these models are not practical for deployment in the real world for mass-screening retinal subjects despite being zero-shot learning models. Contrary to them, the proposed scheme promotes self-supervised learning where it utilizes its prior-learned experiences to perform the classification tasks with remarkable F1 scores of 0.9296, 0.9434, 0.9048, 0.8882, 0.8930, and 0.9298 on average across Zhang [20], BIOMISA [38], Duke-1 [35], Duke-2 [13], Duke-3 [36], and Rabbani [37] datasets, respectively (see Tables II to IX for more details). Furthermore, the proposed scheme provides an optimal trade-off between computational complexity and class performance, and it is a promising choice to be deployed in the clinical setting to perform retinal screening tasks. We further want to highlight that we have tested the proposed scheme in a real clinical setting through a series of blind test experiments against two expert clinicians, where it achieved a high correlation with both clinicians' gradings to screen retinal patients (see Section V-D for more details).

Table IX: Determining the optimal classification method to screen retinal diseases from the OCT scan of each dataset. Bold indicates the best performance, while the second-best scores are underlined. The abbreviations are: ZH: Zhang [20], BO: BIOMISA [38], D1: Duke-1 [35], D2: Duke-2 [13], D3: Duke-3 [36], and RB: Rabbani [37].

Dataset	VLM	Recall	Precision	F1	AUC
ZH [20]	Proposed	**0.9128**	**0.9471**	**0.9296**	**0.9728**
	CLIP [11]	0.8769	0.9052	0.8908	0.9456
	BLIP [56]	0.8941	0.9224	0.9080	0.9683
	FLIP [57]	0.8952	0.9248	0.9097	0.9693
	CLIPS [58]	0.8963	0.9316	0.9136	0.9687
	MedCLIP [59]	0.8785	0.8962	0.8872	0.9473
BO [38]	Proposed	**0.9308**	**0.9564**	**0.9434**	**0.9805**
	CLIP [11]	0.9045	0.9273	0.9157	0.9572
	BLIP [56]	0.9127	0.9368	0.9245	0.9624
	FLIP [57]	0.9153	0.9414	0.9281	0.9658
	CLIPS [58]	0.9169	0.9425	0.9295	0.9664
	MedCLIP [59]	0.9084	0.9286	0.9183	0.9562
D1 [35]	Proposed	**0.8951**	**0.9148**	**0.9048**	**0.9246**
	CLIP [11]	0.8439	0.8625	0.8530	0.9015
	BLIP [56]	0.8674	0.8846	0.8759	0.9238
	FLIP [57]	0.8681	0.8894	0.8786	0.9235
	CLIPS [58]	0.8723	0.8916	0.8818	0.9240
	MedCLIP [59]	0.8496	0.8641	0.8567	0.9102
D2 [13]	Proposed	**0.8729**	**0.9042**	**0.8882**	**0.9182**
	CLIP [11]	0.8239	0.8584	0.8407	0.8759
	BLIP [56]	0.8375	0.8761	0.8563	0.8962
	FLIP [57]	0.8391	0.8789	0.8585	0.9012
	CLIPS [58]	0.8436	0.8842	0.8634	0.9028
	MedCLIP [59]	0.8325	0.8659	0.8488	0.8863
D3 [36]	Proposed	**0.8846**	**0.9017**	**0.8930**	**0.9208**
	CLIP [11]	0.8425	0.8649	0.8535	0.8873
	BLIP [56]	0.8618	0.8857	0.8736	0.9062
	FLIP [57]	0.8675	0.8896	0.8784	0.9103
	CLIPS [58]	0.8721	0.8962	0.8839	0.9126
	MedCLIP [59]	0.8563	0.8758	0.8659	0.8963
RB [37]	Proposed	**0.9217**	**0.9382**	**0.9298**	**0.9408**
	CLIP [11]	0.8752	0.8978	0.8863	0.9164
	BLIP [56]	0.8916	0.9123	0.9018	0.9271
	FLIP [57]	0.9041	0.9286	0.9161	0.9294
	CLIPS [58]	0.9063	0.9245	0.9153	0.9365
	MedCLIP [59]	0.8952	0.9068	0.9009	0.9182

C. Runtime Performance

In addition to comparing the classification performance of the proposed framework with state-of-the-art methods to screen retinal diseases, we also conducted another set of experiments to evaluate the runtime performance of the proposed framework and compared it with state-of-the-art models. The runtime performance of the proposed framework is shown in Table X. From Table X, we can observe that, on average, the proposed framework screens a single scan across each dataset in 22.5μs. Although it lags behind CLIP [11] by 1.33%, the proposed framework still provides a better trade-off between classification performance and computational efficiency compared to CLIP [11] (see Tables IX and X). Moreover, the runtime performance gain achieved by the proposed framework

Table X: Runtime performance comparison between the proposed framework and state-of-the-art methods. Bold indicates the best performance, while the second-best performance is underlined. The runtime performance is reported in microseconds to screen one OCT scan.

Dataset	Proposed	RAG-FW [6]	BLIP [56]	CLIP [11]
Zhang [20]	25.2	36.8	27.6	**24.3**
BIOMISA [38]	**18.5**	31.4	23.1	19.8
Duke-1 [35]	22.4	33.6	23.4	**20.9**
Duke-2 [13]	**21.9**	35.2	37.7	22.6
Duke-3 [36]	**20.8**	29.7	31.3	21.2
Rabbani [37]	26.6	38.1	29.4	**24.5**
Average	22.5	34.1	28.7	**22.2**

compared to the other state-of-the-art models stems from the fact that the proposed framework uses lightweight backbone models as a text and vision encoder, which strikes an optimal balance between classification and runtime performance.

D. Blind Test Experiments

After thoroughly comparing the classification performance of the proposed framework with the state-of-the-art methods, we conducted another series of experiments to see how well it grades the unseen samples in clinical settings compared to the expert ophthalmologists. Here, the proposed framework and the expert clinicians were randomly given 300 unseen OCT samples reflecting each disease symptom. These unseen OCT samples were acquired from Topcon and Spectralis OCT machines, and both the clinicians as well as the proposed framework were given the task to screen them based on their knowledge. Afterward, the results of both the clinicians and the proposed framework were correlated to see its potential and whether it can be deployed in the clinical setting to mass screen retinal diseases irrespective of the scanner specifications. The results of this blind test experiments is shown in Table XI, where the proposed framework achieved a success rate (accuracy) of 0.9473, and both clinicians received a success rate of 0.9122, and 0.9298, respectively. The inter-observer correlation, computed using Eq. 4, between clinician-1 and clinician-2 came out to be 0.9031, and the correlation of the proposed framework with clinician-1 and 2 came out to be 0.9185 and 0.9529 with $p < 0.05$ in both cases, highlighting the fact that the correlation scores between the proposed framework and clinicians' grading are statistically significant. These results evidences the performance of the proposed framework to screen retinal diseases in accordance with the clinical standards and showcase its potential to be deployed in the clinical setting for the accurate mass screening of retinal diseases.

E. Failure Cases

Although the proposed framework is quite robust in screening retinal diseases from the OCT scans, but if we fine-tune the text encoder on an imbalanced set of prompts for each disease type, the overall performance of the proposed framework decreases slightly across each dataset. This performance degradation is due to the fact that the text encoder (BERT [32]) is initially fine-tuned using the conventional cross-entropy loss function, which is vulnerable to class imbalance problems.

Table XI: Results of the blind test experiment on the unseen OCT samples acquired from the Topcon and Spectralis machines. 'IOC' stands for inter-observer correlation. IOC is the correlation computed between the gradings of Clinician-1, and Clinician-2, and it comes out to be 0.9031 with p-value: 7.57×10^{-22}. Moreover, the correlation coefficient r between Clinician-1 and the proposed framework came out to be 0.9185 with p-value: 7.92×10^{-24}. Similarly, the correlation coefficient r between Clinician-2 and the proposed framework is 0.9529 with p-value: 3.54×10^{-30}.

Metric	Clinician-1	Clinician-2	IOC
r	0.9185	0.9529	0.9031
p-value	7.92×10^{-24}	3.54×10^{-30}	7.57×10^{-22}

Metric	Clinician-1	Clinician-2	Proposed
Success Rate	0.9122	0.9298	0.9473

However, we overcame this limitation by generating a balanced set of clinical prompts for each class (retinal disease). But, in the future, this limitation can be remedied in a better way by using the class imbalance resistant loss functions, such as focal loss [60], or balanced affinity [61], to fine-tune the text encoder in generating distinct textual embeddings irrespective of the imbalanced number of samples for each class.

VI. Discussion

This paper presents a novel approach to screen retinal diseases by analyzing the correlation between the textual and visual features generated by the text and vision encoders, respectively, in an angular space using the proposed L_{as} loss function. The main benefit of angular space loss function L_{as} is its ability to constrain the vision encoder toward learning the retinopathy screening tasks in alignment with the feature representations generated by the text encoder from the clinical prompts. Moreover, while constrained by the L_{as} loss function, the proposed framework possesses the capacity to screen retinal diseases by analyzing the clinical manifestations as per the clinical standards irrespective of the dataset or scanner specifications, unlike other state-of-the-art methods. Similarly, the proposed framework possesses a high potential to be deployed in clinics and hospitals to mass screen retinal subjects on-the-fly, easing the loads of the clinicians in their daily routine. To showcase the applicability of the proposed framework to be deployed in clinical settings, we conducted a series of blind test experiments through which we compared the screening performance of the proposed framework with two expert clinicians where it achieved a correlation coefficient r of 0.9185 and 0.9529 having $p < 0.05$ with both clinicians, respectively. Apart from this, the proposed framework has been thoroughly tested on six public datasets, where across each dataset, the proposed framework outperforms state-of-the-art solutions in several metrics. Although, the proposed framework is susceptible to slight performance degradation when the text encoder is fine-tuned on the imbalanced set of clinical prompts for each retinal disease category. This degradation is due to the fact that the text encoder is constrained using conventional categorical cross entropy loss function which is vulnerable to class imbalance problem. We overcame this problem by fine-tuning the text encoder on a balanced set of prompts for each class. But this problem can be handled in a better way in future by fine-tuning the text encoder using

class imbalance resistance loss functions such as focal loss [60], or the balanced affinity [61]. Apart from this, we also believe that the feature resizing methods could be further improved in the future by using an appropriate dimensionality reduction scheme, such as Principal Component Analysis [62] or Fisher's Discriminant Analysis [63], rather than the sub-sampling approach. Such an alteration could result in further improvement in the classification performance of the proposed scheme. In addition to this, we conducted a series of experiments to compare the performance of the proposed L_{as} loss function with the triplet loss function L_t [43], soft nearest neighbor loss function L_{sn} [44], MagFace loss function L_{mf} [45], and the angular loss function L_a [46]. The results for these experiments are reported in Table VI. Although, L_{as} also analyzes the angular relationship between the feature embeddings. But unlike L_a [46] and MagFace L_{mf} [45], the L_{as} loss function is specifically designed to constrain the VLM models to adapt to the new datasets in order to perform the prior learned tasks in a self-supervised manner. The self-supervision here is established when L_{as} minimizes the Euclidean distance between the image features (anchor) and the textual features (positives) in an angular space while maximizing the distances between anchors and negatives (textual features not belonging to the positive class). Moreover, MagFace L_{mf} [45] and L_a [46], although work well for the vision encoders, they are not inherently built to analyze the relationship between image and text features. Due to this reason, their performance deteriorates compared to the proposed L_{as} loss function when constraining the proposed VLM scheme to learn retinopathy screening tasks across different multi-vendor datasets (see Table VI for more details). Furthermore, in future, we also envisage evaluating the proposed framework to recognize other types of diseases across different medical domains.

VII. Conclusion

In this paper, we present a novel framework to screen retinal disabilities from the OCT imagery by analyzing the maximum vision-language correlations generated through the visual and textual features, which inherently are produced by the vision and language encoders, respectively. The proposed framework has been thoroughly tested on six public datasets, where, across each dataset, it outperforms state-of-the-art methods in several classification metrics. The performance gain produced by the proposed framework over state-of-the-art methods is due to its ability to screen retinal diseases from the OCT scans by analyzing their associated clinical manifestations extracted from the text prompts, where the text prompts are carefully crafted within the clinical settings. Furthermore, the proposed framework requires one-time training, and afterward, it can screen retinal diseases from the OCT scans irrespective of the dataset or scanner specifications. This ability of the proposed framework makes it an ideal choice to be deployed in clinical settings for mass screening of retinal subjects. To showcase this capacity of the proposed framework, we conducted a series of blind test experiments through which we compared the screening performance of the proposed framework with the two expert clinicians, where the proposed framework and the expert clinicians were randomly given different unseen OCT scans acquired from different scanners for screening purposes. The proposed framework under blind test experiments achieved a statistically significant correlation coefficient r of 0.9185 and 0.9529 with both clinicians, respectively, which is quite appreciable. In the future, we envisage testing the proposed framework on extreme class imbalance scenarios and evaluating its screening capacity across a wide variety of medical applications.

Acknowledgement

The authors extend their appreciation to the King Salman Center For Disability Research for funding this work through Research Group No. KSRG-2023-540. Furthermore, we thank the clinicians for helping us with the blind test experiments, and also in generating the clinical text prompts to train the proposed framework.

Competing Interests Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The proposed framework has been thoroughly evaluated on six retinal image datasets, and all six of them are publicly available.

References

[1] Y. Niu, L. Gu, Y. Zhao, and F. Lu, "Explainable Diabetic Retinopathy Detection and Retinal Image Generation," IEEE Journal of Biomedical and Health Informatics, 2022.

[2] S. Morales, K. Engan, V. Naranjo, and A. Colomer, "Retinal Disease Screening Through Local Binary Patterns," IEEE Journal of Biomedical and Health Informatics, 2017.

[3] S. Schürer-Waldheim *et al.*, "Robust Fovea Detection in Retinal OCT Imaging Using Deep Learning," IEEE Journal of Biomedical and Health Informatics, 2022.

[4] H.-C. Shao, C.-Y. Chen, M.-H. Chang, C.-H. Yu, C.-W. Lin, and J.-W. Yang, "Retina-TransNet: A Gradient-Guided Few-Shot Retinal Vessel Segmentation Net," IEEE Journal of Biomedical and Health Informatics, 2023.

[5] R. Xu, T. Liu, X. Ye, F. Liu, L. Lin, L. Li, S. Tanaka, and Y.-W. Chen, "Joint Extraction of Retinal Vessels and Centerlines Based on Deep Semantics and Multi-Scaled Cross-Task Aggregation," IEEE Journal of Biomedical and Health Informatics, 2021.

[6] T. Hassan, M. U. Akram, N. Werghi, and N. Nazir, "RAG-FW: A hybrid convolutional framework for the automated extraction of retinal lesions and lesion-influenced grading of human retinal pathology," IEEE Journal of Biomedical and Health Informatics, 2020.

[7] Y. M. Helmy and H. R. A. Allah, "Grading of age-related macular degeneration: Comparison between color fundus photography, fluorescein angiography, and spectral domain optical coherence tomography," Journal of Ophthalmology, Volume 2013, Article ID 85915, 2013.

[8] N. Choudhry and S. K. Sodhi, "Peripheral OCT Imaging in Practice," Retina Today, 2021.

[9] T. Hassan *et al.*, "Angular Contrastive Distillation Driven Self-Supervised Scanner Independent Screening and Grading of Retinopathy," Information Fusion, 2023.

[10] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, and Z. Liu, "Attention to Lesion: Lesion-Aware Convolutional Neural Network for Retinal Optical Coherence Tomography Image Classification," IEEE Transactions on Medical Imaging, 2019.

[11] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," arXiv:2103.00020, 2021.

[12] T. Hassan, M. U. Akram, A. Shaukat, S. G. Khawaja, and B. Hassan, "Structure Tensor Graph Searches Based Fully Automated Grading and 3D Profiling of Maculopathy From Retinal OCT Images," IEEE Access, 2018.

[13] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," Biomedical Optics Express, 2015.

[14] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in oct images of non-exudative amd patients using deep learning and graph search," Biomedical Optics Express, Vol. 8, No. 5, 2017.

[15] A. Butola *et al.*, "Deep learning architecture "LightOCT" for diagnostic decision support using optical coherence tomography images of biological samples," Biomedical Optics Express, 2020.

[16] V. Das, S. Dandapat, and P. K. Bora, "Unsupervised Super-Resolution of OCT Images Using Generative Adversarial Network for Improved Age-Related Macular Degeneration Diagnosis," IEEE Sensors, 2020.

[17] T. Hassan *et al.*, "Incremental Cross-Domain Adaptation for Robust Retinopathy Screening via Bayesian Deep Learning," IEEE Transactions on Instrumentation and Measurement, 2021.

[18] X. Ma, Z. Ji, S. Niu, T. Leng, D. L. Rubin, and Q. Chen, "MS-CAM: Multi-Scale Class Activation Maps for Weakly-Supervised Segmentation of Geographic Atrophy Lesions in SD-OCT Images," IEEE Journal of Biomedical and Health Informatics, Vol. 24, No. 12, 2021.

[19] J. Wang, W. Li, Y. Chen, W. Fang, W. Kong, Y. He, and G. Shi, "Weakly supervised anomaly segmentation in retinal OCT images using an adversarial learning approach," Biomedical Optics Express, 2021.

[20] D. Kermany *et al.*, "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," Elsevier Cell, 2018.

[21] L. F. Desideri *et al.*, "Application and accuracy of artificial intelligence-derived large language models in patients with age related macular degeneration," International Journal of Retina and Vitreous, 2023.

[22] "ChatGPT 3.5,"

[23] Google, "Google Bard (Gemini)," https://gemini.google.com/?hl=en, Accessed: 2024.

[24] Microsoft, "Bing AI," Accessed: 2024.

[25] K. Jin1, L. Yuan, H. Wu, A. Grzybowski, and J. Ye, "Exploring large language model for next generation of artificial intelligence in ophthalmology," Frontiers in Medicine, 2023.

[26] "ChatGPT 4.0,"

[27] A. S. Huang, K. Hirabayashi, L. Barna, *et al.*, "Assessment of a Large Language Model's Responses to Questions and Cases About Glaucoma and Retina Management," JAMA Ophthalmology, 2024.

[28] O. Kleinig *et al.*, "How to use large language models in ophthalmology: from prompt engineering to protecting confidentiality," Eye, 2023.

[29] J. Ong, N. Kedia, S. Harihar, S. C. Vupparaboina, S. R. Singh, *et al.*, "Applying large language model artificial intelligence for retina International Classification of Diseases (ICD) coding," JMAI, 2023.

[30] R. Kianian, D. Sun, E. L. Crowell, and E. Tsui, "The Use of Large Language Models to Generate Education Materials about Uveitis," Ophthalmology Retina, 2024.

[31] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980*, 2014.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," ACL: Human Language Technologies, 2019.

[33] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[34] P. Schober, C. Boer, and L. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation," Anesthesia & Analgesia 126, 5, 2018.

[35] S. Farsiu, S. J. Chiu, R. V. O'Connell, F. A. Folgar, E. Yuan, J. A. Izatt, and C. A. Toth, "Quantitative Classification of Eyes with and without Intermediate Age-related Macular Degeneration Using Optical Coherence Tomography," Ophthalmology, 2014.

[36] P. P. Srinivasan, L. Kim, P. Mettu, S. Cousins, G. Comer, J. Izatt, and S. Farsiu, "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images," Biomedical Optics Express, 2014.

[37] R. Rasti, H. Rabbani, A. Mehri, and F. Hajizadeh, "Macular OCT Classification using a Multi-Scale Convolutional Neural Network Ensemble," IEEE Transactions on Medical Imaging, vol. 37, no. 4, pp. 1024-1034, 2018.

[38] T. Hassan, M. U. Akram, M. F. Masood, and U. Yasin, "BIOMISA Retinal Image Database for Macular and Ocular Syndromes," International Conference on Image Analysis and Recognition (ICIAR), 2018.

[39] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781, 2013.

[40] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," Neural Information Processing Systems (NeurIPS), 2019.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations (ICLR), 2021.

[43] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," Advances in Neural Information Processing Systems (NIPS), 2003.

[44] N. Frosst, N. Papernot, and G. Hinton, "Analyzing and Improving Representations with the Soft Nearest Neighbor Loss," 36th International Conference on Machine Learning (ICML), 2019.

[45] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A Universal Representation for Face Recognition and Quality Assessment," CVPR, 2021.

[46] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep Metric Learning with Angular Loss," ICCV, 2017.

[47] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," Medical Image Analysis, 2019.

[48] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," IEEE International Conference on Computer Vision (ICCV), 2017.

[49] A. Gokaslan, V. Ramanujan, D. Ritchie, K. I. Kim, and J. Tompkin, "Improving shape deformation in unsupervised image-to-image translation," European Conference on Computer Vision (ECCV), 2018.

[50] J. Wang *et al.*, "Weakly supervised anomaly segmentation in retinal OCT images using an adversarial learning approach," Biomedical Optics Express, 2021.

[51] J. Qiu and Y. Sun, "Self-supervised iterative refinement learning for macular OCT volume classification," Computers in Biology and Medicine, 2019.

[52] A. M. Santos *et al.*, "Semivariogram and semimadogram functions as descriptors for AMD diagnosis on SD-OCT topographic maps using support vector machine," Biomedical Signal Processing and Control, Volume 67, 2021.

[53] A. Thomas, P. M. Harikrishnan, A. K. Krishna, P. P., and V. P. Gopi, "A novel multiscale convolutional neural network based age-related macular degeneration detection using OCT images," Biomedical Signal Processing and Control, Volume 67, 2021.

[54] R. Rasti, H. Rabbani, A. Mehridehnavi, and F. Hajizadeh, "Macular OCT classification using a multi-scale convolutional neural network ensemble," IEEE Transactions on Medical Imaging, 2018.

[55] T. Hassan, M. U. Akram, and N. Werghi, "Exploiting the Transferability of Deep Learning Systems Across Multi-modal Retinal Scans for Extracting Retinopathy Lesions," IEEE BIBE, 2020.

[56] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," ICML, 2022.

[57] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He, "Scaling Language-Image Pre-training via Masking," CVPR, 2023.

[58] A. Ahmed *et al.*, "CLIFS: CLIP-Driven Few-shot Learning For Baggage Threat Classification," ICIP, 2024.

[59] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text," EMNLP, 2022.

[60] T. Y. Lin *et al.*, "Focal Loss for Dense Object Detection," CVPR, 2017.

[61] A. Ahmed *et al.*, "Balanced Affinity Loss for Highly Imbalanced Baggage Threat Contour-Driven Instance Segmentation," ICIP, 2022.

[62] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (PCA)," Computers & Geosciences, March 1993.

[63] N. Cristianini, "Fisher Discriminant Analysis (Linear Discriminant Analysis)," Dict. Bioinform. Comput. Biol., April 2014.