This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

SABPI-Net: A Structure-aware Bidirectional Proxy Interaction Network for Infantile Retinal Disease Diagnosis

Shaobin Chen, Xinyu Zhao, Huazhu Fu, *Senior Member, IEEE,* Jiaju Huang, Zhenquan Wu, Behdad Dashtbozorg, *Member, IEEE,* Baiying Lei, *Senior Member, IEEE,* Guoming Zhang, Yue Sun

***Abstract*— Delayed treatment of infantile retinal disease can reduce its effectiveness and may cause severe and irreversible damage. Automated diagnosis of infant retinal diseases faces challenges including subtle early lesions, diverse clinical phenotypes, imaging variations, and imbalanced data. To address these, which cannot be well addressed by existing general foundation models, we propose structure-aware bidirectional proxy interaction network (SABPI-Net) in a universal learning framework. SABPI-Net incorporates a high-frequency mapping branch, and employs a proposed proxy interaction attention module to enable effective interaction between its trunk feature encoding branch and the high-frequency mapping branch, thereby facilitating enhanced perception of retinal detail structures. Domain-agnostic embedding space self-matching, guided by a memory-bank low-frequency component replacement strategy, promotes domain-invariant learning and consistent model performance under diverse image styles. Finally, the tail-aware feature fusion strategy for fine-tuning further enhances the model's diagnostic sensitivity to tailed diseases. In this study, three classification tasks related to infant retinal diseases are implemented on the largest clinical infant retina dataset to date, covering 19 infant retinal diseases or normal conditions. SABPI-Net achieves superior performance compared to 13 SOTA methods, with 95.32% accuracy on mainstream clinical tasks, 73.58% on ROP five-stage classification, and 84.25% on multi-disease classification, representing improvements of 1.57%, 1.88%, and 4.71% respectively over the best competing methods. Extensive experiments demonstrate the effectiveness and superiority of SABPI-Net in diagnosing infant retinal diseases.**

***Index Terms*— Infantile retinal diseases, Structure-aware, Interaction, Tail-aware**

This work was supported partly by Shenzhen Medical Research Fund (Nos. C2301005, D250403004, C2501034, A2403020), Macao Polytechnic University Grant (RP/FCA-16/2025), Science and Technology Development Fund of Macao (0009/2025/ITP1), STCSM (22DZ2229005), National Natural Science Foundation of China (Nos. 82401315, 82271103, 82301269), Sanming Project of Medicine in Shenzhen (SZSM202311018) and Shenzhen Science and Technology Program (JCYJ20240813152703005). We would also like to express our sincere gratitude to Dr. Tao Tan for his invaluable guidance and support throughout this research.

(Corresponding authors: Baiying Lei; Guoming Zhang; Yue Sun)

Shaobin Chen, Jiaju Huang and Yue Sun are with the Intelligent Medical Computing Laboratory, Faculty of Applied Sciences, Macao Polytechnic University, Macao, 999078, China (e-mail: p2314850@mpu.edu.mo; p2316877@mpu.edu.mo; yuesun@mpu.edu.mo).

Xinyu Zhao, Zhenquan Wu and Guoming Zhang are with the Shenzhen Eye Hospital, Shenzhen Eye Medical Center, Southern Medical University, Shenzhen, 518040, China (e-mail: zhaoxinyu1@sz-eyes.com; wuzhenquan@sz-eyes.com; zhangguoming@sz-eyes.com).

Baiying Lei is with the Health Science Center, National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Shenzhen University, Shenzhen, China (e-mail:leiby@szu.edu.cn).

Huazhu Fu is with the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore (e-mail: hzfu@ieee.org)

Behdad Dashtbozorg is with the Netherlands Cancer Institute, Amsterdam, The Netherlands (e-mail: b.dasht.bozorg@nki.nl).

I. Introduction

THE WHO's vision programme has recently prioritized the reduction of childhood blindness as a key objective [1], [2]. Common infantile retinal diseases include retinopathy of prematurity (ROP), retinoblastoma, retinitis pigmentosa, morning glory syndrome and coloboma of the choroid, etc. As the leading cause of visual impairment and irreversible blindness in children worldwide, ROP is a vasoproliferative disease that affects preterm infants of very low birth weight or those who are small for their gestational age [3], [4]. As neonatal care has undergone significant advancements, the survival rate of preterm infants has increased considerably. This leads to an increased demand for ROP screening, as an increasing number of at-risk infants require early detection and management. Although most mild cases of ROP have a favorable prognosis, 5-10% require aggressive intervention to prevent blindness [5]. Currently, clinical diagnosis of ROP relies on binocular indirect ophthalmoscopy or wide-field digital fundus image interpretation by experienced ophthalmologists [6]. However, clinical diagnosis of ROP remains subjective, and a high degree of diagnostic inconsistency has been observed even among ROP specialists [7]. The cost of ROP screening is high on a global scale due to the lack of widespread implementation of cost-effective screening strategies [8]. Given the short window for ROP treatment, there is an urgent need to establish a screening tool that can rapidly identify fundus images that require further attention and rigorous analysis by ophthalmologists, thereby improving the efficiency of early screening for ROP.

Recent studies have leveraged artificial intelligence (AI) to automate the diagnosis of ocular diseases, presenting a

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

promising approach to expanding healthcare access in remote regions where experienced ophthalmologists are unavailable. Multi-level feature fusion strategies have demonstrated effectiveness across various medical imaging domains, including COVID-19 lesion segmentation [9], Alzheimer's disease diagnosis [10], skin lesion recognition [11], and breast abnormality detection [12]–[14]. Despite these advances, current deep learning models for retinal disease analysis still face significant limitations in handling the unique challenges of infantile retinal diseases. Table I provides a comprehensive summary of recent state-of-the-art (SOTA) deep learning approaches for retinal disease analysis, highlighting their key contributions and limitations in addressing challenges such as subtle pathology detection, domain invariance, and long-tail distribution handling. Previous studies have employed traditional regression analyses, utilizing birth characteristics to predict the occurrence of ROP [15], [16]. However, retinal status has also been identified as a contributing factor in the prediction of ROP, yet these prediction models have been shown to be flawed by ignoring retinal status [17]–[19]. Recent studies have shown some encouraging results in the use of AI systems for automated ROP screening using color fundus photography (CFP). Various studies, including those by Huang et al. [20] and Zhang et al. [21], have employed different convolutional neural networks (CNNs) to compare their effectiveness in detecting ROP [22]–[25].

TABLE I
SUMMARY OF EXISTING WORK IN RETINAL DISEASES ANALYSIS

Method	Contribution	Limitation
Bi et al. [26], 2023	Combines ViT global semantics with multi-instance learning (MIL) local features via calibrated attention for instance aggregation.	Relies on MIL for local feature extraction, which may not adequately capture fine-grained texture details within instances.
Wang et al. [27], 2023	Dirichlet-based uncertainty model providing OOD detection scores alongside diagnostic results.	Focuses on OOD rejection rather than improving rare class classification accuracy.
Zhou et al. [28], 2023	Self-supervised ViT foundation model learning universal, generalizable feature representations.	General model lacks infantile pathology-specific feature extraction capabilities.
Zhu et al. [29], 2024	Optimized CNN leveraging inductive bias for small lesion detection.	Limited long-range dependency and global context modeling for complex phenotypes.
Chen et al. [30], 2023	Fused local CNN features with global multi-scale Transformer features, using interactive attention to facilitate cross-scale information exchange.	Fails to explicitly separate high-frequency structural details from style information. Lacks specific mechanisms for long-tail distribution handling.
Wen et al. [31], 2024	Interpretable concept-based approach with cross-attention and retinal foundation model guidance.	Concept bottleneck creates interpretability-accuracy trade-offs; requires lesion annotations.
Singh et al. [32], 2025	Patch-based prototype network comparing test patches with learned disease prototypes.	Patch similarity approach may miss complex global and high-frequency patterns.
Wei et al. [33], 2025	A multi-scale spatial-aware Transformer with dual-path structure using different patch sizes, combining spatial-aware modules and MIL strategies.	Although spatially aware, the model does not separate frequency and style for better domain invariance. MIL may also miss fine details in patches.
Zhang et al. [34], 2025	Uses saliency to detect lesion pixels and builds graph neural networks (GNNs) to model their spatial and topological patterns.	Focuses on spatial patterns of lesions, but may lack robustness to imaging variability and long-tail cases. Graph construction is computationally expensive.
Hou et al. [35], 2025	Quality-aware framework with sample re-weighting loss for mixed noise robustness.	Designed mainly for noise robustness, not for subtle feature detection or long-tail issues.

Several studies have developed deep learning systems to recognize the presence of plus disease, and some have sought to further categorize pre-plus disease [36]–[40]. As the presence and severity of plus disease is determined by vascular abnormalities, the majority of these existing methods aim to measure statistics such as the diameter and curvature of the retinal vessels. Consequently, the accuracy of these measurements is contingent on the outcomes of vessel segmentation in CFP images, implying that errors in segmentation may be amplified in subsequent measurements.

In recent years, more and more studies have begun to focus on other clinical features of ROP. For instance, studies on ROP severity grading seek to subclassify images that are identified as having features consistent with ROP according to the severity of the disease [22], [38], [41]. Hu et al. [22] proposed a two-sub-network architecture interconnected by a feature aggregation operator. The first sub-network extracts high-level features from fundus images, which are then fused by the operator and fed into the second sub-network for class prediction. Tong et al. [38] trained a ResNet architecture and a faster region-based convolutional neural network for classification and recognition of ROP images. Xie et al. [42] introduced an adversarial learning-based multi-level dense knowledge distillation approach for ROP detection. The approach utilizes a pre-trained teacher network to densely transfer knowledge to multiple intermediate and a student network, where each lower-level network is guided by knowledge from all upper-level networks. Feature similarity between adjacent networks is ensured through an adversarial learning module. Subsequent studies incorporated aggressive posterior ROP into the grading system [21], [42]. Peng et al. [43] proposed a network for five-level staging of ROP that combines a multistream parallel feature extractor, a concurrent-based deep feature fusion, and a clinical practice-based sequential classifier.

However, these studies have three critical shortcomings. Firstly, the automatic classification of ROP is not yet comprehensive, as it mainly focuses on the identification of ROP occurrence, plus disease and severity. A complete diagnostic system that can meet the needs of complex clinical analyses is lacking. For example, this system must cover routine ROP, aggressive ROP (A-ROP), laser-treated ROP and ROP refinement staging, etc. Secondly, in addition to ROP, there are other retinal diseases that require attention in infants. Finally, there is still scope for enhancing the performance and disease coverage of existing models based on small datasets. Automated diagnosis of retinal diseases in infants also faces the following challenges:

A) Early-stage infantile retinal diseases are marked by subtle and difficult-to-detect characteristic changes [44], [45]. As shown in Fig. 1, Stage 1 of ROP appears as a thin demarcation line separating vascularized and non-vascularized areas on CFP images. By Stage 2, this demarcation line develops into a ridge.
B) The diversity and complexity of retinal disease phenotypes in infants require a model capable of extracting more distinct features. ROP can be categorized into routine ROP, laser-treated ROP, and A-ROP, with routine ROP further divided into five stages [44]. Additionally, there are over 10 other types of infantile retinal diseases.
C) The imaging heterogeneity in CFP further complicates effective feature extraction. The same disease can exhibit diverse phenotypes under different imaging conditions, and even under the same conditions, lesion manifestations can vary significantly.
D) The long-tailed distribution of disease reduces the model's sensitivity to tailed diseases, leading to lower diagnostic sensitivity for these diseases.

It is essential to finely perceive the detailed structure of retinal images to extract more discriminative features. According to the retinex theory [46], the detailed structure of the retina is concentrated in the high-frequency domain of the retinal image, while the style texture of the image is concentrated in the low-frequency component. In recent years,

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

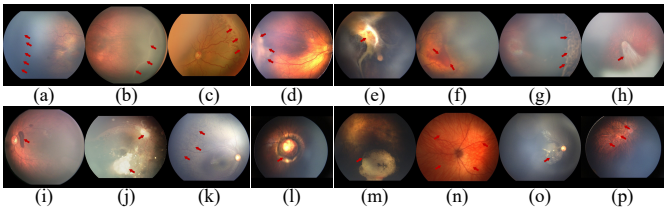


Fig. 1. CFP images of infants taken by RetCam 3. (a) Stage 1 of ROP. (b) Stage 2 of ROP. (c) Stage 3 of ROP. (d) Stage 4 of ROP. (e) Stage 5 of ROP. (f) A-ROP. (g) Laser-treated ROP. (h) Optic disc traction. (i) Hemorrhage. (j) Retinoblastoma. (k) Retinitis pigmentosa. (l) Morning glory syndrome. (m) Coloboma of choroid. (n) Oculocutaneous albinism. (o) Macular disease. (p) Tessellated retina.

spectral knowledge has received more attention in the field of retinal disease diagnosis. Liu et al. [47] analyzed the impact of frequency on inter-domain differences and found that the structural details of retinal images are mainly concentrated in the high-frequency part. Li et al. [48] developed a generic fundus image enhancement network, which utilizes image frequency information and extracts robust structural perception representations from degraded images. According to [49], [50], the domain shift between the source domain and target domain can be reduced by exchanging/integrating low-frequency spectrum. Chen et al. [51] proposed a test-time adaptation method using visual prompts to address domain shift in optic disc and cup segmentation of retinal images. This method leverages low-frequency components to capture image style and texture, and trains specific low-frequency prompts for each test image.

To address the challenges of automatic diagnosis of retinal diseases in infants, we propose a structure-aware bidirectional proxy interaction network (SABPI-Net), which explicitly models the structural information of CFP images and leverages proxy interaction mechanisms to enhance feature representation capabilities, enabling accurate and reliable recognition of retinal diseases in infants. Our main contributions include:

1) To address pathological diversity (challenges A and B), we design a high-frequency mapping branch. This branch is integrated with the trunk encoding branch via a proxy interactive attention module, improving the model's ability to capture detailed retinal structures.
2) To mitigate imaging heterogeneity (challenge C), we propose a domain-agnostic embedding space self-matching strategy, ensuring consistent model performance across different imaging styles.
3) To enhance diagnostic sensitivity for tailed class (challenge D), we adopt a tail-aware feature fusion fine-tuning strategy, which optimizes recognition of these classes.
4) We construct a dataset of over 170,000 CFP images, enabling three classification tasks related to 19 infant retinal diseases or normal conditions. This dataset surpasses existing studies in both scale and disease coverage in infants.

Codes and models of SABPI-Net will be available at https://github.com/SB-Chen/SABPINet-IRD.

II. Related Work

A. Deep neural networks for the diagnosis of ROP

According to the International Classification of Retinopathy of Prematurity [44], [45], the assessment of ROP is to be conducted in accordance with the following criteria: the location of the lesion (zones I-III), the extent of the lesion (12 sectors according to the clock hour), the stage of the disease (stages 1-5), and the presence of plus or pre-plus disease and A-ROP.

Several studies [20]–[23] employed different models for the automatic detection of ROP occurrence or non-occurrence. Zhang et al. [21] integrated a hierarchical bilinear pooling module and a channel attention module into a series of ResNet architectures to enhance the model's feature extraction capability. Wang et al. [23] developed two deep neural networks for ROP identification and severity grading, respectively. Following this, Hu et al. [22] proposed a CNN that connects the two sub-networks via a feature aggregation operator to achieve the same task. Pang et al. [36] developed an automated multi-dimensional platform using retinal images for the classification tasks of the four major dimensions of ROP, including image quality, the presence or absence of ROP, intraocular hemorrhage, and pre-plus/plus disease.

Plus disease is one of the key characteristics of ROP, which manifest on retinal images as dilated and distorted retinal vessels. Consequently, numerous studies were dedicated to the automated identification of plus disease [37], [39], [40]. For instance, Brown et al. [37] proposed a two-stage triple classification diagnostic method for ROP, which firstly extracted retinal vessel map using the U-Net network and then classified the segmented map (i.e. normal, pre-plus, and plus) based on the Inception network. A two-stage strategy was employed by Ramachandran et al. [39], with the focus here being on distinguishing infants with plus disease from healthy infants. The initial stage entailed the utilisation of the YOLO network for the detection of vessel curvature, followed by a subsequent quantification of vessel curvature in a second stage. This quantification was then employed for the assessment of the extent of plus disease.

B. Deep neural networks for the diagnosis of other retinal diseases

Recent advances in deep learning for retinal disease analysis have demonstrated significant progress in adult populations, yet several fundamental limitations remain unaddressed for infantile retinal diseases. Current approaches can be broadly categorized into multi-instance learning methods, foundation models, and hybrid architectures, each with distinct advantages and limitations. Multi-instance and attention-based approaches have attempted to capture both global and local features through various attention mechanisms [26], [32], [34]. While these methods show promise for capturing spatial relationships and prototype-based comparisons, they often struggle with fine-grained texture details critical for subtle infantile pathologies. The reliance on patch-based similarity or graph construction may miss complex global patterns and high-frequency structural details that are crucial for distinguishing early-stage pathological changes in infant retinas. Foundation models such as RETFound have achieved remarkable success

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

in adult retinal disease recognition through self-supervised learning on large-scale unlabeled data [28], [52]. However, these general-purpose models lack specialized mechanisms for capturing infantile pathology-specific features, particularly the subtle structural changes characteristic of early-stage retinal diseases in infants. The anatomical and physiological differences between adult and infant retinas necessitate specialized approaches beyond general foundation models.

Similarly, uncertainty-based frameworks focus primarily on out-of-distribution detection rather than improving classification accuracy for rare diseases [27]. Hybrid CNN-Transformer architectures have emerged to combine local detail extraction with global context modeling [30], [31], [33], [53]. Despite their effectiveness in capturing multi-scale information and providing interpretability through concept bottlenecks, existing methods fail to explicitly separate high-frequency structural details from low-frequency style information, limiting their ability to achieve domain invariance across different imaging conditions. Additionally, most approaches lack specific mechanisms for handling the long-tail distribution challenges inherent in rare infantile retinal disease classification, while some methods designed for noise robustness do not address the fundamental challenges of subtle feature detection [35]. Current literature has three key limitations: (1) ignoring frequency domain characteristics that contain crucial structural information, (2) lacking effective strategies for style-invariant learning across diverse imaging conditions, and (3) inadequately addressing long-tail distribution in tail diseases. To address these issues, we propose SABPI-Net, which models high-frequency structural information, achieves domain invariance through frequency-based style manipulation, and enhances tail disease classification through targeted feature fusion.

III. Methods

A. Overview

The overall framework of our SABPI-Net for infantile retinal disease recognition is shown in Fig. 2. The SABPI-Net is mainly composed of four parts: a high-frequency mapping (HFM) branch, a proxy interaction attention module, domain-agnostic embedding space self-matching learning, and a tail-aware feature fusion fine-tuning strategy. The HFM branch focuses on the high-frequency components of the CFP images, and the trunk feature encoding branch of the CFP images interacts with the HFM branch through the proxy interaction attention module to enhance the model's perception of the detailed structures of the fundus. The domain-agnostic embedding space self-matching learning facilitates the model to maintain consistent performance under various retinal image styles. The tail-aware feature fusion fine-tuning strategy enhances the diagnostic sensitivity of the model to the tailed class. The following section describes each component and its synergistic approach.

B. High frequency mapping

Based on the Retinex theory [46], the high frequency region of the fundus image contains the most important details of the retina. In clinical diagnosis, the characteristic changes of fundus diseases are usually identified by observing the detailed structure of the retina. Therefore, the ability of the model to finely perceive the detailed structure of retinal images is critical to the extraction of discriminative features. Our structure-aware representation explicitly models anatomical structures in the high-frequency domain. After frequency decomposition, critical retinal features (vessel architecture, optic disc boundaries, macular regions, and pathological lesions) are predominantly preserved in high-frequency components. The high-frequency mapping branch isolates and amplifies these structural elements, enabling focused detection of anatomically relevant features such as vessel changes, ridge formations, hemorrhagic patterns, and tissue distortions characteristic of infantile retinal diseases.

In order to effectively extract the high-frequency components of the fundus region and eliminate the interference of irrelevant background regions, it is necessary to fill the median of the image with the invalid region (the part with a mask of 0) to avoid the influence of invalid data on the subsequent calculation. The retinal region retains the original pixel value, which is given by the following equation:

$$x_p = M(x) \times x + (1 - M(x)) \times (\mathit{median}(x) + \alpha), \quad (1)$$

where x represents the input CFP image, *M(x)* represents the corresponding fundus region mask. On the basis of the median, increasing the offset α can avoid the median filling being too smooth and enhance the contrast of high-frequency features. In this study, α is set to 0.2. Subsequently, a Gaussian convolution kernel is employed for the convolution operation to extract low-frequency components. The input is then subtracted from the low-frequency components to obtain high-frequency components, and finally adjusted according to the mask to obtain the final high-frequency components of the input image:

$$\mathcal{F}(x_p) = (clamp(x_p - x_p * g(r, \sigma), -1, 1) + 1) \cdot M(x) - 1, \quad (2)$$

where $\mathcal{F}$ is a high-frequency mapping function, *clamp* is a value clipping function, and *g* is a Gaussian filter with a kernel of (r, σ). This process isolates structural elements such as vessel boundaries, ridge formations, and tissue texture changes while suppressing low-frequency style variations, directing the model's attention toward clinically relevant anatomical features rather than imaging artifacts.

C. Proxy interaction attention module

The trunk feature encoding (TFE) branch of CFP images and the HFM branch interact effectively through the proxy interaction attention module (PIAM), thereby enhancing the model's perception of the detailed structures of the retina. PIAM utilises a small group of high-frequency proxy tokens to aggregate the high-frequency detail information and broadcasts it to the TFE branch. For an input with *N* tokens that can be represented as $x \in \mathbb{R}^{N \times C}$, in each head, the self-attention query, key, and value matrices can be formulated as follows:

$$Q = xW_Q, \quad K = xW_K, \quad \text{and} \quad V = xW_V. \quad (3)$$

where $W_{Q/K/V} \in \mathbb{R}^{C \times d}$ are projection matrices, $Q, K, V \in \mathbb{R}^{N \times C}$. C and d are the embedding dimension of modules

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

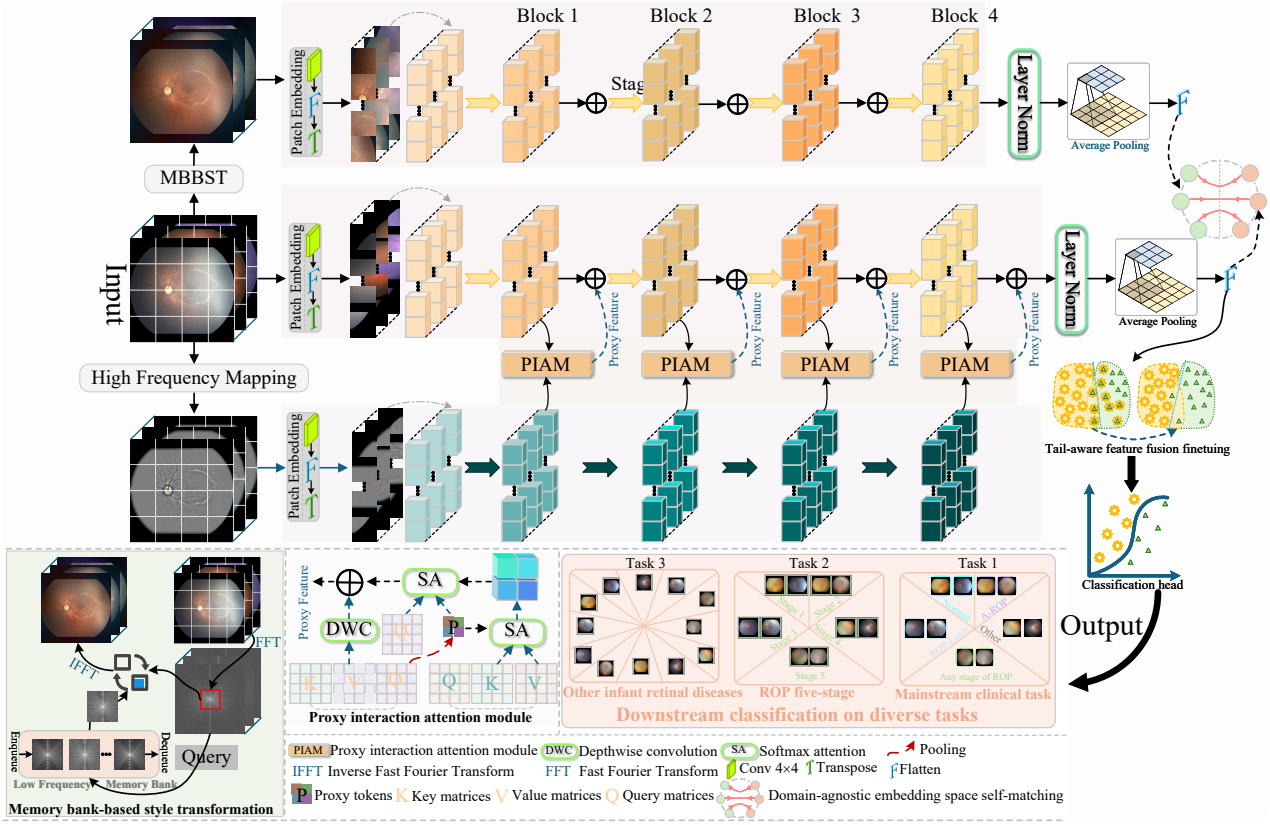


Fig. 2. Overall framework of the proposed SABPI-Net. The architecture incorporates standard hierarchical transformer blocks in Blocks 1 to 4, which produce feature maps with dimensions of $\mathbf{56 \times 56 \times 96}$, $\mathbf{28 \times 28 \times 192}$, $\mathbf{14 \times 14 \times 384}$, and $\mathbf{7 \times 7 \times 768}$. The trunk feature encoding branch interacts with the high-frequency mapping branch via the proxy interaction attention module, enhancing the perception of retinal detail structures. A low-frequency component replacement strategy based on a memory bank is used to simulate appropriate style textures for training images, and domain-agnostic embedding space self-matching learning is used to learn the style invariance of disease. The tail-aware feature fusion fine-tuning strategy enhances the diagnostic sensitivity of the tail classes.

and each head, respectively. In order to facilitate subsequent analysis, the superscripts m and h are used to denote the variables of the TFE branch and the HFM branch, respectively. For a concrete example with input $b \times N \times C$ and H heads, Q^m, K^m, V^m, Q^h, K^h, $V^h \in \mathbb{R}^{b \times H \times N \times d}$ where b is batch size, $d = C/H$. In practice, a straightforward pooling strategy is applied to the query matrix of the TFE branch, thereby yielding the high-frequency proxy tokens that interact with the HFM branch:

$$P = \text{pooling}(Q^m), \tag{4}$$

where $P \in \mathbb{R}^{b \times H \times C_P \times d}$, C_P denotes the number of high-frequency proxy tokens. The PIAM uses high-frequency proxy tokens to aggregate high-frequency detail information from the HFM branch and distribute it to individual image tokens in the TFE branch. Specifically, P is taken as the query and the Softmax attention computation is performed between P, the key K^h of the HFM branch, and the value V^h of the HFM branch to aggregate the high frequency feature F_P, which can be formalized as:

$$F_P = \text{Softmax}(\frac{P(K^h)^T}{\sqrt{d}} + B_1)V^h, \tag{5}$$

where $B_1 \in \mathbb{R}^{b \times H \times C_P \times N}$ is the bias term of high frequency proxy, $F_P \in \mathbb{R}^{b \times H \times C_P \times d}$. Subsequently, utilising P as the key and F_P as the value, in conjunction with the query matrix Q^m of the TFE branch, a secondary Softmax attention computation is performed. This process disseminates the high frequency detail information from the proxy high frequency feature to each query token in the TFE branch. The newly defined P essentially serves as a proxy for Q^m, aggregating global high frequency detail information from the HFM branches K^h and V^h, and then broadcasting it back to Q^m, which can be formalized as:

$$O_{temp} = \text{Softmax}(\frac{Q^m P^T}{\sqrt{d}} + B_2)F_P, \tag{6}$$

where $B_2 \in \mathbb{R}^{b \times H \times N \times C_P}$ is the bias term of high frequency proxy, $O_{temp} \in \mathbb{R}^{b \times H \times N \times d}$. In fact, we set the number of high frequency proxy tokens C_P to a small hyperparameter, and the number of high frequency proxy tokens can be designed to be much smaller than the number of query tokens. In this way, our PIAM achieves a linear complexity of $\mathcal{O}(NC_Pd)$ while maintaining the ability to model global context.

Based on these designs, the proposed PIAM is illustrated in Fig. 2, which can be formalized as:

$$O_P = O_{temp} + DWC(V^m), \tag{7}$$

where *DWC* is the depthwise convolutional module. After

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

reshaping, $O_P \in \mathbb{R}^{b\times N\times C}$. The integration of high-frequency proxy bias terms can make better use of the location information, thereby enabling different high-frequency proxy tokens to concentrate on distinct high-frequency regions through the incorporation of spatial information into Softmax attention. In order to alleviate the effect of insufficient feature diversity [54], the *DWC* module is utilised to introduce the V^m of the TFE branch, thereby enhancing the feature diversity. Ultimately, the features at each block of the TFE branch are fused with the high-frequency proxy features O_P from the output of PIAM before proceeding to the next block of feature extraction.

D. Domain-agnostic embedding space self-matching

In order to enhance the stylistic invariance of retinal disease diagnosis, it is crucial to simulate appropriate stylistic textures for the training images. To this end, a memory bank-based low-frequency component replacement strategy is proposed to simulate the low-frequency components of the transformed training images. The memory bank is constructed on the low-frequency components of preceding training samples and updated using a first-in-first-out strategy. The memory bank, designated as *M*, is responsible for storing the low-frequency components $\{f_{low}^s\}_{s=1}^S$ of *S* training images. The Fast Fourier Transform and Inverse Fast Fourier Transform operations are denoted by $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ respectively, while the amplitude and phase components are denoted by $\mathcal{F}^A(\cdot)$ and $\mathcal{F}^P(\cdot)$. As illustrated in Fig. 2, to attain a transformed image that aligns with the current training image, it is necessary to initially extract the low-frequency component $\mathcal{F}_{low}^A(X_i)$ of the current training image. It is noteworthy that the amplitude component in $\mathcal{F}^A(\cdot)$ has been shifted towards the centre. We use the Euclidean paradigm to compute the similarity between $\mathcal{F}_{low}^A(X_i)$ and each low-frequency component f_{low}^s in the memory bank. The similarity scores are then ranked, and the $\mathcal{F}_{low}^A(X_i)$ of the original image is replaced with the lowest similarity low-frequency component. This process forms a matching alternative stylistic texture image, X_i^*. The process can be formalised as follows:

$$X_i^* = \mathcal{F}^{-1}([\text{swap}(\mathcal{F}^A(X_i), \min(\text{sim}(\mathcal{F}_{low}^A(X_i), \{f_{low}^s\}_{s=1}^S))), \mathcal{F}^P(X_i)]) \quad (8)$$

In order to achieve stable performance of retinal disease diagnosis across imaging styles, we used cosine similarity to calculate cross-stylistic texture correlations:

$$\mathbb{R}_{X_i,X_i^*} = \frac{Z_i \cdot (Z_i^*)^T}{\|Z_i\|_2 \|Z_i^*\|_2}, \quad (9)$$

where Z_i and $Z_i^* \in \mathbb{R}^{b\times 8C}$ denote the matched X_i and X_i^* output features after image encoder respectively, *b* denotes the batch size and *C* is the feature embedding dimension. The objective is to ensure the consistency of the semantic relationship between the image features of X_i and X_i^*. The optimisation objective of self-matching in the domain-agnostic embedding space can be formulated as follows:

$$\mathcal{L}_{match} = 1 - mean(\mathbb{R}_{X_i,X_i^*}). \quad (10)$$

We adopt the cross-entropy loss function as the training objective for the classification task. In the domain-agnostic embedding space self-matching learning, we use cosine similarity to compute the cross-style texture relevance and optimise it using Equation 10. Therefore, the final training objective function is the weighted sum of all loss functions.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}) + \lambda\mathcal{L}_{match}, \quad (11)$$

where λ is the weight of the loss $\mathcal{L}_{match}$, which is set to 0.01 in this paper.

E. Tail-aware feature fusion via fine-tuning

The imbalanced allocation of long-tailed data can result in the model's preferential and accurate classification of the head class, thereby reducing the diagnostic sensitivity of the tail class. A significant contributing factor to this reduced sensitivity is the decision boundary bias, which arises due to an inadequacy of semantic information in the tail class. To address this issue, a tail-aware feature fusion fine-tuning strategy is proposed, which involves the grafting of diversified semantic information from the head class into the tail class. In the classification head adjustment stage, a proportion of the tail class feature map is substituted for the feature map of the head class at a specific ratio. This process enhances the diversity of the tail class features and consequently improves the diagnostic sensitivity of the tail class. Utilising semantically meaningful features to delineate category boundaries can enhance the diversity within the tail class. This refinement can assist in achieving more precise decision boundaries.

During the training process, the incomplete class coverage of mini-batches makes the execution of the fine-tuning strategy cumbersome, so we use two sampling strategies in parallel when loading data. First, we normally sample from the training dataset to obtain sample-wise data Φ^S, which has a high probability of obtaining head class samples, and the features of these data will be used to replace the features of the tail class. Second, we obtain class-balanced data Φ^B through a class-uniform sampling strategy. This sampling strategy first randomly and uniformly selects a target class, and then randomly and uniformly selects a sample from all samples corresponding to the target class to form Φ^B. In the second sampling strategy, the fewer the sample size, the higher the probability of being re-sampled multiple times, because the probability of each category being selected is equal under this strategy. Through this approach, the tail classes' features that are re-sampled will utilize the head classes' features to expand their embedding space distribution with a higher probability. Next, we will utilize the head classes' semantics to enrich the tail classes' features, which is formalized as:

$$\tilde{\mathcal{F}} = \mathcal{F}_S \otimes \mathcal{M}_r + \mathcal{F}_B \otimes \overline{\mathcal{M}_r}, \quad (12)$$

where $\mathcal{F}_S$, $\mathcal{F}_B \in \mathbb{R}^{b\times 8C}$ respectively represent the feature maps from Φ^S and Φ^B, which are the features input to the classification head before fine-tuning. $\mathcal{M}_r \in \mathbb{R}^{1\times 8C}$ is a feature mask matrix with values of 0/1 (The quantity of 1 is $r \times 8C$), which selects features from $\mathcal{F}_S$ at a certain ratio

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

r to replace the corresponding features in $\mathcal{F}_B$. $\overline{\mathcal{M}_r}$ is the complement of $\mathcal{M}_r$. This strategy is fine-tuned with only a small number of epochs.

IV. EXPERIMENTS AND RESULTS

A. Dataset

This study retrospectively collects data from 13,314 patients who underwent ROP screening at the Shenzhen Eye Hospital (from June 29, 2005 to October 31, 2023), amounting to 174,540 CFPs. This study adheres to the principles of the Declaration of Helsinki and was approved by the Medical Ethics Committee of Shenzhen Eye Hospital (ID: 2023KYPJ091). The requirement for informed consent was waived due to the retrospective nature of the study. To ensure confidentiality, all personally identifiable information (such as names, medical record numbers, and contact details) was removed/de-identified prior to data analysis. The CFPs in this study are acquired using three wide-field contact fundus imaging systems: the RetCam (RetCam, Clarity Medical Systems, Inc. USA), ROP Screening System (SW-8000, China), and Nautilus Fundus Imaging System (RS-B002, China). To the best of our knowledge, this dataset represents the largest ROP cohort reported to date[1]

The dataset is divided into training, validation, and test sets at a ratio of approximately 7:2:1 at patient level, ensuring that data from the same patient would not appear in different subsets simultaneously. The overall data distribution is shown in Table II. In order to address real clinical needs, this dataset is primarily used to achieve three classification tasks: Task 1: A mainstream clinical task, including Normal, Laser-treated, Any stage of ROP, A-ROP, and Other (data distribution see Table .II); Task 2: ROP five-stage classification task, which refines 'Any stage of ROP' in Task 1 into five stages (stage 1 to stage 5, data distribution see Table III); Task 3: Other multiple-disease classification, which refines 'Other' in Task 1 into 12 disease categories (see Table IV for details of the data distribution).

B. Implementation details

All experiments are based on Python 3.7.16, PyTorch 1.13.1 and Ubuntu 22.04.5. All of the training procedures are carried out on NVIDIA 4090 GPUs with 22 GB of memory. The initial learning rate has been set to 0.001, and a cosine annealing schedule is employed. The default optimiser employed is SGD, with a momentum setting of 0.9. The weight decay parameter is set to 1e-4. The number of training epochs is 100.

1) *Pre-processing and augmentation strategies:* The image data is uniformly resized to the same pixel size before being fed into the model for training. During the training process, data enhancement techniques such as random cropping, scaling, horizontal flipping and vertical flipping are sequentially applied to the image data.

[1]A representative subset of our dataset has been made publicly available at: https://figshare.com/articles/figure/_b_A_Fundus_Image_Dataset_for_Intelligent_b_b_Retinopathy_of_Prematurity_b_b_System_b_/25514449.

2) *Other configurations:* In domain-agnostic embedding space self-matching learning, the central 3×3 region of the low-frequency component of the training image is replaced with the low-frequency component of another image to generate a matched transformed image. In V, we delve deeper into the impact of substituting different sizes on the performance of the model. For the tail-aware feature fusion fine-tuning stage, we train for 10 epochs with the cross-entropy loss function as the training objective, using only the labels of the class-balanced data Φ^B. Then, we randomly replace 10% of the tail class feature maps with head class feature maps.

C. Task 1: Mainstream clinical task

Table V reports the classification performance of our SABPI-Net with thirteen other SOTA models in the ROP initial screening five classification task. The SABPI-Net outperforms the other methods on all metrics. Among the thirteen SOTA models, which include recently reported methods to ensure comprehensive comparison with SOTA approaches, SABPI-Net achieves improvements over the highest values of the corresponding metrics, with gains of 1.57% in Acc, 0.65% in AUC, 0.61% in Pre, 4.34% in Sen, 3.23% in F1, 0.74% in Spec, and 2.96% in Kappa. In order to provide a comprehensive assessment of the performance of the various methods in each category, an analysis is conducted of the classification results of each model on the test set. These results are then presented visually in the form of confusion matrices, as shown in Fig. 3. The diagonal of the confusion matrix demonstrates the diagnostic sensitivity of the model for each category. The majority of models demonstrate high sensitivities within the Normal, Laser-treated ROP and Any Stage of ROP. However, sensitivities within the A-ROP and Other are significantly lower. For instance, ConvNeXt-S demonstrates a sensitivity of 10.26% on A-ROP, while the highest sensitivities observed in the A-ROP and Other within the comparative methods are 58.33% and 68.65%. It is noteworthy that the SABPI-Net demonstrates significantly higher sensitivity in the A-ROP and Other, with 61.54% and 74.52%, respectively. In order to further enhance the diagnostic sensitivity of the model in the tail categories (A-ROP and Other), we employ the tail-aware feature fusion fine-tuning strategy to refine our SABPI-Net. Consequently, the overall diagnostic sensitivity of the model is enhanced from 85.03% to 89.27%.

To further understand the feature representation quality and interpretability of different methods, we conduct t-SNE visualization analysis for all fourteen methods. As shown in Fig. 4, most methods demonstrate high inter-class separability and intra-class cohesion for Normal, Laser-treated, and Any stage of ROP, while the proposed SABPI-Net achieves superior inter-class discrimination. However, most models struggle with A-ROP and Other, showing poor intra-class cohesion and inter-class separability, often confusing these with other categories. In contrast, our SABPI-Net demonstrates significantly better intra-class cohesion and inter-class discrimination for these challenging tail classes.

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

TABLE II
OVERALL DISTRIBUTION OF THE ROP COHORT DATASET IN THIS STUDY

	Train set		Validation set		Test set		Total	
Category	Image, No.	Patient, No.	Image, No.	Patient, No.	Image, No.	Patient, No.	Image, No.	Patient, No.
Normal	82,505	7,555	19,370	1,889	10,895	1,050	112,770	10,494
Laser-treated	21,268	521	4,899	131	3,275	73	29,442	725
Any stage of ROP	13,425	804	3,859	202	2,229	112	19,513	1,118
A-ROP	1,438	108	613	27	312	15	2,363	150
Other	7,522	596	1,799	149	1,131	82	10,452	827
Total	126,158	9,584	30,540	2,398	17,842	1,332	174,540	13,314

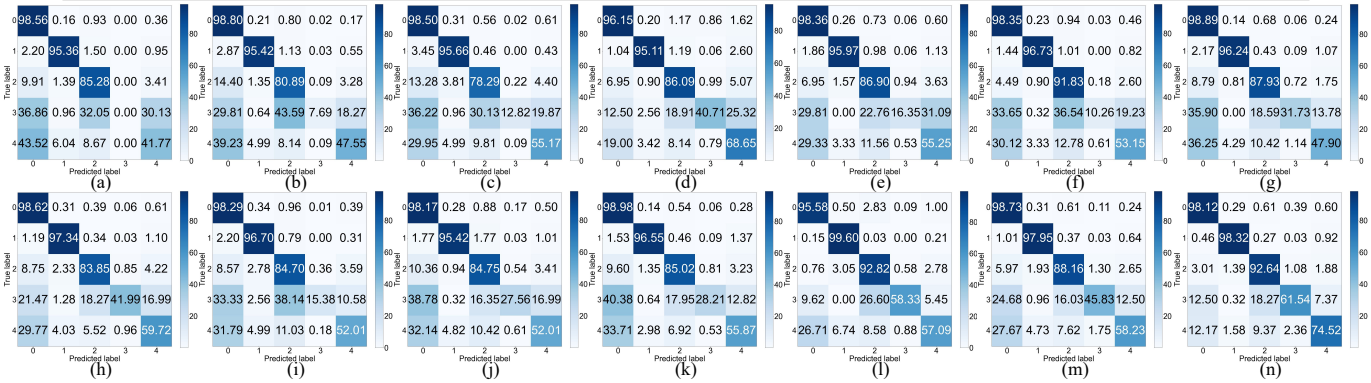


Fig. 3. Comparison of confusion matrix diagram of different methods on Task 1. (a) to (n) respectively represent the results of ResNet50, Tong et al., ViT-B, Conformer-S, RETFound, ConvNeXt-S, FIT-Net, Swin-S, MIL-VT, UIOS, EfficientViM-M3, nnMobileNet, SLAB and the proposed SABPI-Net.

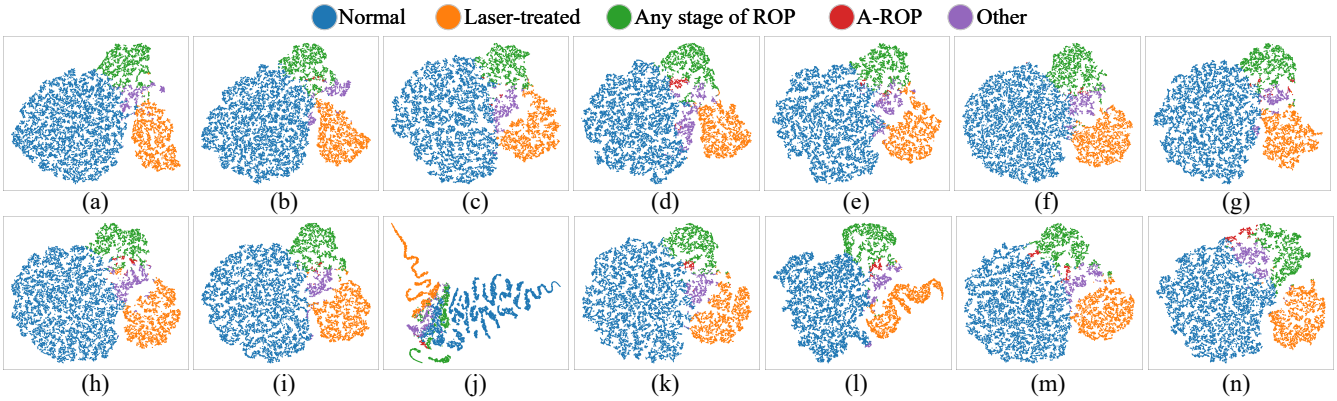


Fig. 4. t-SNE visualization comparison of feature embeddings on Task 1.(a) to (n) respectively represent the results of ResNet50, Tong et al., ViT-B, Conformer-S, RETFound, ConvNeXt-S, FIT-Net, Swin-S, MIL-VT, UIOS, EfficientViM-M3, nnMobileNet, SLAB and the proposed SABPI-Net.

TABLE III
DISTRIBUTION OF ROP 5 STAGES DATASET

Category	Train set	Validation set	Test set	Total
Stage 1	1,187	268	193	1,648
Stage 2	4,988	1,325	811	7,124
Stage 3	6,286	1,592	762	8,640
Stage 4	961	303	68	1,332
Stage 5	583	101	85	769
Total	14,005	3,589	1,919	19,513

D. Task 2: ROP five-stage classification task

In light of the clinical necessity to stage any stage of ROP as delineated in Task 1, a refinement of the staging system is undertaken, resulting in the delineation of 5 stages, from stage 1 to stage 5. As illustrated in Table VI, this study compares the classification performance of our SABPI-Net with that of thirteen other SOTA models in Task 2. The proposed SABPI-Net demonstrates superior comprehensive performance, achieving a favorable balance across all evaluation metrics when compared to the competing methods. Among the thirteen SOTA models, SABPI-Net achieves improvements over the highest values of the corresponding metrics, with gains of 1.88% in Acc, 0.36% in AUC, 3.50% in Sen, 4.54% in F1, 0.85% in Spec, and 4.59% in Kappa. The classification results of each model on the test set are analysed and presented visually with the confusion matrix shown in Fig. 5. Overall, the diagnostic sensitivity of the models is higher on Stage 5, which, despite having the least amount of data in this category, is characterised by a distinct pathology with a clinical presentation of complete retinal detachment. Conversely, in the initial stages of ROP, the distinction between Stage 1 and Stage 2 symptoms remains ambiguous. As Stage 1 is a tail

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

TABLE IV
DISTRIBUTION OF OTHER 12 CATEGORIES OF LESIONS

Category	Train set	Validation set	Test set	Total
Retinoblastoma	1,626	844	147	2,617
Retinitis pigmentosa	97	9	16	122
Morning glory syndrome	52	47	13	112
Cataract	137	49	16	202
Coloboma of choroid	198	29	36	263
Oculocutaneous albinism	124	16	41	181
Retinal detachment	139	29	33	201
Macular disease	191	15	21	227
Tessellated retina	177	57	15	249
Retinal traction or fold	873	210	273	1,356
Hemorrhage	2,829	817	503	4,149
Peripheral retinal exudation	593	126	54	773
Total	7,036	2,248	1,168	10,452

TABLE V
COMPARISON RESULTS ON TASK 1

Method	Acc	AUC	Pre	Sen	F1	Spec	Kappa
ResNet50 [55]	90.96	96.85	67.92	64.21	65.48	96.50	83.58
Tong et al. [38]	91.08	97.17	85.73	66.09	69.24	96.44	83.74
ViT-B [56]	91.19	97.34	85.80	68.09	71.48	96.61	84.05
Conformer-S [57]	91.98	97.42	78.50	77.41	77.84	97.50	86.01
RETFound [52]	92.30	97.74	81.03	70.57	73.25	97.23	86.23
ConvNeXt-S [58]	92.81	98.05	84.24	70.06	72.31	97.42	87.16
FIT-Net [30]	92.59	97.57	86.03	72.50	76.89	97.05	86.60
Swin-S [59]	93.07	98.15	86.51	76.30	80.17	97.41	87.59
MIL-VT [26]	91.89	97.43	86.32	69.42	72.83	96.98	85.43
UIOS [27]	91.81	96.42	83.57	71.58	75.40	96.90	85.26
EfficientViM-M3 [60]	92.80	98.08	86.07	72.92	77.19	97.13	86.99
nnMobileNet [29]	92.86	98.42	86.68	80.69	82.80	97.80	87.58
SLAB [61]	93.75	98.58	86.61	77.78	81.28	97.69	88.86
$Our^{\dagger}$	94.40	99.20	82.22	**89.27**	85.08	**98.57**	90.43
Our	**95.32**	**99.23**	**87.22**	85.03	**86.03**	98.54	**91.82**

$Our^{\dagger}$ denotes the result obtained by our SABPI-Net following the fine-tuning of the tailed-class features.

category of data in this study, most models have a lower diagnostic sensitivity for it, with a maximum of 33.16% in the comparison methods. However, the tail-aware feature fusion fine-tuning strategy employed by our SABPI-Net enhanced the diagnostic sensitivity of Stage 1 to 68.39%, thereby attaining the maximum average diagnostic sensitivity in Stage 1 and Stage 2. This enhancement fulfils the clinical requirement for the recognition of early-stage ROP.

Fig. 6 reveals that Stage 1 and Stage 2 exhibit high feature overlap in most models due to their similar clinical manifestations. The proposed SABPI-Net demonstrates better inter-class discrimination for these early stages. Stage 4 and Stage 5, representing later ROP stages with more pronounced pathological features, show better separability from other stages in most models. However, some models still exhibit feature overlap between Stage 4 and Stage 5 due to their shared retinal detachment characteristics, while the proposed SABPI-Net achieves superior intra-class cohesion and inter-class discrimination for these categories.

E. Task 3: Other infantile retinal diseases classification

In order to analyse the Other category in Task 1 more finely, we further classify it into 12 common lesion types. Subsequently, we evaluate the classification performance of SABPI-Net against thirteen other SOTA classification models in Task 3, and the results are shown in Table VII. Among the thirteen SOTA models, SABPI-Net achieves improvements over the highest values of the corresponding metrics, with

TABLE VI
COMPARISON RESULTS ON TASK 2

Method	Acc	AUC	Pre	Sen	F1	Spec	Kappa
ResNet50 [55]	69.88	90.51	67.03	61.66	63.46	90.28	52.17
Tong et al. [38]	69.83	90.68	67.51	64.40	65.36	90.36	52.57
ViT-B [56]	69.52	90.04	69.24	58.93	61.78	89.97	51.04
Conformer-S [57]	68.84	90.71	64.26	60.46	60.85	90.09	50.81
RETFound [52]	68.79	90.58	66.95	61.73	63.07	89.95	50.67
ConvNeXt-S [58]	69.78	90.30	66.70	63.93	65.10	90.62	53.16
FIT-Net [30]	70.51	90.63	67.55	57.05	59.23	90.21	52.15
Swin-S [59]	70.04	92.21	71.11	63.69	65.67	90.25	52.42
MIL-VT [26]	70.14	90.80	71.95	55.32	57.99	89.92	51.04
UIOS [27]	69.52	87.30	66.56	60.66	62.21	90.11	51.48
EfficientViM-M3 [60]	70.40	90.91	70.65	56.81	59.34	90.07	51.70
nnMobileNet [29]	71.70	91.25	**74.67**	57.66	57.42	90.46	53.75
SLAB [61]	70.19	91.94	70.85	63.59	65.85	90.27	52.60
$Our^{\dagger}$	71.76	**92.68**	70.87	**74.60**	**72.00**	**91.72**	57.92
Our	**73.58**	92.57	74.57	67.90	70.39	91.47	**58.34**

$Our^{\dagger}$ denotes the result obtained by our SABPI-Net following the fine-tuning of the tailed-class features.

gains of 4.71% in Acc, 1.49% in AUC, 7.68% in Pre, 5.27% in Sen, 7.17% in F1, 0.46% in Spec, and 6.56% in Kappa. The classification results of each model on the test set are analysed and visualised through the confusion matrix shown in Fig. 7. The majority of models manifest elevated diagnostic sensitivity within the head and middle categories (namely, retinoblastoma, retinal traction or fold, haemorrhage, and peripheral retinal exudation). However, the diagnostic sensitivity is conspicuously inadequate in the majority of tail categories. Our SABPI-Net, through the tail-aware feature fusion fine-tuning strategy, has increased the average diagnostic sensitivity of the remaining 8 tail classes to 59.97%, which is 14.39% higher than the best performance among the thirteen SOTA classification models (Fig. 13). This significant improvement demonstrates the effectiveness of the proposed SABPI-Net in handling rare infantile retinal diseases, outperforming recent methods like SLAB (79.54% Acc) and nnMobileNet (79.28% Acc) with our achieved 84.25% Acc.

Fig. 8 demonstrates that head categories like Retinoblastoma and Hemorrhage exhibit good inter-class discrimination across most models, with the proposed SABPI-Net achieving superior intra-class cohesion. For middle categories such as Retinal traction or fold and Peripheral retinal exudation, most models show reduced intra-class cohesion, while for the remaining tail categories, inter-class discrimination becomes significantly worse. Our SABPI-Net model substantially alleviates these issues, providing better feature representation for rare diseases.

TABLE VII
COMPARISON RESULTS ON TASK 3

Method	Acc	AUC	Pre	Sen	F1	Spec	Kappa
ResNet50 [55]	72.35	90.64	31.52	32.94	31.36	96.80	60.73
Tong et al. [38]	72.43	90.44	33.65	33.61	32.67	96.70	60.49
ViT-B [56]	70.63	93.11	39.33	43.87	39.83	97.16	60.92
Conformer-S [57]	73.89	91.13	45.84	48.03	45.57	97.40	65.03
RETFound [52]	70.98	90.71	38.00	42.37	39.14	97.21	61.34
ConvNeXt-S [58]	76.71	95.65	52.85	49.75	47.69	97.69	68.71
FIT-Net [30]	78.34	95.22	49.18	57.54	50.60	98.01	71.43
Swin-S [59]	79.37	95.81	48.98	52.17	48.50	98.00	72.20
MIL-VT [26]	77.65	95.18	50.52	47.78	46.65	97.71	69.60
UIOS [27]	74.40	90.10	48.45	48.23	45.87	97.41	65.35
EfficientViM-M3 [60]	71.83	90.82	45.08	44.67	42.21	97.11	61.97
nnMobileNet [29]	79.28	95.48	52.84	55.77	52.21	97.75	71.68
SLAB [61]	79.54	94.26	54.96	50.20	48.18	97.90	72.16
$Our^{\dagger}$	83.73	97.11	**63.01**	**68.89**	**62.21**	**98.50**	78.35
Our	**84.25**	**97.30**	62.64	62.81	59.38	98.47	**78.76**

$Our^{\dagger}$ denotes the result obtained by our SABPI-Net following the fine-tuning of the tailed-class features.

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

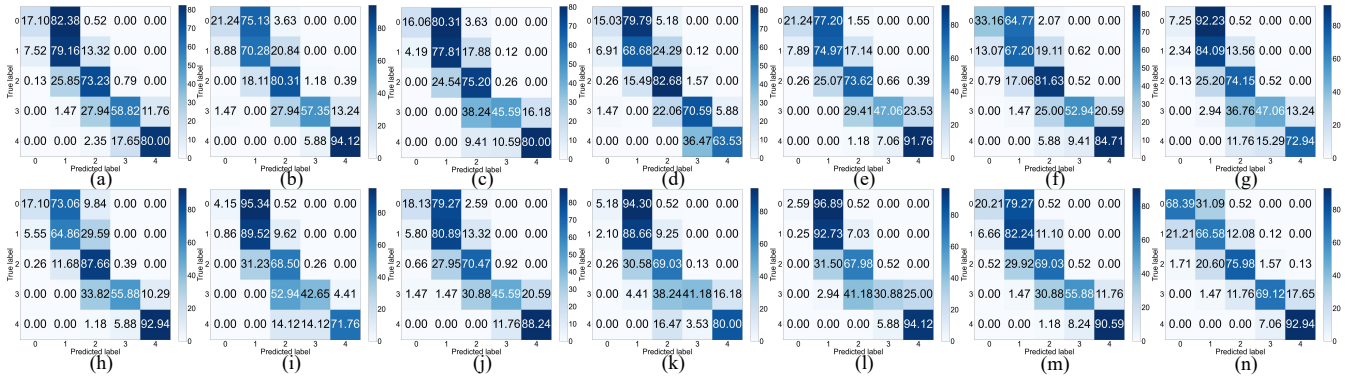


Fig. 5. Comparison of confusion matrix diagram of different methods on Task 2. (a) to (n) respectively represent the results of ResNet50, Tong et al., ViT-B, Conformer-S, RETFound, ConvNeXt-S, FIT-Net, Swin-S, MIL-VT, UIOS, EfficientViM-M3, nnMobileNet, SLAB and the proposed SABPI-Net.

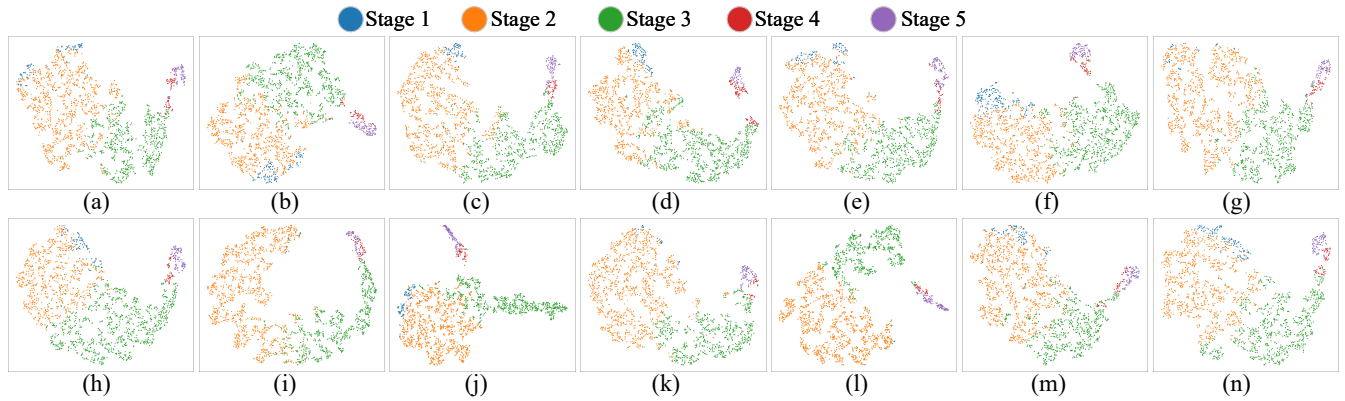


Fig. 6. t-SNE visualization comparison of feature embeddings on Task 2. (a) to (n) respectively represent the results of ResNet50, Tong et al., ViT-B, Conformer-S, RETFound, ConvNeXt-S, FIT-Net, Swin-S, MIL-VT, UIOS, EfficientViM-M3, nnMobileNet, SLAB and the proposed SABPI-Net.

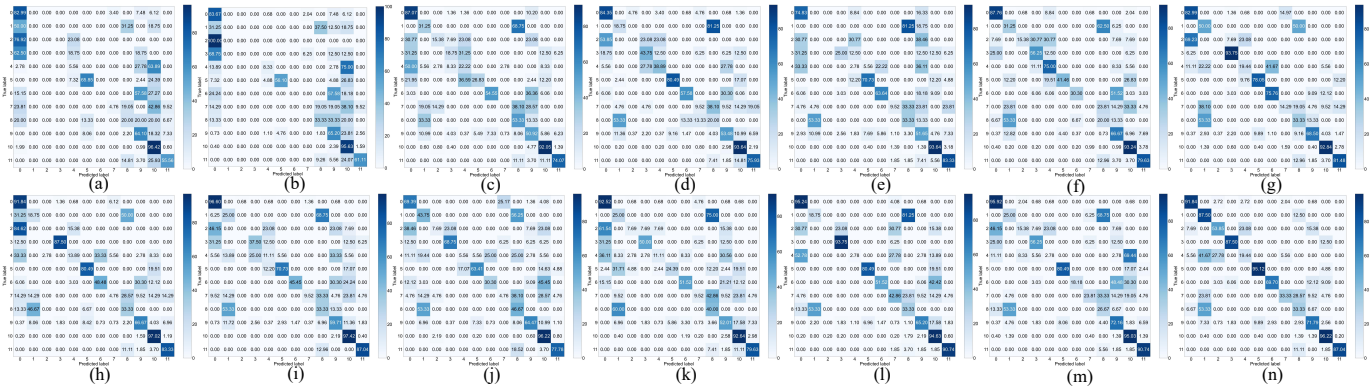


Fig. 7. Comparison of confusion matrix diagram of different methods on Task 3. (a) to (n) respectively represent the results of ResNet50, Tong et al., ViT-B, Conformer-S, RETFound, ConvNeXt-S, FIT-Net, Swin-S, MIL-VT, UIOS, EfficientViM-M3, nnMobileNet, SLAB and the proposed SABPI-Net.

F. Module ablation study

Table VIII shows the results of the ablation studies of different modules in SABPI-Net. For simplicity, we use Acc as the default evaluation metric. We first examine the impact of domain-agnostic embedding space self-matching (DAESSM) learning on the model. It is found that the introduction of DAESSM improves the average Acc of the three classification tasks by 1.1%, 2.82%, and 2.65%. Next, we introduce the HFM branch and implement the interaction between the HFM branch and the TFE branch through the PIAM. We find that the introduction of HFM and PIAM can improve the overall performance of the three tasks by 1.23%, 2.92%, and 2.74%. Finally, when DAESSM, HFM and PIAM are combined, we find that the overall performance of the three tasks can be further improved, with Acc improvements of 2.18%, 3.39% and 3.51%.

To directly evaluate the effectiveness of the tail-aware feature fusion strategy (TFFS), we conduct specific ablation experiments as shown in Table IX. The results demonstrate significant improvements in diagnostic sensitivity across all three tasks when TFFS is applied. Specifically, TFFS enhances the overall sensitivity by 4.24% for Task 1, 6.70% for Task

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

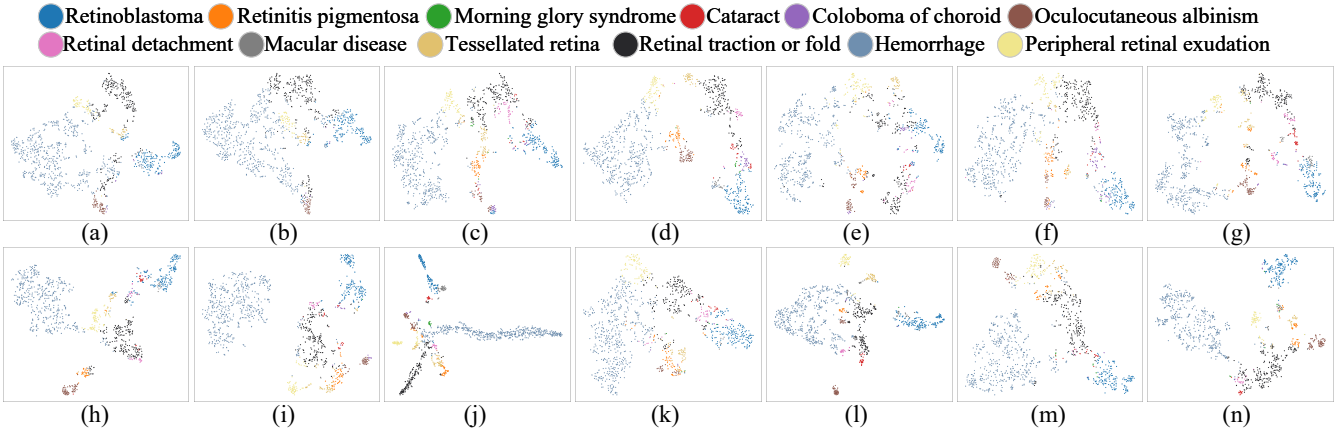


Fig. 8. t-SNE visualization comparison of feature embeddings on Task 3.(a) to (n) respectively represent the results of ResNet50, Tong et al., ViT-B, Conformer-S, RETFound, ConvNeXt-S, FIT-Net, Swin-S, MIL-VT, UIOS, EfficientViM-M3, nnMobileNet, SLAB and the proposed SABPI-Net.

TABLE VIII
MODULE ABLATION EXPERIMENTS

Module	HFM	PIAM	DAESSM	Acc	Diff.
Task 1				93.14	-2.18
			✓	94.24	-1.08
	✓	✓		94.37	-0.95
	✓	✓	✓	**95.32**	**0**
Task 2				70.19	-3.39
			✓	73.01	-0.57
	✓	✓		73.11	-0.47
	✓	✓	✓	**73.58**	**0**
Task 3				80.74	-3.51
			✓	83.39	-0.86
	✓	✓		83.48	-0.77
	✓	✓	✓	**84.25**	**0**

2, and 6.08% for Task 3. These substantial improvements validate the critical role of TFFS in addressing the long-tail distribution challenge inherent in rare infantile retinal disease classification, thereby enhancing the model's capability to diagnose tail classes effectively.

TABLE IX
ABLATION STUDY OF TAIL-AWARE FEATURE FUSION STRATEGY (TFFS) ACROSS THREE CLASSIFICATION TASKS

Method	Task 1	Task 2	Task 3
SABPI-Net w/o TFFS	85.03	67.90	62.81
SABPI-Net w TFFS	89.27	74.60	68.89

V. DISCUSSION

Retinal image foundation models, such as RETFound, have significantly advanced clinical AI in ophthalmology, particularly for adult retinal disease recognition. However, adult and infant color fundus images differ considerably in anatomical structure, physiological traits, and pathological manifestations, stemming from variations in developmental stages, retinal maturity, and ocular physiology. Infantile retinal images differ significantly from adult images across multiple dimensions. Structurally, infant retinas exhibit incomplete vascularization, smaller optic discs, and immature macular architecture, while adult retinas display fully developed vascular networks [62]. In imaging characteristics, infant fundus photography often

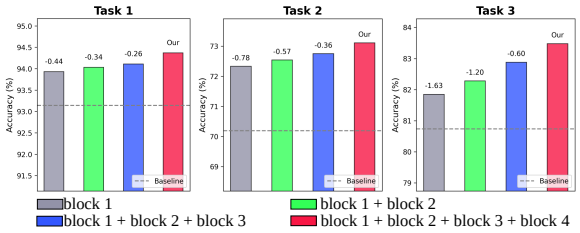


Fig. 9. Applying PIAM to different blocks of the Baseline.

yields lower quality due to smaller pupils, eye movement, and limited cooperation, resulting in reduced contrast compared to adult imaging [63]. Pathologically, infantile diseases manifest unique features: ROP presents as demarcation lines and ridge formations, A-ROP shows rapid posterior changes [44], [45]. These differences necessitate specialized approaches for infant retinal analysis. We propose a PIAM to enable effective interaction between the TFE and HFM branches. The objective of this is to enhance the model's perception of the detailed structure of the retina. We conduct an ablation study to examine the impact of PIAM across different blocks. As demonstrated in Fig. 9, the model's Acc undergoes a progressive enhancement as the number of blocks employing PIAM increases. The application of PIAM to facilitate interaction between the two branches at all four blocks results in the model achieving optimal performance on all three tasks. Specifically, the Acc is improved by 1.23% for Task 1, 2.92% for Task 2, and 2.74% for Task 3 relative to the baseline. These results further verify that the TFE branch of retinal images can enhance the model's representation ability for the detailed structure of the retinal through effective information interaction with the HFM branch via PIAM, thereby improving the model's diagnostic performance.

In order to validate the effectiveness of the HFM branch, a comparison is conducted for testing the performance of four common foundation models with and without its integration. As shown in Table X, the inclusion of the HFM branch improves the Acc of all four common foundation models. However, the proposed method still achieves the best performance. This is because simply adding the HFM branch is insufficient to enhance the model's perception of retinal

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

detail structures, whereas the PIAM can effectively address this limitation. To further elucidate the model's decision-making process, we employ Grad-CAM to generate heatmaps highlighting the regions in the input images that contribute most significantly to a specific class prediction. As shown in Fig. 10, the addition of the HFM branch enables the model to better perceive detailed structures in the retinal region and focus more precisely on disease lesion locations. Conversely, without the HFM branch, the model's sensitivity to subtle details diminishes, leading to an inability to focus on the actual lesion location for certain diseases.

To provide deeper insights into model interpretability, we present Grad-CAM visualizations for challenging misclassified cases in Fig. 11. The second column shows a typical case where regular fundus hemorrhage is misclassified as A-ROP, which occurs because A-ROP clinically manifests as capillary abnormalities accompanied by vitreous hemorrhage, leading to occasional misclassification of regular hemorrhage cases. The fourth column demonstrates a case where ROP Stage 4 is misclassified as Stage 5. This confusion arises because both stages share retinal detachment features, with Stage 5 presenting complete retinal detachment while Stage 4 shows partial retinal detachment. These visualizations reveal that even with advanced models, distinguishing subtle pathological differences remains challenging, highlighting the clinical complexity of infantile retinal disease diagnosis.

TABLE X
COMPARISON OF COMMON FOUNDATION MODELS WITH HFM BRANCH

Method	w/o HFM branch	w HFM branch
ViT-B	91.19	92.62
Conformer-S	91.98	93.06
ConvNeXt-S	92.81	93.80
Swin-S	93.07	93.90
Our	94.24	**95.32**

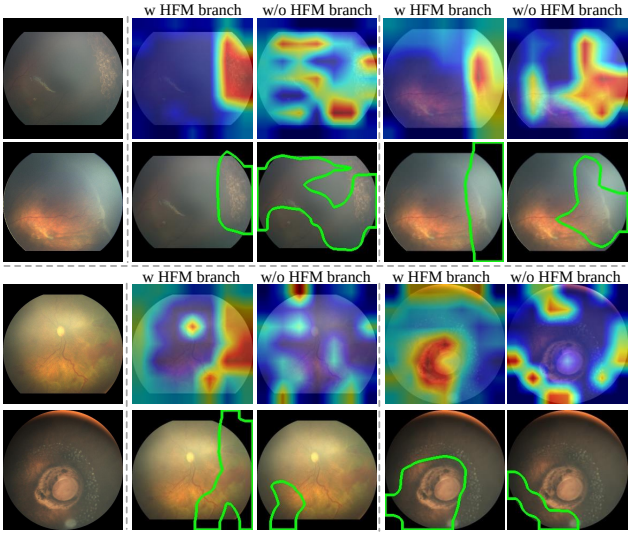


Fig. 10. Comparison of Grad-CAM visualisation results with and without HFM branch.

A low-frequency component replacement strategy is designed on the basis of a memory bank, with the aim of

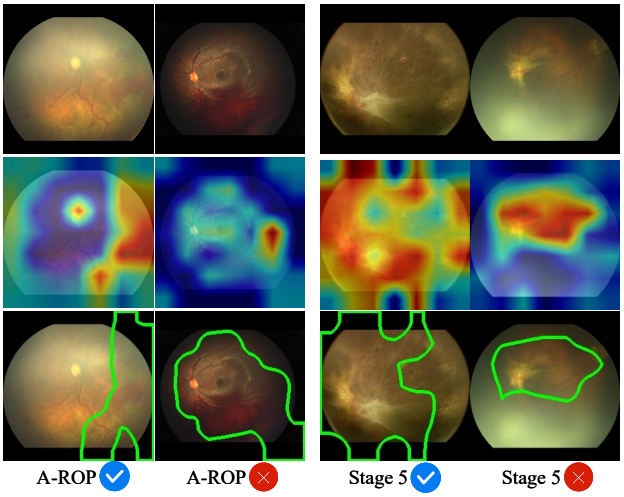


Fig. 11. Grad-CAM visualization analysis for misclassified cases.

enhancing the style invariance of retinal disease diagnosis by simulating suitable style textures for training images. To evaluate the impact of different sizes of low-frequency component replacement on the model, we further conduct ablation experiments. As demonstrated in Table XI, the model's overall accuracy exhibits a slight decline as the size of the low-frequency component replacement increases. The model achieves the highest accuracy in all three tasks when the size of the low-frequency component replacement is 3×3, with an increase of 1.10% in Task 1, 2.82% in Task 2, and 2.65% in Task 3. In order to further explore the impact of replacing low-frequency components of different sizes on style textures, the results are plotted as shown in Fig. 12. The analysis demonstrates that when the replacement size is designated as S3, the style texture transformation effect is optimal. However, as the size increases, the image detail information becomes gradually blurred and may result in alterations to the original image structure (for example, the optic disc area may shift), which is not the expected outcome. Consequently, S3 is selected as the optimal replacement size. These results further verify that our domain-agnostic embedding space self-matching learning strategy can achieve stable performance in retinal disease diagnosis under various imaging styles to a certain extent.

TABLE XI
ABLATION EXPERIMENTS WITH DIFFERENT SIZE OF LOW-FREQUENCY SUBSTITUTION COMPONENT IN DAESSM LEARNING

Method	Task 1		Task 2		Task 3	
	Acc	Diff.	Acc	Diff.	Acc	Diff.
Baseline	93.14	-1.10	70.19	-2.82	80.74	-2.65
S12	93.45	-0.79	72.38	-0.63	82.19	-1.20
S6	93.60	-0.64	72.64	-0.37	82.79	-0.60
S3	**94.24**	**0**	**73.01**	**0**	**83.39**	**0**

To further evaluate the impact of feature fusion with respect to fine-tuning strategies on the tail class diagnostic sensitivity, we compare the tail class diagnostic sensitivity of different models in three tasks. As shown in Fig. 13, even without

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

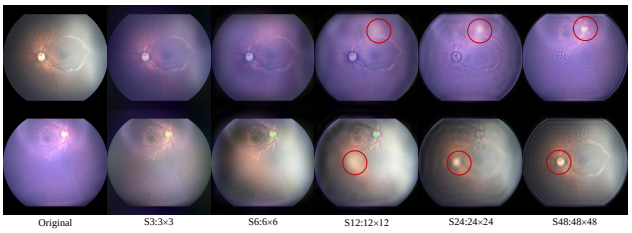


Fig. 12. Transformation of the image style texture after replacing the low-frequency components of different sizes.

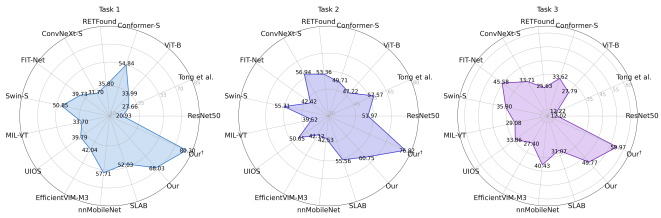


Fig. 13. Comparison of diagnostic sensitivity performance of tail classes.

the use of fine-tuning strategies, the proposed method still performed best in terms of tail class diagnosis sensitivity in the three tasks. After adopting the tail-aware feature fusion fine-tuning strategy, the tail class diagnosis sensitivity are further improved in the three tasks, with the increases of 12.27%, 16.07% and 10.20%.

TABLE XII
COMPUTATIONAL COMPLEXITY AND PERFORMANCE COMPARISON

Method	Params↓ (M)	FLOPs↓ (GMac)	Inference Time↓ (ms)	Task 1 Acc↑ (%)	Task 2 Acc↑ (%)	Task 3 Acc↑ (%)	Avg Acc↑(%)
ResNet50 [55]	23.52	4.13	0.060	90.96	69.88	72.35	77.73
Tong et al. [38]	42.51	7.86	0.120	91.08	69.83	72.43	77.78
ViT-B [56]	86.00	16.86	0.045	91.19	69.52	70.63	77.11
Conformer-S [57]	37.70	10.60	0.175	91.98	68.84	73.89	78.24
RETFound [52]	303.10	59.68	0.087	92.30	68.79	70.98	77.36
ConvNeXt-S [58]	50.00	8.70	0.074	92.81	69.78	76.71	79.77
FIT-Net [30]	95.06	6.70	0.445	92.59	70.51	78.34	80.48
Swin-S [59]	50.00	8.70	0.151	93.07	70.04	79.37	80.83
MIL-VT [26]	21.80	4.28	0.054	91.89	70.14	77.65	79.89
UIOS [27]	23.52	4.13	0.069	91.81	69.52	74.40	78.58
EfficientViM-M3 [60]	16.60	0.66	0.115	92.80	70.40	71.83	78.34
nnMobileNet [29]	30.89	3.45	0.126	92.86	71.70	79.28	81.28
SLAB [61]	49.79	8.56	0.172	93.75	70.19	79.54	81.16
Our	**28.36**	**4.37**	**0.089**	**95.32**	**73.58**	**84.25**	**84.38**

To evaluate the practical deployment feasibility of the proposed SABPI-Net, we conduct a comprehensive computational complexity analysis comparing our method with thirteen SOTA approaches. Table XII presents the detailed comparison of computational requirements, including the number of parameters (Params), floating-point operations (FLOPs), average inference time, and Acc performance across the three classification tasks. The computational complexity analysis reveals several key insights regarding the practical deployment of SABPI-Net.

First, our method achieves the highest average Acc (84.38%) across all three tasks while maintaining competitive computational efficiency. With 28.36M parameters and 4.37 GMac FLOPs, SABPI-Net demonstrates a favorable balance between model complexity and performance. Notably, our method requires significantly fewer parameters than heavy-weight models like RETFound (303.10M), ViT-B (86.00M), and FIT-Net (95.06M), while achieving superior diagnostic accuracy. The inference time analysis shows that SABPI-Net processes images in 0.089ms on average, which is faster than many competing methods including Conformer-S (0.175ms), SLAB (0.172ms), and Swin-S (0.151ms).

While some lighter models like ResNet50 (0.060ms) and ViT-B (0.045ms) show faster inference times, they significantly lag behind our method in diagnostic accuracy (77.73% and 77.11% vs. our 84.38%). Particularly noteworthy is the comparison with methods of similar computational complexity. For instance, MIL-VT and UIOS have comparable parameter counts and FLOPs to our method, but achieve substantially lower average accuracies (79.89% and 78.58% respectively). This demonstrates that the proposed SABPI-Net design effectively utilizes computational resources to achieve superior diagnostic performance.

VI. CONCLUSION

This paper proposes a diagnostic model for infant retinal diseases called SABPI-Net, which is trained on a large dataset of over 170,000 color fundus photographs, the largest in scale and disease coverage to date. The SABPI-Net can identify 19 types of infant retinal diseases. These extensive experiments, showcasing SABPI-Net's strong performance in a mainstream clinical task, ROP five-stage classification, and other multi-disease classification, highlight the broad applicability of CFP-based AI models to early screening and diagnosis of infant retinal diseases. Despite the effectiveness of SABPI-Net, there is often a compromise in the diagnostic performance of the head diseases when improvements in the sensitivity of the tail diseases are pursued. Future work will address this shortcoming and apply it to a broader range of retinal diseases.

Despite the promising results, several limitations exist. Our approach currently relies exclusively on color fundus photographs, which may constrain comprehensive pathological assessment compared to multimodal approaches. Additionally, the inherent trade-offs between head and tail class performance in long-tail distribution learning remain challenging. Future work will focus on incorporating complementary imaging modalities including OCT, OCTA, and clinical text reports to enhance diagnostic sensitivity for rare diseases. Longitudinal disease progression modeling, and real-time clinical integration represent additional promising directions for advancing AI-driven infant retinal disease diagnosis.

REFERENCES

[1] WHO, *World report on vision*. World Health Organization, 2019.
[2] World Health Organization, *Global patient safety action plan 2021-2030: towards eliminating avoidable harm in health care*. World Health Organization, 2021.
[3] J. Wang, J. Ji *et al.*, "Automated Explainable Multidimensional Deep Learning Platform of Retinal Images for Retinopathy of Prematurity Screening," *JAMA Network Open*, vol. 4, no. 5, pp. e218 758–e218 758, 05 2021.
[4] Q. Wu, Y. Hu *et al.*, "Development and Validation of a Deep Learning Model to Predict the Occurrence and Severity of Retinopathy of Prematurity," *JAMA Network Open*, vol. 5, no. 6, pp. e2 217 447–e2 217 447, 06 2022.
[5] A. Ramanathan, S. E. Athikarisamy, and G. C. Lam, "Artificial intelligence for the diagnosis of retinopathy of prematurity: A systematic review of current algorithms," *Eye*, vol. 37, no. 12, pp. 2518–2526, Aug 2023.

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

[6] W. M. Fierson, M. F. Chiang *et al.*, "Screening examination of premature infants for retinopathy of prematurity," *Pediatrics*, vol. 142, no. 6, p. e20183061, 12 2018.

[7] A. Gschließer, E. Stifter *et al.*, "Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity," *American Journal of Ophthalmology*, vol. 160, no. 3, pp. 553–560.e3, 2015.

[8] T.-Y. Yu, T. Donovan *et al.*, "Retinopathy of prematurity: the high cost of screening regional and remote infants," *Clinical & Experimental Ophthalmology*, vol. 46, no. 6, pp. 645–651, 2018.

[9] I. Bakkouri and K. Afdel, "Mlca2f: Multi-level context attentional feature fusion for covid-19 lesion segmentation from ct scans," *Signal, Image and Video Processing*, vol. 17, no. 4, pp. 1181–1188, Jun 2023.

[10] I. Bakkouri, K. Afdel *et al.*, "Bg-3dm2f: Bidirectional gated 3d multi-scale feature fusion for alzheimer's disease diagnosis," *Multimedia Tools and Applications*, vol. 81, no. 8, pp. 10 743–10 776, Mar 2022.

[11] I. Bakkouri and K. Afdel, "Computer-aided diagnosis (cad) system based on multi-layer feature fusion network for skin lesion recognition in dermoscopy images," *Multimedia Tools and Applications*, vol. 79, no. 29, pp. 20 483–20 518, Aug 2020.

[12] K. Afdel and I. Bakkouri, "Multi-scale cnn based on region proposals for efficient breast abnormality recognition," *Multimedia Tools and Applications*, vol. 78, no. 10, pp. 12 939–12 960, May 2019.

[13] J. Huang, T. Tan *et al.*, "C2maot: Cross-modal complementary masked autoencoder wit optimal transport for cancer segmentation in pet-ct images," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. Cham: Springer Nature Switzerland, 2026, pp. 87–97.

[14] T. Tan, J. Huang *et al.*, "Synergistic perception: Fusing expert knowledge and foundation models for semi-supervised mammogram segmentation," *Biomedical Signal Processing and Control*, vol. 112, p. 108633, 2026.

[15] A. Pivodic, A.-L. Hård *et al.*, "Individual Risk Prediction for Sight-Threatening Retinopathy of Prematurity Using Birth Characteristics," *JAMA Ophthalmology*, vol. 138, no. 1, pp. 21–29, 01 2020.

[16] S. Chen, X. Zhao *et al.*, "Multi-risk factors joint prediction model for risk prediction of retinopathy of prematurity," *EPMA Journal*, vol. 15, no. 2, pp. 261–274, Jun 2024.

[17] Q. Wu, Y. Hu *et al.*, "Development and Validation of a Deep Learning Model to Predict the Occurrence and Severity of Retinopathy of Prematurity," *JAMA Network Open*, vol. 5, no. 6, pp. e2 217 447–e2 217 447, 06 2022.

[18] J. Sun, G. Zhang *et al.*, "An ultra-wide-field fundus image dataset for intelligent diagnosis of intraocular tumors," *Scientific Data*, vol. 12, no. 1, p. 1521, Aug 2025.

[19] H. Xie, B. Lei *et al.*, "But-net: Boundary-aware u-net structure with two-path transformers for lesion segmentation in mcnv using oct images," *Expert Systems with Applications*, vol. 297, p. 129405, 2026.

[20] Y.-P. Huang, S. Vadloori *et al.*, "Deep learning models for automated diagnosis of retinopathy of prematurity in preterm infants," *Electronics*, vol. 9, no. 9, 2020.

[21] R. Zhang, J. Zhao *et al.*, "Automatic diagnosis for aggressive posterior retinopathy of prematurity via deep attentive convolutional neural network," *Expert Systems with Applications*, vol. 187, p. 115843, 2022.

[22] J. Hu, Y. Chen *et al.*, "Automated analysis for retinopathy of prematurity by deep neural networks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 269–279, 2019.

[23] J. Wang, R. Ju *et al.*, "Automated retinopathy of prematurity screening using deep neural networks," *EBioMedicine*, vol. 35, pp. 361–368, 2018.

[24] S. Chen, X. Zhao, B. Lei *et al.*, "Dual-scale Swin Transformer via Feature Alignment and Adversarial Discrimination for Retinopathy of Prematurity Diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 12, pp. 9013–9026, 12 2025.

[25] X. Zhao, S. Chen, G. Zhang *et al.*, "A fundus image dataset for intelligent retinopathy of prematurity system," *Scientific Data*, vol. 11, no. 1, p. 543, May 2024.

[26] Q. Bi, X. Sun *et al.*, "Mil-vit: A multiple instance vision transformer for fundus image classification," *Journal of Visual Communication and Image Representation*, vol. 97, p. 103956, 2023.

[27] M. Wang, T. Lin *et al.*, "Uncertainty-inspired open set learning for retinal anomaly identification," *Nature Communications*, vol. 14, no. 1, p. 6757, 2023.

[28] Y. Zhou, M. A. Chia, S. K. Wagner *et al.*, "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, no. 7981, pp. 156–163, Oct 2023.

[29] W. Zhu, P. Qiu *et al.*, "nnmobilenet: Rethinking cnn for retinopathy research," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, pp. 2285–2294.

[30] S. Chen, Z. Wu *et al.*, "Fit-net: Feature interaction transformer network for pathologic myopia diagnosis," *IEEE Transactions on Medical Imaging*, vol. 42, no. 9, pp. 2524–2538, 2023.

[31] C. Wen, M. Ye *et al.*, "Concept-based lesion aware transformer for interpretable retinal disease diagnosis," *IEEE Transactions on Medical Imaging*, vol. 44, no. 1, pp. 57–68, 2025.

[32] M. Singh, V. B. S, J. Gubbi, and R. V. Babu, "Protopatchnet: An interpretable patch-based prototypical network," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, June 2025, pp. 721–728.

[33] X. Wei, Y. Liu *et al.*, "Mstnet: Multi-scale spatial-aware transformer with multi-instance learning for diabetic retinopathy classification," *Medical Image Analysis*, vol. 102, p. 103511, 2025.

[34] P. Zhang, Y. Li *et al.*, "Sigraph: Saliency image-graph network for retinal disease classification in fundus image," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, pp. 10 049–10 057, Apr. 2025.

[35] J. Hou, J. Xu, R. Feng, and H. Chen, "Qmix: Quality-aware learning with mixed noise for robust retinal disease diagnosis," *IEEE Transactions on Medical Imaging*, vol. 44, no. 8, pp. 3345–3355, 2025.

[36] J. Wang, J. Ji *et al.*, "Automated Explainable Multidimensional Deep Learning Platform of Retinal Images for Retinopathy of Prematurity Screening," *JAMA Network Open*, vol. 4, no. 5, pp. e218 758–e218 758, 05 2021.

[37] J. M. Brown, J. P. Campbell *et al.*, "Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks," *JAMA Ophthalmology*, vol. 136, no. 7, pp. 803–810, 07 2018.

[38] Y. Tong, W. Lu *et al.*, "Automated identification of retinopathy of prematurity by image-based deep learning," *Eye and Vision*, vol. 7, no. 1, p. 40, Aug 2020.

[39] S. Ramachandran, P. Niyas *et al.*, "A deep learning framework for the detection of plus disease in retinal fundus images of preterm infants," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 362–375, 2021.

[40] S. Ramachandran, N. Strisciuglio *et al.*, "U-cosfire filters for vessel tortuosity quantification with application to automated diagnosis of retinopathy of prematurity," *Neural Computing and Applications*, vol. 32, no. 16, pp. 12 453–12 468, Aug 2020.

[41] Y.-P. Huang, W.-C. Wu *et al.*, "Automated detection of early-stage rop using a deep convolutional neural network," *British Journal of Ophthalmology*, vol. 105, no. 8, pp. 1099–1103, 2021.

[42] H. Xie, B. Lei *et al.*, "Adversarial learning-based multi-level dense-transmission knowledge distillation for ap-rop detection," *Medical Image Analysis*, vol. 84, p. 102725, 2023.

[43] Y. Peng, W. Zhu *et al.*, "Automatic staging for retinopathy of prematurity with deep feature fusion and ordinal classification strategy," *IEEE Transactions on Medical Imaging*, vol. 40, no. 7, pp. 1750–1762, 2021.

[44] M. F. Chiang, G. E. Quinn *et al.*, "International classification of retinopathy of prematurity, third edition," *Ophthalmology*, vol. 128, no. 10, pp. e51–e68, 2021.

[45] G. A. Gole, G. Holmstrom *et al.*, "The International Classification of Retinopathy of Prematurity Revisited," *Archives of Ophthalmology*, vol. 123, no. 7, pp. 991–999, 07 2005.

[46] E. H. Land, "The retinex theory of color vision," *Scientific American*, vol. 237, no. 6, pp. 108–129, 1977.

[47] H. Li, J. Liu *et al.*, "Frequency-mixed single-source domain generalization for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Cham: Springer Nature Switzerland, 2023, pp. 127–136.

[48] H. Li, H. Liu *et al.*, "A generic fundus image enhancement network boosted by frequency self-supervised representation learning," *Medical Image Analysis*, vol. 90, p. 102945, 2023.

[49] Q. Xu, R. Zhang *et al.*, "A fourier-based framework for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 14 383–14 392.

[50] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[51] Z. Chen, Y. Pan, Y. Ye, M. Lu, and Y. Xia, "Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 11 184–11 193.

[52] Y. Zhou, M. A. Chia *et al.*, "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, no. 7981, pp. 156–163, Oct 2023.

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. 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/TMI.2026.3650968

[53] S. Chen, Y. Sun *et al.*, "Bi-branch bidirectional coupled interaction fusion network for multi-retinal diseases diagnosis," *Knowledge-Based Systems*, vol. 326, p. 114069, 2025.

[54] D. Han, X. Pan, Y. Han *et al.*, "Flatten transformer: Vision transformer using focused linear attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 5961–5971.

[55] K. He, X. Zhang *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[56] A. Dosovitskiy, L. Beyer *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[57] Z. Peng, W. Huang, S. Gu *et al.*, "Conformer: Local features coupling global representations for visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 367–376.

[58] Z. Liu, H. Mao *et al.*, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11 976–11 986.

[59] Z. Liu, Y. Lin *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 012–10 022.

[60] S. Lee, J. Choi, and H. J. Kim, "Efficientvim: Efficient vision mamba with hidden state mixer based state space duality," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, June 2025, pp. 14 923–14 933.

[61] J. Guo, X. Chen, Y. Tang, and Y. Wang, "SLAB: Efficient transformers with simplified linear attention and progressive re-parameterized batch normalization," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235. PMLR, 21–27 Jul 2024, pp. 16 802–16 812.

[62] A. M. Dubis, D. M. Costakos *et al.*, "Subclinical macular findings in infants screened for retinopathy of prematurity with spectral-domain optical coherence tomography," *Ophthalmology*, vol. 120, no. 8, pp. 1665–1671, 2013.

[63] M. P. Byrne, K. R. McMillan, and B. Coats, "Morphological analysis of retinal microvasculature to improve understanding of retinal hemorrhage mechanics in infants," *Investigative Ophthalmology & Visual Science*, vol. 61, no. 3, pp. 16–16, 03 2020.

Authorized licensed use limited to: FUDAN UNIVERSITY. Downloaded on January 07,2026 at 06:24:18 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.