Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01849-y

Diagnosing pathologic myopia by identifying morphologic patterns using ultra widefield images with deep learning

Check for updates

Yang Liu[1,11], Keming Zhao[1,2,11], Lihui Luo[1], Ziheng Zhang[1], Zhenghang Qian[1,3], Cenk Jiang[1], Zhicheng Du[1], Simin Deng[2], Chengming Yang[4], Duanpo Wu[5], Shuai Wang[5], Xingru Huang[5], Chenggang Yan[5], Yingting Zhu[6], Yehong Zhuo[6], Chunsheng Qu[7], Jiaqi Chen[7], Zhenqiang Huang[7], Chenying Lu[8], Minjiang Chen[8], Dongmei Yu[9], Jiantao Wang[2] ✉, Peiwu Qin[1,10] ✉ & Jiansong Ji[8] ✉

Pathologic myopia is a leading cause of visual impairment and blindness. While deep learning-based approaches aid in recognizing pathologic myopia using color fundus photography, they often rely on implicit patterns that lack clinical interpretability. This study aims to diagnose pathologic myopia by identifying clinically significant morphologic patterns, specifically posterior staphyloma and myopic maculopathy, by leveraging ultra-widefield (UWF) images that provide a broad retinal field of view. We curate a large-scale, multi-source UWF myopia dataset called PSMM and introduce RealMNet, an end-to-end lightweight framework designed to identify these challenging patterns. Benefiting from the fast pretraining distillation backbone, RealMNet comprises only 21 million parameters, which facilitates deployment for medical devices. Extensive experiments conducted across three different protocols demonstrate the robustness and generalizability of RealMNet. RealMNet achieves an F1 Score of 0.7970 (95% CI 0.7612–0.8328), mAP of 0.8497 (95% CI 0.8058–0.8937), and AUROC of 0.9745 (95% CI 0.9690–0.9801), showcasing promise in clinical applications.

The increasing prevalence of myopia worldwide is a significant public health concern[1]. It is projected that by 2050, nearly 50% of the global population will be affected. Myopia, defined by a spherical equivalent ≤ -0.5 diopters, can lead to visual impairments that greatly reduce patients' quality of life and impose substantial economic burdens[2]. All degrees of myopia pose potential risks for adverse changes in ocular tissues, especially at high levels of myopia (defined as spherical equivalent worse than -5.0 or -6.0 diopters) and pathologic myopia (resulting in irreversible visual impairment or blindness due to pathological retinal changes secondary to high myopia)[3]. Ophthalmic examinations, typically involving fundus imaging, are necessary for detecting and diagnosing relevant fundus lesions. While traditional color fundus photography (CFP) captures the retina within 30–60 degrees, novel imaging modalities such as ultra-widefield (UWF) imaging with a field of view ranging from 100 to 200 degrees[4], can capture retinal lesions missed by CFP, leading to improved screening accuracy and early detection. Despite the increasing use of advanced retinal imaging in ophthalmic practices, publicly available UWF datasets remain scarce, hindering the development of diagnostic and support systems necessary to help clinicians interpret these advanced imaging modalities.

Advancements in deep learning have made it possible to automatically process medical images for various tasks, achieving performance comparable to clinical experts[5–9]. In the case of retinal diseases, deep learning models not only accurately diagnose and monitor conditions such as diabetic retinopathy and age-related macular degeneration from retinal images[10–12],

[1]Institute of Biopharmaceutics and Health Engineering, Tsinghua Shenzhen International Graduate School, Shenzhen, China. [2]Shenzhen Eye Hospital, Jinan University, Shenzhen Eye Institute, Shenzhen, China. [3]Department of Automation, Tsinghua University, Beijing, China. [4]Southern University of Science and Technology Hospital, Shenzhen, China. [5]Hangzhou Dianzi University, Hangzhou, Zhejiang, China. [6]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, China. [7]Clinical Laboratory of Lishui People's Hospital, First Affiliated Hospital of Lishui College, Wenzhou Medical College Lishui Hospital, Lishui, Zhejiang, China. [8]Zhejiang Key Laboratory of Imaging and Interventional Medicine, Department of Radiology, Lishui Central Hospital, The Fifth Affiliated Hospital of Wenzhou Medical University, Lishui, Zhejiang, China. [9]School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai, China. [10]Center of Precision Medicine and Healthcare, Tsinghua-Berkeley Shenzhen Institute, Shenzhen, China. [11]These authors contributed equally: Yang Liu, Keming Zhao. ✉e-mail: wangjiantao65@126.com; pwqin@sz.tsinghua.edu.cn; jjstcty@wmu.edu.cn

but also assist in developing personalized treatment plans[13]. In addition, deep learning has been applied to myopia-related screening[14,15], assessing the risk of myopia progression by analyzing retinal images and enabling early intervention. Although these methods are robust, there is a need for further investigation into sophisticated morphologic patterns. A system called the Meta-Analysis of Pathologic Myopia (META-PM)[16] categorizes myopic atrophic components into five classes: no myopic retinal lesions (Grade 0), tessellated fundus only (Grade 1), diffuse chorioretinal atrophy (Grade 2), patchy chorioretinal atrophy (Grade 3), and macular atrophy (Grade 4). Pathologic myopia is now defined as myopic maculopathy (according to META-PM criteria: grade 2 or above) or posterior staphyloma[17]. Posterior staphyloma appears as an outpouching of the ocular wall, with a curvature radius smaller than that of the surrounding sclera. Posterior staphyloma often leads to changes in the retina, choroid, and nerve fiber layer, subsequently affecting the patient's vision. Early identification of posterior staphyloma is crucial because it can lead to severe complications, such as retinal detachment, macular hemorrhage, and choroidal neovascularization, all of which may cause irreversible vision loss. Myopic maculopathy is one of the primary causes of vision deterioration in individuals with high myopia, as it directly affects the macula. Since the macula is the area of the retina that provides the highest visual acuity, any damage to this area can significantly affect vision quality. Early diagnosis and management of these conditions can help slow or prevent disease progression and reduce the risk of vision loss. While the association between high myopia and peripheral retinal lesions is well-established, automated identification of clinically relevant morphologic patterns using ultra-widefield imaging remains a technical and practical challenge. This study bridges that gap by developing and validating a dedicated deep learning model on a large-scale and expert-annotated dataset. This advancement allows for scalable screening and risk assessment in real-world settings where access to retinal specialists and multimodal imaging may be limited.

Existing research has some limitations despite advancements. Firstly, less attention is given to myopic maculopathy and posterior staphyloma. This is possibly due to the difficulty in identifying their complete contour on CFP accurately. In contrast, UWF imaging allows for precise diagnosis of peripheral lesions and the edges of staphyloma, appearing as a dark gray band-shaped ring with twisted retinal and choroidal vessels. However, the high equipment cost, intricate operational procedures, and data acquisition expenses make large-scale UWF data collection challenging for many studies. Additionally, existing research lacks UWF datasets that feature lesion-wise labeling. This is largely due to the necessity for accurate localization and identification of each lesion, which requires extensive clinical case support and expert review. In imaging for pathologic myopia, the lesions are complicated and widely distributed, particularly in the peripheral regions of the retina, making it significantly challenging to label them accurately. The scarcity of datasets with lesion-wise labels restricts the depth of research in pathologic myopia and hinders the improvement of clinical diagnosis and treatment. Secondly, previous studies often use balanced data, ignoring the significant data imbalance in real-world scenarios[18]. Retinal lesions in pathologic myopia are highly heterogeneous and often coexist with other types of retinal lesions. This leads to imbalanced data, making it more challenging to accurately distinguish posterior staphyloma from myopic maculopathy. Thirdly, detecting posterior staphyloma and myopic maculopathy involves complex multi-label learning tasks, which pose higher demands on algorithm models. Many existing studies focus on identifying a single lesion or simpler pathologies and cannot handle multiple complex coexisting lesions. Thus, traditional imaging data and diagnostic tools may not provide precise classifications, limiting the exploration of these specific lesions. Lastly, there has been a strong focus on building large and complex models[19]. Although these models are powerful, they lack flexibility and incur high costs when applied to medical devices, particularly in resource-constrained clinical environments. Reducing the number of trainable parameters results in faster training times and lower computational costs, which are crucial for rapidly adapting models to specific medical applications. Lightweight models that maintain high accuracy with fewer parameters are particularly advantageous in medical settings, where computational resources may be limited or where devices require constant iteration.

In this work, we present a detailed and efficient workflow (Fig. 1) tailored to meet the clinical need for early and large-scale identification of pathologic myopia. We developed a specialized deep learning model aimed at detecting its key morphologic features: posterior staphyloma and myopic maculopathy. Our approach emphasizes the distinct structural abnormalities observed in UWF images of patients with high myopia, offering a focused diagnostic tool in this context. We compile a dataset containing UWF images of pathologic myopia with clinically significant lesions from multiple medical sources. Experienced ophthalmologists label images related to posterior staphyloma, myopic maculopathy, and peripheral lesions under the guidance of META-PM and double-check annotations to ensure accuracy. With the support of this curated dataset, we are able to identify clinically significant morphologic patterns by developing an end-to-end framework called RealMNet that embraces Real-world Myopia diagnosis. The name RealMNet is particularly chosen to reflect its focus on real-world applications and its ability to handle complex, real-world data more effectively than its predecessors. With the adoption of a compact and efficient vision transformer[20] as our backbone, the framework is light enough to be applied to medical devices. We approach this challenge as a multi-label learning task for two reasons: first, posterior staphyloma may be present with myopic maculopathy, jointly indicating pathologic myopia, and second, peripheral lesions could coexist. We comprehensively evaluate RealMNet's performance using three distinct experimental protocols: centralized inference, main-source robustness, and cyclic-source generalizability. Under the centralized inference protocol, we compare the inference performance of RealMNet on the PSMM dataset against four pretrained benchmark approaches (DeiT[21], ConvNeXt[22], EfficientNet[23], and Swin Transformer[24]) and two recent foundation models (DINOv2[25] and VisionFM[26]). The other two protocols are used to assess the robustness and generalizability of the model for lesion identification, which is crucial for clinical use. We evaluate labeling efficiency using RealMNet with increasing resolutions and interpret parameters at different stages of the backbone. We demonstrate the effectiveness of regularization techniques used in the proposed method with extensive evaluation experiments. We examine the potential negative impact of physical device boundaries in images captured by ultra-widefield imaging, as these boundaries may obstruct essential information. We demonstrate the advantages of the UWF modality by comparing them with fake CFP images. Finally, we investigate transfer learning on identifying peripheral lesions (Supplementary Fig. 6): no peripheral lesion (NoPL), lattice degeneration or cystic retinal tuft (LDoCRT), holes or tears (HoT), rhegmatogenous retinal detachment (RRD), and postoperative cases (PC), which can co-occur in high myopic eyes and lead to significant visual impairment.

Results

Multi-source curated UWF myopia dataset provides a solid foundation for multi-lesion identification

We gathered a specialized dataset called PSMM derived from five distinct hospital sources for identifying posterior staphyloma and myopic maculopathy that could assist clinicians in diagnosing pathologic myopia. The PSMM dataset comprised 43,371 ultra-widefield images of 4560 patients who sustained high myopia or pathologic myopia after data filtering for quality assurance. We also separately managed the five sub-sources that integrated the PSMM dataset to facilitate characteristic research. Generally, the PSMM dataset provided a competitive scale considering the expense of ultra-widefield imaging that captured a broader retinal field of view compared to color fundus photography (Supplementary Fig. 1a). Experienced ophthalmologists labeled posterior staphyloma with binary annotations to indicate its presence (NoPS or PS) and myopic maculopathy with five categories: no myopic retinal lesions (NoMRL), tessellated fundus only (TFO), diffuse chorioretinal atrophy (DCA), patchy chorioretinal atrophy (PCA), and

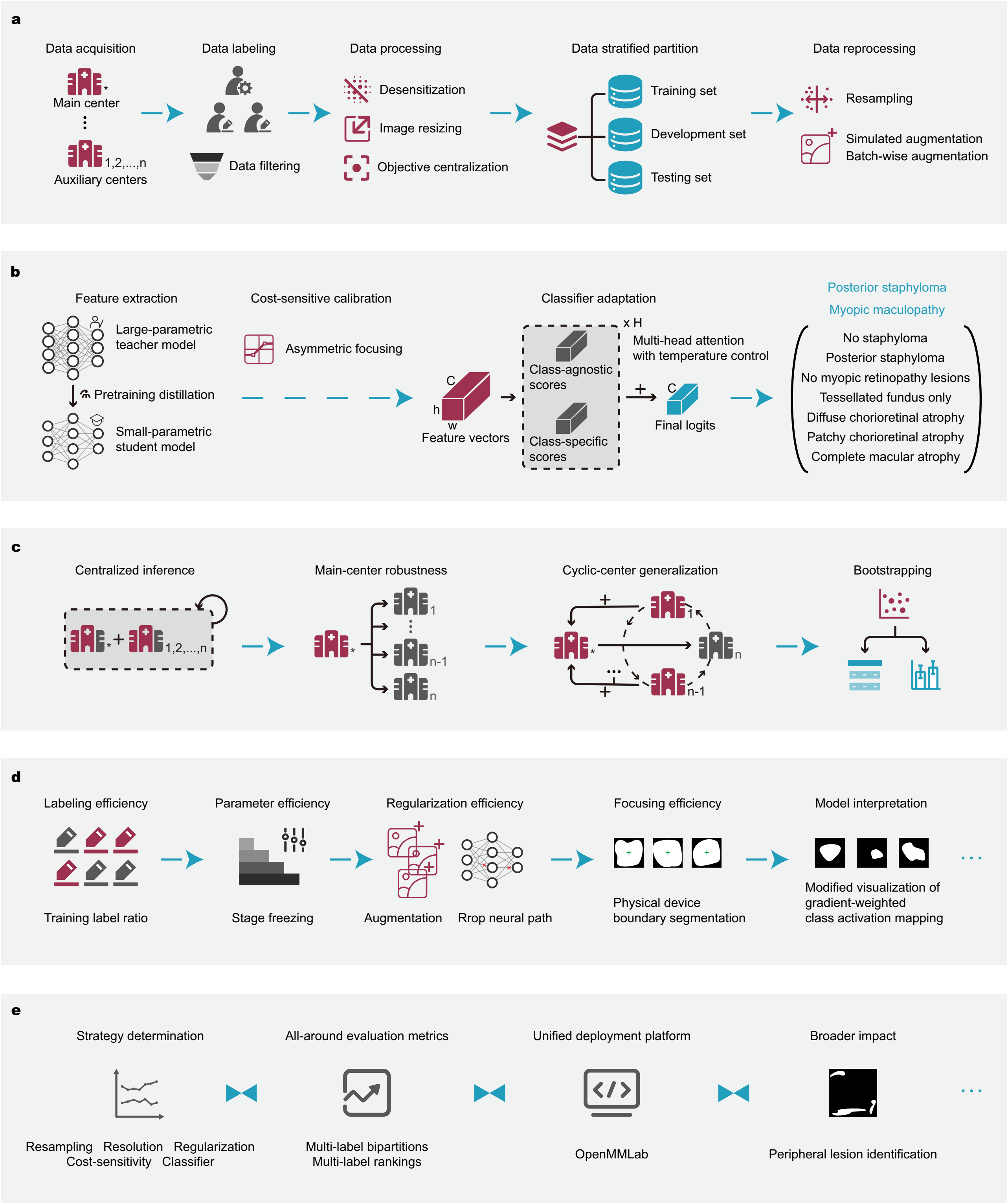

Fig. 1 | General overview of the study. **a** Data machining: data are collected from one main center and four auxiliary centers. After double-checking labeling, quality filtering, and essential processing, a stratified partition is implemented to ensure that the distribution of lesions remains similar across sets. Resampling and augmentation techniques are then used to alleviate label imbalance. **b** Model training and inference: the pretraining-distilled small parametric model is task-specifically fine-tuned with asymmetric focusing and classifier adaptation, which complementarily mitigate label imbalance. **c** Experimental protocols: three protocols are designed to demonstrate precise inference, robustness, and generalizability of the proposed method. All experiments are implemented by bootstrapping the testing set 1000 times. **d** Reasoned workflow: model efficiencies of dataset labeling, training parameters, regularization techniques, and focusing regions are extensively examined. Visualizations of gradient-weighted class activation mapping are provided for intuitive interpretations. **e** Model development and assessment: models are progressively developed through strategy determination, and their performance is assessed on a unified deployment platform using all-around evaluation metrics.

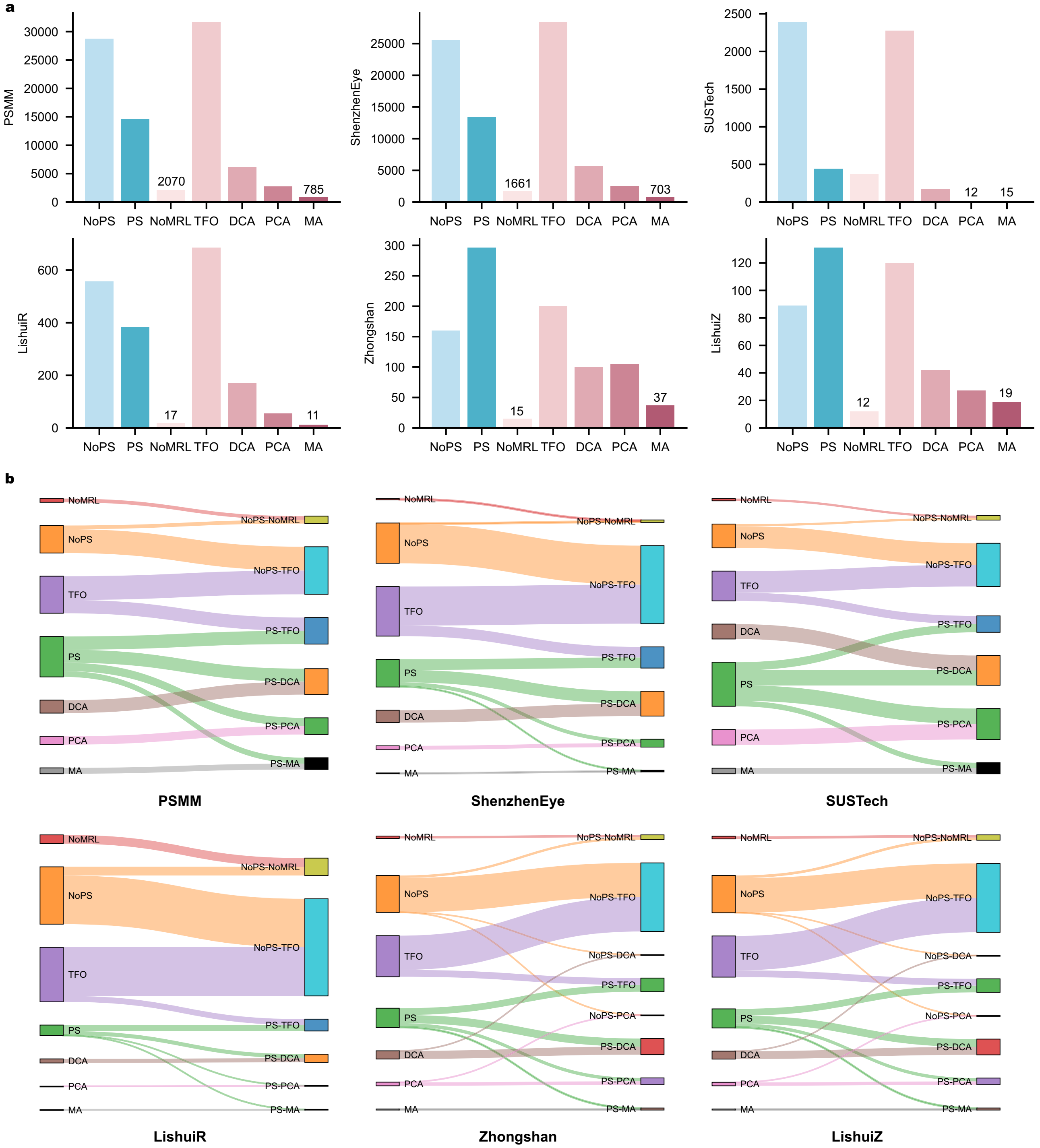

Fig. 2 | Statistics and complications associated with lesions of posterior staphyloma and myopic maculopathy. **a** Statistical analysis of the seven categories in the PSMM dataset and its subsets, with specific values assigned to the minimum two categories of each dataset. **b** Illustrations of complications arising from posterior staphyloma and myopic maculopathy. Sankey diagrams are plotted to illustrate the distribution of these complications in the PSMM dataset and its subsets.

macular atrophy (MA). An intuitive illustration of these morphologic patterns can be found in Supplementary Fig. 1b and 1c. Notably, posterior staphyloma and myopic maculopathy may appear simultaneously (Fig. 2b), forming multi-label datasets. The PSMM dataset exhibits an imbalanced distribution (Fig. 2a), posing a significant challenge to method development. Overall, the PSMM dataset is well-curated on fine-grained multi-lesion recognition and the diagnosis of pathologic myopia, which also provides convenience for those developing deep learning models for recognizing retinal diseases, as well as empowering large-parametric deep learning techniques like foundation models to discern retinal diseases requiring ultra-widefield images.

End-to-end lightweight hybrid framework with optimization mitigates multi-label imbalance issue

The imbalance present in multi-label datasets significantly impacts the model's performance, leading to biased learning and inadequate knowledge acquisition. This study presented three techniques to address the imbalance issue in the PSMM dataset (Supplementary Table 2): resampling methods, classifier adaptation, and cost-sensitive calibration. Cost-sensitive calibration addressed the multi-label imbalance by developing the loss function based on Binary Cross-Entropy (BCE) Loss[27]. This approach recognizes that multi-label learning decomposes the multi-label task into several binary tasks, each aimed at distinguishing samples within a target class category. We gradually introduced configurable parameters for BCE Loss to reduce the imbalance issue on the PSMM dataset. To begin, we trained the model using BCE Loss. We implemented a commonly used weighting factor $\alpha \in [0, 1]$ to form an α-balanced BCE Loss for class imbalance. In our experiments, we discovered that the model performed better when using an α value of 0.75 (Supplementary Table 5), which aligned with its original use in the dense detection task. We introduced a focusing parameter, γ, to adjust the loss function and concentrate training on difficult negative samples by reducing the impact of easy samples[28]. We tested different α values for each candidate focusing parameter within the list of $[0, 0.1, 0.2, 0.5, 1, 2, 5]$, as recommended in the original literature. We found that increasing the focusing parameter did not yield any benefits (Supplementary Table 6), possibly due to the elimination of gradients from rare positive samples while devaluing the contribution from easy negatives. To address this issue, we utilized γ_+ and γ_- to separate the focusing levels of positive and negative samples, allowing the model to emphasize the positive samples while minimizing the influence of easy negative samples[29]. The experimentally determined cost-sensitive calibration helps the model learn from balanced samples (Supplementary Fig. 4a), ultimately leading to optimal performance with $\gamma_+ = 3$ and $\gamma_- = 4$. We introduced a probability-shifting mechanism to assess the influence of very easy and mislabeled negative samples. The results showed that adjusting the shifted probability did not improve the model's performance, indicating that our dataset was well-curated and had minimal errors. We also studied a state-of-the-art approach called Two-way Loss[30], which is exclusively designed for multi-label learning. This method uses relative comparison with the softmax function. We adjusted the margins between positive and negative logits using positive temperature T_P and negative temperature T_N. We evaluated different values for T_P and T_N within the list of $[0.5, 1, 2, 4]$. The results (Supplementary Table 7) showed a similar trend to the original study, but the best-performing choice still did not outperform our implementation using asymmetric focusing. Classifier adaptation involves residual attention, combining class-specific and class-agnostic features during the inference stage[31]. We introduced a configurable parameter λ to leverage these two types of features, searching within the range of [0.2, 1.4] with a step of 0.2, as done in the original literature using Vision Transformer as the backbone on the MS-COCO dataset. The residual attention was extended in a multi-head (H) manner, initially set at $H = 8$. The model with $\lambda = 1.2$ and $H = 2$ achieved better mAP compared to other settings while maintaining similar performance on other evaluation metrics (Supplementary Fig. 4a).

Multi-protocol experiments demonstrate valued inference with robustness and generalizability

We devised three distinct experimental protocols (Fig. 1c) to explore the model's inference capacity, robustness, and generalizability (see detailed strategies in "Experimental protocols" section). The results (Fig. 3) under the centralized inference protocol revealed that RealMNet-Max outperformed ($P < 0.001$) four benchmark approaches on F1 Score with 0.7970 (95% CI 0.7612–0.8328), mAP with 0.8497 (95% CI 0.8058–0.8937), and AUROC with 0.9745 (95% CI 0.9690–0.9801). Notably, RealMNet-Max outperformed ($P < 0.001$) two foundation models, DINOv2[25] and VisionFM[26] by (mean estimate) 5.73% and 8.02% on F1 Score, 5.94% and 7.95% on mAP, 1.10% and 1.71% on AUROC, respectively. Unless otherwise noted, these three metrics were considered the primary criteria for measuring the model's performance. We presented other evaluation metrics for comprehensive analysis (see "Evaluation metrics" section). Specifically, RealMNet-Max achieved the lowest Coverage of 2.2522 (95% CI 2.2157–2.2888), significantly surpassing ($P < 0.001$) other models, indicating that the proposed model could better approximate the realistic situation. Precision and Recall were two opposite measures, with one tending to be high and the other low. In our case, we preferred a superior Recall for developing a discrimination model that would identify as many potential positive samples as possible to aid in screening. Guided by the main-source robustness protocol, we discovered that the model trained exclusively on the main subset could reliably identify posterior staphyloma and myopic maculopathy on auxiliary subsets in general (Fig. 4a). On the other hand, it illustrated abundant task-specific knowledge implied in the primary source data. RealMNet represented robustness on the SUSTech subset, achieving an F1 Score of 0.7956 (95% CI 0.7187-0.8724), mAP of 0.8927 (95% CI 0.8211-0.9642), and AUROC of 0.9869 (95% CI 0.9830-0.9908). Even when tested on the Zhongshan subset whose hard negative samples may hinder model inference, our model still maintained acceptable performance (mean value) with an F1 score of over 70%, mAP over 80%, and AUROC over 95%. When examined under the cyclic-source generalizability protocol, RealMNet exhibited similar performance to that under the main-source robustness protocol (Fig. 4b), reflecting its stable performance when additional information was introduced. On the Zhongshan subset, the model displayed difficulty in correctly distinguishing a small fraction of label pairs, as evidenced by a Hamming Loss of 0.0985 (95% CI 0.0898-0.1072) and a Ranking Loss of 0.0530 (95% 0.0467-0.0593). This could be attributed to a relatively high Coverage value, indicating that the model required more steps to infer all relevant labels for the samples of posterior staphyloma and myopic maculopathy.

Reasoned workflow facilitates convincing diagnosis of pathologic myopia in clinical application

Even though deep learning methods offer powerful capacities, they are commonly known as black boxes due to their intricate inference mechanisms[32]. To be useful in clinical applications, these methods need to be not only efficient but explainable and trustworthy. Labeling efficiency refers to the amount of training data and labels required to achieve a certain level of performance for a given task, which shows the annotation workload for medical experts[19]. RealMNet achieves precise identification even with only half of the training resources in data ablation studies (Fig. 5a), demonstrating its capability to capture clinically significant morphologic patterns at a low-level feature space. RealMNet-384 exemplified a remarkable improvement (mean value) in F1 Score by 10%, mAP by 10%, and AUROC by 1%, despite an increase in labeling from 20% to 50%. Although the RealMNet-Min and the RealMNet performed similarly as more training data was used, RealMNet-Max consistently achieved superior performance, demonstrating the non-trivial benefits of abundant information involved in higher resolution. The model could have gained even slightly higher performance when using ninety percent of the training resources; we insisted that the model trained on all available data eliminate the variability and produce unbiased results. We aimed to assess the contribution of each stage of the used backbone by measuring parameter efficiency (Fig. 5b). Freezing the first one or two layers of the model did not decrease performance, indicating that the model retained low-level general features from pre-training distillation on large-scale natural image datasets (e.g., ImageNet-21k). However, the performance of RealMNet dropped significantly when the first three or four layers were frozen, indicating that the model still required high-level features related to morphologic patterns. Furthermore, we observed the efficacy of the regularization used in this study. Augmentation is a crucial regularization technique widely adopted in deep learning-based approaches to augment training data to avoid overfitting, especially when the amount of training data is not large enough in many tasks of medical fields. We explored the impact of the proposed simulated augmentation and batch-wise augmentation (Fig. 5c) with ablation studies and found that employing these two types of augmentation techniques brought a gain of 3.5% on F1 Score, 6.7% on mAP, and 3.2% on AUROC, respectively (w/o Augmentation vs. Augmentation). The simulated augmentation was used to mirror real-world situations. The model's performance decreased significantly when the simulated augmentation was removed (w/o S-Augmentation vs. Augmentation). This suggested that the model was trained with overly optimistic and simplistic objectives because the training data did not represent real-world scenarios. Batch-wise augmentation

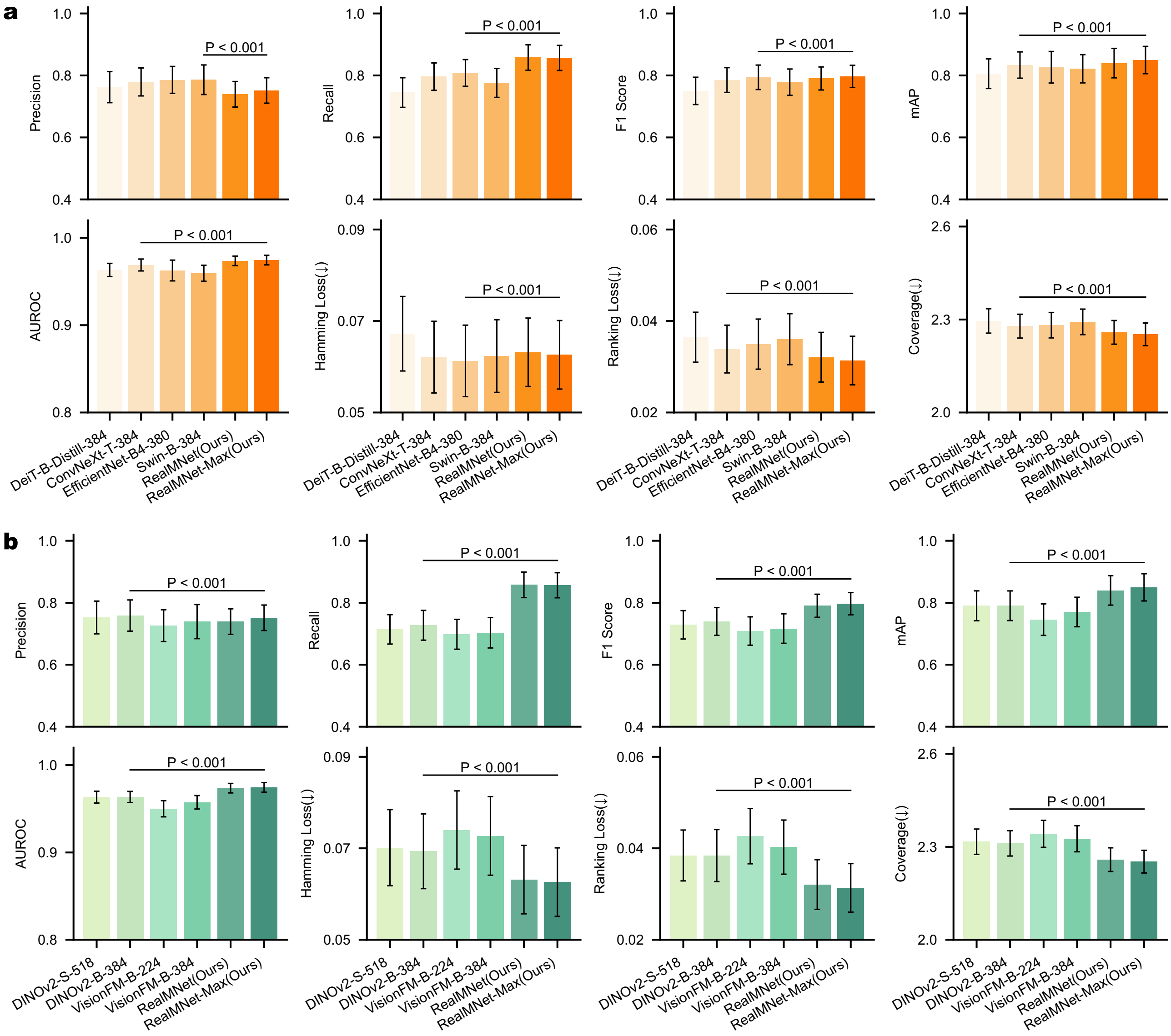

Fig. 3 | Model performance under the centralized inference protocol. **a** The proposed models are compared to four well-known benchmarks: DeiT, ConvNeXt, EfficientNet, and Swin Transformer. **b** The proposed models are compared to two recent foundation models: DINOv2 and VisionFM. The error bars represent the 95% confidence interval of the estimates, and the bar center represents the mean estimate of the displayed metric. The estimates are computed by generating a bootstrap distribution with 1000 bootstrap samples for corresponding testing sets with $n = 1000$ samples. All P-values are computed with a two-sided t-test between RealMNet-Max and the most competitive comparison model to determine if there are statistically significant differences.

involved enhancing synthetic samples by interweaving two samples. Removing batch-wise augmentation did not cause a significant loss (w/o B-Augmentation vs. Augmentation), indicating that the model had inherently been adequate to build intra- and inter-affinities between morphologic patterns. A slight decrease in Ranking Loss and Coverage suggested that batch-wise augmentation helped the model learn more accurate label distributions. Drop path[33] is another regularization technique that markedly circumvents the overfitting issue by randomly dropping the neural path of the network. We used the drop path because of the overfitting hazard caused by a relatively small scale of training data (Supplementary Fig. 5c). To interpret the panoramic focusing capacity of RealMNet, we considered the potential negative impact of the physical device boundaries inevitably imaged along with the imaging targets by ultra-widefield imaging, which may occlude essential information. The comparative experimental results (Fig. 6a) showed that there was no significant difference (mean estimate 0.15% on F1 Score, −0.21% on mAP, −0.02% on AUROC) between the performance of models trained on data with and without boundary segmentation, which suggested that the model distinguished instrumental regions and focused on the field of view within the boundaries of the physical devices. We conducted experiments to specifically measure the benefits of using UWF images over CFP images (Supplementary Fig. 7b) for identifying posterior staphyloma and myopic maculopathy. Experimental results (Fig. 6b) indicated that the model trained with UWF images significantly outperformed the model trained with fake CFP images by (mean estimate) 6.03% on F1 Score, 4.84% on mAP, and 0.93% on AUROC, respectively. This enhanced performance can be attributed to the higher resolution, superior imaging quality, and broader retinal field of view afforded by the UWF modality. Visual interpretability has been widely recognized as an intuitive representation of the decision-making process in deep learning techniques. We adopted an improved version of gradient-weighted class activation mapping[34] that mapped objects' morphology better and explained occurrences of multiple objects of a class in a single image. We generated visualizations of random samples for each category using RealMNet (Fig. 7). The outputs are in the exploratory phase and require rigorous quantitative validation or assessment by experts to ensure their reliability and effectiveness. These heatmaps qualitatively revealed

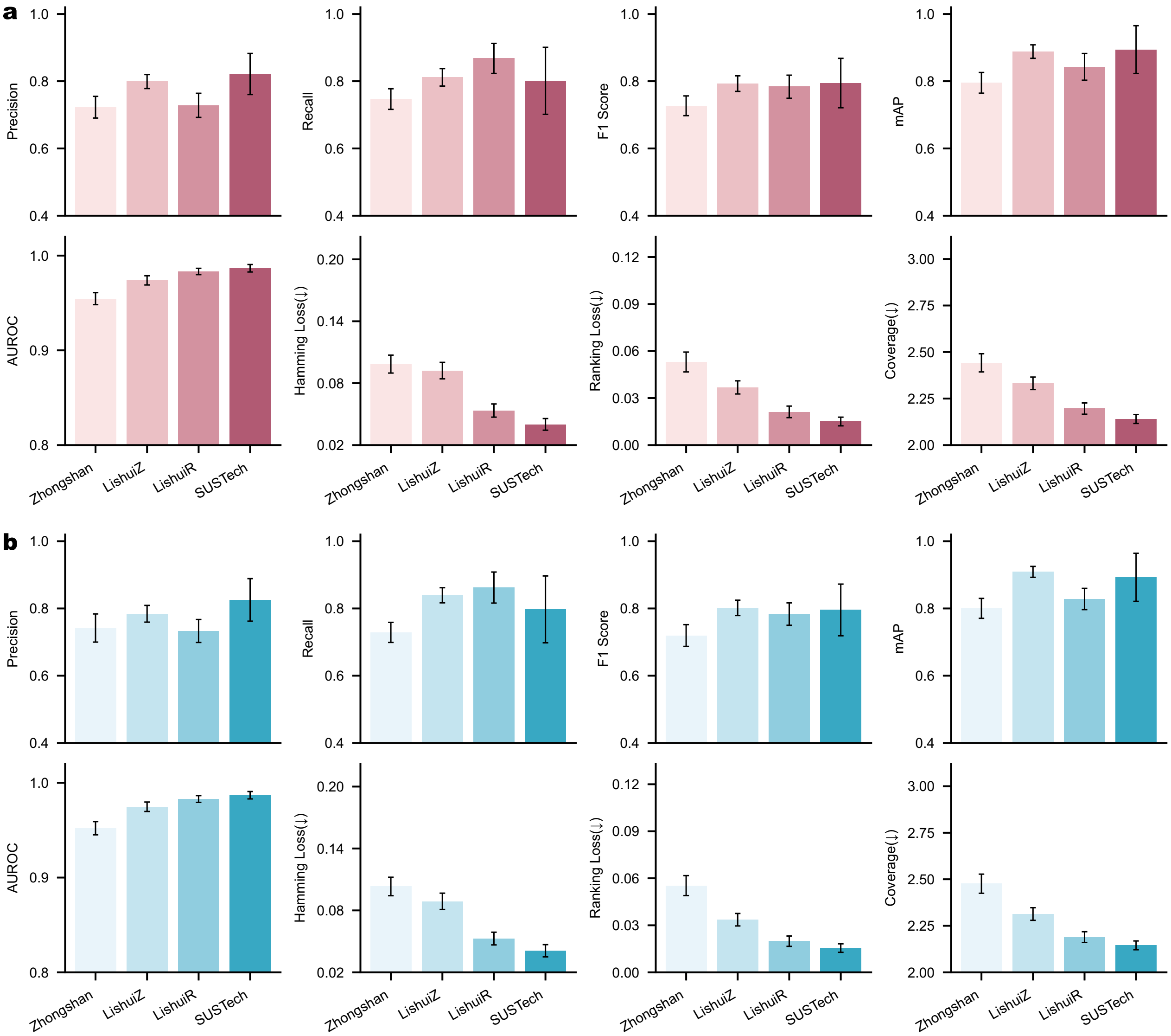

Fig. 4 | Model performance under the main-source robustness protocol and cyclic-source generalizability protocol. **a** Assessing model robustness by training on the main source subset and testing on four auxiliary source subsets under the main-source robustness protocol. **b** Assessing model generalizability by training on the main source subset combined with three of the four auxiliary source subsets and testing on the remaining subset under the cyclic-source generalizability protocol. The error bars represent the 95% confidence interval of the estimates, and the bar center represents the mean estimate of the displayed metric. The estimates are computed by generating a bootstrap distribution with 1000 bootstrap samples for corresponding testing sets with *n*=1000 samples.

irregular attentive regions corresponding to diffused morphologic patterns embodied in different lesion categories. We observed that the highlighted regions often correspond to clinically relevant areas of the retina. For example, the heatmaps for MA and PCA tended to focus on the macular region, which aligns with the typical distribution of lesions seen in clinical diagnoses. Likewise, in cases of DCA, the heatmaps commonly covered the mid-peripheral retina, consistent with known patterns of atrophic progression. These spatial relationships suggested that the model implicitly learns to concentrate on anatomically significant regions when predicting different subtypes of myopic maculopathy, thereby enhancing both interpretability and clinical relevance. We also visualized class-wise confusion matrices for identifying morphologic patterns on the PSMM dataset (Supplementary Fig. 10 and 11).

The inherent patterns of the model developed in this study make it easy to use for tasks concerning concurrent lesion identification. In this study, we emphasized the significance of identifying peripheral retinal lesions in highly myopic eyes. We observed (Fig. 8b) that the fine-tuned model generally performed well, with an AUROC of 0.8642 (95% CI 0.8405–0.8880) in discerning concurrent peripheral retinal regions with the proposed off-the-shelf workflow without bells and whistles. We found that the fine-tuned model could accurately perceive PC with an F1 Score of 0.8394 (95% CI 0.8033–0.8754), mAP of 0.8894 (95% CI 0.8580–0.9208), and AUROC of 0.9029 (95% CI 0.8721–0.9336). We inferred an inferior capacity to distinguish RRD and HoT, possibly due to the scarcity of real-world data. Notably, we used consistent training settings for the intuitive perception of transfer capacity, which signified the potential for improved performance with further investigation. The success of our workflow in identifying peripheral retinal lesions highlights its broader utility for enhancing the diagnosis of retinal diseases and other complex medical scenarios Tables 1–3.

Discussion

In this study, we introduced a novel perspective for assisting in diagnosing pathologic myopia by means of identifying posterior staphyloma and myopic maculopathy using ultra-widefield images with deep learning. We found that there have been many studies dedicated to the application of deep learning to assist myopia diagnosis[35,36]. However, the majority of these

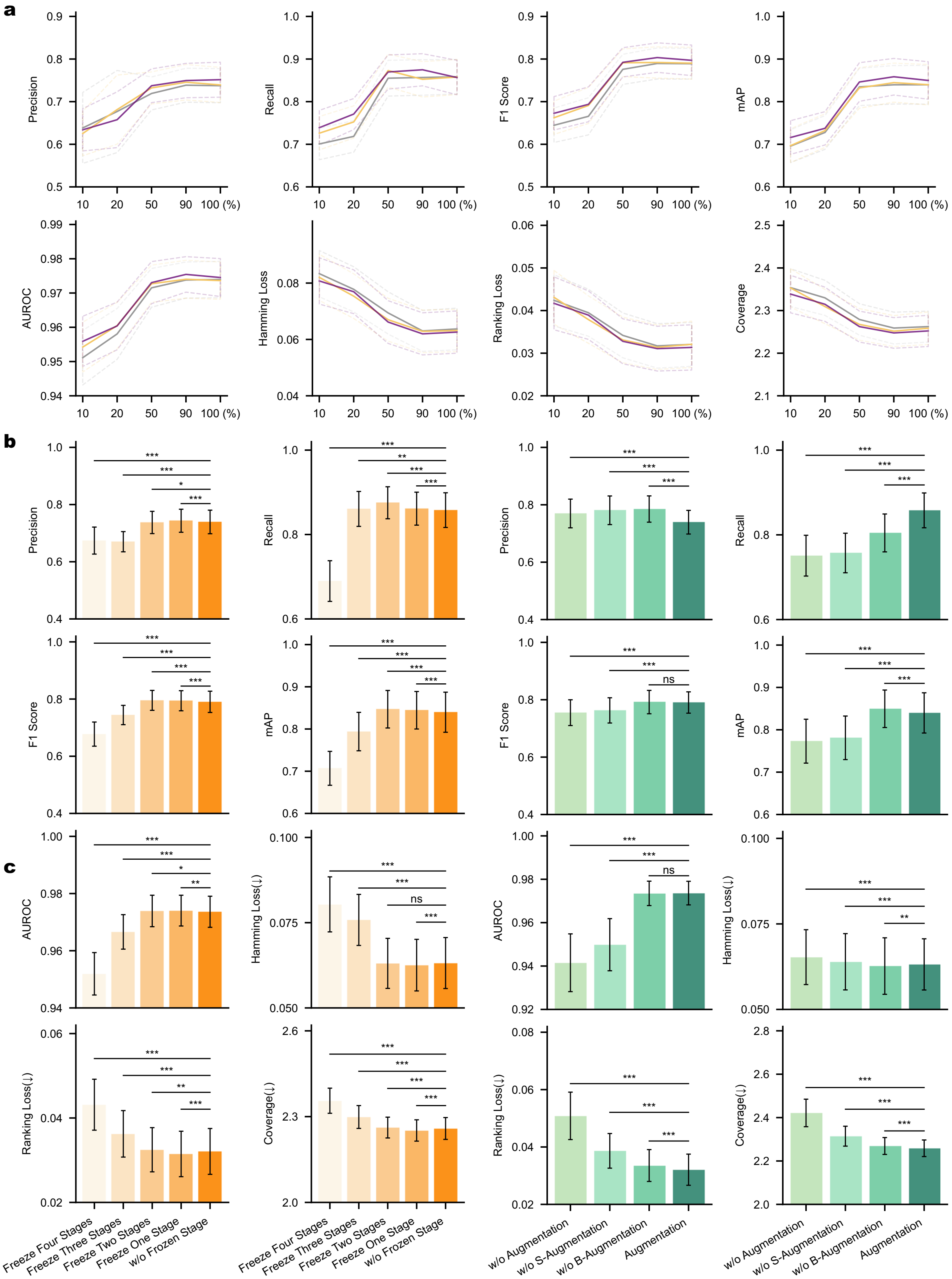

Fig. 5 | Efficiency of RealMNet in identifying posterior staphyloma and myopic maculopathy on the PSMM dataset. a Labeling efficiency: we progressively increase the amount of training data and labels to achieve precise and stable performance. **b** Parameter efficiency: we freeze training parameters from different stages to observe the contribution of each stage. **c** Augmentation efficiency: we ablate two types of augmentation techniques, namely simulated augmentation (S-Augmentation) and batch-wise augmentation (B-Augmentation), to observe the performance gains that RealMNet gets as a result of these techniques. The error bars represent 95% CI of the estimates, and the bar center represents the mean estimate of the displayed metric. The estimates are computed by generating a bootstrap distribution with 1000 bootstrap samples for corresponding testing sets with n=1000 samples. All P values are computed with a two-sided t-test between the original model and its variants to determine if there are statistically significant differences. The bars marked with `ns' are not significant. The asterisks indicate statistically significant differences: $*P < 0.05$; $**P < 0.01$; $***P < 0.001$.

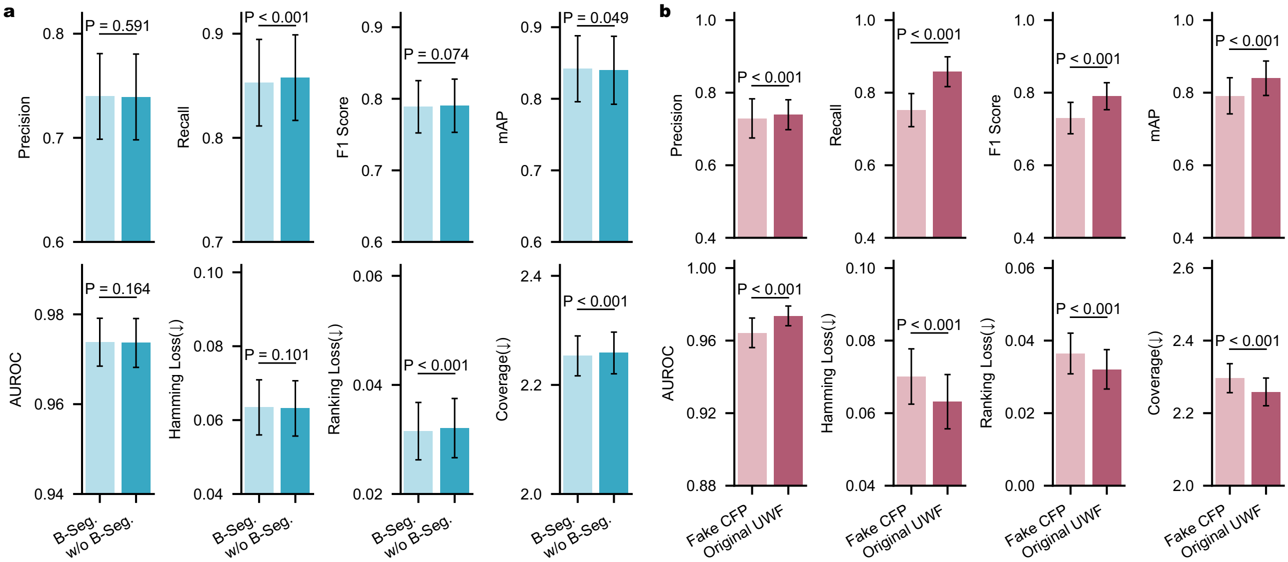

Fig. 6 | Investigating the superiority of UWF modality. a Comparing the performance of models trained on UWF images with and without boundary segmentation. **b** Comparing the performance of models trained on data with original UWF and fake CFP images. The error bars represent the 95% confidence interval of the estimates, and the bar center represents the mean estimate of the displayed metric. The estimates are computed by generating a bootstrap distribution with 1000 bootstrap samples for corresponding testing sets with n=1000 samples. All P values are computed with a two-sided t-test between two comparison models to determine if there are statistically significant differences.

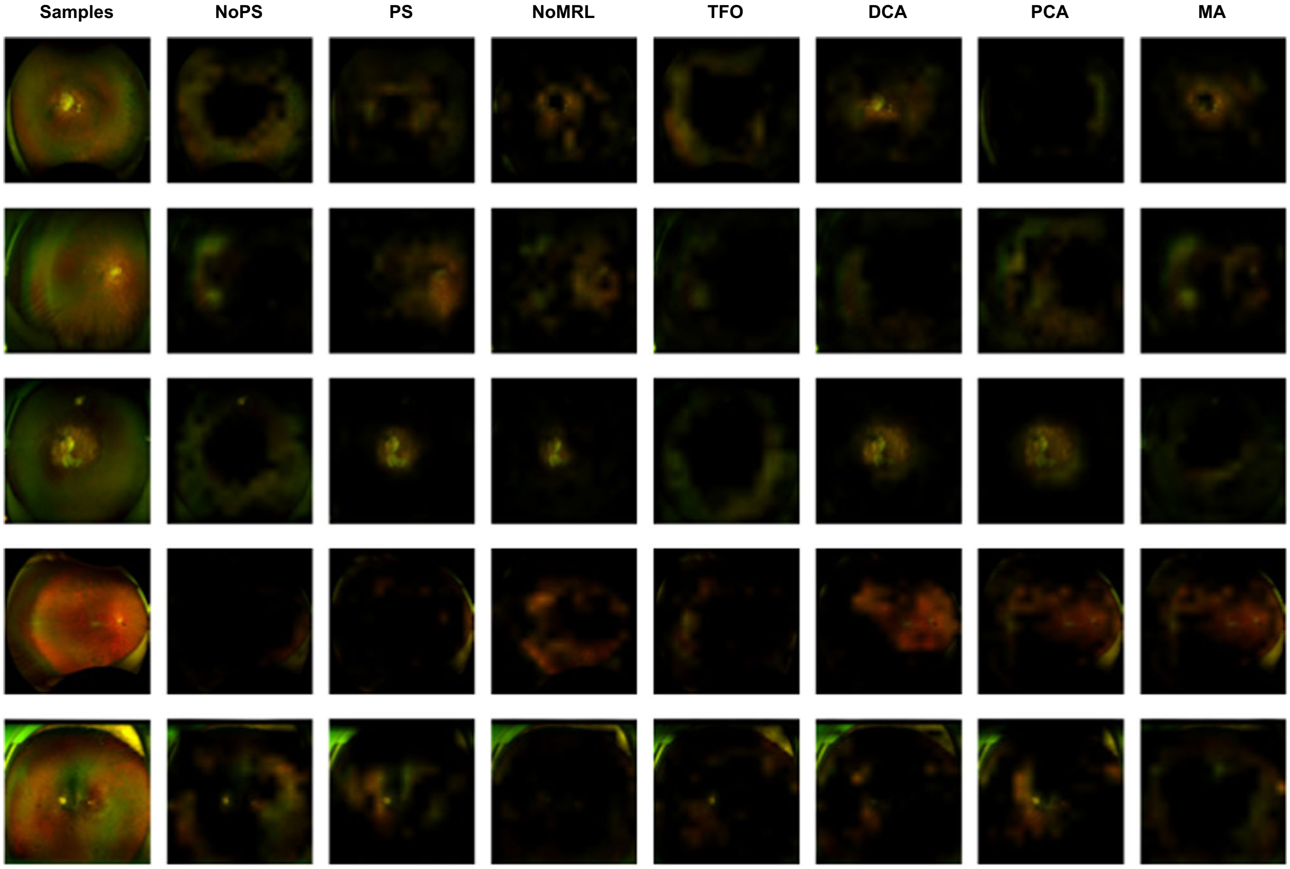

Fig. 7 | We generated visualizations using an improved version of gradient-weighted class activation mapping. These visualizations show the qualitative predictions of RealMNet for presence of posterior staphyloma (NoPS or PS) and myopic maculopathy with five categories: no myopic retinal lesions (NoMRL), tessellated fundus only (TFO), diffuse chorioretinal atrophy (DCA), patchy chorioretinal atrophy (PCA), and macular atrophy (MA). By merging the heatmaps with the original images, we highlighted irregular attentive regions that correspond to diverse morphologic patterns found in different lesion categories when the model made decisions. These heatmaps provided a qualitative reference for clinicians when making further diagnoses.

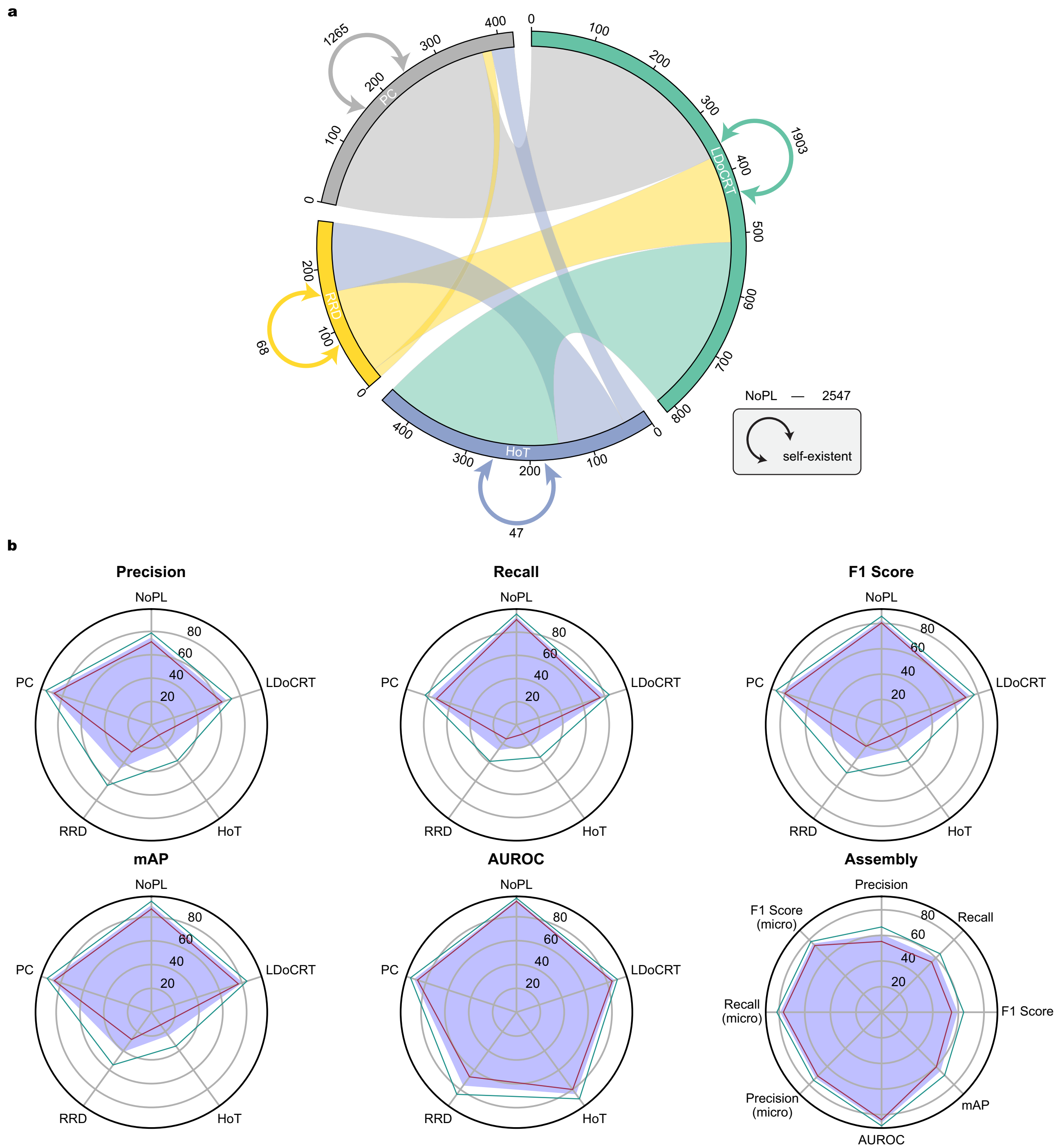

Fig. 8 | Identifying complicated peripheral lesions. a Concurrent distribution of peripheral retinal lesions: no peripheral lesion (NoPL), lattice degeneration or cystic retinal tuft (LDoCRT), holes or tears (HoT), rhegmatogenous retinal detachment (RRD), and postoperative cases (PC). Peripheral lesions may have different concurrent relationships with each other, or they may occur separately. **b** Model performance on peripheral lesion identification. The blue facecolor represents the mean of the results, and the green outer and red inner boundaries represent the upper and lower bounds of the 95% confidence interval, respectively. All radar plots display class-wise performance on specific metrics, with the last radar plot representing the average performance on all evaluated metrics.

studies overlooked exclusive discrimination mechanisms due to a lack of specialized datasets built on ophthalmological expertise. Pathologic myopia has been broadly recognized as myopic maculopathy with meticulously defined categories or with the presence of posterior staphyloma[37]. Nonetheless, to our knowledge, limited research has thoroughly examined these lesions, and there are no publicly available datasets for this purpose. While Pathologic myopia currently lacks curative treatment, timely detection is still clinically significant. Early identification of complications like posterior staphyloma and myopic maculopathy aids risk assessment and close monitoring, allowing for interventions such as anti-VEGF therapy for secondary choroidal neovascularization or refractive surgery consultations. Our workflow identifies high-risk eyes for further evaluation and ongoing surveillance, aligning with current clinical management strategies. Specifically, we gathered a large-scale dataset comprising ultra-widefield images from five distinct hospital sources (Fig. 1a). We sought experienced ophthalmologists to label posterior staphyloma with binary annotations to indicate its presence: NoPS and PS, and myopic maculopathy with five categories: NoMRL, TFO, DCA, PCA, and MA. We built an end-to-end

Table 1 | Model class-wise performance on the evaluation metric of F1 Score

Model	NoPS	PS	NoMRL	TFO	DCA	PCA	MA
DeiT	92.9080 ± 0.0220	84.9378 ± 0.0465	65.8115 ± 0.1929	93.1385 ± 0.0216	64.9436 ± 0.1112	64.0049 ± 0.1537	59.4071 ± 0.3663
ConvNeXt	93.4954 ± 0.0216	86.6145 ± 0.0443	67.4836 ± 0.1821	93.3462 ± 0.0205	69.4887 ± 0.1009	69.1593 ± 0.1467	70.0695 ± 0.3148
EfficientNet	93.3104 ± 0.0220	86.3812 ± 0.0440	67.8674 ± 0.1824	93.4917 ± 0.0202	68.5025 ± 0.1026	74.8105 ± 0.1361	71.4768 ± 0.3003
Swin Transformer	93.2645 ± 0.0213	85.9297 ± 0.0436	64.8995 ± 0.2008	93.6117 ± 0.0207	69.4373 ± 0.1047	67.9535 ± 0.1534	69.7311 ± 0.3287
DINOv2	92.3681 ± 0.0231	83.9888 ± 0.0473	60.9115 ± 0.2047	93.1201 ± 0.0207	66.1795 ± 0.1114	62.1230 ± 0.1636	59.0787 ± 0.3724
VisionFM	92.5725 ± 0.0224	84.3148 ± 0.0471	58.5730 ± 0.2083	92.4366 ± 0.0227	62.4830 ± 0.1157	58.5566 ± 0.1653	52.8073 ± 0.4006
RealMNet(Ours)	**93.7815** ± 0.0208	**86.6268** ± 0.0427	**68.8711** ± 0.1658	**93.5031** ± 0.0209	**71.4770** ± 0.0929	**72.1523** ± 0.1373	**66.7895** ± 0.3134
RealMNet-Max(Ours)	**93.8404** ± 0.0204	**86.2816** ± 0.0427	**70.5547** ± 0.1658	**93.5852** ± 0.0206	**71.8504** ± 0.0909	**72.2685** ± 0.1367	**69.5145** ± 0.3117

Reported values are the mean estimate with the standard error of the targeted metric. The estimates are computed by generating a bootstrap distribution with 1000 bootstrap samples for corresponding testing sets with n=1000 samples. The prominent results of the proposed methods are highlighted in bold.

Table 2 | Model class-wise performance on the evaluation metric of mAP

Model	NoPS	PS	NoMRL	TFO	DCA	PCA	MA
DeiT	98.3357 ± 0.0082	90.0271 ± 0.0595	72.0752 ± 0.2092	98.0086 ± 0.0121	70.5901 ± 0.1388	66.5143 ± 0.1985	68.4258 ± 0.3871
ConvNeXt	98.4937 ± 0.0078	90.1852 ± 0.0607	75.1091 ± 0.2010	98.5105 ± 0.0105	73.1993 ± 0.1250	70.9387 ± 0.1844	77.0458 ± 0.3123
EfficientNet	98.5506 ± 0.0080	92.6208 ± 0.0424	74.7456 ± 0.1943	98.2751 ± 0.0115	74.2411 ± 0.1148	74.2415 ± 0.1849	65.8337 ± 0.4204
Swin Transformer	98.1808 ± 0.0108	89.8026 ± 0.0575	69.9827 ± 0.2277	98.0135 ± 0.0140	71.0696 ± 0.1376	70.6761 ± 0.1823	77.4035 ± 0.3282
DINOv2	98.2004 ± 0.0086	90.6269 ± 0.0502	68.4695 ± 0.2151	98.2794 ± 0.0097	68.9774 ± 0.1330	65.6704 ± 0.1957	63.0445 ± 0.4232
VisionFM	97.9320 ± 0.0133	90.6120 ± 0.0467	64.9242 ± 0.2290	97.6936 ± 0.0143	67.2612 ± 0.1374	61.0721 ± 0.2025	59.6629 ± 0.4034
RealMNet(Ours)	**98.7762** ± 0.0063	**93.3474** ± 0.0390	**76.8974** ± 0.1809	**98.7188** ± 0.0071	**76.2035** ± 0.1202	**74.0739** ± 0.1749	**69.8559** ± 0.4057
RealMNet-Max(Ours)	**98.7822** ± 0.0063	**93.0462** ± 0.0432	**78.0920** ± 0.1822	**98.7352** ± 0.0072	**75.7264** ± 0.1222	**76.1924** ± 0.1705	**74.2466** ± 0.3623

Reported values are the mean estimate with the standard error of the targeted metric. The estimates are computed by generating a bootstrap distribution with 1000 bootstrap samples for corresponding testing sets with n=1000 samples. The prominent results of the proposed methods are highlighted in bold.

Table 3 | Model class-wise performance on the evaluation metric of AUROC

Model	NoPS	PS	NoMRL	TFO	DCA	PCA	MA
DeiT	96.1189 ± 0.0173	96.1254 ± 0.0172	97.7149 ± 0.0238	94.8871 ± 0.0234	94.7713 ± 0.0240	95.2838 ± 0.0433	99.3335 ± 0.0096
ConvNeXt	96.4324 ± 0.0170	96.4378 ± 0.0170	98.1800 ± 0.0227	96.0292 ± 0.0188	95.3681 ± 0.0220	96.3815 ± 0.0382	99.3870 ± 0.0102
EfficientNet	96.7455 ± 0.0158	96.7423 ± 0.0158	98.1674 ± 0.0185	95.5410 ± 0.0209	94.8309 ± 0.0247	94.1538 ± 0.0759	97.6156 ± 0.0868
Swin Transformer	95.9789 ± 0.0188	95.9643 ± 0.0189	96.7591 ± 0.0371	95.1124 ± 0.0227	94.7008 ± 0.0249	94.5249 ± 0.0592	98.5647 ± 0.0364
DINOv2	95.9202 ± 0.0172	95.9195 ± 0.0172	97.3954 ± 0.0228	95.2212 ± 0.0201	94.9398 ± 0.0228	96.3000 ± 0.0246	98.7755 ± 0.0243
VisionFM	95.6557 ± 0.0190	95.6629 ± 0.0190	96.4368 ± 0.0292	94.0472 ± 0.0254	94.2133 ± 0.0252	95.3203 ± 0.0317	98.8797 ± 0.0202
RealMNet(Ours)	**97.1399** ± 0.0139	**97.1671** ± 0.0139	**98.3832** ± 0.0145	**96.3504** ± 0.0170	**95.8280** ± 0.0194	**97.5641** ± 0.0199	**99.1062** ± 0.0160
RealMNet-Max(Ours)	**97.1369** ± 0.0140	**97.1838** ± 0.0140	**98.5153** ± 0.0140	**96.4073** ± 0.0170	**95.9304** ± 0.0186	**98.0075** ± 0.0164	**98.9830** ± 0.0248

Reported values are the mean estimate with the standard error of the targeted metric. The estimates are computed by generating a bootstrap distribution with 1000 bootstrap samples for corresponding testing sets with n=1000 samples. The prominent results of the proposed methods are highlighted in bold.

lightweight framework called RealMNet on the basis of the unified platform to identify these concurrent lesions with multi-label learning (Fig. 1b). We have provided detailed comparisons of model complexity, specifically focusing on the number of parameters and FLOPs, benchmarked against well-established architectures (Table 4). Our model has only 21 million parameters and requires approximately 13.77G FLOPs per inference, which has a smaller scale but is more efficient than other similarly lightweight models. Notably, RealMNet outperforms two powerful foundation models while containing four times fewer parameters (~21M compared to ~86M) and up to three times fewer FLOPs (~27G compared to ~78G). We progressively determined resampling approaches (Supplementary Fig. 5a), cost-sensitive calibration (Supplementary Fig. 4a), and classifier adaptation (Supplementary Fig. 4b) with the development set for mitigating negative impacts caused by imbalanced label distributions (Fig. 2). Hence, the proposed model was functionally reliable by identifying these clinically significant lesions and was objectively instrumental by alleviating multi-label imbalance issues. We formulated a multi-faceted strategy (Supplementary Note: Clinical practice deployment) that emphasizes clinician engagement and seamless integration into daily clinical practice. While our study did not directly examine disease pathogenesis or progression, it offers a validated UWF imaging-based workflow for the early identification of structural complications, thereby supporting surveillance and timely referrals within myopia management frameworks Tables 5–7.

We devised three experimental protocols (Fig. 1c) to demonstrate the model's inference capacity, robustness, and generalizability. We observed that the proposed model outperformed ($P < 0.001$) all benchmark approaches (Fig. 3a). In particular, RealMNet demonstrated significantly better performance compared to two recent foundation models designed for

Table 4 | Information of model backbones

Model	Architecture	Implementation	Version	Image Size	#Params(M)	FLOPs(G)
DeiT[21]	Transformer	Distillation	Base	384	86.10	55.65
ConvNeXt[22]	ConvNet	Hierarchy	Tiny	384	27.83	13.14
EfficientNet[23]	ConvNet	Scaling	B4	380	17.56	4.51
Swin Transformer[24]	Transformer	Hierarchy	Base	384	86.89	47.19
DINOv2[25]	Transformer	Foundation Model	Base	384	86.14	78.46
VisionFM[26]	Transformer	Foundation Model	Base	384	86.46	55.54
RealMNet-Min (Ours)	Hybrid	Hierarchy Pretraining Distillation	21M	224	20.63	4.28
RealMNet (Ours)	Hybrid	Hierarchy Pretraining Distillation	21M	384	20.66	13.77
RealMNet-Max (Ours)	Hybrid	Hierarchy Pretraining Distillation	21M	512	20.70	27.02

Table 5 | Data overview of the centralized inference protocol (CIP)

Protocol	Training set		Development set		Testing set	
	Patients	Images	Patients	Images	Patients	Images
CIP	3192 (r. 3138)	30,420 (r. 24,683)	684	6377	684	6574

The numbers with prefix r. mean resampling results.

natural and ophthalmic images (Fig. 3b). This improvement can be attributed to the hierarchical design of our backbone and its effective management of multi-label data imbalance. Traditional foundation models typically focus on fundus images with a smaller field of view to identify various retinal diseases. In contrast, our proposed model excels in the fine-grained diagnosis of pathologic myopia, using lesion-wise labeling in UWF images. Meanwhile, our model exhibited good robustness (Fig. 4a) and generalizability (Fig. 4b), even when assessed on challenging subsets. To verify that the developed model has a broader application impact, we carried out a transfer learning on peripheral lesion discrimination, which could simultaneously exist in high myopic eyes (Fig. 8a) and give rise to severe visual impairment. The results (Fig. 8b) obtained from transfer learning for RealMNet demonstrated promise in detecting peripheral lesions and distinguishing postoperative cases.

Our model exhibited good labeling efficiency, taking different ratios of training data as input (Fig. 5a). As a transformer-based architecture with hierarchical design[24], each stage of RealMNet maintained helpful knowledge for lesion identification (Fig. 5b). The simulated and batch-wise augmentation jointly helped the model avoid over-fitting (Fig. 5c). From the heatmaps of the final results, we observed that the model's attention presented a diverse region of interest for different categories. We noticed that ultra-widefield images contained boundaries of physical imaging devices, which might hinder models from effectively capturing essential information. We constructed the dataset based on the scale of the two imaging types in the PSMM dataset (Supplementary Table 3). The processed data without boundaries (Supplementary Fig. 7a) was then used to re-train RealMNet. Experimental results (Fig. 6a) showed that our model was not affected by these physical boundaries, demonstrating the model's prominent capacity to capture informative regions. We conducted experiments to specifically measure the benefits of using UWF images over CFP images (Supplementary Fig. 7b) for identifying posterior staphyloma and myopic maculopathy. Experimental results (Fig. 6b) indicated that the model trained with UWF images significantly outperformed the model trained with fake CFP images. UWF imaging allows for improved visualization of structural alterations in the posterior pole and peripheral retina, including chorioretinal atrophy and optic disc abnormalities. These benefits of the UWF modality guarantee significance in the diagnosis of pathologic myopia.

Although this work starts from the essential and exclusive discrimination mechanisms of diagnosing pathologic myopia based on the workflow with deep learning, there are still some limitations and challenges to address in the follow-up work. Our model cannot currently recognize "plus" lesions[37], namely, lacquer cracks, myopic choroidal neovascularization, and fuchs spot, primarily due to insufficient high-quality data. We acknowledge that the gold-standard diagnosis of posterior staphyloma typically requires a multimodal assessment, including optical coherence tomography (OCT), axial length measurement, and choroidal imaging. The annotations in our study were derived from clinical diagnoses recorded in the electronic medical records by experienced retinal specialists, who had access to multimodal data during their evaluations. However, the absence of strict prospective revalidation of the labeled data with OCT and biometric measurements may have introduced some label noise. Hence, future efforts in dataset construction will incorporate standardized multimodal confirmation to enhance label reliability and generalizability. Results on peripheral lesion discrimination exposed limited performance on lesions with very few training data (e.g., RRD and HoT). Similar to most deep learning models, the model developed still lacks structural explainability, despite its strong inference capabilities. In light of these challenges, we propose to gather qualified data on "plus" lesions from additional medical sources and integrate clinical textual data such as axial length to improve the model's identification performance and consider post hoc methods (e.g., attention maps[38]) for approximating the attention to input tokens given attention weights. Quantitative evaluations and structured clinician feedback should be included to determine whether these visual outputs improve diagnostic trust or decision-making. It is also essential to quantify and analyze the clinical benefits of this lightweight design to ensure its feasibility for deployment and reliable application in real-world scenarios. We are optimistic that the model developed would receive excellent transfer ability when pretrained on large-scale UWF images and other informative modalities.

In summary, we offer a dataset comprising high-quality ultra-widefield images and introduce a helpful and reliable workflow for identifying morphologic patterns to aid in diagnosing pathologic myopia. Through comprehensive evaluation metrics on the hand-crafted PSMM dataset, we have verified the efficiency of RealMNet relative to competitive benchmark models. RealMNet has demonstrated superior robustness and generalizability, offering novel perspectives for deep learning-based fine-grained clinical diagnosis of pathologic myopia. In future work, we will further expand the data on scarce lesions and investigate the performance of the current model in multimodal situations.

Methods

Dataset construction

We show details about the course of data acquisition and labeling. We perform essential data processing and stratified data partitioning to facilitate model training.

The PSMM dataset consisted of five sub-sources: ShenzhenEye, SUSTech, LishuiR, Zhongshan, and LishuiZ. The study followed the guidelines of the World Medical Association Declaration of Helsinki 1964, updated in

Table 6 | Data overview of the main-source robustness protocol (MRP)

Protocol	Training set			Testing set		
	Subset	Patients	Images	Subset	Patients	Images
MRP	ShenzhenEye	4003 (r. 3944)	38,922 (r. 31,575)	SUSTech	226	2835
				LishuiR	155	938
				Zhongshan	85	456
				LishuiZ	91	220

The numbers with prefix r. mean resampling results.

Table 7 | Data overview of the cyclic-source generalizability protocol (CGP)

Protocol	Training set			Testing set		
	Subset	Patients	Images	Subset	Patients	Images
CGP	PSMM (w/o SUSTech)	4334 (r. 4256)	40,536 (r. 32,888)	SUSTech	226	2835
	PSMM (w/o LishuiR)	4405 (r. 4330)	42,433 (r. 34,408)	LishuiR	155	938
	PSMM (w/o Zhongshan)	4475 (r. 4398)	42,915 (r. 34,871)	Zhongshan	85	456
	PSMM (w/o LishuiZ)	4469 (r. 4389)	43,151 (r. 35,009)	LishuiZ	91	220

The numbers with prefix r. mean resampling results.

October 2013, and was conducted after approval by the Ethics Committees of Shenzhen Eye Hospital (2023KYPJ087), the Ethics Committee of the Zhongshan Ophthalmic Center (No.2022KYPJ105-2), the Institutional Review Board and Human Ethics Committee of the Fifth Affiliated Hospital of Wenzhou Medical University, the Ethics Committee of the Southern University of Science and Technology Hospital, and the Ethics Committee of Lishui People's Hospital. The review board waived the requirement for informed consent based on the retrospective study design and de-identification of the images. The ShenzhenEye subset contained 38,922 UWF images of 4003 patients collected from Shenzhen Eye Hospital of China between January 1st, 2019 and December 31st, 2023. The SUSTech subset contained 2835 UWF images of 226 patients collected from the Southern University of Science and Technology Hospital of China between January 1st, 2023 and June 31st, 2023. The LishuiR subset contained 938 UWF images of 155 patients collected from Lishui People's Hospital of China between January 1st, 2021 and December 31st, 2023. The Zhongshan subset contained 456 UWF images of 85 patients collected from Zhongshan Ophthalmic Center, Sun Yat-sen University of China. The LishuiZ subset contained 220 UWF images of 91 patients collected from Lishui Central Hospital of China between January 1st, 2021 and December 31st, 2023. Ultimately, we integrated these resources to establish the PSMM dataset that contained 43,371 UWF images of 4560 patients. Two UWF scanning laser ophthalmoscopy imaging devices captured these images, Daytona (P200T) and California (P200DTx). We retrieved these images by the keywords of ⟨HighMyopia, PathologicMyopia⟩. We were prone to partially retrieve severe samples from the hospital to form the Zhongshan subset as a challenging subset. Fewer samples were collected in the LishuiR and LishuiZ subsets due to certain limitations in the medical record management of the two hospitals, despite retrieving them over a long period. The ShenzhenEye subset naturally served as the main subset in proportion, and the other fours as auxiliary subsets. Two junior ophthalmologists labeled these UWF images, and one senior specialist then double-checked the labeled images by discarding distorted or damaged images for rigorous quality assurance. Specifically, each UWF image was independently annotated by two board-certified ophthalmologists with prior experience in retinal disease diagnosis. The annotations covered both the presence of posterior staphyloma and the identification of myopic maculopathy into five defined categories. To ensure annotation reliability and reduce subjectivity, a senior retinal specialist subsequently reviewed the annotations from both junior annotators. In cases of disagreement, the senior specialist made the final determination based on clinical judgment and established diagnostic criteria[16]. Certain retinal findings that may mimic or confound pathologic myopia diagnosis, such as laser scars, paving stone degeneration, white-without-pressure, and choroidal nevi, were retained in the dataset. These findings frequently occur in highly myopic eyes and reflect the complexities found in real-world clinical practice. These findings were not categorized as separate diagnostic features since they do not meet the established clinical definitions of pathologic myopia (e.g., META-PM[16]). During the annotation process, retinal specialists were specifically directed to focus on the hallmark lesions associated with pathologic myopia, such as diffuse atrophy, patchy atrophy, and myopic choroidal neovascularization. They were instructed to differentiate these incidental or benign findings from true pathological changes. We did not exclude images with various retinal findings, artifacts, comorbidities, or borderline cases to reflect more realistic conditions. We acknowledge that this approach may introduce some label noise and decrease specificity, while models designed for real-world applications should be able to handle these complexities effectively. To evaluate inter-rater reliability, Cohen's kappa coefficients[39] were calculated for each lesion label between the two junior annotators before senior review. The kappa values ranged from 0.78 to 0.86 across categories (Supplementary Table 9), indicating substantial agreement. The composition of the hand-crafted PSMM dataset and its integral subsets are presented in Supplementary Table 1.

We desensitized all the data to prevent privacy exposure. We centralized the objective (photographing area) by removing futile black outer boundaries and then resized images beforehand to facilitate model training. For ease of application and adaptation, we structured the dataset following the format of the PASCAL Visual Object Classes Challenge (PASCAL VOC) 2007 dataset[40], which is a well-known dataset in the computer vision field developed to recognize objects in realistic scenes. Due to a limited amount of data, an increasing number of published methods are trained on the training set and evaluated on the testing set directly to showcase optimal performance presentation regardless of fair comparison. However, in real-world scenarios, researchers need to develop reliable methods in various situations. This means it is crucial to evaluate these methods on a separate development set for convincing model validation. In order to support our claim, we divided the PSMM dataset into three separate parts: training, development, and testing sets with a distribution of 7:1.5:1.5. This allowed us to assess the research using the development set and then finalize the method and evaluate it using the unseen testing set. While dividing data into different sets is common in deep learning tasks, it becomes more complex when dealing with the clinical challenge presented in this study. Notably, each patient typically has multiple

UWF images, which can occur in two scenarios: multiple images are taken in a single examination to ensure an accurate diagnosis, or images are taken at different times during multiple examinations. To ensure reliable photography, several UWF images are captured at the same time for each patient, and many patients undergo examinations at different times. As a result, it's not feasible to split UWF images from the same patient into different sets during data partitioning. Furthermore, as mentioned earlier, our objective involves a multi-label learning task, which further complicates the data partitioning process. To address this, we adopted an approach where we assigned a single-class label for each patient and employed a stratified strategy to ensure independent and identically distributed partitioning[41]. Specifically, we assigned a pseudo single-class label that was quantitatively dominant over all labels of UWF images for each patient and then stratified the patient image groups into training, development, and testing sets.

End-to-end lightweight framework

We present details about the feature extraction backbone and optimized designs with cost-sensitive calibration and classifier adaptation for multi-label imbalance alleviation.

We harness TinyViT[20] as the fundamental backbone of the proposed RealMNet to ensure the model achieves excellent performance while remaining lightweight. TinyViT is favored for its application of distillation during pretraining for knowledge transfer. We employ a hierarchical design to address the need for multi-scale features in identifying morphologic patterns. This architecture comprises four stages, each featuring a gradual reduction in resolution akin to the Swin Transformer[24] and LeViT[42]. The patch embedding block incorporates two convolutions with a 3 × 3 kernel, a stride of 2, and a padding of 1. In the initial stage, we implement lightweight and efficient MBConvs[43] and downsampling blocks, recognizing that convolutions at earlier layers can proficiently learn low-level representations due to their strong inductive biases. The subsequent three stages are constructed with transformer blocks, leveraging window attention to mitigate computational costs. To capture local information, we introduce attention biases and a 3 × 3 depth-wise convolution between attention and MLP. Each block in the initial stage, as well as attention and MLP blocks, is complemented by a residual connection. High-performing neural network activation functions GELU[44] are used for smoothing model training. The normalization layers for convolution and linear operations are BatchNorm[45] and LayerNorm[46], respectively. The embedded dimensions in each stage of the adopted backbone are 96, 192, 384, and 576. Furthermore, the number of blocks in each stage of the backbone corresponds to that of Swin-T: 2, 2, 6, and 2. Depending on the input resolutions that the model could accept, three model variants are defined: RealMNet-Min for 224 × 224 inputs, RealMNet for 384 × 384 inputs, and RealMNet-Max for 512 × 512 inputs. We differentiate RealMNet from the vanilla TinyViT backbone and other general-purpose foundation models to highlight the superiority of the proposed method in handling multi-label imbalance while identifying morphologic patterns for diagnosing pathologic myopia. RealMNet encompasses strategic enhancements like cost-sensitive calibration and classifier adaptation and is fine-tuned with unique training processes tailored to address real-world challenges that are not fully captured by existing models.

Cost-sensitive methods are practical and efficient techniques that take into account the costs resulting from prediction mistakes made by the model. When dealing with the complication of lesions in terms of posterior staphyloma and myopic maculopathy, we aim to explore cost-sensitive approaches suitable for multi-label learning. We begin by using the BCE Loss, based on cross-entropy in information theory. In this context, cross-entropy of the distribution q relative to a distribution p over a given set is defined as follows:

$$\mathcal{H}(p,q) = -\mathbb{E}_p[\log q] \tag{1}$$

where $\mathbb{E}_p[\cdot]$ is the expected value operator regarding the distribution p. Cross-entropy can be utilized to create a loss function in machine learning and optimization:

$$\mathcal{H}(p,q) = -\sum_i p_i \log q_i = -\left[y \log \hat{y} + (1-y)\log(1-\hat{y})\right] \tag{2}$$

where y means the ground-truth and $\hat{y}$ means the predictions from the model. Next, we introduce a weight factor $\alpha \in [0,1]$ to help address class imbalance and a modulating factor $(1-p)^{\gamma}$ to reshape the loss function, thereby reducing the emphasis on easy examples and focusing training on challenging negatives[28]. Till now, we define the cost-sensitive calibration (CSC) as follows:

$$CSC = -\alpha\left[p^{\gamma}\log p + (1-p)^{\gamma}\log(1-p)\right] \tag{3}$$

where $p = \sigma(z)$ is the prediction probability given output logits z and γ is the focusing parameter. We also separate the focusing levels of positive and negative samples to avoid eliminating gradients from rare positive samples when setting a high value for γ. Additionally, we examine the effects of asymmetric probability shifting, achieved by setting a probability margin $m \geq 0$ to reject mislabeled negative samples[29]. Therefore, the ultimate CSC is defined as follows:

$$CSC = -\alpha\left[(p_m)^{\gamma_-}\log p + (1-p)^{\gamma_+}\log(1-p)\right] \tag{4}$$

where $p_m = \max(p-m, 0)$ is the shifted probability, γ_+ and γ_- are positive and negative focusing parameters, respectively. Furthermore, we evaluate the effectiveness of a state-of-the-art cost-sensitive method called Two-way Loss[30], specially designed for multi-label learning. We follow the original computational formula:

$$\ell = \text{softplus}\left[T_{\mathcal{N}}\log\sum_{n\in\mathcal{N}} e^{\frac{x_n}{T_{\mathcal{N}}}} + T_{\mathcal{P}}\log\sum_{p\in\mathcal{P}} e^{-\frac{x_p}{T_{\mathcal{P}}}}\right] \tag{5}$$

where $\text{softplus}(\cdot) = \log[1+\exp(\cdot)]$, $\mathcal{P}$ means positive labels, $\mathcal{N}$ means negative labels, $T_{\mathcal{N}}$ and $T_{\mathcal{N}}$ are two temperatures applied to negative and positive logits, respectively. We fine-tune temperature parameters through grid search for optimal performance.

Classifier adaptation is technically complex but helpful for addressing multi-label imbalance issues by adjusting the model's classifier design. The design of the implemented classifier is inspired by a simple and efficient module called class-specific residual attention[31] that achieves state-of-the-art results on multi-label recognition.

Given an input image $\mathcal{I}$ with the scale of $H \times W$, the backbone as a feature extractor $\mathcal{F}$ transforms the input image into a feature tensor $\boldsymbol{x} \in \mathbb{R}^{d\times h\times w}$ by $\boldsymbol{x} = \mathcal{F}(\mathcal{I};\theta)$, where θ represents parameters of the backbone. The feature tensor is decoupled as $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_P$, where $\boldsymbol{x}_p \in \mathbb{R}^d$ indicates the p-th feature tensor in positions $P = h \times w$. The class-specific attention scores are presented by $s_p^i = \frac{\exp(\mathcal{T}\boldsymbol{x}_p^{\top}\boldsymbol{c}_i)}{\sum_{l=1}^{P}\exp(\mathcal{T}\boldsymbol{x}_l^{\top}\boldsymbol{c}_i)}$, where Here, s_p^i can be regarded as the probability of i-th class appearing at the position p with $\sum_{p=1}^{P} s_p^i = 1$ and $\mathcal{T}$ stands for the temperature controlling the sharpness of the scores. The class-specific feature vector for i-th class is $\boldsymbol{v}_{spec}^i = \sum_{p=1}^{P} s_p^i \boldsymbol{x}_p$. The class-agnostic feature vector for the entire image is $\boldsymbol{v}_{agno} = \frac{1}{P}\sum_p^P \boldsymbol{x}_p$. The final feature vector for the i-th class is $\boldsymbol{v}^i = \boldsymbol{v}_{agno} + \lambda\boldsymbol{v}_{spec}^i$. The classifier produces $\hat{\boldsymbol{y}} \triangleq (y^1, y^2, \cdots, y^n) = (\boldsymbol{c}_1^{\top}\boldsymbol{v}^1, \boldsymbol{c}_2^{\top}\boldsymbol{v}^2, \cdots, \boldsymbol{c}_n^{\top}\boldsymbol{v}^n)$, where n stands for the number of classes. The final prediction is produced with multi-head extension to the residual attention by $\hat{\boldsymbol{y}} = \sum_{h=1}^{H}\hat{\boldsymbol{y}}_{\mathcal{T}_h}$, where $\hat{\boldsymbol{y}}_{\mathcal{T}_h} \in \mathbb{R}^n$ represents the logits of head h.

Experimental protocols

We introduced three distinct experiment protocols that naturally empowered both the internal and external validation of the model, quantitatively demonstrating that the proposed model was efficient with good robustness and generalizability.

The centralized inference protocol aimed to demonstrate the inference capacity of models directly on the intact PSMM dataset. Models were trained on the training set of the PSMM dataset and tested on the testing set of the PSMM dataset. Models learned task-specific knowledge from all available training resources and were developed on the development set of the PSMM dataset, eventually inferring all available unseen testing resources. In our experiments, we compared our method, RealMNet, with four widely recognized models under the centralized inference protocol, in which models were sufficiently motivated for optimal identification performance.

The main-source robustness protocol aimed to demonstrate the robustness of models on the separate PSMM dataset. Models were trained solely on the main subset and tested on four auxiliary subsets, the averaged performances of which were provided. All data from the main-source dataset comprised the training set, and each auxiliary-center dataset served as the testing set separately. In our experiments, we implemented our method, RealMNet, under the main-source robustness protocol for robustness verification.

The cyclic-source generalizability protocol aimed to demonstrate the generalizability of models on the separate PSMM dataset. Models were trained on the main-source dataset combined with three auxiliary-center datasets and tested on the rest of the auxiliary dataset. The performances of four cyclic experiments were provided. In our experiments, we implemented our method, RealMNet, under the cyclic-source generalizability protocol for generalization verification.

Evaluation metrics

Cutting-edge artificial intelligence models frequently excel based on a single or a few evaluation metrics. However, this can introduce bias into the results and impact the perception of their scientific objectivity[47]. This issue is particularly relevant in multi-label learning, which is more intricate than single- and multi-class learning[48]. In our study, we opted for comprehensive measures to assess both bipartitions and rankings, considering the characteristics of multi-label data[49].

Considering a development set that has multi-label samples $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ where $i = 1, \ldots, N$ and N means the number of samples. The labelset of i-th sample $\boldsymbol{y}_i \subseteq \mathcal{L}$ where $\mathcal{L} = \{\lambda_l : j = 1, \ldots, L\}$ is the set of all ground-truth labels and L means the number of labels. For each label λ, the rank is termed as $r_i(\lambda)$. The predictions made by the Multi-Label Classifier are defined as $\hat{\boldsymbol{y}}_i$. Let tp_λ, fp_λ, tn_λ, and fn_λ be the number of true positives, false positives, true negatives, and false negatives after binary evaluation for a label λ.

For the evaluation of bipartitions, we use Precision $= \frac{tp}{tp+tp}$ to reflect the ability not to label as positive a sample that is negative. We use Recall $= \frac{tp}{tp+tp}$ (also called Sensitivity) to reflect the ability to find all positive samples. A good discrimination model should be sensitive in identifying as many potential positive samples as possible to help screen in medical scenarios. The F-measure is the harmonic mean of the Precision and Recall that symmetrically represents Precision and Recall in one metric. We use F1 Score $= \frac{2\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$ to reveal the balanced ability of the model to both capture positive cases (Recall) and be accurate with the cases it does capture (Precision), which is exceptionally able to measure performance objectively when the class balance is skewed. We use mean Average Precision (mAP) to reflect the average fraction of relevant labels ranked higher than one other relevant label, which is calculated by:

$$\text{mAP} = \frac{1}{L}\sum_{\lambda=1}^{L}\sum_{n}\left(R_n - R_{n-1}\right)P_n \tag{6}$$

where R_n and P_n stand for Precision and Recall at the n-th threshold, respectively. The AUROC (Area Under the Receiver Operating Characteristic Curve) indicates the level of separability of a model. This metric is calculated as the area under the Receiver Operating Characteristic Curve (ROC). A larger AUROC indicates that the model can achieve a high true positive rate while maintaining a low false positive rate. Essentially, it demonstrates the model's ability to differentiate between classes. The measures above can be calculated using two types of averaging operations: macro-averaging and micro-averaging. Specifically, given a bipartition-based measure $\mathcal{B}$,

$$\mathcal{B}_{\text{macro}} = \frac{1}{L}\sum_{\lambda=1}^{L}\mathcal{B}\left(tp_\lambda, fp_\lambda, tn_\lambda, fn_\lambda\right) \tag{7}$$

$$\mathcal{B}_{\text{micro}} = \mathcal{B}\left(\sum_{\lambda=1}^{L} tp_\lambda, \sum_{\lambda=1}^{L} fp_\lambda, \sum_{\lambda=1}^{L} tn_\lambda, \sum_{\lambda=1}^{L} fn_\lambda\right) \tag{8}$$

We use Hamming Loss to measure the proportion of incorrectly classified instance-label pairs, which is defined as follows:

$$\text{Hamming Loss} = \frac{1}{NL}\sum_{i=1}^{N}\left|\boldsymbol{y}_i \neq \hat{\boldsymbol{y}}_i\right| \tag{9}$$

For the evaluation of rankings, we use Coverage to assess the average number of steps required to encompass all relevant labels in the ranked label list for each example, which is defined as follows:

$$\text{Coverage} = \frac{1}{N}\sum_{i=1}^{N}\max_{\lambda \in \boldsymbol{y}_i} r_i(\lambda) - 1 \tag{10}$$

We use Ranking Loss to evaluate the fraction of reversely ordered label pairs, which is defined as follows:

$$\text{Ranking Loss} = \frac{1}{N|\boldsymbol{y}_i||\overline{\boldsymbol{y}_i}|}\sum_{i=1}^{N}\left|\left\{\left(\lambda_a, \lambda_b\right) : r_i\left(\lambda_a\right) > r_i\left(\lambda_b\right), \left(\lambda_a, \lambda_b\right) \in \boldsymbol{y}_i \times \overline{\boldsymbol{y}_i}\right\}\right| h \tag{11}$$

where $\overline{\boldsymbol{y}_i}$ is the complementary set of $\boldsymbol{y}_i$ with respect to $\mathcal{L}$.

Implementation details

Our model remained lightweight due to pretraining distillation techniques and utilized hierarchical transformer architectures that incorporated convolution operations. Therefore, we chose various widely used benchmark counterparts: DeiT[21], ConvNeXt[22], EfficientNet[23], and Swin Transformer[24]. Specifically, DeiT is a convolution-free transformer trained with a distillation procedure. ConvNeXt is a pure ConvNet that is modernized toward the design of a vision transformer. EfficientNet is a ConvNet designed using neural architecture search to enable model scaling with significantly fewer parameters. Swin Transformer is a hierarchical transformer that can be modeled at various scales. Foundation models designed for multiple purposes can perform tasks even if they haven't been explicitly pretrained for them. They can adapt to various clinical applications and demonstrate generalizability. In this context, we chose two recent foundation models trained using different types of images: DINOv2[25], which was trained with natural images, and VisionFM[26], which specialized in ophthalmic images. We compared RealMNet to these benchmark approaches concerning model development in Table 4.

We approached the problem in this study as a multi-label learning task to account for the complex relationships between morphologic patterns and explore their underlying interdependencies. We chose TinyViT-21m as the feature extractor backbone of RealMNet and initialized it with weights pretrained on ImageNet-21k using pretraining distillation. The image size was set at 384 × 384 for model development and 512 × 512 for optimal performance. The model was optimized using Adam with decoupled weight decay with an initial learning rate of 1e-4 and a weight decay of 0.05, trained with a batch size of 16 per graphics processing unit. We implemented warmup for 10% of the total 50 epochs, with a starting factor of 1e-2, followed by a cosine annealing schedule with a learning rate of 1e-6. A drop path rate of 0.5 was used to prevent over-fitting. We employed two types of augmentation techniques: simulated and batch-wise. Simulated augmentation was intended to mirror real-world scenarios by means of spatial-level and pixel-level transformation. For spatial-level transformation, we used a random affine, random flip, and random erasing. For pixel-level transformation, we used a Gaussian blur, Gauss noise, and Color jitter. The batch-wise transformation involved Mixup[50] and CutMix[51]. For simplicity, we used the same parameter settings as in the previous study[41] for UWF images. We leveraged asymmetric focusing as a cost-sensitive calibration with configurable parameters ($\gamma_+ = 3$ and $\gamma_- = 4$). We harnessed classifier adaptation with the leveraging parameter $\lambda = 1.2$ and $H = 2$ multi-head attention. In the centralized inference protocol, the entire PSMM dataset is divided into a training set, a development set, and a test set at a ratio of 7:1.5:1.5 using stratified partitioning. In the main-source robustness protocol, the ShenzhenEye subset is utilized as the training set, while the remaining four source subsets take turns as the test set. In the cyclic-source generalizability protocol, the ShenzhenEye subset and three of the remaining four sources are used as the training set, and testing is conducted on the subset of the last source. In all experimental protocols, the ML-RUS[52] resampling method was applied to the training set only, with an undersampling ratio of 0.2. We trained benchmark approaches using a consistent setup on a unified platform. The weights from the teacher network of VisionFM were utilized as the encoder for VisionFM. VisionFM maintained a three-layer MLPs as the decoder for optimal performance. DINOv2 initially set the LayerScale value to 1e-5. We utilized transfer learning by initializing the backbone with weights that were trained on the PSMM dataset, followed by fine-tuning the model with data specific to peripheral retinal lesions. Experiments were deterministic and reproducible, with a fixed seed of 42. We conducted the training and testing on the OpenMMLab platform using 4 NVIDIA GeForce RTX 4090 GPUs.

UWF imaging investigation

We investigated the potential negative impact of physical device boundaries in images captured by UWF imaging. We demonstrated the advantages of the UWF modality by comparing them with fake CFP images.

Modern ultra-widefield imaging inevitably captures the boundaries of the physical devices along with the imaging targets, which can obscure essential information. To determine if these boundaries negatively impact the model's inference capability, we segmented out these boundaries and retrained our model using data without them. We found that nearly three-quarters of the images in the PSMM dataset contain significant black borders, and the remaining images, although lacking black borders, still exhibit considerable interference from the device boundaries.

To create a segmentation dataset, we randomly sampled 1% of the data from the two imaging types, selecting at the patient level to avoid information leakage that could arise from stratified partitioning. We enlisted the expertise of professional physicians to annotate the dataset at the pixel level. The resulting segmentation dataset consisted of 412 images, comprising 303 images with black borders and 109 images without them. We divided this dataset into training, development, and testing sets in an 8:1:1 ratio.

For segmentation, we employed ResNet-50[53] as the backbone and used DeepLab-v3[54] as the segmentation model, utilizing weights pretrained on the PASCAL dataset. We utilized the SGD optimizer with a batch size of 4, a learning rate of 0.01, and a momentum of 0.9 for 2000 epochs, implementing early stopping. After fine-tuning, we applied the model to segment the boundaries of the physical devices and then retrained RealMNet with these segmented images.

In comparison to conventional CFP, UWF imaging provides a broader field of view and captures more detailed essential information. To investigate the benefits of using UWF for diagnosing pathologic myopia by detecting fundus lesions, we utilized existing UWF data to create fake CFP images. Specifically, we generated fake CFP images by center-cropping original UWF images with a crop ratio of 2.5, which is in line with the ratio of the retinal field of view between UWF and CFP images. We used the same training strategies as in the original UWF data to ensure a fair comparison.

Data availability

The datasets that support the findings of this study are publicly accessible at https://github.com/yo3nglau/RealMNet.

Code availability

The codes and trained models developed for this study are publicly available at https://github.com/yo3nglau/RealMNet. Labeling is completed with the open-source graphical image annotation software labelme v5.3.1 (https://github.com/labelmeai/labelme). The codes are built on three open-source projects, MMCV v2.0.1, MMEngine v0.8.4, and MMPretrain v1.0.2 of the OpenMMLab codebase (https://github.com/open-mmlab). The evaluation metrics are implemented with scikit-learn v1.3.2 (https://github.com/scikit-learn/scikit-learn). All experiments are supervised on Weights & Biases v0.16.4 (https://github.com/wandb/wandb). All histograms and line charts are drawn by Matplotlib v3.9.1 (https://github.com/matplotlib/matplotlib). The Sankey map and Chord diagram are drawn by Plotly v5.22.0 (https://github.com/plotly/plotly.py) and pyCirclize v1.6.0 (https://github.com/moshi4/pyCirclize), respectively.

Received: 9 November 2024; Accepted: 29 June 2025;
Published online: 13 July 2025

References

1. Baird, P. N. et al. Myopia. *Nat. Rev. Dis. Prim.* **6**, 99 (2020).
2. Dolgin, E. A myopia epidemic is sweeping the globe. here's how to stop it. *Nature* **629**, 989–991 (2024).
3. Morgan, I. G., Ohno-Matsui, K. & Saw, S.-M. Myopia. *Lancet* **379**, 1739–1748 (2012).
4. Choudhry, N., Golding, J., Manry, M. W. & Rao, R. C. Ultra-widefield steering-based spectral-domain optical coherence tomography imaging of the retinal periphery. *Ophthalmology* **123**, 1368–1374 (2016).
5. Burlina, P. M. et al. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol.* **135**, 1170–1176 (2017).
6. Peng, B. et al. Practical guidelines for cell segmentation models under optical aberrations in microscopy. *Computational Struct. Biotechnol. J.* **26**, 23–39 (2024).
7. Andrianov, A. M., Shuldau, M. A., Furs, K. V., Yushkevich, A. M. & Tuzikov, A. V. Ai-driven de novo design and molecular modeling for discovery of small-molecule compounds as potential drug candidates targeting sars-cov-2 main protease. *Int. J. Mol. Sci.* **24**, 8083 (2023).
8. meysa Duygun, R. et al. Classification of patients with dementia with lewy bodies by event-related oscillations. In *Alzheimer's Association International Conference* (ALZ, 2024).
9. Sahin, Z., Ozkan Vardar, D., Erdogmus, E., Calamak, S. & Koçer Gumusel, B. Monomer release, cytotoxicity, and surface roughness of temporary fixed prosthetic materials produced by digital and conventional methods. *Odontology* 1–16 (2025).
10. Dai, L. et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. *Nat. Commun.* **12**, 3242 (2021).
11. Bora, A. et al. Predicting the risk of developing diabetic retinopathy using deep learning. *Lancet Digital Health* **3**, e10–e19 (2021).
12. Yim, J. et al. Predicting conversion to wet age-related macular degeneration using deep learning. *Nat. Med.* **26**, 892–899 (2020).
13. Romo-Bucheli, D., Erfurth, U. S. & Bogunović, H. End-to-end deep learning model for predicting treatment requirements in neovascular amd from longitudinal retinal oct imaging. *IEEE J. Biomed. Health Inform.* **24**, 3456–3465 (2020).
14. Qi, Z. et al. A deep learning system for myopia onset prediction and intervention effectiveness evaluation in children. *npj Digital Med.* **7**, 206 (2024).

15. Li, Y. et al. Development and validation of a deep learning system to screen vision-threatening conditions in high myopia using optical coherence tomography images. *Br. J. Ophthalmol.* **106**, 633–639 (2022).
16. Ohno-Matsui, K. et al. International photographic classification and grading system for myopic maculopathy. *Am. J. Ophthalmol.* **159**, 877–883 (2015).
17. Ohno-Matsui, K., Lai, T. Y., Lai, C.-C. & Cheung, C. M. G. Updates of pathologic myopia. *Prog. retinal eye Res.* **52**, 156–187 (2016).
18. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
19. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
20. Wu, K. et al. Tinyvit: Fast pretraining distillation for small vision transformers. In *European Conference on Computer Vision*, 68–85 (Springer, 2022).
21. Touvron, H. et al. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357 (PMLR, 2021).
22. Liu, Z. et al. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986 (2022).
23. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114 (PMLR, 2019).
24. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022 (2021).
25. Oquab, M. et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal* 1–31 (2024).
26. Qiu, J. et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI* **1**, Aloa2300221 (2024).
27. De Boer, P.-T., Kroese, D. P., Mannor, S. & Rubinstein, R. Y. A tutorial on the cross-entropy method. *Ann. Oper. Res.* **134**, 19–67 (2005).
28. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988 (2017).
29. Ridnik, T. et al. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 82–91 (2021).
30. Kobayashi, T. Two-way multi-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7476–7485 (2023).
31. Zhu, K. & Wu, J. Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 184–193 (2021).
32. Marcondes, D., Simonis, A. & Barrera, J. Back to basics to open the black box. *Nat. Mach. Intell.* **6**, 498–501 (2024).
33. Larsson, G., Maire, M. & Shakhnarovich, G. Fractalnet: Ultra-deep neural networks without residuals. In *International Conference on Learning Representations* (2022).
34. Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-cam$++$: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–847 (IEEE, 2018).
35. Dai, S., Chen, L., Lei, T., Zhou, C. & Wen, Y. Automatic detection of pathological myopia and high myopia on fundus images. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6 (IEEE, 2020).
36. Babenko, B. et al. Detection of signs of disease in external photographs of the eyes via deep learning. *Nat. Biomed. Eng.* **6**, 1370–1383 (2022).
37. Ruiz-Medrano, J. et al. Myopic maculopathy: current status and proposal for a new classification and grading system (atn). *Prog. retinal eye Res.* **69**, 80–115 (2019).
38. Abnar, S. & Zuidema, W. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4190–4197 (2020).
39. Cohen, J. A coefficient of agreement for nominal scales. *Educ. psychological Meas.* **20**, 37–46 (1960).
40. Everingham, M., Van Gool, L., Williams, C. K., Winn, J. & Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Computer Vis.* **88**, 303–338 (2010).
41. Engelmann, J. et al. Detecting multiple retinal diseases in ultra-widefield fundus imaging and data-driven identification of informative regions with deep learning. *Nat. Mach. Intell.* **4**, 1143–1154 (2022).
42. Graham, B. et al. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12259–12269 (2021).
43. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520 (2018).
44. Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).
45. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 448–456 (2015).
46. Ba, J. L. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
47. Roberts, M., Hazan, A., Dittmer, S., Rudd, J. H. & Schönlieb, C.-B. The curious case of the test set auroc. *Nat. Mach. Intell.* **6**, 373–376 (2024).
48. Wu, X.-Z. & Zhou, Z.-H. A unified view of multi-label performance measures. In *international conference on machine learning*, 3780–3788 (PMLR, 2017).
49. Tsoumakas, G., Katakis, I. & Vlahavas, I. Mining multi-label data. *Data mining and knowledge discovery handbook* 667–685 (2010).
50. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations* (2018).
51. Yun, S. et al. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6023–6032 (2019).
52. Charte, F., Rivera, A. J., del Jesus, M. J. & Herrera, F. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing* **163**, 3–16 (2015).
53. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
54. Chen, L.-C. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

Acknowledgements

We thank the support from Shenzhen Medical Research Funds, D2301002; the National Natural Science Foundation of China 32350410397; Science, Technology, Innovation Commission of Shenzhen Municipality, JCYJ20240813112016022, JCYJ20220530143014032, JCYJ20230807113017035, KCXFZ20211020163813019, Tsinghua Shenzhen International Graduate School Cross-disciplinary Research and Innovation Fund Research Plan, JC2022009; and Bureau of Planning, Land and Resources of Shenzhen Municipality (2022) 207. We thank Cuiyi Peng and Ziyan Chen for helping organize the initial data. We thank Xiaoyun Zhong for providing valuable insights and suggestions.

Author contributions

Y.L., K.Z., J.W., P.Q. and J.J. conceptualized and designed the study. Y.L., K.Z., L.L., Z.Z., Z.D. and Z.Q. participated in the experimental design. Y.L.

wrote the code for model development, experimental analyses, and visualizations. Y.L., K.Z., C.Yang, Y.Zhu, D.W., S.W., X.H., C.Yan, Y.Zhuo, C.Q., J.C., Z.H., C.L., M.C. and D.Y. were involved in data curation. Y.L., K.Z., L.L., Z.Z., Z.Q. and S.D. conducted data investigation and formal analysis. J.W., P.Q. and J.J. supervised the study. Y.L. and K.Z. drafted the initial manuscript. C.J. and L.L. edited the initial manuscript. All reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01849-y.

Correspondence and requests for materials should be addressed to Jiantao Wang, Peiwu Qin or Jiansong Ji.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.