Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Diagnosing Pathologic Myopia by Identifying Posterior Staphyloma and Myopic Maculopathy Using Ultra-Widefield Images with Deep Learning

Peiwu Qin

pwqin@sz.tsinghua.edu.cn

Tsinghua University

Yang Liu

Tsinghua University

Keming Zhao

Shenzhen Eye Hospital, Jinan University, Shenzhen Eye Institute

Lihui Luo

Institute of Biopharmaceutics and Health Engineering, Tsinghua Shenzhen International Graduate School

Ziheng Zhang

Institute of Biopharmaceutics and Health Engineering, Tsinghua Shenzhen International Graduate School

Zhenghang Qian

Department of Automation, Tsinghua University

Cenk Jiang

Institute of Biopharmaceutics and Health Engineering, Tsinghua Shenzhen International Graduate School

Zhicheng Du

Institute of Biopharmaceutics and Health Engineering, Tsinghua Shenzhen International Graduate School

Simin Deng

Shenzhen Eye Hospital, Jinan University, Shenzhen Eye Institute

Chengming Yang

Southern University of Science and Technology Hospital

Duanpo Wu

Hangzhou Dianzi University

Shuai Wang

Hangzhou Dianzi University

Xingru Huang

Hangzhou Dianzi University

Chenggang Yan

Hangzhou Dianzi University

Yingting Zhu

State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology Visual Science, Guangdong Provincial Clinical Researc

Yehong Zhuo

State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology Visual Science, Guangdong Provincial Clinical Researc

Chunsheng Qu

Clinical Laboratory of Lishui People's Hospital, First Affiliated Hospital of Lishui College, Wenzhou Medical College Lishui Hospital

Jiaqi Chen

Clinical Laboratory of Lishui People's Hospital, First Affiliated Hospital of Lishui College, Wenzhou Medical College Lishui Hospital

Zhenqiang Huang

Clinical Laboratory of Lishui People's Hospital, First Affiliated Hospital of Lishui College, Wenzhou Medical College Lishui Hospital

Chenying Lu

Zhejiang Key Laboratory of Imaging and Interventional Medicine, Department of Radiology, Lishui Central Hospital, The Fifth Affiliated Hospital of Wenzhou Medical University

Minjiang Chen

Zhejiang Key Laboratory of Imaging and Interventional Medicine, Department of Radiology, Lishui Central Hospital, The Fifth Affiliated Hospital of Wenzhou Medical University

Dongmei Yu

School of Mechanical, Electrical \& Information Engineering, Shandong University

Jiantao Wang

Shenzhen Eye Hospital, Jinan University, Shenzhen Eye Institute

Jiansong Ji

Zhejiang Key Laboratory of Imaging and Interventional Medicine, Department of Radiology, Lishui Central Hospital, The Fifth Affiliated Hospital of Wenzhou Medical University

---

Article

Keywords:

Posted Date: November 29th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5421907/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at npj Digital Medicine on July 13th, 2025.
See the published version at https://doi.org/10.1038/s41746-025-01849-y.

Diagnosing Pathologic Myopia by Identifying Posterior Staphyloma and Myopic Maculopathy Using Ultra-Widefield Images with Deep Learning

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

[1]Institute of Biopharmaceutics and Health Engineering, Tsinghua Shenzhen International Graduate School, Shenzhen, 518055, China.
[2]Shenzhen Eye Hospital, Jinan University, Shenzhen Eye Institute, Shenzhen, 518055, China.
[3]Center of Precision Medicine and Healthcare, Tsinghua-Berkeley Shenzhen Institute, Shenzhen, 518055, China.
[4]Department of Automation, Tsinghua University, Beijing, 100084, China.
[5]Southern University of Science and Technology Hospital, Shenzhen, 518055, China.
[6]Hangzhou Dianzi University, Hangzhou, Zhejiang, 310018, China.
[7]School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai, 264209, China.
[8]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, 510060, China.
[9]Clinical Laboratory of Lishui People's Hospital, First Affiliated Hospital of Lishui College, Wenzhou Medical College Lishui Hospital, Lishui, Zhejiang, 323000, China.
[10]Zhejiang Key Laboratory of Imaging and Interventional Medicine, Department of Radiology, Lishui Central Hospital, The Fifth Affiliated Hospital of Wenzhou Medical University, Lishui, Zhejiang, 323000, China.

*Corresponding author(s). E-mail(s): wangjiantao65@126.com; pwqin@sz.tsinghua.edu.cn; jjstcty@wmu.edu.cn;
Contributing authors: lyang22@mails.tsinghua.edu.cn; 542807062@qq.com;
†These authors contributed equally to this work.

Abstract

Pathologic myopia (PM) has long been a leading cause of visual impairment and blindness. While numerous deep learning-based approaches have improved the efficiency and accuracy of recognizing PM, few have thoroughly investigated clinically significant pathological patterns due to the scarcity of datasets with lesion-wise labeling, particularly those comprising ultra-widefield (UWF) images that encompass a broader retinal field of view. In this study, we gather a large-scale multi-source ultra-widefield imaging myopia dataset, PSMM, labeled with posterior staphyloma (PS) and myopic maculopathy (MM). Compared with traditional colored fundus photography, UWF images exhibit informative characteristics concerning peripheral lesions caused by axial elongation and structural deformation in eyes with pathologic myopia. The labels obtained from the dataset can substantially assist in the progression diagnosis of pathologic myopia and guide prognosis. We introduce an end-to-end lightweight framework called RealMNet, which precisely identifies these challenging pathological patterns underpinned by a well-curated dataset. RealMNet is more adaptable to medical devices with only 21 million parameters compared to existing approaches. Through extensive experiments on a unified platform using all-around metrics regarding bipartitions and rankings across three experimental protocols, we demonstrate the robustness and generalizability of RealMNet, showcasing promising merit in clinical applications.

1 Introduction

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 (SE) $\leq$ -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, which hinders the development of diagnostic and support systems needed to help clinicians interpret these advanced imaging modalities.

Recent advancements in deep learning (DL) have made it possible to automatically process medical images for various tasks, achieving performance comparable to human experts. In the case of retinal diseases, DL models not only accurately diagnose

and monitor conditions such as diabetic retinopathy and age-related macular degeneration from retinal images [5–7], but also assist in developing personalized treatment plans. In addition, DL has been applied to myopia-related screening, assessing the risk of myopia progression by analyzing retinal images and enabling early intervention. Although these methods are robust, there is a need for more investigation into sophisticated pathological patterns. A system called the Meta-Analysis of Pathologic Myopia (META-PM) [8] 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 [9]. Posterior staphyloma manifests as an outpouching of the ocular wall, with a curvature radius less than that of the surrounding sclera. PS often leads to changes in the retina, choroid, and nerve fiber layer, subsequently affecting the patient's vision. Early identification of PS 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. MM is one of the primary causes of vision deterioration in patients with high myopia because the macula is the area of the retina with the highest visual acuity, and any damage there can significantly impact vision quality. Early diagnosis and management of these conditions can help slow or prevent disease progression and reduce the risk of vision loss. Existing research has some limitations despite advancements. Firstly, more attention should be given to myopic maculopathy and posterior staphyloma. This is due to the difficulty in identifying their complete contour on CFP accurately, hence identifying these lesions requires high-quality ultra-widefield (UWF) imaging data. 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, complex operations, and data acquisition expenses make large-scale UWF data collection challenging for many studies. Secondly, previous studies often use balanced data, ignoring the significant data imbalance in real-world scenarios [10]. Retinal lesions in pathologic myopia are highly heterogeneous and often coexist with other types of retinal lesions, creating imbalanced data and making it more challenging to distinguish PS from MM accurately. Thirdly, detecting PS and MM 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 [11]. While these models are powerful, due to their size and complexity, they need to be more adaptable for use in medical devices, especially in resource-constrained clinical environments. Therefore, the creation of the PSMM dataset fills these gaps, providing a high-quality data source that supports the precise identification of multiple lesions and clinical research, thereby improving patient diagnosis and treatment outcomes.

Previous studies employing DL models for myopia detection often rely on CFP, assessing only a narrow range of the posterior pole of the retina [12]. However, with

the elongation of the eyeball in highly myopic eyes, the likelihood of peripheral retinal lesions increases significantly, necessitating the use of UWF imaging for comprehensive evaluation [13]. With this in mind, we suggest adopting a recognition system for 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). When combined with UWF imaging, this system allows for a thorough assessment of retinal health in myopic patients. For instance, peripheral lattice degeneration, seen as a white lattice pattern on UWF images due to retinal microvascular occlusion, may develop into various-sized circular atrophic holes over time. These changes are closely associated with rhegmatogenous retinal detachment, potentially giving rise to severe visual impairment [14]. Evaluating peripheral retinal lesions significantly enhances our ability to comprehensively monitor and treat myopic retinal changes by enabling the earlier detection and management of such sight-threatening complications.

In this work, we present a detailed and efficient workflow (Fig. 1) for identifying challenging lesions. We compile a dataset containing UWF images of pathologic myopia with clinically significant lesions from multiple medical sources. Experienced physicians 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 pathological patterns by developing an end-to-end framework called RealMNet that embraces **Real**-world **M**yopia diagnosis. Thanks to the adoption of a compact and efficient vision transformer [15] as our backbone, the framework is lightweight enough to be applied to modern 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 comparison models: DeiT [16], EfficientNet [17], ConvNeXt [18], and Swin Transformer [19]. 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. Furthermore, we investigate the potential negative impact of the physical device boundaries present in images captured by modern ultra-widefield imaging, which may impede peripheral information. The boundaries have been segmented out to make sure that the learning process of the model is not compromised. To visually interpret the model's decision-making for inference, we utilize an improved version of gradient-weighted class activation mapping called Grad-CAM++ that better localizes objects and explains occurrences of multiple objects of a class in a single image [20]. The model performance is reported with all-around measures (details are listed in

Evaluation metrics/Supplemental materials) for evaluating both bipartitions and rankings concerning the ground truth of multi-label data. All P values are calculated with a two-sided t-test between RealMNet and the other comparison model to check for significance.

2 Results

2.1 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 (PS) and myopic maculopathy (MM) that could assist clinicians in diagnosing pathologic myopia. The PSMM dataset comprised 43,371 ultra-widefield images of 4,560 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 (Fig. A1a). Experienced clinicians 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 macular atrophy (MA). An intuitive illustration of these pathological patterns can be found in (Fig. A1b). Notably, posterior staphyloma and myopic maculopathy may appear simultaneously, forming multi-label datasets (MLDs). The PSMM dataset exhibits an imbalanced distribution (Fig. 2), as exposed in other retinal diseases, 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.

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

The imbalance present in multi-label datasets (MLDs) significantly impacts the model's performance, leading to biased learning and inadequate knowledge acquisition. This study presented three techniques to tackle the imbalance issue: resampling methods, classifier adaptation, and cost-sensitive calibration. Cost-sensitive calibration addressed the multi-label imbalance by developing the loss function from Binary Cross-Entropy (BCE) Loss [21], considering that multi-label learning involves decomposing the multi-label task into multiple binary tasks, each focusing on distinguishing samples within a target class category. We gradually introduced tunable parameters for BCE Loss to alleviate the imbalance issue on the PSMM dataset. Initially, We attempted to train the model using Binary Cross-Entropy (BCE) Loss, and to address

class imbalance, we implemented a commonly used weighting factor $\alpha \in [0, 1]$ to form an α-balanced BCE Loss. In our experiments, we discovered that the model performed better when using an α value of 0.75 (Table A5), 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 [22]. 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 (Table A6), 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 [23]. The experimentally determined cost-sensitive calibration helps the model learn from balanced samples (Fig. A4a), 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 [24], 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 (Table A7) 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 [25]. We introduced a tunable 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 (ViT) 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 mean Average Precision (mAP) compared to other settings while maintaining similar performance on other evaluation metrics A4.

2.3 Multi-protocol experiments demonstrate valued inference with robustness and generalizability

We devised three distinct experimental protocols 1c to excavate the model's inference capacity, robustness, and generalizability (see detailed strategies in 'Experimental protocols'). The results (Fig. 3a) under the centralized inference protocol revealed that RealMNet outperformed ($P < 0.001$) all other benchmark approaches on F1 Score with 0.7903 (95% CI 0.7531-0.8275), mAP with 0.8398 (95% CI 0.7923-0.8873), and AUROC with 0.9736 (95% CI 0.9682-0.9791). Unless otherwise noted, these three metrics were considered the primary criteria for measuring the model's performance. We additionally presented other evaluation metrics for complementary analysis ('Evaluation metrics' in Methods). RealMNet achieved the lowest Coverage of 2.2586 (95% CI 2.2204-2.2968), 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. 3b). 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 impeded 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. 3c), reflecting its stable exertion 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 (posterior staphyloma and myopic maculopathy).

2.4 Interpretable 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 [26]. 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 [11]. RealMNet achieves precise identification even with only half of the training resources (Fig. 4a), demonstrating its capability to capture clinically significant pathological 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-224 and the RealMNet-384 performed similarly as more training data was used, RealMNet-512 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. 4b). Freezing the first one or two layers of the model did not decrease performance, indicating that the model retained low-level general features from pretraining 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 pathological 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. 4c) 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 Aug vs. Aug). The simulated augmentation was used to mirror real-world situations. The model's performance decreased significantly when the simulated augmentation was removed (w/o SA vs. Aug). 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 involved enhancing synthetic samples by interweaving two samples. Removing batch-wise augmentation did not cause a significant loss (w/o BA vs. Aug), indicating that the model had inherently been adequate to build intra- and inter-affinities between pathological patterns. A slight decrease in Ranking Loss and Coverage suggested that batch-wise augmentation helped the model learn more accurate label distributions. Drop path [27] 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 (Fig. A5c). 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 modern ultra-widefield imaging, which may occlude partial information. The comparative experimental results (Fig. A6b) showed that RealMNet was not affected by these barriers, demonstrating its outstanding focusing ability. 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 (Grad-CAM++) [20] that localized objects 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. 5). These heatmaps revealed irregular attentive regions corresponding to diffused pathological patterns embodied in different lesion levels, manifesting the explainable learning of the proposed model.

3 Discussion

In this study, we introduced a novel perspective for assisting in diagnosing pathological 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 [28, 29]. However, the majority of these 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 [30]. Nonetheless, to our knowledge, limited research has thoroughly examined these lesions, and there are no publicly available datasets for this purpose. To tackle this, we gathered a large-scale dataset comprising ultra-widefield images from five distinct hospital sources (Fig. 1a). We sought experienced clinicians to label 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 macular atrophy (MA). We built an end-to-end lightweight framework called RealMNet on the basis of the unified platform to identify these concurrent lesions with multi-label learning (Fig. 1b). We progressively determined resampling approaches (Fig. A5a), cost-sensitive calibration (Fig. A4a), and classifier adaption (Fig. A4b) with the development set for mitigating negative impacts caused by imbalanced label distributions (Fig. 2). Hence, the proposed model was functionally interpretable by identifying these clinically significant lesions and objectively instrumental by alleviating multi-label imbalance issues. 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 other benchmark approaches (Fig. 3a). Meanwhile, our model exhibited good robustness (Fig. 3b) and generalizability (Fig. 3c), even when assessed on challenging subsets. For deep learning-based applications in the medical field, interpretability is critical when developing convincing workflows. Our model exhibited good labeling efficiency, taking different ratios of training data as input (Fig.). As a transformer-based architecture with hierarchical design [19], each stage of RealMNet maintained helpful knowledge for lesion identification (Fig. 4b). The simulated and batch-wise augmentation jointly helped the model avoid over-fitting (Fig. 4c). 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 impede models from effectively capturing helpful information. We constructed the dataset based on the scale of the two imaging types in the PSMM dataset (Table A3). We employed ResNet-50 as the segmentation backbone and DeepLab-v3 as the segmentation model to remove these boundaries accurately (Fig. A6a and Table A4). The processed data without boundaries was then used to re-trained RealMNet with processed data. Results (Fig. A6b) showed that our model was not affected by these physical boundaries, demonstrating the model's prominent capacity to capture informative regions. In order 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. 6a) and give rise to severe visual impairment. The results (Fig. 6b) obtained from transfer learning for RealMNet demonstrated promise in detecting peripheral lesions and distinguishing postoperative cases (PC).

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. First, our model cannot currently recognize "plus" lesions [30], namely, lacquer cracks, myopic

choroidal neovascularization, and Fuchs spot, primarily due to insufficient high-quality data. Second, although our model performed well with UWF images alone, we have not yet incorporated multimodal data (e.g., axial length) to improve performance further. Finally, results on peripheral lesion discrimination exposed limited performance on lesions with very few training data (e.g., RRD and HoT). 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 identification performance. We are optimistic that the developed model would receive excellent transfer ability when pretrained on large-scale UWF images instead of natural ones.

In summary, we offer a dataset comprising high-quality ultra-widefield images and introduce a powerful and reliable workflow for identifying clinically significant lesions to aid in diagnosing pathologic myopia. Through comprehensive evaluation metrics on the hand-crafted PSMM dataset, we have verified the efficacy and 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 decisions.

4 Methods

4.1 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.

4.1.1 Acquisition and labeling

The PSMM dataset consisted of five sub-sources: ShenzhenEye, SUSTech, LishuiR, Zhongshan, and LishuiZ. The ShenzhenEye subset contained 38,922 UWF images of 4,003 patients collected from Shenzhen Eye Hospital of China between January 1^{st}, 2019 and December 31^{st}, 2023. The SUSTech subset contained 2,835 UWF images of 226 patients collected from the Southern University of Science and Technology Hospital of China between January 1^{st}, 2023 and June 31^{st}, 2023. The LishuiR subset contained 938 UWF images of 155 patients collected from Lishui People's Hospital of China between January 1^{st}, 2021 and December 31^{st}, 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 1^{st}, 2021 and December 31^{st}, 2023. Ultimately, we integrated these resources to establish the PSMM dataset that contained 43,371 UWF images of 4,560 patients. Two UWF scanning laser ophthalmoscopy imaging devices captured these images, Daytona (P200T) and California (P200DTx). We retrieved these images by the keywords of $\langle \text{High Myopia}, \text{Pathologic Myopia} \rangle$. 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 clinicians labeled these UWF images, and one senior clinician then double-checked the labeled images by discarding distorted or damaged images for rigorous quality assurance. The composition of the hand-crafted PSMM dataset and its integral subsets are presented in Table A1.

4.1.2 Data processing and stratified partition

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 [31], 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 [32]. 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.

4.2 End-to-end lightweight hybrid framwork

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

4.2.1 Lightweight pretraining distillation backbone

We harness TinyViT [15] as the fundamental backbone to ensure the model achieves excellent performance while retaining 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 pathological patterns. This architecture comprises four stages, each featuring a gradual reduction in resolution akin to the Swin Transformer [19] and LeViT [33]. The patch embedding block incorporates two convolutions with a 3x3 kernel, a stride of 2, and a padding of 1. In the initial stage, we implement lightweight and efficient MBConvs [34] 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 3x3 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. The activation functions adhere to the GELU model, and the normalization layers for convolution and linear operations are BatchNorm and LayerNorm, 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.

4.2.2 Cost-sensitive calibration

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 Binary Cross-Entropy (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:

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

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:

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

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 tackle 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 [22]. Till now, we define the cost-sensitive calibration (CSC) as follows:

$$CSC = -\alpha[p^\gamma \log p + (1 - p)^\gamma \log(1 - p)]$$

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 [23]. Therefore, the ultimate CSC is defined as follows:

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

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 [24], specially designed for multi-label learning. We follow the original computational formula:

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

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.

4.2.3 Classifier adaptation

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 [25] 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 $s_p^i = \frac{\exp\left(\mathcal{T} \boldsymbol{x}_p^\top \boldsymbol{c}_i\right)}{\sum_{l=1}^{P} \exp\left(\mathcal{T} \boldsymbol{x}_l^\top \boldsymbol{c}_i\right)}$, 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 $\hat{\boldsymbol{y}} \triangleq \left(y^1, y^2, \cdots, y^n\right) = \left(\boldsymbol{c}_1^\top \boldsymbol{v}^1, \boldsymbol{c}_2^\top \boldsymbol{v}^2, \cdots, \boldsymbol{c}_n^\top \boldsymbol{v}^n\right)$, 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.

4.3 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.

4.3.1 Centralized inference

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.

4.3.2 Main-source robustness

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.

4.3.3 Cyclic-source generalization

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.

4.4 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 [35]. This issue is particularly relevant in multi-label learning, which is more intricate than single- and multi-class learning [36]. In our study, we opted for comprehensive measures to assess both bipartitions and rankings, considering the characteristics of multi-label data [37].

Considering a development set that has multi-label samples $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ where $i = 1, ..., N$ and N means the number of samples. The labelset of i-th sample $\boldsymbol{y}_i \subseteq \mathcal{L}$ where $\mathcal{L} = \{\lambda_l : j = 1, ..., L\}$ 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 (MLC) 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:

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

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 $\mathcal{B}$,

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

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

We do not use popular Accuracy as an evaluation metric, which overestimates models that only predict well for the majority class by simplistically measuring the absolute amount of correct predictions. We use Hamming Loss to measure the proportion of incorrectly classified instance-label pairs, which is defined as follows:

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

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:

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

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

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

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

4.5 Implementation details

4.5.1 Benchmark approaches

Our model retained lightweight thanks to pretraining distillation techniques and leveraged hierarchical transformer architectures that incorporated convolution operations. Therefore, we selected various widely used benchmark counterparts: DeiT [16], ConvNeXt [18], EfficientNet [17], and Swin Tranformer [19]. 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. We compare RealMNet to these benchmark approaches with respect to model development in Table 4.

4.5.2 Training and testing

We approached the problem in this study as a multi-label learning task to account for the complex relationships between pathological 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 (AdamW) [38] 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 [39] and CutMix [40]. For simplicity, we

used the same parameter settings as in the previous study [32] for UWF images. We leveraged asymmetric focusing as a cost-sensitive calibration with tunable 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 [41] resampling method was applied to the training set only, with an undersampling ratio of 0.2. 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.

4.6 Extensibility

4.6.1 Broader impact statement

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. To resolve this challenge, we employed our model by initializing the backbone with weights trained on the PSMM dataset and then fine-tuning the model on data specific to peripheral retinal lesions. We observed 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 postoperative cases (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 rhegmatogenous retinal detachment (RRD) and holes or tears (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.

Acknowledgements. We thank the support from the National Natural Science Foundation of China 32350410397; Science, Technology, Innovation Commission of Shenzhen Municipality, JCYJ20220530143014032, JCYJ20230807113017035, KCXFZ20211020163813019, Shenzhen Medical Research Funds, D2301002; Department of Chemical Engineering-iBHE special cooperation joint fund project, DCE-iBHE-2022-3; 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.

Declarations

4.7 Competing interests

The authors declare no competing interests.

4.8 Ethics approval

The study followed the guidelines of the World Medical Association Declaration of Helsinki 1964, updated in October 2013, and was conducted after approval by the Ethics Committees of Shenzhen Eye Hospital (2023KYPJ087).

4.9 Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

4.10 Data availability

All benchmark datasets we assembled are publicly available at Zenodo/GitHub/-figshare/Kaggle (choose the optimal alternative).

4.11 Code availability

The codes and trained models developed for this study are publicly available at https://github.com/yo3nglau (updated with the particular project website). Labeling was completed with the open-source graphical image annotation software labelme v5.3.1 (https://github.com/labelmeai/labelme). The code is 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.

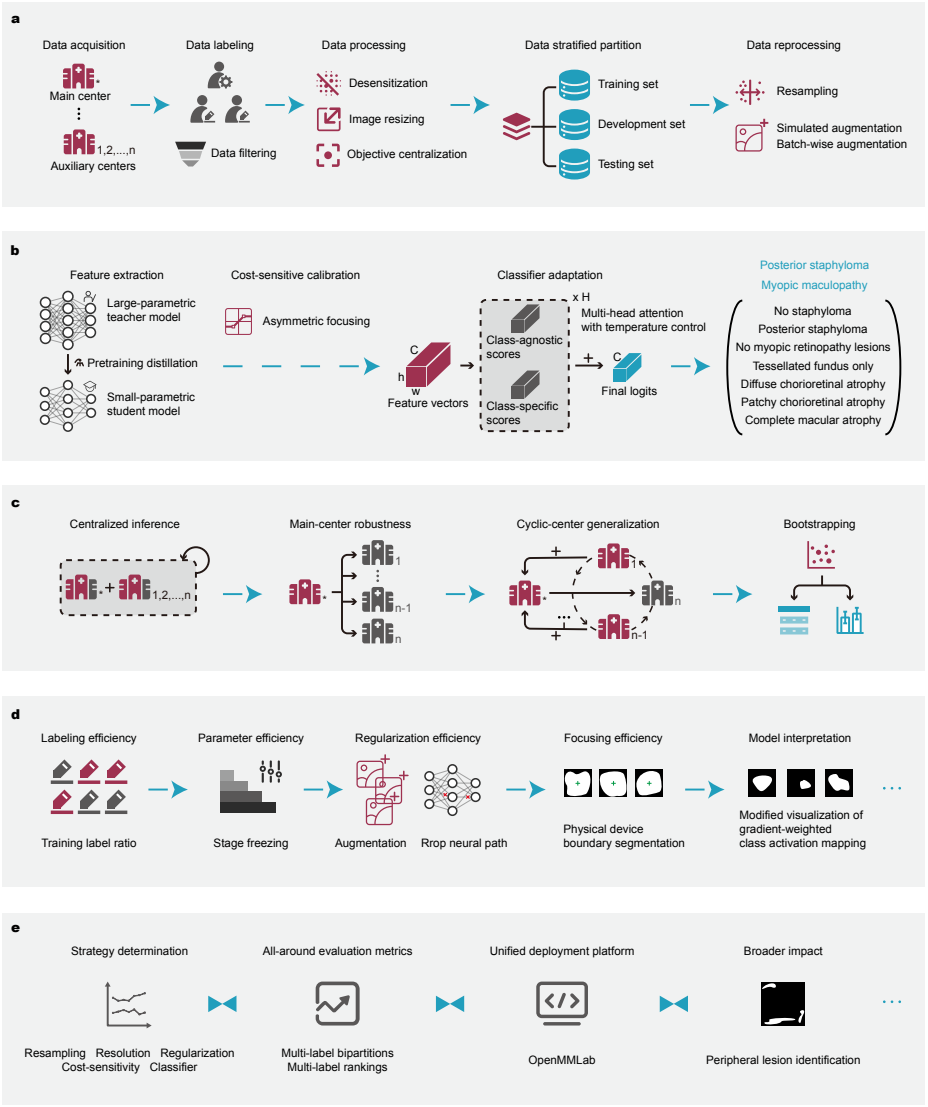


Fig. 1: **General overview of the study. a**, Data machining: data are collected from one main center and four auxiliary centers. After double-checking labeling, quality filtering, and essential processing, a stratified partition is implemented to ensure that the distribution of lesions remains similar 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 complementally 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 1,000 times. **d**, Interpretable 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.

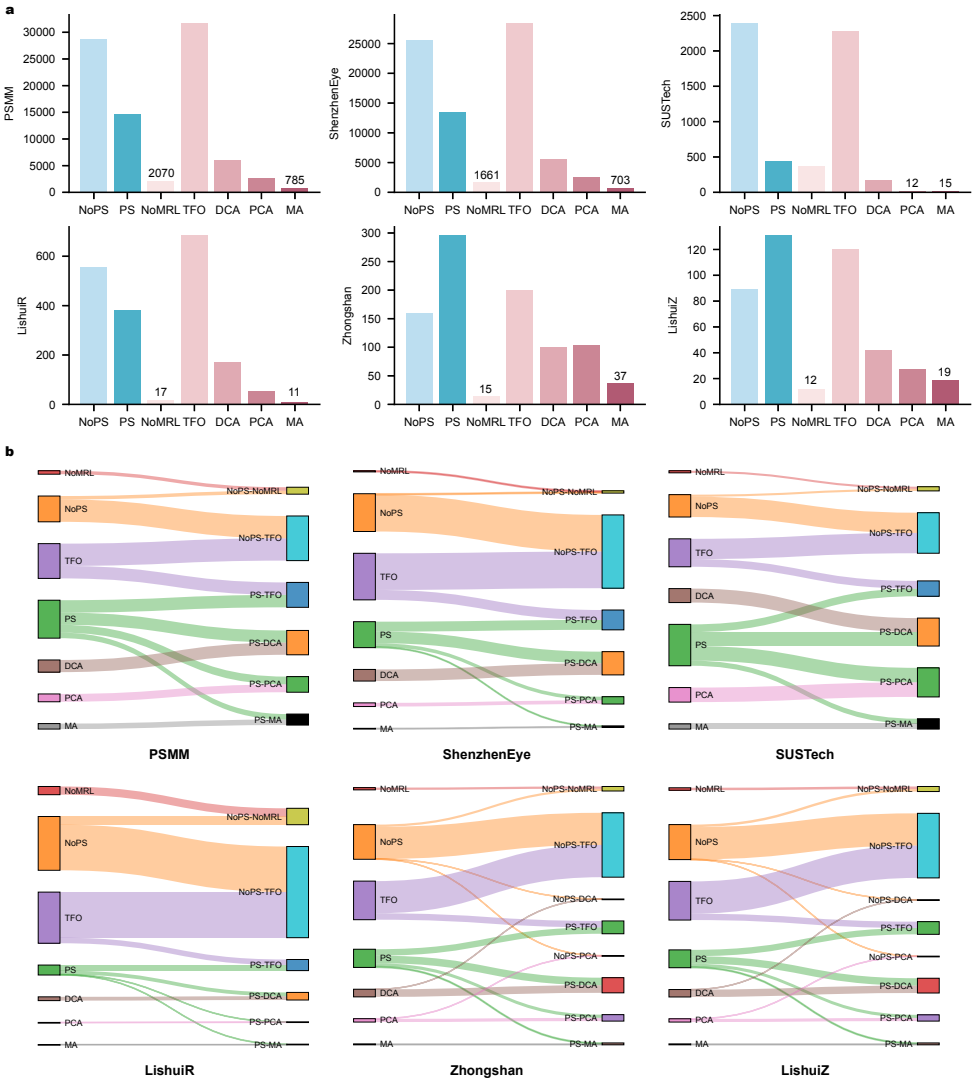


Fig. 2: **Statistics and complications associated with lesions of posterior staphyloma and myopic maculopathy. a**, Statistical analysis of the seven categories in the PSMM dataset and its subsets, with specific values assigned to the minimum two categories of each dataset. **b**, Illustrations of complications 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.

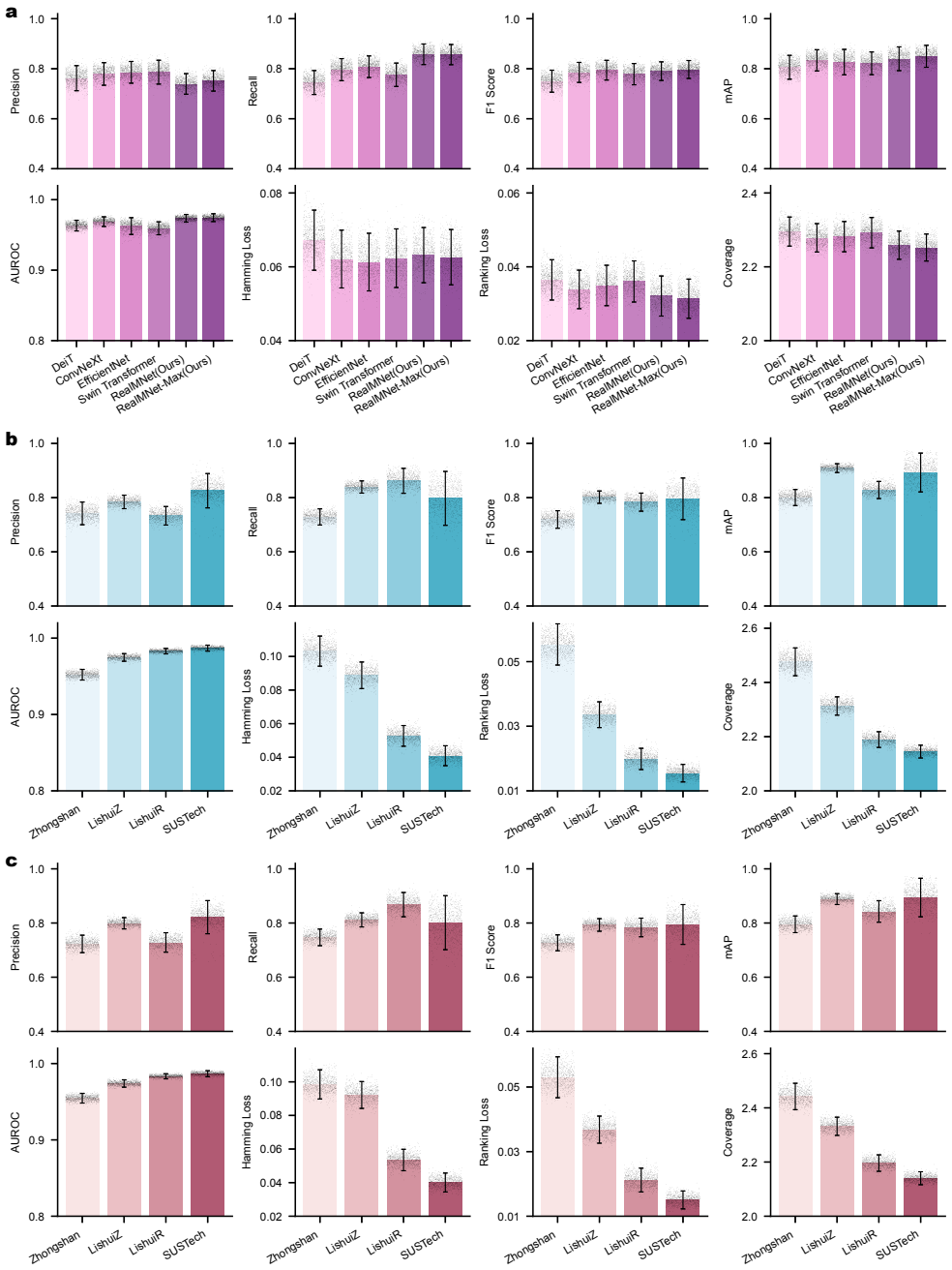


Fig. 3: **Model performance under three experimental protocols. a**, Evaluating model inference capability using the centralized inference protocol. The proposed models are compared to four well-known benchmarks: DeiT, ConvNeXt, EfficientNet, and Swin Transformer. **b**, Assessing model robustness by training on the main source subset and testing on four auxiliary source subsets under the main-source robustness protocol. **c**, 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 (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 1,000 bootstrap samples for corresponding testing sets with n=1,000 samples.

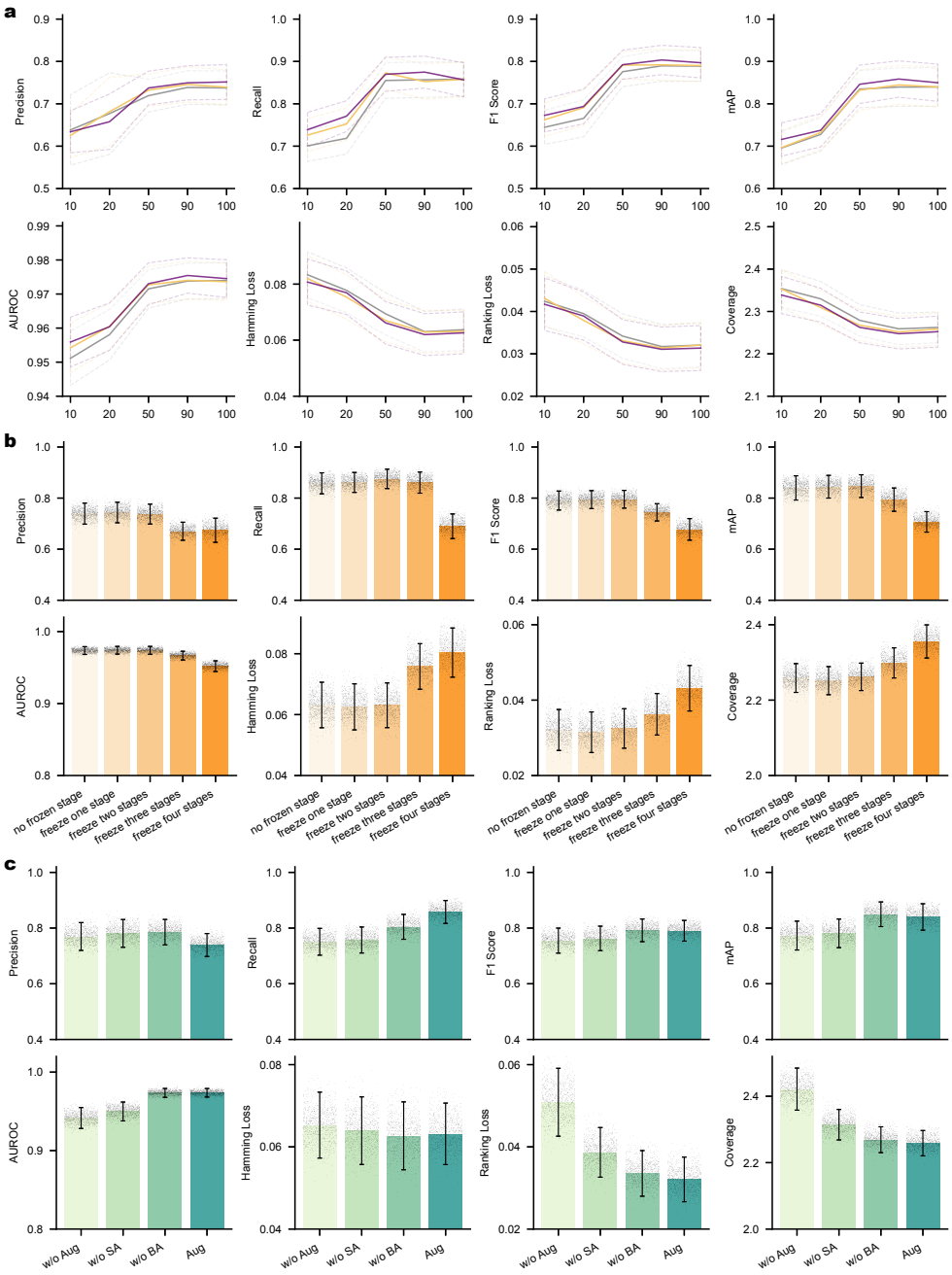


Fig. 4: **Efficiency of RealMNet in identifying posterior staphyloma and myopic maculopathy on the PSMM dataset. a**, Labeling efficiency: we progressively increase the amount of training data and labels to achieve precise and stable performance. The 95% confidence interval (CI) of the displayed metrics are plotted in dotted lines, and the central lines indicate the mean value. **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 (SA) and batch-wise augmentation (BA), 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 1,000 bootstrap samples for corresponding testing sets with n=1,000 samples.

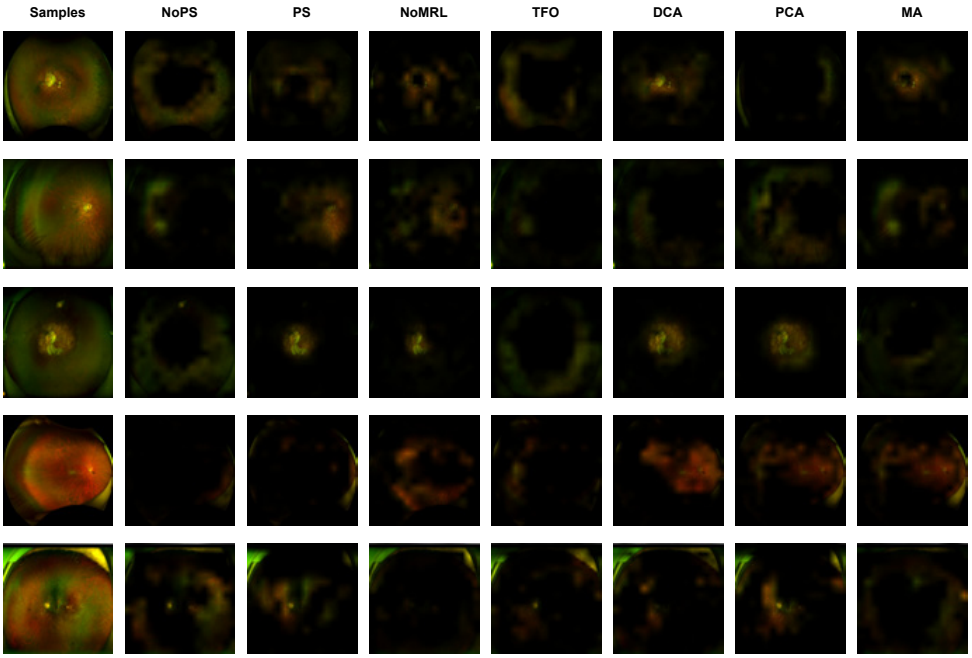


Fig. 5: We generated visualizations using an improved version of gradient-weighted class activation mapping (Grad-CAM++). These visualizations show the predictions of RealMNet for each category of posterior staphyloma and myopic maculopathy. By merging the heatmaps with the original images, we highlight the dispersed regions that are associated with lesions related to posterior staphyloma and myopic maculopathy.

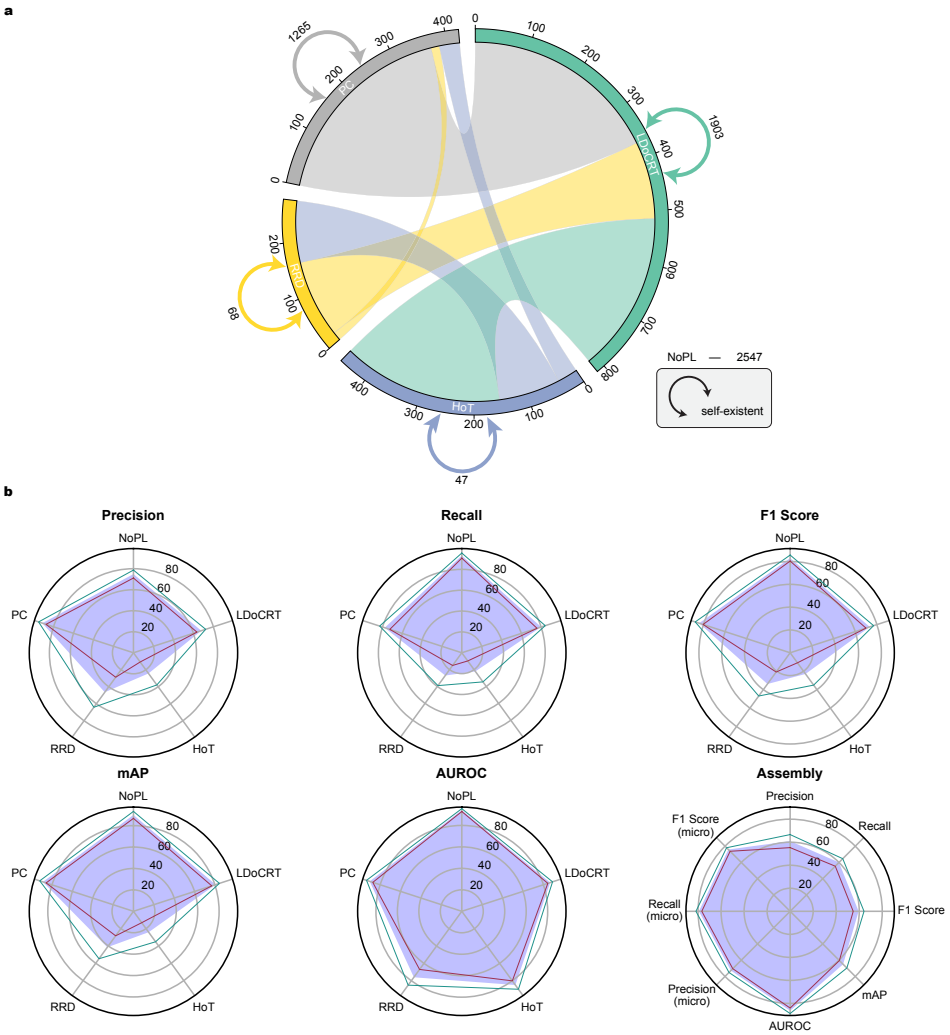


Fig. 6: **Identifying complicated peripheral lesions. a**, Concurrent distribution of peripheral lesions. Peripheral lesions may have different concurrent relationships with each other, or they may occur separately. **b**, Model performance on peripheral lesion identification. The blue facecolor represents 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.

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
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

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

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
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

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

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
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

Table 4: Model information.

Model	Architecture	Implementation	Scale	Image Size	#Params(M)	FLOPs(G)
DeiT	Transformer	Distillation	Base	384	87.63	55.65
ConvNeXt	ConvNet	Hierarchy	Tiny	384	28.59	13.14
EfficientNet	ConvNet	Scaling	B4	380	19.34	4.66
Swin Transformer	Transformer	Hierarchy	Base	384	87.90	44.49
RealMNet (Ours)	Hybrid	Hierarchy Pretraining Distillation	21M	384	21.23	13.85
RealMNet (Ours)	Hybrid	Hierarchy Pretraining Distillation	21M	512	21.27	27.15

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

Protocol	Training set		Development set		Testing set	
	Patients	Images	Patients	Images	Patients	Images
CIP	3,192 (r. 3,138)	30,420 (r. 24,683)	684	6,377	684	6,574

The numbers with prefix r. mean resampling results.

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

Protocol	Training set			Testing set		
	Subset	Patients	Images	Subset	Patients	Images
MRP	ShenzhenEye	4,003 (r. 3,944)	38,922 (r. 31,575)	SUSTech	226	2,835
				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)	4,334 (r. 4,256)	40,536 (r. 32,888)	SUSTech	226	2,835
	PSMM (w/o LishuiR)	4,405 (r. 4,330)	42,433 (r. 34,408)	LishuiR	155	938
	PSMM (w/o Zhongshan)	4,475 (r. 4,398)	42,915 (r. 34,871)	Zhongshan	85	456
	PSMM (w/o LishuiZ)	4,469 (r. 4,389)	43,151 (r. 35,009)	LishuiZ	91	220

The numbers with prefix r. mean resampling results.

Appendix A Extended Data

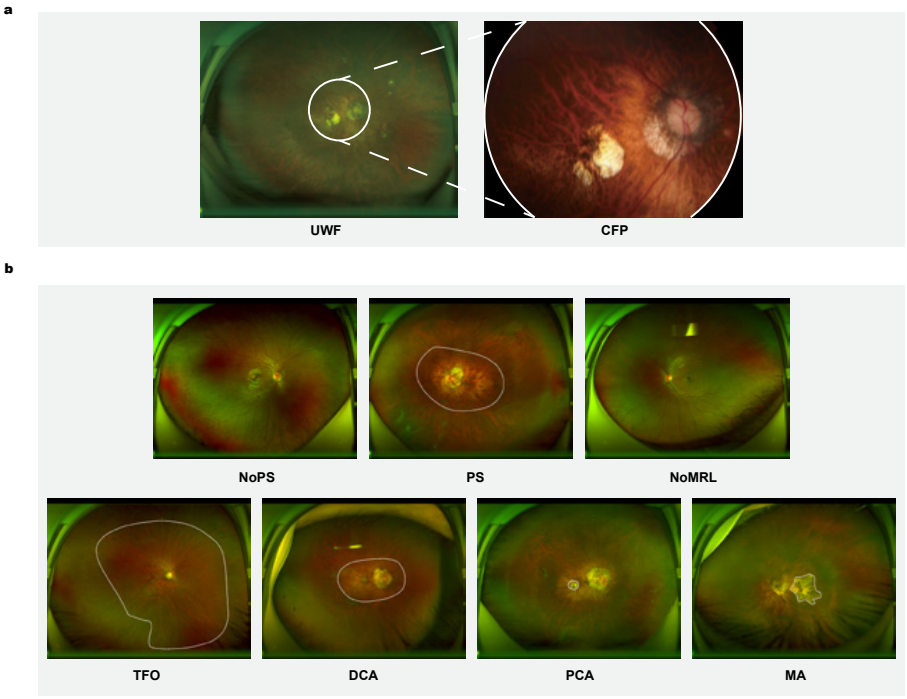


Fig. A1: **Illustration of ultra-widefield imaging and lesion types on the PSMM dataset. a**, Retinal field of view comparison between ultra-widefield (UWF) imaging and color fundus photography (CFP). We present UWF and CFP images from the same patient to illustrate the expanded field of view provided by UWF imaging. **b**, Lesion regions of posterior staphyloma and myopic maculopathy. We show the presence of posterior staphyloma (NoPS or PS) and five categories of myopic maculopathy: no myopic retinal lesions (NoMRL), tessellated fundus only (TFO), diffuse chorioretinal atrophy (DCA), patchy chorioretinal atrophy (PCA), and macular atrophy (MA).

We introduce the seven categories of posterior staphyloma and myopic maculopathy annotated in the PSMM dataset, along with their corresponding lesion regions of clinical interest. Different types of lesions require different areas of concern, making accurate segmentation challenging and posing a subsequent challenge for further work.

It can be observed that the data collected are mainly concentrated on young adults requiring timely diagnosis and treatment. Data on younger and older patients have also been collected to provide a more comprehensive perception. While some datasets often collect data that are balanced in terms of age and gender, we prioritize gathering real-world data to support the development of models that can handle imbalanced

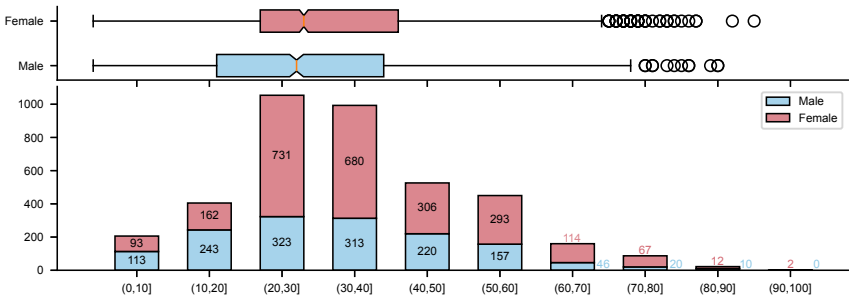


Fig. A2: **Statistics of age and gender of the ShenzhenEye subset.** The histogram presents the number of males and females within each ten-year age interval, while the box plot illustrates the distribution of ages.

data. Although this approach may lead to lower model performance, it is essential to have the courage to confront these challenges.

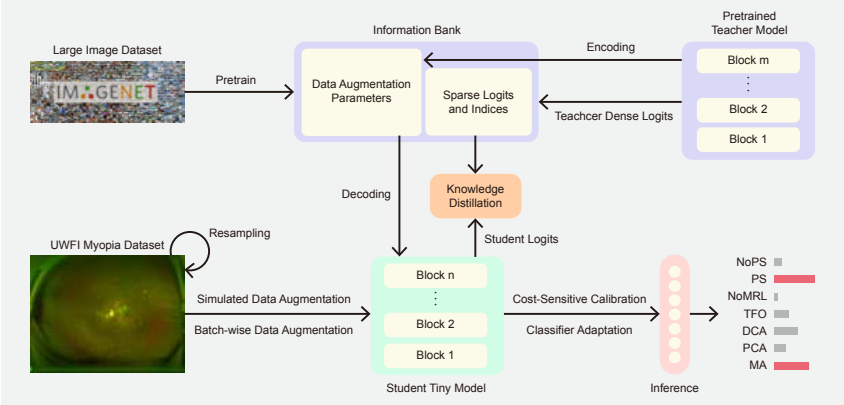


Fig. A3: **Architecture details of RealMNet.** RealMNet harnesses the TinyViT as the feature extraction backbone. The process of pretraining distillation is explained, and the resulting distilled student model is employed for fine-tuning to tackle the challenge of identifying posterior staphyloma and myopic maculopathy.

Resampling methods are essential for addressing the imbalance issue in multi-label datasets (MLDs). Researchers have developed various algorithms to tackle different MLDs and minimize the potential adverse effects of imbalanced data distributions. In this study, we examine six widely adopted approaches (Fig. A5a). LP-ROS (Label Powerset Random Over Sampling) [42] is a method that oversamples multi-label datasets by cloning random samples of minority label sets until the dataset is $r\%$ larger than the original. LP-RUS (Label Powerset Random UnderSampling) [42] is a

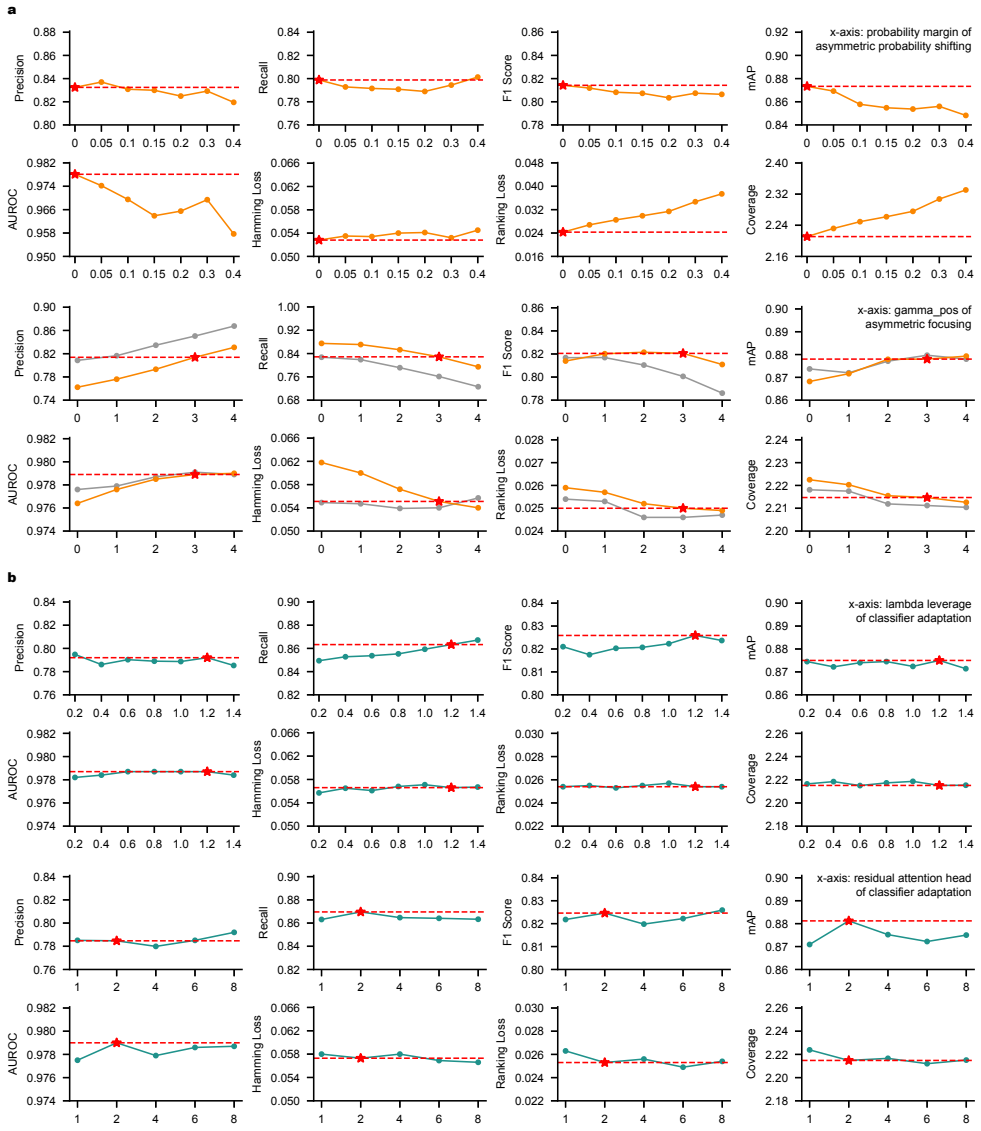


Fig. A4: Researching the advancement of cost-sensitive calibration and classifier adaptation. **a**, We entail an exploration of asymmetric probability shifting and asymmetric focusing, with a search for the probability margin m. In the illustrated results, the gray lines denote the negative focusing parameter $\gamma_- = 2$, while the other colored lines represent $\gamma_- = 4$. **b**, We progressively examine the leveraging parameter λ and the quantity of residual attention head. The determined choice is highlighted with a red star, accompanied by a horizontal line to facilitate comparison.

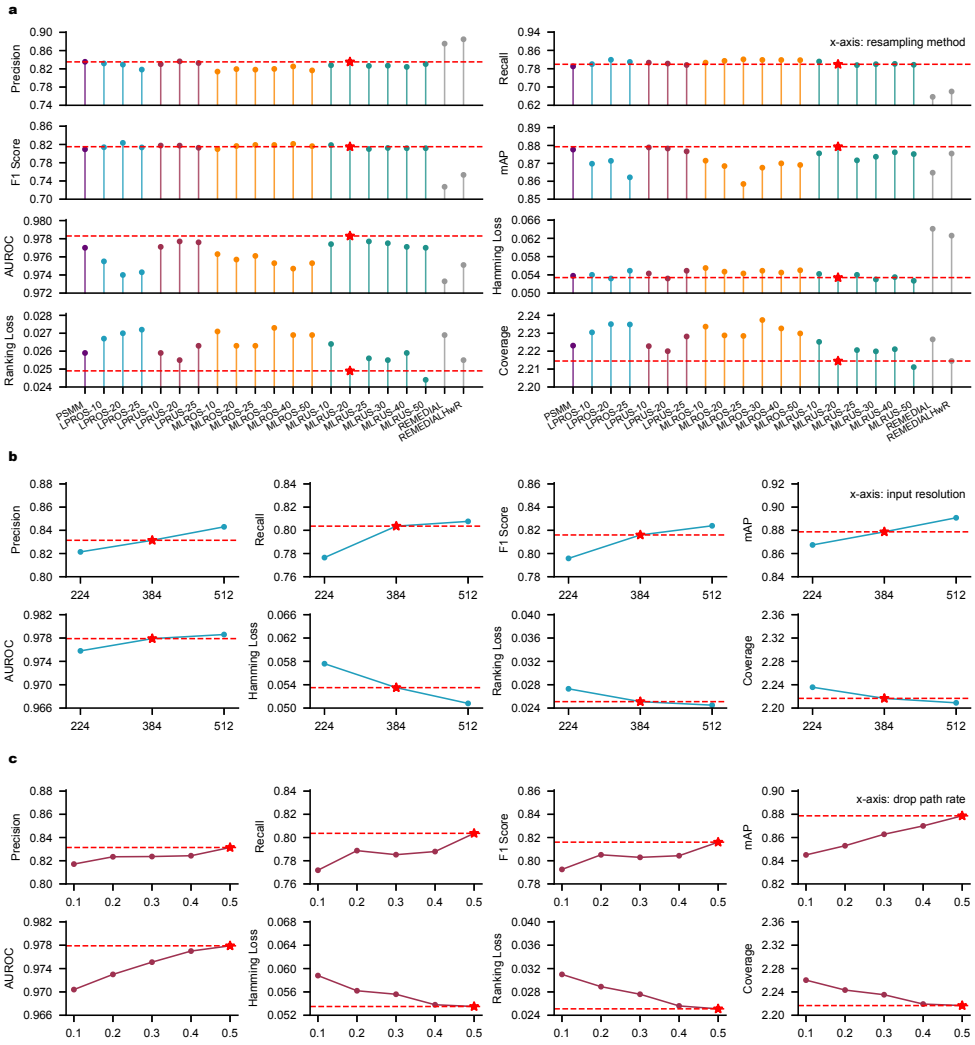


Fig. A5: **Comparing model performance using various resampling methods, input resolutions, and drop path rates. a**, Investigating the resampling methods. We resample the training set using multiple resampling methods: LP-ROS, LP-RUS, ML-ROS, ML-RUS, REMEDIAL, and REMEDIAL-HwR. We explore these methods with various resampling ratios denoted as r. **b**, Investigating the resolutions of input images. We assess common resolutions of 224, 384, and 512 using a development set. **c**, We set a maximum drop path rate of 0.5 with an increment of 0.1 to observe the impact of different drop path rates. The determined choice is highlighted with a red star, accompanied by a horizontal line to facilitate comparison.

method that undersamples multi-label datasets by deleting random samples of majority label sets until the dataset is reduced to (1-r%) of its original size. LP-ROS and LP-RUS evaluate the frequency of complete label sets during preprocessing. ML-ROS (Multi-Label Random Over Sampling) [41] identifies samples with minority labels and clones them, while ML-RUS (Multi-Label Random Under Sampling) [41] identifies samples with majority labels and deletes them. ML-ROS and ML-RUS evaluate the frequency of individual labels, isolating samples in which one or more minority labels appear. REMEDIAL (REsampling MultilabEl datasets by Decoupling highly ImbAlanced Labels) [43] is an algorithm that edits and oversamples by decoupling frequent and rare classes appearing in the same sample and adding new samples to the original dataset.

Let N be the sample number of a multi-label dataset, L the full set of labels, λ the label being analyzed, and $\boldsymbol{y}_i$ the label set of i-th sample. We use the LRlbl (Imbalance Ratio per Label) measure that is calculated individually for each label:

$$\mathrm{IRLbl}\left(\lambda\right) = \frac{\max\limits_{\lambda' \in L}\left(\sum_{i=1}^{N} [\![\lambda' \in \boldsymbol{y}_i]\!]\right)}{\sum_{i=1}^{N} [\![\lambda \in \boldsymbol{y}_i]\!]}$$

where the symbol $[\![]\!]$ denotes the Iverson bracket, which returns 1 if the expression inside it is true or 0 otherwise. The higher the IRLbl, the larger would be the imbalance, which helps identify minority or majority labels. Then, we calculate the MeanIR (Mean Imbalance Ratio) measure by averaging IRLbl to estimate the global imbalance level:

$$\mathrm{MeanIR} = \frac{1}{L}\sum_{\lambda \in L}\mathrm{IRLbl}\left(\lambda\right)$$

The REMEDIAL algorithm is calculated relying on the SCUMBLE (Score of ConcUrrence among iMBalanced LabEls) measure that aims to quantify the imbalance variance among the labels present in each data sample. SCUMBLE is based on the Atkinson index and the IRLbl measure. The SCUMBLE value of each sample in a multi-label dataset D is calculated as follows:

$$\mathrm{SCUMBLE}_{sample}\left(s\right) = 1 - \frac{1}{\overline{\mathrm{IRLbl}_s}}\left(\prod_{\lambda=1}^{L}\mathrm{IRLbl}_s\left(\lambda\right)\right)^{(1/L)}$$

where $\mathrm{IRLbl}_s\left(\lambda\right) = \mathrm{IRLbl}\left(\lambda\right)$ if the label λ is present in the sample s, otherwise 0. $\overline{\mathrm{IRLbl}_s}$ stands for the average imbalance level of the labels appearing in sample s. We average all scores of samples to obtain the final SCUMBLE value:

$$\mathrm{SCUMBLE}_{dataset}\left(D\right) = \frac{1}{L}\sum_{\lambda=1}^{L}\mathrm{SCUMBLE}_{sample}\left(\lambda\right)$$

We also harness the SCUMBLELbl to leverage the difficulty of labels:

$$\text{SCUMBLELbl}\left(\lambda\right) = \frac{\sum_{i=1}^{N} [\![\lambda \in \boldsymbol{y}_i]\!] \cdot \text{SCUMBLE}_{sample}\left(s\right)}{\sum_{i=1}^{N} [\![\lambda \in \boldsymbol{y}_i]\!]}$$

Based on our evaluation, ML-ROS and ML-RUS outperform LP-ROS and LP-RUS in terms of mAP and AUROC, despite having similar F1 Score results. Therefore, we investigate the performance of ML-ROS and ML-RUS with more compact resampling ratios. Our findings indicate that ML-RUS surpasses ML-ROS in both mAP and AUROC, while also exhibiting lower Hamming Loss, Ranking Loss, and Coverage. We also observe that both the REMEDIAL algorithm and its adapted version, REMEDIAL-HwR, do not yield better performance. This confirms that REMEDIAL performs poorly on multi-label datasets with a low SCUMBLE level, which in our case is 0.0741. As a result, we opt for the ML-RUS algorithm with a resampling ratio of $r = 20$, as it consistently excels across all evaluation metrics.

The choice of resolution directly impacts the quality of features the model can learn. Most neural networks use resolutions like 224, 256, and 384. We test different resolutions on a development set to see how they affect model performance (Fig. A5b). Our backbone is designed to work with a resolution of 512, which is larger than typical backbones. When we fine-tune the model using higher resolutions, we increase the window size of each self-attention layer to match the input resolution. Our results show that higher resolutions lead to more accurate results, but they require more training time and computational resources. After considering performance and resource requirements, we choose 384 as our main resolution for model development. Ultimately, we also scale up the resolution to 512 to demonstrate model capability.

Drop path [27] is a critical regularization technique that involves randomly dropping entire neural paths to prevent model over-fitting. Since the size of the collected dataset is still relatively small compared to those in computer vision, this technique plays a significant role in constraining the model to fit our tasks. Experimental results (Fig. A5c) show that using a higher drop path rate benefits the model by effectively preventing over-fitting. Therefore, we decide to use a drop path rate of 0.5 for the rest of the experiments in this study.

Modern ultra-widefield imaging inevitably photos the boundaries of the physical devices along with the imaging targets, which occlude partial information. To assess whether these boundaries negatively affect the model's inference capability, we intend to segment out these boundaries and re-train our model using data without them. We discover that nearly three-quarters of the images in the PSMM dataset have significant black borders, and the remaining images, while lacking black borders, still show considerable device boundary interference. We randomly sample 1% of the data from the two imaging types to construct a segmentation dataset. We select at the patient level to circumvent the information leakage emphasized in the stratified partitioning. We seek the expertise of professional physicians to annotate the dataset at the pixel level. The resulting segmentation dataset comprises 412 images, involving 303 images with black borders and 109 images without black borders. We divide the dataset into

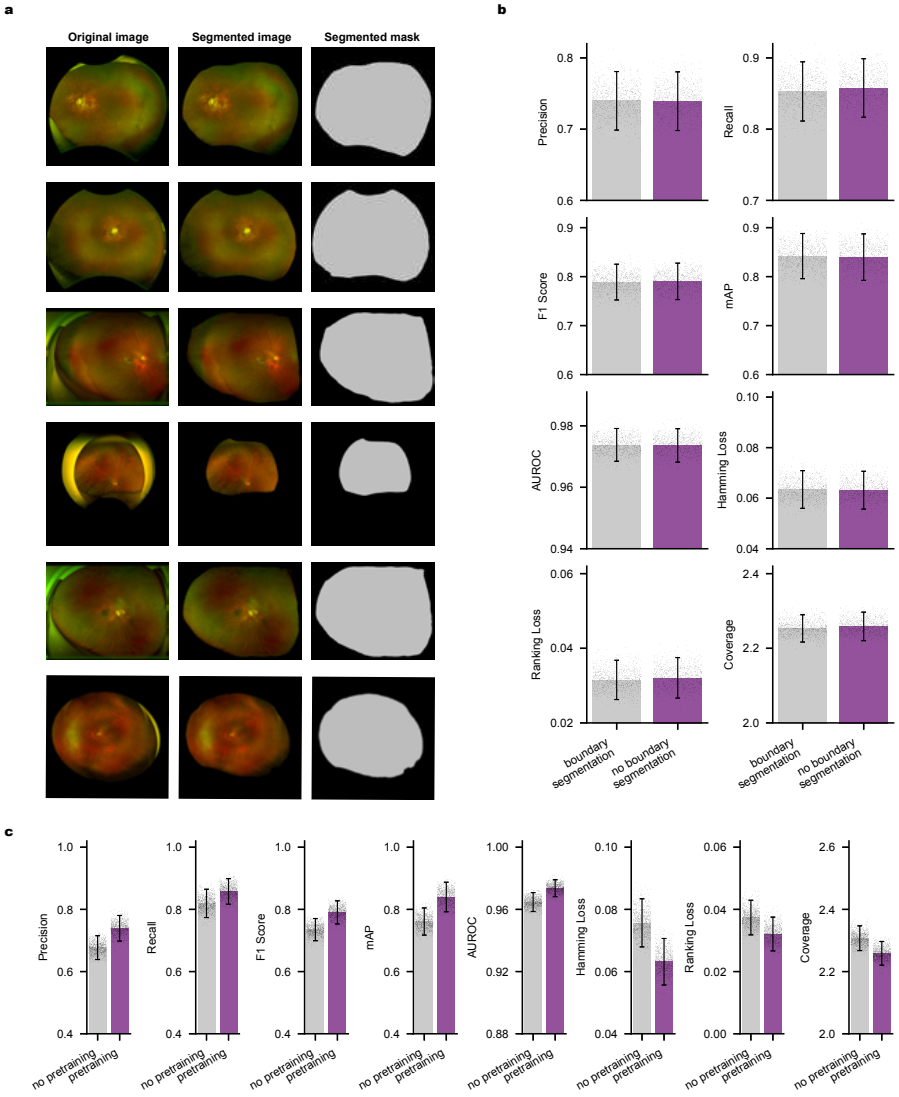


Fig. A6: **Investigating the impact of physical device boundaries and pretraining on the model performance. a**, Visualizing the results of boundary segmentation. We present original images, segmented images, and segmented masks, respectively. **b**, Comparing the performance of models trained on data with and without boundary segmentation. **c**, Comparing the performance of models trained with and without weights derived from large-scale natural image datasets (e.g., ImageNet-21k).

a training set, development set, and testing set in an 8:1:1 ratio. We employ ResNet-50 as the segmentation backbone and DeepLab-v3 as the segmentation model with weights trained on the PASCAL Visual Object Classes Challenge (PASCAL VOC) 2012 dataset [44]. We utilize the SGD optimizer with a batch size of 4, a learning rate of 0.01, and a momentum of 0.9 for 2,000 epochs with early stopping. After fine-tuning, we harness the model to segment out the boundaries of the physical devices and re-train RealMNet with these images. Ultimately, we find that there is hardly any difference between the performance of models trained on data with and without boundary segmentation, which suggests that the model distinguishes instrumental regions and focuses the field of view within the boundaries of the physical devices.

We demonstrate the effectiveness of pretraining by observing the advantages gained from distilling the pretraining of the backbone on large-scale natural image datasets, such as ImageNet-21k [45]. Training a model from scratch requires a large dataset and a significant amount of time. Pretraining allows for the transfer of knowledge to downstream tasks, improving performance and reducing the need to start training from scratch. It can also conserve computational resources by utilizing the already learned representations. We find that the performance of the model improves significantly when initialized with weights that encompass the abundant knowledge from the large-parametric model (in our case, CLIP-ViT-L/14-21k [46]), which demonstrates the superiority of utilizing the power of pertaining.

Table A1: Overview of the PSMM dataset and its subsets.

Dataset	Source	Patients	UWF Images
PSMM	Integrated	4,560	43,371
ShenzhenEye	Main	4,003	38,922
SUSTech	Auxiliary	226	2,835
LishuiR	Auxiliary	155	938
Zhongshan	Auxiliary	85	456
LishuiZ	Auxiliary	91	220

Table A2: Imbalance levels of the PSMM dataset.

Measure	NoPS	PS	NoMRL	TFO	DCA	PCA	MA
IRLbl	1.1145	2.0999	15.4234	1.	5.0833	11.5897	38.9930
SCUMBLELbl	0.0394	0.1393	0.4986	0.0124	0.0915	0.2853	0.5596
w/ REMEDIAL	0.0018	0.0606	0.	0.0124	0.0915	0.	0.
MeanIR	10.7577						
SCUMBLE	0.0741 (w/ REMEDIAL 0.0174)						

Table A3: Black border statistics for image data of the PSMM dataset and its subsets.

Dataset	PSMM	ShenzhenEye	SUSTech	LishuiR	Zhongshan	LishuiZ
w/ Black Border	31,244	28,409	2,835	0	0	0
w/o Black Border	12,127	10,513	0	938	456	220

Table A4: Boundary segmentation results.

Boundary Segmentation	Accuracy	mIoU
w/ Black Border	0.9813	0.9586
w/o Black Border	0.9796	0.9482

Table A5: Varying α for Cross-Entropy Loss ($\gamma = 0$).

α	Precision	Recall	F1 Score	mAP	AUROC	Hamming Loss↓	Ranking Loss↓	Coverage↓
.10	0.9071	0.6071	0.7105	0.8778	0.9766	0.0659	0.0255	2.2233
.25	0.8720	0.7325	0.7922	0.8783	0.9774	0.0554	0.0249	2.2145
.50	0.8309	0.8031	0.8158	0.8794	0.9781	0.0541	0.025	2.2156
.75	**0.7805**	**0.8716**	**0.8229**	**0.8787**	**0.9782**	**0.0575**	**0.0253**	**2.2148**
.90	0.7163	0.9347	0.8082	0.8781	0.9783	0.0709	0.0247	2.2108
.99	0.5193	0.9914	0.6569	0.8492	0.9747	0.1389	0.0254	2.2194
.999	0.3981	0.9961	0.5186	0.7532	0.9607	0.2545	0.0314	2.2614

Table A6: Varying γ for Focal Loss (with optimal α).

γ	α	Precision	Recall	F1 Score	mAP	AUROC	Hamming Loss↓	Ranking Loss↓	Coverage↓
0	.75	**0.7805**	**0.8716**	**0.8229**	**0.8787**	**0.9782**	**0.0575**	**0.0253**	**2.2148**
0.1	.75	0.7822	0.8693	0.8230	0.8764	0.9787	0.0571	0.0251	2.2134
0.2	.75	0.7826	0.8688	0.8230	0.8779	0.9788	0.0569	0.0248	2.2109
0.5	.50	0.8296	0.7981	0.8122	0.8774	0.9786	0.0537	0.0249	2.2133
1.0	.25	0.8628	0.7207	0.7806	0.8747	0.9774	0.0566	0.0252	2.2156
2.0	.25	0.8660	0.7180	0.7802	0.8768	0.9783	0.0564	0.0251	2.2142
5.0	.25	0.8695	0.7097	0.7760	0.8791	0.9790	0.0572	0.0253	2.2155

Table A7: Varying $T_{\mathcal{P}}$ and $T_{\mathcal{N}}$ for Two-way Loss.

$T_{\mathcal{P}}$	$T_{\mathcal{N}}$	Precision	Recall	F1 Score	mAP	AUROC	Hamming Loss↓	Ranking Loss↓	Coverage↓
0.5	0.5	0.7493	0.8883	0.8124	0.8803	0.9774	0.0668	0.0261	2.2241
0.5	1	0.7095	0.922	0.8006	0.8821	0.978	0.0773	0.0254	2.2153
0.5	2	0.6787	0.9453	0.7871	0.8837	0.9788	0.0878	0.0246	2.2075
0.5	4	0.6531	0.9591	0.77	0.8851	0.9785	0.0948	0.0246	2.2075
1	0.5	0.7729	0.8694	0.8178	0.8736	0.9764	0.0607	0.0262	2.2272
1	1	0.7244	0.9072	0.804	0.8771	0.9772	0.0699	0.025	2.2159
1	2	0.6938	0.9388	0.7949	0.8812	0.9783	0.0797	0.0241	2.2073
1	4	0.6548	0.9571	0.7704	0.8802	0.9781	0.0918	0.0245	2.2101
2	0.5	0.798	0.837	0.8162	0.8699	0.9752	0.0568	0.0261	2.2288
2	1	0.7536	0.8835	0.8123	0.8734	0.9769	0.0629	0.025	2.2186
2	2	0.714	0.9226	0.8025	0.8761	0.9777	0.0712	0.0243	2.2103
2	4	0.6691	0.9532	0.7778	0.8771	0.9781	0.0817	0.0238	2.2043
4	0.5	0.8294	0.7918	0.806	0.8675	0.9745	0.0557	0.0263	2.2293
4	1	**0.8017**	**0.8371**	**0.8164**	**0.8702**	**0.9761**	**0.0578**	**0.0259**	**2.2241**
4	2	0.7475	0.885	0.809	0.8708	0.9763	0.0639	0.0252	2.2161
4	4	0.6853	0.9424	0.7852	0.8688	0.9759	0.073	0.025	2.2117

References

[1] Baird, P.N., Saw, S.-M., Lanca, C., Guggenheim, J.A., Smith III, E.L., Zhou, X., Matsui, K.-O., Wu, P.-C., Sankaridurg, P., Chia, A., *et al.*: Myopia. Nature reviews Disease primers **6**(1), 99 (2020)

[2] Dolgin, E.: A myopia epidemic is sweeping the globe. here's how to stop it. Nature **629**(8014), 989–991 (2024)

[3] Morgan, I.G., Ohno-Matsui, K., Saw, S.-M.: Myopia. The Lancet **379**(9827), 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**(6), 1368–1374 (2016)

[5] Dai, L., Wu, L., Li, H., Cai, C., Wu, Q., Kong, H., Liu, R., Wang, X., Hou, X., Liu, Y., *et al.*: A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nature communications **12**(1), 3242 (2021)

[6] Bora, A., Balasubramanian, S., Babenko, B., Virmani, S., Venugopalan, S., Mitani, A., Oliveira Marinho, G., Cuadros, J., Ruamviboonsuk, P., Corrado, G.S., *et al.*: Predicting the risk of developing diabetic retinopathy using deep learning. The Lancet Digital Health **3**(1), 10–19 (2021)

[7] Yim, J., Chopra, R., Spitz, T., Winkens, J., Obika, A., Kelly, C., Askham, H., Lukic, M., Huemer, J., Fasler, K., *et al.*: Predicting conversion to wet age-related macular degeneration using deep learning. Nature Medicine **26**(6), 892–899 (2020)

[8] Ohno-Matsui, K., Kawasaki, R., Jonas, J.B., Cheung, C.M.G., Saw, S.-M., Verhoeven, V.J., Klaver, C.C., Moriyama, M., Shinohara, K., Kawasaki, Y., *et al.*: International photographic classification and grading system for myopic maculopathy. American journal of ophthalmology **159**(5), 877–883 (2015)

[9] Ohno-Matsui, K., Lai, T.Y., Lai, C.-C., Cheung, C.M.G.: Updates of pathologic myopia. Progress in retinal and eye research **52**, 156–187 (2016)

[10] De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., *et al.*: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine **24**(9), 1342–1350 (2018)

[11] Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., *et al.*: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)

[12] Tan, T.-E., Anees, A., Chen, C., Li, S., Xu, X., Li, Z., Xiao, Z., Yang, Y., Lei, X., Ang, M., *et al.*: Retinal photograph-based deep learning algorithms for myopia and a blockchain platform to facilitate artificial intelligence medical research: a retrospective multicohort study. The Lancet Digital Health **3**(5), 317–329 (2021)

[13] Yang, D., Li, M., Wei, R., Xu, Y., Shang, J., Zhou, X.: Optomap ultrawide field imaging for detecting peripheral retinal lesions in 1725 high myopic eyes before implantable collamer lens surgery. Clinical & Experimental Ophthalmology **48**(7), 895–902 (2020)

[14] Lewis, H.: Peripheral retinal degenerations and the risk of retinal detachment. American journal of ophthalmology **136**(1), 155–160 (2003)

[15] Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L.: Tinyvit: Fast pretraining distillation for small vision transformers. In: European Conference on Computer Vision, pp. 68–85 (2022). Springer

[16] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357 (2021). PMLR

[17] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR

[18] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)

[19] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

[20] 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), pp. 839–847 (2018). IEEE

[21] De Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Annals of operations research **134**, 19–67 (2005)

[22] 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, pp. 2980–2988 (2017)

[23] Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 82–91 (2021)

[24] Kobayashi, T.: Two-way multi-label loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7476–7485 (2023)

[25] 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, pp. 184–193 (2021)

[26] Marcondes, D., Simonis, A., Barrera, J.: Back to basics to open the black box. Nature Machine Intelligence, 1–4 (2024)

[27] Larsson, G., Maire, M., Shakhnarovich, G.: Fractalnet: Ultra-deep neural networks without residuals. In: International Conference on Learning Representations (2022)

[28] 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), pp. 1–6 (2020). IEEE

[29] Babenko, B., Mitani, A., Traynis, I., Kitade, N., Singh, P., Maa, A.Y., Cuadros, J., Corrado, G.S., Peng, L., Webster, D.R., *et al.*: Detection of signs of disease in external photographs of the eyes via deep learning. Nature biomedical engineering **6**(12), 1370–1383 (2022)

[30] Ruiz-Medrano, J., Montero, J.A., Flores-Moreno, I., Arias, L., García-Layana, A., Ruiz-Moreno, J.M.: Myopic maculopathy: current status and proposal for a new classification and grading system (atn). Progress in retinal and eye research **69**, 80–115 (2019)

[31] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010)

[32] Engelmann, J., McTrusty, A.D., MacCormick, I.J., Pead, E., Storkey, A., Bernabeu, M.O.: Detecting multiple retinal diseases in ultra-widefield fundus imaging and data-driven identification of informative regions with deep learning. Nature Machine Intelligence **4**(12), 1143–1154 (2022)

[33] Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M.: Levit: a vision transformer in convnet's clothing for faster inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12259–12269 (2021)

[34] 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, pp. 4510–4520 (2018)

[35] Roberts, M., Hazan, A., Dittmer, S., Rudd, J.H., Schönlieb, C.-B.: The curious case of the test set auroc. Nature Machine Intelligence **6**(4), 373–376 (2024)

[36] Wu, X.-Z., Zhou, Z.-H.: A unified view of multi-label performance measures. In: International Conference on Machine Learning, pp. 3780–3788 (2017). PMLR

[37] Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. Data mining and knowledge discovery handbook, 667–685 (2010)

[38] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)

[39] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018)

[40] Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032 (2019)

[41] Charte, F., Rivera, A.J., Jesus, M.J., Herrera, F.: Addressing imbalance in multilabel classification: Measures and random resampling algorithms. Neurocomputing **163**, 3–16 (2015)

[42] Charte, F., Rivera, A., Jesus, M.J., Herrera, F.: A first approach to deal with imbalance in multi-label datasets. In: Hybrid Artificial Intelligent Systems: 8th International Conference, HAIS 2013, Salamanca, Spain, September 11-13, 2013. Proceedings 8, pp. 150–160 (2013). Springer

[43] Charte, F., Rivera, A.J., Jesus, M.J., Herrera, F.: Dealing with difficult minority labels in imbalanced mutilabel data sets. Neurocomputing **326**, 39–53 (2019)

[44] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

[45] Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: Imagenet-21k pretraining for the masses. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021)

[46] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR