Deep Hierarchies and Invariant Disease-Indicative Feature Learning for Computer Aided Diagnosis of Multiple Fundus Diseases

Yuxin Lin[1], Wei Wang[1, †], Xiaoling Luo[2], Zhihao Wu[1], Chengliang Liu[3], Jie Wen[1], Yong Xu[1]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
[2]College of Computer Science and Software Engineering, Shenzhen University
[3]Computer Science and Engineering, Hong Kong University of Science and Technology

linyuxin6688@gmail.com; wangwei2019@hit.edu.cn; xiaolingluoo@outlook.com; horatio_ng@163.com; liucl1996@163.com; jiewen_pr@126.com; laterfall@hit.edu.cn

Abstract

With the advancement of computer vision, numerous models have been proposed for screening of fundus diseases. However, the recognition of multiple fundus diseases is often hampered by the simultaneous presence of multiple disease types and the confluence of lesion types in fundus images. This paper addresses these challenges by conceptualizing them as multi-level feature fusion and self-supervised disease-indicative feature learning problems. We decode fundus images at various levels of granularity to delineate scenarios wherein multiple diseases and lesions co-occur. To effectively integrate these features, we introduce a hierarchical vision transformer (HVT) that adeptly captures both inter-level and intra-level dependencies. A novel forward-attention module is proposed to enhance the integration of lower-level semantic information into higher semantic layers, thereby enriching the representation of complex features. Additionally, we introduce a novel self-supervised mask-consistent feature learner (MCFL). Unlike traditional mask-autoencoders that reconstruct original images using encoder-decoder structures, MCFL utilizes a teacher-student framework to reconstruct mask-consistent feature maps. In this setup, exponential moving averaging is employed to derive classification-guided features, serving as labels for reconstruction rather than merely reconstructing the original images. This innovative approach facilitates the extraction of disease-indicative features. Extensive experiments demonstrate that our method significantly outperforms existing state-of-the-art models.

Introduction

A recent report by the World Health Organization (WHO) indicates that approximately 216.6 million people worldwide suffer from moderate to severe vision impairment, with 36 million individuals being blind. Studies suggest that at least 45% of these cases could be prevented through early identification and treatment(Flaxman et al., 2017; Yen & Leong, 2008). Fundus diseases (FD), including diabetic retinopathy (DR), glaucoma (G), cataracts (C), myopia retinopathy (MR), and age-related macular degeneration (AMD), are leading causes of blindness globally(N. Li et al.,[1] 2021). Early detection and timely treatment of these diseases are critical in preventing irreversible vision loss. However, the effective implementation of such screening programs is often hindered by a shortage of qualified medical professionals capable of managing the growing number of at-risk patients(J. Lin et al., 2021; J. Wang et al., 2020). In response to these challenges, this paper proposes an automated system for the detection of multiple FDs, aiming to enhance the efficiency and reach of diagnostic services.

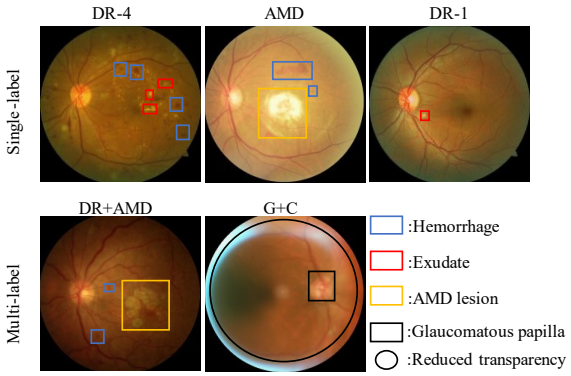

Figure 1: The lesion complexity across various ocular conditions. DR-4: DR in stage 4; DR-1: DR in stage 1; DR+AMD: a fundus image with both DR and AMD; G+C: a fundus image with both G and C.

In the task of multi-FD recognition, models face significant challenges in feature extraction(F. Chen et al., 2023), as illustrated in Figure 1. Key issues include: (1) Diverse Lesion Types: FDs manifest distinct lesion types (G. Ali et al., 2023); for instance as shown in figure 1, AMD presents

specific alterations in the macula, while DR-1 is characterized by small exudative lesions. Moreover, there is an overlap of lesion types across different diseases, such as in DR-4 and AMD, where both display hemorrhagic lesions in the fundus. This overlap complicates the model's ability to differentiate between diseases(Das et al., 2023; Sun et al., 2022). (2) Complex Cases with Multiple Lesions: In scenarios involving fundus images with multiple concurrent lesions (e.g. in figure. 1, DR+AMD and G+C), the images depict lesions specific to each disease. This requires the model to not only distinguish between different lesions but also to accurately identify and decode multiple disease-specific lesions within a single image. (3) Variation Across Disease Stages: Fundus images from different stages of the same disease can exhibit substantial feature variability(Luo et al., 2021, 2023). For an example in figure 1, DR-4 images display complex exudative and hemorrhagic lesions, whereas DR-1 images may only show a minor exudative spot. This variability poses a significant challenge as the model might misclassify different stages of the same disease as distinct diseases. To address these challenges, we propose a HVT and a self-supervised MCFL.

Hierarchical Vision Transformer (HVT): The HVT model functions as a novel hierarchical representation method, effectively capturing and decoding complex lesion conditions. Within the HVT framework, the complex lesion condition in a fundus image is initially decoded by extracting multi-grained details at various depths within the CNN architecture. To integrate inter-level and intra-level features, we employ a novel forward-attention module alongside traditional self-attention. Traditional self-attention captures short-range patterns within a specific scale level, while our forward-attention module models relationships between inter-level features by computing attention scores between lower-level query vectors and higher-level key vectors. This mechanism ensures that essential details captured by lower layers are enhanced by the contextual understanding provided by higher layers.

Self-Supervised Mask-Consistent Feature Learner (MCFL): In response to challenge (3) - Variation Across Disease Stages, where models may misclassify different stages of the same disease as distinct diseases due to lesion variations, the MCFL is introduced. This model is designed to consistently extract disease-indicative features even from partially masked fundus images, akin to manually scaled-down or number-reduced lesions. The MCFL comprises two sub-networks: a Global Feature Learner (GFL) to output disease-indicative features, and a Local Feature Learner (LFL) to reconstruct these features from masked images. This dual approach significantly enhances robustness in diseases classification by maintaining consistent feature representations even from randomly masked images. The contribution of this paper is:

(1) We design a novel hierarchical transformer that effectively captures both inter-level and intra-level dependencies, enabling enhancement and interaction of semantic feature representations across various granularities, thereby achieving improved classification of fundus diseases.
(2) We pioneer the MCFL, a self-supervised feature learning approach using a teacher-student framework to reconstruct mask-consistent feature maps. Leveraging EMA, it surpasses standard techniques by focusing on mask-irrelevant features for enhancing disease-specific feature extraction.
(3) We introduce a novel hybrid architecture that blends the strengths of transformers and CNNs, significantly enhancing the recognition accuracy of multiple fundus diseases by adeptly integrating both local and global information for a more comprehensive analysis.

Related Works

Hierarchical feature extraction form CNN

Multi-level feature maps, extracted from varying depths within CNN architectures, capture distinct structural components of images, reflecting a progression from basic to complex representations(Ghahremani et al., 2024; S. Guo, 2021; Z. Wu et al., 2024). This hierarchical extraction process is particularly advantageous in biomedical image segmentation, where precise delineation of diverse image regions is required(Y. Li et al., 2023; X. Zhang et al., 2022; Z. Zhang et al., 2022). For instance, Guo (S. Guo et al., 2019) successfully implemented a multi-level feature-based framework for the segmentation of multiple lesions in fundus images, demonstrating the practical benefits of this approach. Similarly, another work (X. Wang et al., 2022) applied these principles, using features drawn from a pre-trained backbone to effectively segment and grade diabetic retinopathy lesions.

Feature representation learning

Feature representation learning is pivotal in enhancing the accuracy and robustness of deep learning models(L. Wu et al., 2023), particularly in medical signal analysis(Y. Lin et al., 2023; Zhu et al., 2022). Recently, feature representation learning has delved into self-supervised and semi-supervised learning.

Self-Supervised Learning: Self-supervised feature learning has emerged as a formidable technique in deep learning, showing substantial promise in ophthalmic disease detection due to its capability to utilize unlabeled data(Shi et al., 2024; Zhou et al., 2023). This approach enables models to acquire rich, generalizable features critical for diagnostic

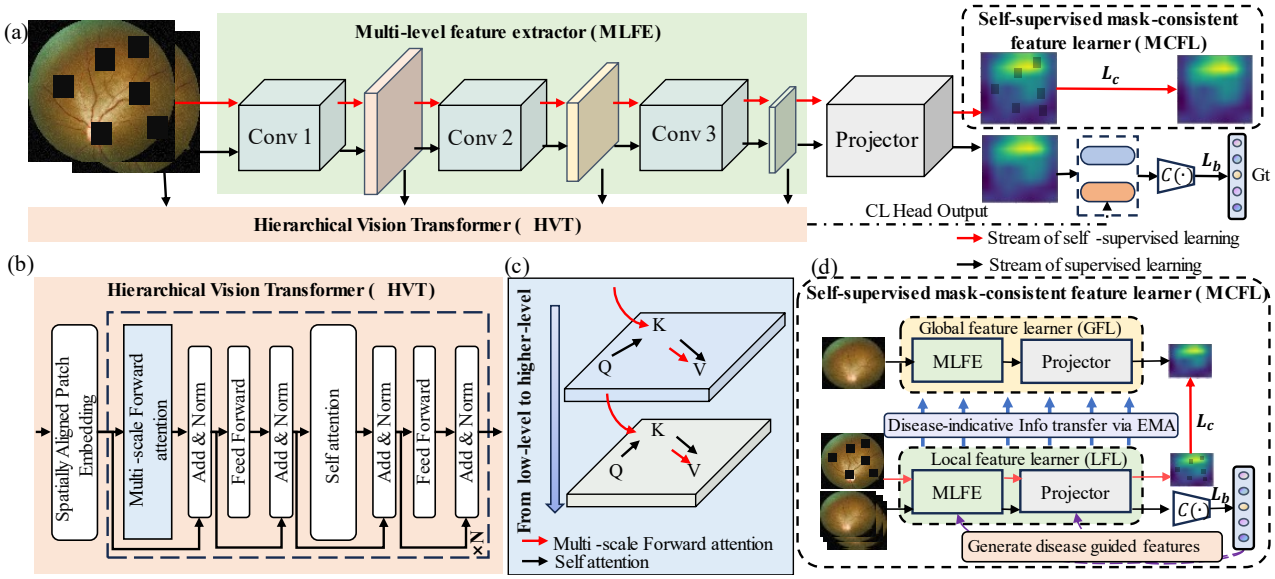

Figure 2: The proposed framework: (a). workflow: Multi-level features is extracted from the multi-level feature extractor (MLFE) and the outputs of the HVT and MCFL is integrated for the final predictions. (b)HVT: It models the inter- and the intra-level information by the proposed forward-attention and the self-attention. (c). The proposed forward-attentions are hierarchically applied to the multi-level feature maps for gradually infusing the information from lower to higher levels. (d). A novel self-supervised masked mean teacher structure is utilized for disease-indicative feature maps extraction.

tasks(Huang et al., 2024). Contrastive learning has been proposed for an effective feature learning in biomedical image processing(Y. Zhang et al., 2024). Wang proposed a two-level causal contrastive model for fundus classification(W. Wang et al., 2024). Another recently developed method in self-supervised learning, the mask-autoencoder(MAE) (He et al., 2022), captures detailed features by reconstructing input images from partially masked versions, demonstrating its effectiveness in feature extraction. Yang proposed a ViT-MAE structured model for DR classification(Yang et al., 2024).

Semi-Supervised Learning: By leveraging both a small amount of labeled data and a large amount of unlabeled data, semi-supervised learning allows models to learn richer and more effective feature representations. Mean teacher method, a popular semi-supervised learning approach, has proven highly effective across various fields(Gu et al., 2023; E. Guo et al., 2023). This method employs two neural networks: a student model trained through standard backpropagation and a teacher model representing an exponential moving average of the student model's weights. This setup stabilizes the learning process and minimizes training signal noise, which is particularly beneficial in handling complex medical imaging data(Tang et al., 2023; Xu et al., 2022).

Challenges and Innovations: Despite their advantages, both the MAE and the mean teacher method have notable limitations. For the MAE, a primary concern is its typical use as a pre-trained feature extraction model; trained without disease-specific information, the extracted features may not directly benefit disease classification tasks in fundus disease detection. Moreover, the MAE requires training both an encoder and a decoder, incurring higher computational costs. On the other hand, the mean teacher method's principal limitation is its requirement for additional data during training, presenting logistical challenges. In response to these limitations, our work introduces the MCFL. It aggregates the advantages of the mean teacher and MAE and is able to learn masked robust feature representations.

Proposed Methods

Figure 2 illustrates the framework of our method. We design two novel models: the HVT and the MCFL. The HVT is employed for fusing the multi-level features and the MCFL is employed for enhancing the feature representation.

Multi-level feature extraction from MLFE

Classic pretrained networks have been employed as the feature extraction backbone, such as Res-Net and VGG net(Luo et al., 2021, 2023), in recent works. Here, we employ the VGG-19, pretrained on ImageNet, as the backbone network. For multi-level analysis of fundus images, we retain the first three convolutional blocks of VGG-19 de-

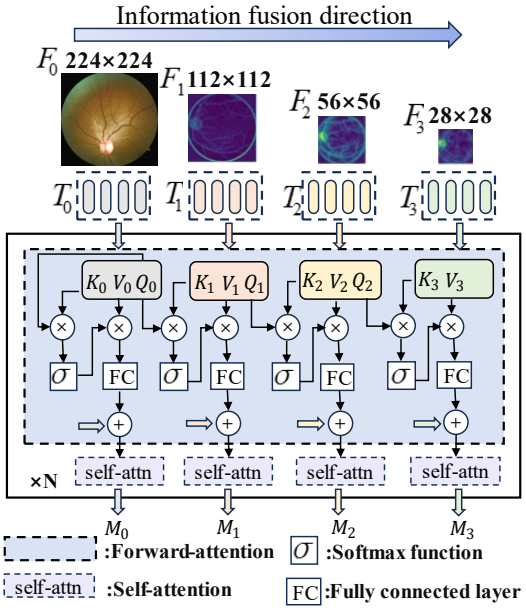

Figure 3: The fusion process of the HVT.

noted as "Conv 1", "Conv 2", and "Conv 3". This modification is intended to preserve more spatial information in the input image for our model.

Given a fundus image F_0 with shape of $224\times224\times3$. Then the multi-level feature maps can be obtained by:

$$F_1 = Conv1(F_0) \in R^{112\times112\times64} \tag{1}$$

$$F_2 = Conv2(F_1) \in R^{56\times56\times128} \tag{2}$$

$$F_3 = Conv3(F_2) \in R^{28\times28\times256} \tag{3}$$

Hierarchical Vision Transformer(HVT)

In this work, we propose a HVT model to integrate the multi-level features, specifically F_0, F_1, F_2 and F_3. The specific fusion process is shown in the figure 3.

The first step is to tokenize the feature maps. Given the varying scales of the feature maps, it is necessary to meticulously design the patch sizes for each feature to ensure that these patches correspond to identical regions across the different levels of feature maps. Therefore, for the feature maps $F_i \in R^{m\times n\times h}$, the patch size P_i is defined as:

$$P_i = \left(\frac{p}{2^i}\right)\times\left(\frac{q}{2^i}\right) \tag{4}$$

where p and q denote pre-defined parameters of the model. Then, the tokens for the multi-level features can be obtained:

$$T_i = \left[T_{i,1};T_{i,2};\cdots;T_{i,L}\right] + E_{pos} \tag{5}$$

where L and E_{pos} denotes the number of patches and the position embedding, respectively. It is worth noting that while the multi-level features F_i vary in scale, the tokens $T_{*,j}$ and $T_{\bullet,j}$ consistently hold semantic information from the same spatial region across different levels.

Forward-attention: In the context of image processing, self-attention focuses on capturing dependencies and relationships within different parts of an image at the same feature level. While it allows the model to refine its internal representations, attention across different levels can adaptively recalibrate which features are more relevant in a given context. To address the issue of reduced accuracy in disease recognition tasks resulting from the neglect of lower-level feature information, we propose the forward-attention. This approach incrementally integrates information from low-level features into high-level features. By doing so, forward-attention ensures that essential details captured by lower layers are preserved and utilized in conjunction with the broader contextual understanding provided by higher layers, leading to improved diagnostic performance.

Here, the token sequence T_i is projected into a query matrix Q_i, a key matrix K_i and a value matrix V_i. Then the forward-attention is computed by:

$$FA_i = \begin{cases} T_i + FC\left(Soft\max\left(\dfrac{Q_i K_i}{\sqrt{D}}\right)V_i\right) & i = 0 \\ T_i + FC\left(Soft\max\left(\dfrac{Q_{i-1} K_i}{\sqrt{D}}\right)V_i\right) & i > 0 \end{cases} \tag{6}$$

As shown in the equation (6) and the figure 3, the traditional self-attention is employed on the F_0. It is because that F_0 is the original image and contains the sparest information, modeling long-term dependencies within F_0 is essential. To capture cross-level relationships within different parts of an image, attention scores are computed by multiplying the queries from the lower level with the keys from the deeper level feature maps. The aggregated attention map is then used to identify high-response regions in the next level feature maps, thereby infusing lower-level information into the deeper levels. The final output is obtained by applying a fully connected layer and adding the original T_i to it. In this study, we utilize the multi-head version of the proposed forward-attention module. Since multi-head attention follows the original self-attention mechanism and can be easily derived, its detailed implementation is omitted in this paper.

As shown in figure 2(c), to balance the modeling of inter- and intra-level feature relationships, self-attention is employed following forward-attention. This dual application ensures that the model can refine its internal representations within each feature level while simultaneously integrating important information across different levels. By doing so, the model maintains a comprehensive understanding of both localized and global features. Finally, the output of the HVT M_i is obtained through N-layer stacked forward-attention and self-attention modules.

Self-supervised Mask-Consistent Feature Learner

Figure 2(d) shows the structure of the self-supervised MCFL, which consists of a global feature learner (GFL) and a local feature learner (LFL). Both the GFL and the LFL share the same MLFE-projector architecture. The MCFL serves as the following three important roles:

(1) Learning the feature map for final classification: The LFL is responsible for learning detailed feature maps that are ultimately used for the final disease classification.

(2) Transferring classification information from LFL to the GFL: The LFL transfers the classification information it learns to the GFL using exponential moving average (EMA) of its weights. This process helps the GFL to aggregate classification-guided information, enabling it to output more disease-indicative features.

(3) Learning mask-consistent feature representations: The LFL, input with random masked images, is optimized to predict consistent feature representations from the pseudo label that obtained from GFL. We employ the mask with the purpose that we can obtain relatively stable features unsensitive for insignificant variants of the input information.

By doing so, it learns to maintain robust feature representations even when parts of the input are masked.

Masking: Given a fundus image F_0, we first divide it into regular non-overlapping patches. Then the randomly masked image $\hat{F}_0$ can be obtained by sampling a subset of patches and randomly masking the remaining ones, as shown in Figure 2(d). This masking process simulates variations in lesion scale, thus creating a task that predicting the disease-indicative feature representations that are irrelevant to the scale of lesion.

Self-supervised consistent loss: More formally, we define the feature representation consistent loss L_c as the expected distance between the prediction of the GFL and the prediction of the LFL:

$$L_c(\theta) = \left\| GFL(F_0, \theta^{\bullet}) - LFL(\hat{F}_0, \theta) \right\|^2 \tag{7}$$

where θ and $\theta^{\bullet}$ denotes the network weights of the LFL and GFL, respectively. It is noting that the weight of the GFL $\theta^{\bullet}$ is not optimized in this process.

Exponential moving average (EMA): To update $\theta^{\bullet}$, we define $\theta_t^{\bullet}$ at training step t as the EMA of successive weights:

$$\theta_t^{\bullet} = \alpha\theta_{t-1}^{\bullet} + (1-\alpha)\theta_t \tag{8}$$

where α is a smoothing coefficient hyperparameter. In this way, the GFL's prediction of each example is formed by an ensemble of the model's current version and those earlier versions that evaluated the same example. This ensemble work improves the quality of the label. By leveraging the ensemble of predictions over time, the GFL aggregates classification-guided information, which enhances the model's capability to output more accurate and disease-indicative features. This approach ensures that the LFL learns from a robust and consistent set of pseudo labels, thereby improving the overall robustness and effectiveness of the feature representations.

After the consistent feature representation learning, F_0 is input to the LFL, obtaining final representation $R_c = LFL(F_0, \theta)$ for disease detection.

Algorithm 1: The training process of our method

Input: Fundus images F_0.

Parameters: HVT patches parameters p and q, masking ratio, masking patch scale, training epochs *Epo*

Output: A trained model.

1: **for** k=1 **to** *Epo* **do**
2: Extract multi-level feature maps by MLFE
3: Input the feature maps into HVT to obtain the lower-level enhanced feature map M_4.
4: Input F_0 to LFL to obtain the consistent feature representation R_c
6: Compute the final prediction scores from R_c and M_4, obtaining *pred*
5: Input masked images $\hat{F}_0$ into LFL to obtain mask-consistent feature maps and input F_0 into EMA updated GFL to obtain reconstructing labels.
6: Compute the consistent loss and the BCEloss
7: Update gradient.
8: **end for**

Multiple fundus disease classification

Since forward-attention aims to propagate information from lower levels to the M_4, the classification head *CL* of M_4 is employed for disease detection. To fuse the consistent feature representation R_c with the *CL*, R_c is flattened and passed through a fully connected layer to ensure it has the same dimension as *CL*, resulting in R_c^f .The final classification results are obtained by concatenating R_c^f and *CL*, taking the maximum value across the concatenated features, and passing the result through a multi-layer perceptron (MLP) followed by a sigmoid activation function:

$$pred = sigmoid(MLPs(\max(concat(R_c^f, Cl)))) \tag{9}$$

This process combines the consistent features learned through the self-supervised method with the hierarchically fused features obtained from HVT, resulting in accurate and reliable disease detection.

Since our task is a multi-label classification problem(C. Liu et al., 2023), the most applicable loss function binary cross entropy is employed as the loss function:

$$L_b = BCEloss(pred, label) \tag{10}$$

The final loss is obtained by:

$$Loss = L_b + L_c \tag{11}$$

The training process of our method can be seen in Algorithm 1.

Experiments

Experimental Setups

Datasets: We conduct experiments on multi-disease fundus dataset (ODIR), which contains 10,000 fundus images from both eyes of 5,000 patients. The dataset can be obtained from (N. Li et al., 2021). For multiple disease detection, diabetic retinopathy (DR), glaucoma (G), cataracts (C), myopia retinopathy (MR), age-related macular degeneration (AMD), and normal (N) fundus, totaling six classes, are employed for classification in this study. The dataset is divided into a training set, an off-site testing set, and an on-site testing set by the data provider. To expand the training dataset and mitigate issues related to overfitting and data imbalance, we employ various augmentation strategies, such as rotation, horizontal flipping, and vertical flipping.

Model details: The backbone of the CNN is VGG19, pretrained on ImageNet. Both self-attention and forward-attention employ eight heads. The depth of HVT is six. The projector of the MCFL is inherited from the last two convolution blocks in VGG19. The scale of the masking non-overlapping patches is 16×16 and the masking ratio is set to 50%.

Evaluation metrics: Evaluation metrics for multiple FD classification include commonly-agreed metrics such as the F1 score (F1), accuracy (Acc.), and Kappa score (Ka.).

Compared methods: We conduct exhaustive comparison experiments on the proposed methods. For a fair comparison, other methods based on techniques directly related with our methods are compared. Since the proposed method is a hybrid ViT-CNN approach, we compare the proposed model to CNN-only methods (ResNet50(Pan et al., 2023), EfficientNet_b7(Tan & Le, 2019), and VGG16(Qassim et al., 2018)), ViT-only methods (ViT_L_16(Dosovitskiy et al., 2021), Swin_B(Z. Liu et al., 2021), and BEiT(Bao et al., 2022)), and hybrid ViT-CNN methods (Pvt (W. Wang et al., 2022) and Cvt_13(H. Wu et al., 2021)). Given that the proposed MCFL draws on the concept of MAE, we also compare it to MAE-related models (Hiera-mae-base and Hiera-mae-tiny)(He et al., 2022; Ryali et al., 2023). Finally, since our proposed method includes a cross-attention-based module, the forward-attention, we also compare it to other open-sourced models with various cross-attention (CA) modules (xcit(A. Ali et al., 2021) and Crossvit (C.-F. (Richard) Chen et al., 2021)).

Types	Models	Off-site testing set			On-site testing set		
		F1	Acc.	Ka.	F1	Acc.	Ka.
CNN	Res-Net50	90.4	84.0	42.4	90.6	84.4	43.8
	Effcient-net_b7	92.6	87.7	56.6	91.8	86.5	51.9
	VGG19	94.6	91.7	70.3	94.2	89.6	64.1
ViT	Vit_l_16	92.6	87.6	54.3	92.5	87.4	53.4
	Swin_b	94.9	91.6	70.3	94.5	90.8	67.2
	BEIT	93.4	89.0	60.9	92.6	87.7	56.2
CNN+ViT	Pvt	95.0	91.8	71.1	94.6	90.9	66.9
	Cvt_13	94.7	91.3	69.1	94.4	90.7	67.0
ViT+MAE	Hiera - mae-tiny	93.5	89.3	61.2	93.4	88.9	60.1
	Hiera - mae-base	94.1	90.2	65.2	93.8	89.6	63.2
ViT+CA	xcit	94.6	91.1	68.5	94.2	90.3	65.6
	Crossvit	94.4	90.7	66.9	94.0	90.0	64.3
The proposed method		**95.9**	**93.1**	**75.5**	**94.8**	**91.3**	**67.7**

Table 1: Comparison of other state-of-art open models.

Main results

In this section, we first compare the performance of our proposed method with other models. To identify strengths and weaknesses specific to each disease condition, we also present a comparison across different fundus diseases. Additionally, to determine the contribution and impact of each component, we provide an ablation study and a hyperparameters analysis.

Comparison with open-sourced state-of-art models: We evaluate the models on both the off-site and on-site testing sets, as presented in Table 1. The results indicate that models combining CNN and ViT architectures outperform those utilizing either ViT or CNN alone. Our proposed method, which fuses multi-level features from both CNN and ViT, achieves superior performance in terms of F1, Acc., and Ka. compared to other fusion CNN-ViT models. When compared to the MAE-enhanced ViT models, our method shows notable improvements. These results demonstrate the effectiveness of our MCFL module in extracting robust and disease-indicative feature representations, surpassing the capabilities of MAE. Furthermore, our method also exceeds the performance of ViT models that incorporate other cross-attention mechanisms, such as those employing cross-attention across different scales of input images(C.-F. (Richard) Chen et al., 2021). This is attributable to our HVT's ability to effectively integrate and utilize hierarchical visual features across CNN layers, enabling more efficient handling of complex visual tasks compared to methods focusing

Model types	Model	N		AMD		G		DR		C		MR		Mean Acc	Mean F1
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1		
CNN	ResNet50	60.5	58.1	94.9	48.6	94.0	41.6	65.4	50.2	94.9	44.1	94.5	16.3	84.0	43.2
	Effcient-net_b7	68.7	69.5	95.0	51.9	93.7	42.0	77.7	63.4	96.0	62.5	95.4	46.9	87.8	56.0
	VGG19	77.2	77.1	97.9	78.4	95.0	47.9	82.1	68.6	99.3	94.5	99.2	92.7	91.8	76.5
ViT	Vit_l_16	66.2	57.0	96.2	62.2	92.1	52.0	73.2	58.0	99.1	92.9	99.1	91.8	87.7	69.0
	Swin_b	78.4	77.0	96.9	74.2	96.0	62.5	83.1	73.0	98.8	90.7	99.3	94.1	92.1	78.6
	BEIT	71.3	66.4	94.5	58.6	94.9	58.7	75.6	63.1	98.9	91.7	99.3	94.1	89.1	72.1
CNN+ ViT	Pvt	77.4	77.5	96.8	63.6	**96.5**	**70.4**	81.6	69.6	99.6	96.8	99.1	90.9	91.8	78.1
	Cvt_13	76.2	74.3	97.9	79.5	95.2	51.4	80.4	69.8	99.1	92.8	99.3	94.1	91.4	76.9
ViT+ MAE	Hiera - mae-tiny	71.3	68.2	96.5	67.5	94.4	47.5	75.4	60.6	99.1	92.8	99.2	92.8	89.3	71.6
	Hiera - mae-base	74.2	71.2	96.6	69.1	95.6	59.2	77.2	65.0	98.8	90.9	99.3	**94.2**	90.3	74.9
ViT+ CA	xcit	76.1	73.7	97.9	79.5	95.2	58.1	79.5	69.9	99.1	92.3	99.1	91.4	91.2	77.5
	Crossvit	74.0	73.9	**98.0**	**82.4**	95.7	61.9	77.7	59.9	99.5	**95.7**	99.2	93.5	90.7	77.9
The proposed method		**81.7**	**81.1**	97.6	78.0	95.4	54.0	**85.6**	**76.3**	**99.6**	94.7	**99.3**	91.6	**93.2**	**79.3**

Table 2: Comparison of the models across the diseases. The best results are highlighted in bold.

Ablation Studies	Off-site testing set			On-site testing set		
	F1	Acc.	Ka.	F1	Acc	Ka.
MLFE	95.0	91.7	70.0	94.2	89.6	64.1
MLFE+MCFL	95.5	92.6	73.8	94.8	89.8	64.8
MLFE+HVT	95.3	92.2	72.2	94.5	90.6	65.6
MLFE+MCFL + HVT	**95.9**	**93.1**	**75.5**	**94.8**	**91.3**	**67.7**

Table 3: Ablation studies in the proposed method.

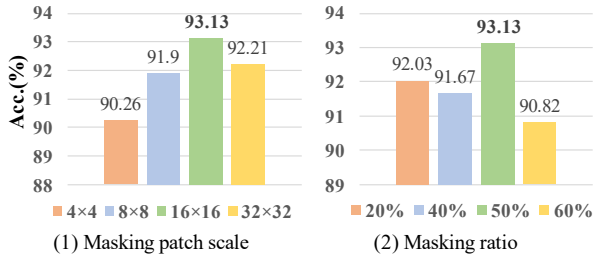

Figure 4: Hyperparameter analysis of the MCFL

solely on single-scale or same-layer attention. From the table, it is evident that the proposed model achieves a significantly higher Ka. score compared to other methods. A higher Ka. indicates superior agreement between the predicted classifications and the true labels, beyond what would be expected by chance. In particular, Table 2 provides the Acc. and F1 score for each FD. For a comprehensive analysis, the mean values of F1 and Acc. are also provided. It is evident that the proposed method achieves the highest mean F1 and Acc. scores among all methods.

Ablation Studies and Hyperparameter Evaluations: Through the experiment, we scrutinize the contribution of each component to the proposed model, as described in Table 3. It can be seen that employing the MCFL and HVT modules separately or together will both improves the performance of FD detection. The best classification results are achieved when both modules are included. We analyze the influence of mask ratio and scale in the proposed MCFL. Figure 4 shows that a mask scale of 16×16 achieves the highest accuracy at 93.13%, indicating optimal feature learning balance. A mask ratio of 50% also yields the best accuracy, suggesting it effectively balances information retention and masking.

Conclusions

In this paper, we introduce an effective framework for classification of multiple FD, which fully exploits the rich knowledge about disease-indicative features and hierarchical structure within fundus. We successfully tame the model with the proposed HVT to integrate the hierarchical semantic information from higher to lower levels. Both the inter-level and intra-level semantic representations are enhanced at various granularities. Furthermore, we introduce a self-supervised MCFL module to extract the most disease-relevant feature maps. By predicting the invariant disease-indicative feature from random-masked fundus images, the proposed model is able to maintain the disease-related representations despite changes in lesion scale. Extensive experiments demonstrate our method can significantly improves the classification performance of multiple fundus diseases and achieve most accurate FD recognition among other state-of-art models. In the future, we will focus on enhancing our multi-disease recognizing system's functionality, including the segmentation of various FD lesions, thus further aiding in diagnosis.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62371157, Shenzhen Science and Technology Program under Grant KJZD20230923114600002 and JCYJ20240813105135047, the project of Shenzhen Science and Technology Innovation Committee under Grant JCYJ20240813141424032 and the Foundation for Young innovative talents in ordinary universities of Guangdong under Grant 2024KQNCX042.

References

Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., & Jegou, H. (2021). XCiT: Cross-Covariance Image Transformers. Advances in Neural Information Processing Systems, 34, 20014–20027.

Ali, G., Dastgir, A., Iqbal, M. W., Anwar, M., & Faheem, M. (2023). A Hybrid Convolutional Neural Network Model for Automatic Diabetic Retinopathy Classification From Fundus Images. IEEE Journal of Translational Engineering in Health and Medicine, 11, 341–350. https://doi.org/10.1109/JTEHM.2023.3282104

Bao, H., Dong, L., Piao, S., & Wei, F. (2022). BEiT: BERT Pre-Training of Image Transformers (arXiv:2106.08254). arXiv. https://doi.org/10.48550/arXiv.2106.08254

Chen C F R, Fan Q, Panda R. Crossvit: Cross-attention multi-scale vision transformer for image classification[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 357-366.

Chen, F., Ma, S., Hao, J., Liu, M., Gu, Y., Yi, Q., Zhang, J., & Zhao, Y. (2023). Dual-Path and Multi-Scale Enhanced Attention Network for Retinal Diseases Classification Using Ultra-Wide-Field Images. IEEE Access, 11, 45405–45415. https://doi.org/10.1109/ACCESS.2023.3273613

Das, D., Nayak, D. R., & Pachori, R. B. (2023). CA-Net: A Novel Cascaded Attention-Based Network for Multistage Glaucoma Classification Using Fundus Images. IEEE Transactions on Instrumentation and Measurement, 72, 1–10. https://doi.org/10.1109/TIM.2023.3322499

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Flaxman, S. R., Bourne, R. R. A., Resnikoff, S., Ackland, P., Braithwaite, T., Cicinelli, M. V., Das, A., Jonas, J. B., Keeffe, J., Kempen, J. H., Leasher, J., Limburg, H., Naidoo, K., Pesudovs, K., Silvester, A., Stevens, G. A., Tahhan, N., Wong, T. Y., Taylor, H. R., … Zheng, Y. (2017). Global causes of blindness and distance vision impairment 1990–2020: A systematic review and meta-analysis. The Lancet Global Health, 5(12), e1221–e1234. https://doi.org/10.1016/S2214-109X(17)30393-5

Ghahremani M, Khateri M, Jian B, et al. H-ViT: A Hierarchical Vision Transformer for Deformable Image Registration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 11513-11523.

Gu, R., Zhang, J., Wang, G., Lei, W., Song, T., Zhang, X., Li, K., & Zhang, S. (2023). Contrastive Semi-Supervised Learning for Domain Adaptive Segmentation Across Similar Anatomical Structures. IEEE Transactions on Medical Imaging, 42(1), 245–256. https://doi.org/10.1109/TMI.2022.3209798

Guo, E., Fu, H., Zhou, L., & Xu, D. (2023). Bridging Synthetic and Real Images: A Transferable and Multiple Consistency Aided Fundus Image Enhancement Framework. IEEE Transactions on Medical Imaging, 42(8), 2189–2199. https://doi.org/10.1109/TMI.2023.3247783

Guo, S. (2021). Fundus image segmentation via hierarchical feature learning. Computers in Biology and Medicine, 138, 104928. https://doi.org/10.1016/j.compbiomed.2021.104928

Guo, S., Li, T., Kang, H., Li, N., Zhang, Y., & Wang, K. (2019). L-Seg: An end-to-end unified framework for multi-lesion segmentation of fundus images. Neurocomputing, 349, 52–63. https://doi.org/10.1016/j.neucom.2019.04.019

He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16000-16009.

Huang, Y., Lyu, J., Cheng, P., Tam, R., & Tang, X. (2024). SSiT: Saliency-Guided Self-Supervised Image Transformer for Diabetic Retinopathy Grading. IEEE Journal of Biomedical and Health Informatics, 28(5), 2806–2817. https://doi.org/10.1109/JBHI.2024.3362878

Li, N., Li, T., Hu, C., Wang, K., & Kang, H. (2021). A Benchmark of Ocular Disease Intelligent Recognition: One Shot for Multi-disease Detection. In F. Wolf & W. Gao (Eds.), Benchmarking, Measuring, and Optimizing (pp. 177–193). Springer International Publishing. https://doi.org/10.1007/978-3-030-71058-3_11

Li, Y., Zhang, Y., Liu, J.-Y., Wang, K., Zhang, K., Zhang, G.-S., Liao, X.-F., & Yang, G. (2023). Global Transformer and Dual Local Attention Network via Deep-Shallow Hierarchical Feature Fusion for Retinal Vessel Segmentation. IEEE Transactions on Cybernetics, 53(9), 5826–5839. https://doi.org/10.1109/TCYB.2022.3194099

Lin, J., Cai, Q., & Lin, M. (2021). Multi-Label Classification of Fundus Images With Graph Convolutional Network and Self-Supervised Learning. IEEE Signal Processing Letters, 28, 454–458. https://doi.org/10.1109/LSP.2021.3057548

Lin, Y., Wing-Kuen Ling, B., Wang, W., Hu, L., Xu, N., & Zhou, X. (2023). Fusion of electroencephalograms at different channels and different activities via multivariate quaternion valued singular spectrum analysis for intellectual and developmental disorder recognition. Biomedical Signal Processing and Control, 79, 104256. https://doi.org/10.1016/j.bspc.2022.104256

Liu, C., Wen, J., Luo, X., & Xu, Y. (2023). Incomplete Multi-View Multi-Label Learning via Label-Guided Masked View- and Category-Aware Transformers (arXiv:2303.07180). arXiv. http://arxiv.org/abs/2303.07180

Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

Luo, X., Liu, C., Wong, W., Wen, J., Jin, X., & Xu, Y. (2023). MVCINN: Multi-view diabetic retinopathy detection using a deep cross-interaction neural network. Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial

Intelligence, 37, 8993–9001. https://doi.org/10.1609/aaai.v37i7.26080

Luo, X., Pu, Z., Xu, Y., Wong, W. K., Su, J., Dou, X., Ye, B., Hu, J., & Mou, L. (2021). MVDRNet: Multi-view diabetic retinopathy detection by combining DCNNs and attention mechanisms. Pattern Recognition, 120, 108104. https://doi.org/10.1016/j.patcog.2021.108104

Pan, Y., Liu, J., Cai, Y., Yang, X., Zhang, Z., Long, H., Zhao, K., Yu, X., Zeng, C., Duan, J., Xiao, P., Li, J., Cai, F., Yang, X., & Tan, Z. (2023). Fundus image classification using Inception V3 and ResNet-50 for the early diagnostics of fundus diseases. Frontiers in Physiology, 14. https://doi.org/10.3389/fphys.2023.1126780

Qassim, H., Verma, A., & Feinzimer, D. (2018). Compressed residual-VGG16 CNN model for big data places image recognition. 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), 169–175. https://doi.org/10.1109/CCWC.2018.8301729

Ryali, C., Hu, Y.-T., Bolya, D., Wei, C., Fan, H., Huang, P.-Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., Malik, J., Li, Y., & Feichtenhofer, C. (2023). Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles. Proceedings of the 40th International Conference on Machine Learning, 29441–29454.

Shi, D., Zhang, W., Chen, X., Liu, Y., Yang, J., Huang, S., Tham, Y. C., Zheng, Y., & He, M. (2024). EyeFound: A Multimodal Generalist Foundation Model for Ophthalmic Imaging (arXiv:2405.11338). arXiv. https://doi.org/10.48550/arXiv.2405.11338

Sun, K., He, M., Xu, Y., Wu, Q., He, Z., Li, W., Liu, H., & Pi, X. (2022). Multi-label classification of fundus images with graph convolutional network and LightGBM. Computers in Biology and Medicine, 149, 105909. https://doi.org/10.1016/j.compbiomed.2022.105909

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning, 6105–6114.

Tang, L., Li, K., He, C., Zhang, Y., & Li, X. (2023). Source-Free Domain Adaptive Fundus Image Segmentation with Class-Balanced Mean Teacher. In H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, & R. Taylor (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 (pp. 684–694). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43907-0_65

Wang, J., Yang, L., Huo, Z., He, W., & Luo, J. (2020). Multi-Label Classification of Fundus Images With EfficientNet. IEEE Access, 8, 212499–212508. https://doi.org/10.1109/ACCESS.2020.3040275

Wang, W., Quan, X., Huang, W., Cheng, Y., & Zhang, H. (2024). TL-CCL: Two-level causal contrastive learning for multi-label ocular disease diagnosis with fundus images. Biomedical Signal Processing and Control, 95, 106308. https://doi.org/10.1016/j.bspc.2024.106308

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2022). PVT v2: Improved baselines with Pyramid Vision Transformer. Computational Visual Media, 8(3), 415–424. https://doi.org/10.1007/s41095-022-0274-8

Wang, X., Xu, M., Zhang, J., Jiang, L., Li, L., He, M., Wang, N., Liu, H., & Wang, Z. (2022). Joint Learning of Multi-Level Tasks for Diabetic Retinopathy Grading on Low-Resolution Fundus Images. IEEE Journal of Biomedical and Health Informatics, 26(5), 2216–2227. https://doi.org/10.1109/JBHI.2021.3119519

Wu H, Xiao B, Codella N, et al. Cvt: Introducing convolutions to vision transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 22-31.

Wu, L., Xu, Y., Hou, J., Chen, C. L. P., & Liu, C.-L. (2023). A Two-Level Rectification Attention Network for Scene Text Recognition. IEEE Transactions on Multimedia, 25, 2404–2414. https://doi.org/10.1109/TMM.2022.3146779

Wu, Z., Xu, Y., Yang, J., & Li, X. (2024). Misclassification in Weakly Supervised Object Detection. IEEE Transactions on Image Processing, 33, 3413–3427. https://doi.org/10.1109/TIP.2024.3402981

Xu, Z., Lu, D., Luo, J., Wang, Y., Yan, J., Ma, K., Zheng, Y., & Tong, R. K.-Y. (2022). Anti-Interference From Noisy Labels: Mean-Teacher-Assisted Confident Learning for Medical Image Segmentation. IEEE Transactions on Medical Imaging, 41(11), 3062–3073. https://doi.org/10.1109/TMI.2022.3176915

Yang, Y., Cai, Z., Qiu, S., & Xu, P. (2024). Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image. PLOS ONE, 19(3), e0299265. https://doi.org/10.1371/journal.pone.0299265

Yen, G. G., & Leong, W.-F. (2008). A Sorting System for Hierarchical Grading of Diabetic Fundus Images: A Preliminary Study. IEEE Transactions on Information Technology in Biomedicine, 12(1), 118–130. https://doi.org/10.1109/TITB.2007.910453

Zhang, X., Tian, Y., Huang, W., Ye, Q., Dai, Q., Xie, L., & Tian, Q. (2022). HiViT: Hierarchical Vision Transformer Meets Masked Image Modeling (arXiv:2205.14949). arXiv. https://doi.org/10.48550/arXiv.2205.14949

Zhang, Y., Ma, X., Huang, K., Li, M., & Heng, P.-A. (2024). Semantic-oriented Visual Prompt Learning for Diabetic Retinopathy Grading on Fundus Images. IEEE Transactions on Medical Imaging, 1–1. https://doi.org/10.1109/TMI.2024.3383827

Zhang, Z., Zhang, H., Zhao, L., Chen, T., Arik, S. Ö., & Pfister, T. (2022). Nested Hierarchical Transformer: Towards Accurate, Data-Efficient and Interpretable Visual Understanding. Proceedings of the AAAI Conference on Artificial Intelligence, 36(3), Article 3. https://doi.org/10.1609/aaai.v36i3.20252

Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., Liu, T., Xu, M., Lozano, M. G., Woodward-Court, P., Kihara, Y., Altmann, A., Lee, A. Y., Topol, E. J., Denniston, A. K., Alexander, D. C., & Keane, P. A. (2023). A foundation model for generalizable disease detection from retinal images. Nature, 622(7981), 156–163. https://doi.org/10.1038/s41586-023-06555-x

Zhu H, Ke W, Li D, et al. Dual cross-attention learning for fine-grained visual categorization and object re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 4692-4702.