Scientific Reports

https://doi.org/10.1038/s41598-026-47858-z

Article in Press

Long-tailed multi-label retinal disease classification using alternate group training and gradient-based re-weighting

Received: 11 November 2025

Accepted: 3 April 2026

Published online: 24 April 2026

Cite this article as: Jian Y., Jia X., Zhang H. *et al.* Long-tailed multi-label retinal disease classification using alternate group training and gradient-based re-weighting. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-47858-z

Yingying Jian, Xiaoyan Jia, Han Zhang, Qian Zhou & Canhua Xu

Long-tailed Multi-label Retinal Disease Classification using Alternate Group Training and Gradient-based Re-weighting

Yingying Jian[1], Xiaoyan Jia[1], Han Zhang[2], Qian Zhou[3*], Canhua Xu[1*]

[1*]Department of Biomedical Engineering, Fourth Military Medical University, Xi'an, 710032, Shaanxi, China.
[2]Department of Epidemiology, School of Public Health, Fourth Military Medical University, Xi'an, 710032, Shaanxi, China.
[3]School of Computer Science, Wuhan University, Wuhan, 430072, Hubei, China.

*Corresponding author(s). E-mail(s): zhouqian@whu.edu.cn; canhuaxu@fmmu.edu.cn;
Contributing authors: jianyingying2025@163.com; jiaxiaoyan_0919@163.com; zhanghan021695@163.com;

Abstract

Ocular diseases have emerged as the leading causes of blindness and low vision, necessitating timely detection and treatment. However, computer-aided approaches face significant challenges in accurately diagnosing these diseases. Specifically, ocular diseases often exhibit a long-tailed distribution, leading to a complex class-imbalanced scenario. Moreover, the coexistence of multiple diseases in a single patient gives rise to a problematic issue of label co-occurrence. In this study, we propose a novel alternate group training strategy as an effective approach to tackle the multi-label long-tailed data distribution problem. Firstly, we partition the long-tailed data into several groups based on semantic feature relations. This division helps reduce the challenges of class imbalance and label co-occurrence. With these groups established, we employ a gradient-based self-weighted loss to train a teacher network in an alternate way. Furthermore, a student model is trained on the original dataset under the guidance of the teacher network, utilizing a weighted class-balanced distillation loss. The class-balanced distillation loss also alleviates the class-wise imbalanced distribution and instance-wise label co-occurrence. Extensive experimental results have

demonstrated the superiority of our proposed method which achieves promising performance on the publicly available dataset. In addition, our approach achieves promising performance when expanding the single-teacher model to multiple-teacher models.

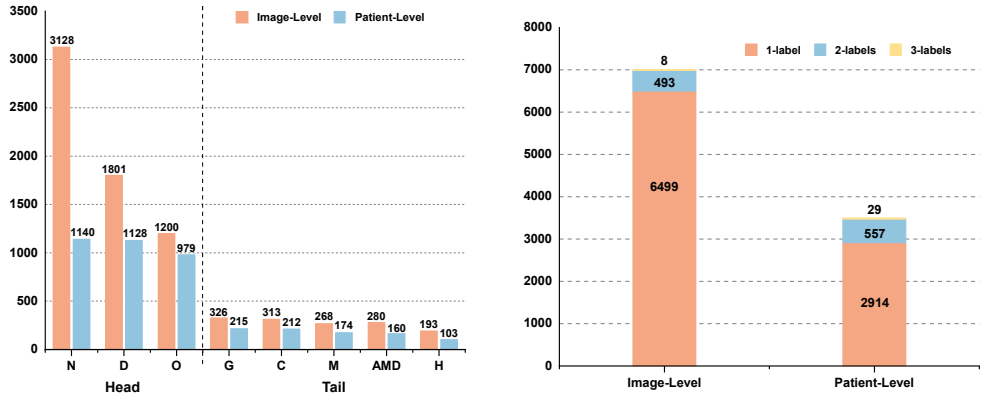
Keywords: Ocular disease recognition, Long-tailed classification, Knowledge distillation, Alternate group training, Re-weighting.

1 Introduction

Retinal diseases, such as uncorrected refractive errors and cataracts, significantly contribute to vision impairment and blindness [1], underscoring the critical need for their timely detection and effective treatment. Manual diagnosis of these conditions, however, is time-consuming and requires the expertise of skilled ophthalmologists to analyze retinal fundus images meticulously. This scenario has led to a surge in the research of automatic retinal disease screening methods, especially those using deep learning techniques. Notable advancements have been made in the detection of prevalent diseases like diabetic retinopathy (DR) and age-related macular degeneration (AMD), with considerable improvements in diagnostic accuracy [2, 3].

For instance, the automatic detection of DR, a leading retinal disease, has seen remarkable progress. Sungheetha *et al.* [4] developed a framework utilizing dense deep feature extraction for identifying Hard Executes (HE) spots in retinopathy images, aiding in DR severity prediction. Oh *et al.* [5] introduced a DR detection system leveraging deep learning for segmenting ultra-wide-field fundus photographs with U-Net [6], followed by classification using ResNet [7]. Similarly, Sugeno *et al.* [8] applied transfer learning with EfficientNet [9], pre-trained on ImageNet [10], for DR grading and lesion detection. The detection of AMD, a major cause of vision loss among the elderly, has also been addressed with innovative approaches, such as a multiscale CNN for classifying OCT images by Thomas *et al.* [11] and an automated detection system combining CNN and handcrafted features by Kadry *et al.* [12]. He *et al.* [13] proposed a novel AMD detection method using deep learning and an outlier detection algorithm.

Despite these strides, the diagnosis of rare retinal diseases remains a formidable challenge. The predominant focus of existing studies on common retinal diseases has resulted in effective performance under ideal, class-balanced binary classification scenarios. However, real-world clinical settings are far more intricate, often involving patients with multiple retinal diseases [14, 15] whose class-frequency distribution follows a long-tailed pattern [16, 17], meaning that a small number of disease categories contain a large proportion of samples while the majority of categories have significantly fewer instances. As illustrated in Fig. 1(a), this severe class imbalance results in a heavy-tailed distribution across disease categories. This scenario poses a unique challenge known as long-tailed multi-label classification, which most existing methods have yet to address adequately.



(a) (b)

Fig. 1: Statistics for the ODIR-5K [18] dataset. (a) is the number of eight diseases, *i.e.*, normal (N), diabetes (D), glaucoma (G), cataract (C), age-related macular degeneration (AMD), hypertension (H), myopia (M), and other diseases (O); (b) is the number of multi-label patients. Overall, the whole dataset exhibits a long-tailed distribution and an image may contain multiple labels.

Existing studies on long-tailed multi-label classification have proposed various techniques, including loss re-weighting [19], subset-based learning [20], and knowledge distillation [21], to mitigate class imbalance and label co-occurrence. While these approaches are effective from a modeling perspective, they are typically developed in a modular manner, with limited discussion of how different components interact during optimization, especially when deployed within a shared model. As a result, the training stability and optimization dynamics induced by combining these techniques remain insufficiently understood.

By assigning correlated labels to the same subset, multi-label samples are transformed into low-co-occurrence cases within each subset, while the effective class imbalance is reduced due to the narrower class-frequency range. However, this decomposition fundamentally alters the learning objective. Instead of optimizing the model with respect to the original joint label distribution, training is performed on a sequence of partial distributions, each corresponding to a subset-specific conditional view of the data. As a result, the effective optimization target becomes time-varying rather than stationary. Parameters that are optimal under one partial distribution may be suboptimal under another, leading to structured gradient interference during learning. In this setting, gradients computed from one subset may actively suppress parameters that are critical for other subsets, giving rise to systematic forgetting rather than random performance fluctuations.

This analysis reveals that long-tailed multi-label recognition is not merely a problem of data imbalance, but a structured optimization problem characterized by sequential learning and class-wise gradient interference. Without explicitly regulating these gradient interactions, techniques designed to reduce label co-occurrence

or rebalance classes may introduce new and unavoidable failure modes, ultimately undermining training stability. Motivated by this, we argue that subset grouping, loss re-weighting, and knowledge distillation should be designed as a coherent optimization framework, rather than as an incremental combination of independent components. Following this principle, we propose an alternate group training (AGT) and gradient-based re-weighting framework for long-tailed multi-label retinal disease classification, extending our previous work [22].

Our framework operates in three stages. First, the original long-tailed multi-label dataset is partitioned into multiple relational subsets using K-means clustering on semantic feature representations. Each subset corresponds to a partial view of the original joint label distribution, within which label co-occurrence is reduced and the effective class imbalance is alleviated due to the narrower class-frequency range. Unlike shot-based grouping strategies that rely solely on class frequency, relational grouping explicitly accounts for label dependency and semantic similarity, which are critical in multi-label long-tailed settings. In multi-label learning, labels are not independent: co-occurring labels induce coupled gradient updates during optimization, such that learning signals for one label directly influence the representation of others. Shot-based grouping ignores this coupling and may group labels with similar frequencies but conflicting semantic or visual patterns, thereby amplifying gradient interference rather than alleviating it. By contrast, clustering-based grouping aligns labels that frequently co-occur or share semantic characteristics into the same subset, ensuring that gradient updates within each subset are more coherent. This alignment reduces cross-label gradient conflict during alternate optimization and provides a more principled decomposition of the original joint label distribution. From a clinical perspective, this strategy is also well aligned with the observation that many retinal diseases exhibit correlated manifestations. For example, DR and AMD have been reported to share pathological associations [23, 24]. Grouping such correlated diseases into the same subset encourages the model to capture shared visual characteristics while suppressing spurious cross-label interference, leading to more stable and effective representation learning under alternate optimization.

Second, we train a single teacher model using an AGT strategy. The teacher model shares a common feature extractor across all subsets, while its classifier is optimized alternately on one subset at a time. In each training phase, the model is updated using samples drawn from the current subset, optionally augmented with a small proportion of samples from other subsets to preserve cross-group awareness. As a result, the training objective becomes non-stationary, since the model is sequentially optimized with respect to different partial label distributions. To stabilize optimization under this setting, we introduce a gradient-based re-weighting mechanism that explicitly balances positive and negative gradients at the class level. This mechanism prevents dominant classes from suppressing minority-class updates during alternate optimization and mitigates the objective inconsistency induced by subset-wise training.

Finally, the knowledge learned by the teacher model is transferred to a student model trained on the original long-tailed dataset via a class-balanced knowledge distillation loss. This distillation process aggregates subset-specific knowledge into a unified model while compensating for the class imbalance and label co-occurrence

inherent in the original data distribution. Through this design, relational grouping defines the optimization subspaces, AGT governs how these subspaces are explored, gradient-based re-weighting stabilizes the training dynamics, and knowledge distillation consolidates the learned knowledge into a single deployable model.

The key contributions of this work are as follows:

- We introduce a teacher-student model framework to effectively address the intricate challenge of long-tailed multi-label retinal disease recognition. This approach, building on our previous work [22], enhances training efficiency and explores disease relationships for improved representation learning.
- We utilize a simple clustering algorithm to identify relational subsets, offering an adaptable method for other domains without needing extensive prior knowledge.
- Our novel approach to mitigate catastrophic forgetting in alternate training strategies through sampling 'others' and applying a gradient-based self-weighted loss offers a potential solution in the realm of lifelong learning.
- Our extensive experiments demonstrate the effectiveness of our methodology, with theoretical and practical proofs of its superiority. Specifically, our single-teacher model trained on the complete dataset achieves a Kappa score of 0.707 ± 0.014 on the ODIR-5K dataset with a ResNet-50 backbone, which is comparable to the multi-teacher RLKD framework (0.712 ± 0.011) while requiring substantially lower training overhead. Furthermore, the proposed method can be seamlessly integrated with ensemble learning strategies, achieving an improved Kappa score of 0.722 ± 0.013 on the same dataset.

2 Related Work

2.1 Retinal Diseases Classification

Significant efforts have been dedicated to retinal disease recognition; however, most of these works primarily focus on a single or a few common diseases, with a particular emphasis on DR. Chen *et al.* [25] proposed a recognition pipeline based on deep convolutional neural networks to detect DR. Zhou *et al.* [26] proposed a sub-divisional algorithm of DR degree to identify the severity of DR based on fluorescein fundus angiography (FFA) images. Eladawi *et al.* [27] introduced a new computer-aided diagnosis system for detecting early-stage DR using optical coherence tomography angiography images. Some recent works take advantage of multi-task learning, in which the segmentation results of vessels or lesions are combined with the original image for a more accurate diagnosis [28–30]. Nair *et al.* [28] proposed a two-stage method for blood vessel segmentation and DR recognition, the classification results are obtained based on the features extracted from the segmented blood vessel.

In clinical practice, patients may suffer from multiple diseases and some of these diseases may be very rare, hence the aforementioned methods may be limited by poor performance in rare diseases when applied to multi-label disease classification. Several approaches have focused on multi-label classification. For instance, He *et al.* [31] proposed a patient-level multi-label ocular disease classification model based on convolutional neural networks taking both left and right color fundus images as input.

Ju *et al.* [32] presented a novel framework that leverages the prior knowledge in retinal diseases for training a more robust representation of the model under a hierarchy-sensible constraint. Cheng *et al.* [33] proposed a multi-label classification method based on the graph convolutional network, to detect eight types of fundus lesions in color fundus images. While these existing approaches achieve promising performance in multi-label disease classification, they still have limitations. The consideration of prior knowledge or relations among retinal diseases is not well-incorporated. Even if some methods like [32, 34] leverage the prior knowledge or relations among diseases, they require medical experts with extensive clinical experience to preprocess the data, such as manually dividing the original dataset into related subsets. In contrast, our approach takes an automated way to achieve this.

2.2 Long-Tailed Recognition

2.2.1 Re-sampling

As the dominant paradigm for long-tailed classification, re-sampling methods usually over-sample the tail classes and under-sample the head classes to make the account of all categories more balanced [35]. However, the over-sampling strategy may overfit the minority classes, and the under-sampling tends to harm the performance of the majority classes [16]. Ren *et al.* [36] presented balanced meta-softmax, an elegant unbiased extension of softmax, to estimate the optimal sampling rates of different classes for long-tailed learning. Ju *et al.* [32] developed an instance-wise class-balanced sampling technique to handle the prediction bias in long-tailed multi-label recognition. Re-sampling methods explicitly adjust the sampling frequency of different classes to achieve a more balanced distribution. Nevertheless, the artificially designed sampling strategy is not always effective and may even harm the performance of the network.

2.2.2 Cost-sensitive Learning

Cost-sensitive learning aims to re-balance classes with a well-designed loss function, also called re-weighting in some papers. Focal loss [37] was originally proposed for dense object detection, but could also be applied for long-tailed classification. Wu *et al.* [38] extended focal loss to a new loss function called Distribution-Balanced Loss for the multi-label recognition problems. Tan *et al.* [39] proposed equalization loss to solve long-tailed object detection by simply ignoring those gradients for rare categories. Equalization loss v2 [40] was further proposed to maintain balanced gradients between positives and negatives, with which a balanced classifier can be learned. Re-weighting approaches assign higher weights to minority classes, which can lead to the overfitting of these classes [17]. Moreover, they merely assign weights to samples based on class frequencies without considering the underlying characteristics or relationships between classes. As a result, these methods may not effectively capture the inherent complexities and correlations within the dataset.

2.2.3 Ensemble Learning

Utilizing ensemble learning in long-tailed classification is emerging. Ensemble learning approaches usually train multiple experts to solve long-tailed visual learning problems. Li *et al.* [41] proposed BAGS to balance the classifiers within the detection frameworks. Specifically, the classes are divided into several sub-groups according to the number of samples of each class. Then, different classification heads (*i.e.*, experts) and a shared feature extractor are trained on different data sub-groups. Without dataset division, Wang *et al.* [42] proposed RoutIng Diverse Experts (RIDE) to train multiple experts on the whole dataset through a KL-divergence loss, in which the model variance and model bias can be both reduced. Zhang *et al.* [43] developed Test-time Aggregating Diverse Experts (TADE) to train experts on different data distributions with skill-diverse expertise-guided losses so that the model can handle various unknown test distributions. However, ensemble-based methods generally lead to higher computational costs due to the use of multiple experts [42].

2.2.4 Large-Scale Pretraining and Vision-Language Models

Recent advances in large-scale pretraining and vision-language models (VLMs) have influenced medical image analysis. Contrastive learning frameworks such as MedCLIP [44] and BiomedCLIP [45] align medical images with textual descriptions to learn transferable representations from large-scale corpora. Instruction-tuned multimodal systems, including LLaVA-Med [46], A^2M-Diff [47], and Med-Flamingo [48], further extend such models to medical image understanding and visual question answering tasks. By leveraging semantic priors from large datasets, these approaches aim to improve generalization under limited supervision.

In long-tailed scenarios, large-scale pretraining may alleviate class imbalance by transferring high-level semantic knowledge to rare categories. However, their effectiveness relies on the availability of large image–text corpora, while specialized retinal datasets often lack sufficiently rich paired annotations, particularly for rare diseases. Moreover, these approaches primarily enhance representation learning and do not explicitly regulate class-wise gradient dynamics during downstream optimization. As a result, gradient imbalance under long-tailed multi-label distributions may still persist after fine-tuning on limited datasets.

Compared with large multimodal models that typically require substantial computational resources, lightweight optimization-driven strategies remain important for practical deployment in clinical environments. In contrast to large-scale pretraining paradigms, our method focuses on explicitly organizing label relationships and regulating gradient imbalance under long-tailed multi-label settings.

3 Method

3.1 Problem Formulation

Denote the long-tailed dataset as $D = \{(X, Y)\}$, where $X = \{x_1, x_2, \cdots, x_n\}$ is the training images, $Y = \{y_1, y_2, \cdots, y_n\}$ is the corresponding labels, and n is the number of training samples. In our case, the dataset is multi-label, meaning that each sample

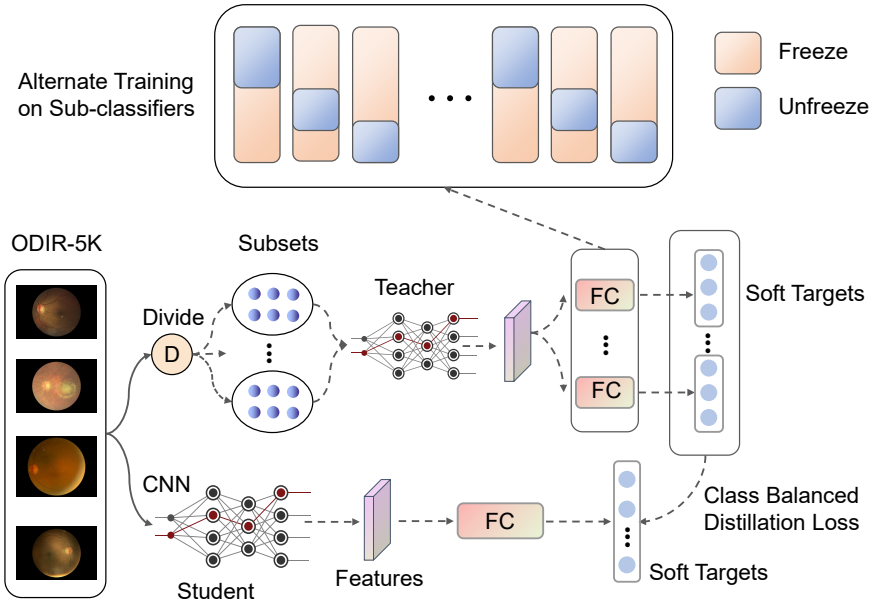


Fig. 2: Pipeline of the proposed framework. The teacher model is trained on divided subsets alternately and will be distilled into a unified student model with the class-balanced distillation loss.

can have multiple labels. Let C represent the number of classes. For each label vector y_i, we have $y_i = \{y_i^1, y_i^2, \cdots, y_i^C\}$, where each element $y_i^k \in \{0, 1\}$. In this binary representation, $y_i^k = 1$ indicates the presence of label k for i-th image, while $y_i^k = 0$ indicates its absence. It is worth noting that the number of training samples is less than or equal to the total number of samples across all categories, since one sample may be counted several times if it belongs to multiple classes.

3.2 Framework Overview

As shown in Fig. 2, our framework contains a teacher model and a student model. The teacher model is trained on the relational subsets (groups) using the AGT strategy and the gradient-based re-weighting loss. The teacher model aims to learn general representations and balanced classifiers for relational classes. The student model is guided by the teacher model and trained on the original long-tailed multi-label dataset. During the teacher model training, the classifier is divided into several parts with respect to the relational subsets and trained alternately. This means that while training one part, the others are kept frozen. To mitigate the impact of catastrophic forgetting caused by the integration of data from different subsets, we introduce "other" samples, which represent samples from subsets other than the current one. Furthermore, we re-weight the logits (*i.e.*, outputs of classifier) according to the gradients. Finally, the teacher model will transfer the learned knowledge to the student model using knowledge distillation. In such a design, the teacher model achieves a general and balanced

performance across all classes, but may not excel in the head classes. On the other hand, the student model performs well in the head classes as it is directly trained on the original long-tailed dataset. Additionally, benefiting from the guidance of the teacher model, the student model shows enhanced performance in the tail classes.

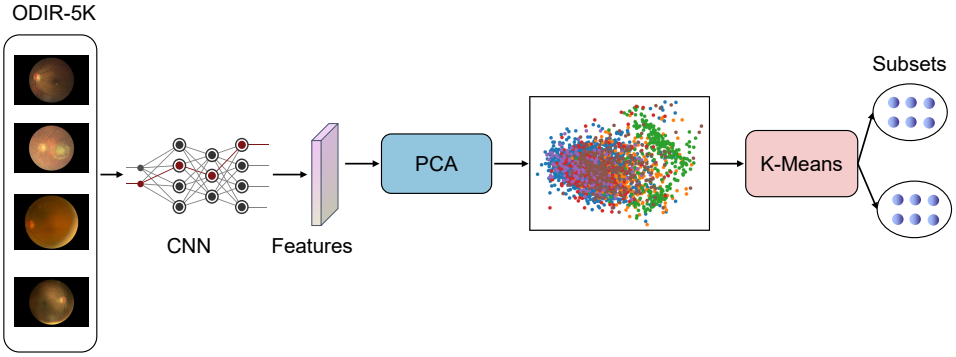


Fig. 3: Diagram of relational subsets generation. We only perform this pipeline on six labels (*i.e.*, D, G, C, AMD, H, M).

3.3 Relational Subsets Division

Many ocular diseases share similar semantic features. For example, diabetic retinopathy and glaucoma can both lead to abnormal retinal blood vessels. Therefore, there are strong relations among ocular diseases. As can be seen in Fig. 3, an automated algorithm is utilized to group the diseases which share similar semantics into the same subsets. And then we train the teacher model on these subsets so that both common shared and unique features can be learned.

Specifically, we use a ResNet-50 [7] network pre-trained on ImageNet as the feature extractor and obtain 2048-dimensional feature representations for each sample. To stabilize the subsequent clustering process, we apply principal component analysis (PCA) as a preprocessing step to reduce noise and redundancy in the high-dimensional feature space while preserving its dominant variance structure. PCA is fitted once on the full set of extracted features to ensure a globally consistent transformation. We retain the smallest number of principal components that preserve at least 95% of the total explained variance, which corresponds to 50 principal components in our implementation.

The dimension-reduced features are then clustered using the K-Means algorithm [49]. Clustering is performed on the entire dataset to obtain a global and consistent grouping of diseases based on their feature distributions. To verify the robustness of the relational subset division, we repeat K-Means with different random initializations and observe highly consistent grouping results. We also vary the number of retained principal components within a reasonable range (*i.e.*, 30 to 100) and

find that the overall clustering groups remains stable, indicating that the subset division is not sensitive to specific PCA configurations. Finally, based on the clustering results in the latent space, diseases are grouped into relational subsets for subsequent alternate group training.

The experiments are conducted on ODIR-5K [18], which includes normal (N), diabetes (D), glaucoma (G), cataract (C), age-related macular degeneration (AMD), hypertension (H), myopia (M), and other diseases (O). Since label 'N' (normal) contains no diseases and label 'O' (other) contains more than one disease, moreover, they all belong to the head classes, we group 'N' and 'O' into one subset. The left six diseases are divided into two subsets according to the clustering results, and the final subsets are {D, AMD, H, M}, {G, C}, {N, O}, respectively.

Long-tailed multi-label classification poses two main challenges: label co-occurrence and class imbalance. The training of the teacher model on relational subsets helps alleviate these difficulties. For example, for a retinal image x with label $y_1 = 1$ and $y_2 = 1$, if y_1 and y_2 are assigned to different subsets, x becomes a single-label image in each subset. This approach effectively addresses the issue of label co-occurrence. Moreover, the class imbalance ratio of each subset (defined as the ratio between the largest and smallest number of samples: $\rho = \frac{N_{max}}{N_{min}}$) can be less than or equal to the original long-tailed dataset.

Importantly, this relational grouping strategy explicitly accounts for semantic similarity and label dependency among diseases, which are critical in multi-label learning. In multi-label settings, co-occurring labels induce coupled gradient updates during optimization; grouping semantically related diseases into the same subset aligns these gradients and reduces cross-label interference. This provides a more principled alternative to shot-based grouping strategies that rely solely on class frequency and ignore label dependency.

3.4 Teacher Model Training

Fig. 4 provides an overview of the proposed teacher model training framework. The training process proceeds in a cyclic group-wise manner. After the original dataset is divided into multiple relational subsets $\{g_1, g_2, \ldots, g_n\}$, each subset is associated with a dedicated sub-classifier, while all sub-classifiers share a common feature extraction backbone. Each sub-classifier is responsible for predicting the disease labels belonging to its corresponding subset. During training, the subsets are processed sequentially. At a given training stage, one subset g_i is selected as the current group. A mini-batch is then constructed by sampling $(1 - \alpha)$ of the samples from g_i and α of the samples from all remaining subsets. The samples from g_i are used to update the corresponding sub-classifier c_i, while the samples from other subsets provide auxiliary supervision for the remaining sub-classifiers and help preserve previously learned knowledge. For the current mini-batch, only the parameters of the shared backbone and the active sub-classifier c_i are updated. All other sub-classifiers are kept frozen and do not receive parameter updates at this stage. The model then moves to the next subset and repeats the same procedure, forming a cyclic training process over all subsets. To further stabilize training under this sequential optimization scheme, gradient-based re-weighting is applied when computing the loss. For each class, the contributions of positive and

negative samples in the current mini-batch are measured, and a class-specific weight is assigned to balance their gradients. These weights are incorporated into a weighted binary cross-entropy loss, which is used to update the active sub-classifier and the shared backbone. By explicitly controlling gradient magnitudes at the class level, this mechanism mitigates interference between subsets and reduces catastrophic forgetting during cyclic training.

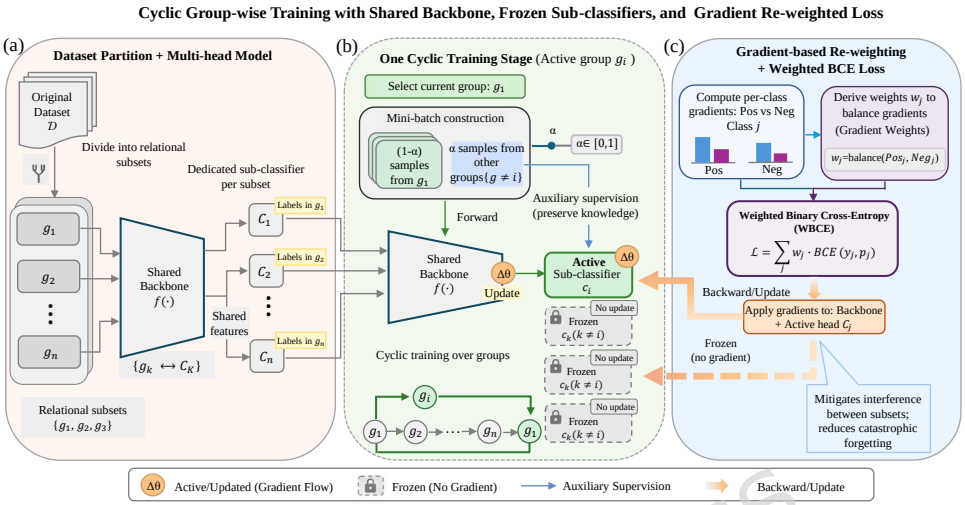


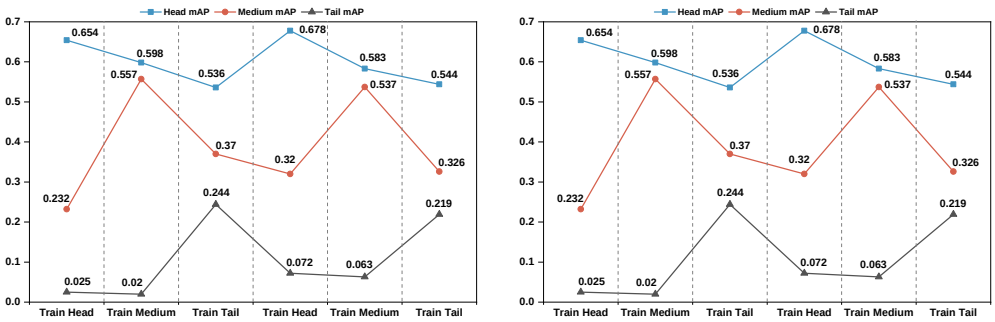
Fig. 4: Cyclic group-wise teacher model training framework. (a) Dataset partition and multi-classifier model with a shared backbone. (b) One cyclic training stage with an active subset and frozen inactive sub-classifiers. (c) Gradient-based re-weighting with weighted loss for stable optimization.

3.4.1 Alternate Group Training

To train the teacher model efficiently, we adopt only one expert as the teacher model instead of multiple experts used in other approaches [22, 34, 42, 43, 50]. After relational subsets are obtained, the classifier of the teacher model is partitioned into multiple sub-classifiers, each corresponding to one relational subset, while all sub-classifiers share a common backbone for feature extraction.

Formally, given relational subsets $\{g_1, g_2, \cdots, g_n\}$, we construct sub-classifiers $\{c_1, c_2, \cdots, c_n\}$, where each c_i is responsible for predicting the disease labels belonging to subset g_i. AGT proceeds in a cyclic manner. At each training stage, one subset g_i is selected as the current group, and only the corresponding sub-classifier c_i is activated. The shared backbone and c_i are updated using samples from g_i, while all other sub-classifiers remain frozen.

The backbone is not frozen during training, as data imbalance primarily affects the classifier layers and does not significantly hinder the learning of shared representations [51]. Through cyclic optimization over all subsets, the teacher model is expected to learn generalized feature representations together with more balanced subset-specific classifiers.



(a) (b)

Fig. 5: Validation accuracy on the CIFAR100-LT dataset when using AGT. (a) is AGT without sampling others and (b) is AGT with sampling others. mAP is the mean Average Precision.

However, optimizing the model solely with samples from the current group introduces a sequential optimization effect, where updating the model for one subset can adversely affect the performance on previously learned subsets. To verify this, we conduct diagnostic experiments on the CIFAR100-LT dataset [52]. CIFAR100-LT provides a standard long-tailed classification benchmark where classes can be naturally divided into head, medium, and tail groups according to their sample frequencies. In this experiment, the classification model is trained by alternately optimizing on head, medium, and tail class groups, following the same alternate training strategy used for relational subsets. As shown in Fig. 5a, this phenomenon leads to severe performance fluctuations during training, which is commonly referred to as catastrophic forgetting. This observation motivates the introduction of additional stabilization mechanisms, described in the following subsections.

Algorithm 1 illustrates the base version of AGT without additional stabilization mechanisms. In practice, this procedure is extended by incorporating other-group sampling (Algorithm 2) and gradient-based re-weighting (Algorithm 3), as summarized in Fig. 4.

3.4.2 Sampling "Others"

Motivated by BAGS [41], we introduce samples from "other" subsets into each training mini-batch to alleviate catastrophic forgetting induced by alternate group training. Let $\alpha \in [0, 1]$ denote the sampling ratio and b the batch size. For the current subset

Algorithm 1 Alternate Group Training (Base Version)

Input: Long-tailed dataset D, teacher model T with backbone B and classifier C, total epochs E

Generate relational subsets $\{g_1, g_2, \cdots, g_n\}$
Partition classifier C into sub-classifiers $\{c_1, c_2, \cdots, c_n\}$
for *epoch* $\leftarrow 1$ **to** n **do**
  for each subset $g_i \in \{g_1, g_2, \cdots, g_n\}$ **do**
    Freeze all sub-classifiers $\{c_k\}_{k=1}^{n}$
    Activate sub-classifier c_i (unfreeze c_i)
    Update backbone B and c_i using samples from g_i
  end for
end for

Output: Trained teacher model T

g_i, we sample $b \times (1-\alpha)$ instances, while the remaining $b \times \alpha$ instances are sampled from all other subsets using class-balanced sampling.

During alternate group training, samples from the current subset g_i provide primary supervision for the active sub-classifier c_i. In contrast, samples from other subsets are not used to update their corresponding sub-classifiers, which remain frozen, but instead act as auxiliary constraints through the shared backbone. By introducing these auxiliary samples, the model is discouraged from excessively suppressing features and decision boundaries associated with previously learned subsets.

From an optimization perspective, this sampling strategy partially counteracts the sequential bias of alternate training by maintaining gradient signals related to inactive subsets at the representation level. As a result, cross-group awareness is preserved during cyclic training, leading to more stable optimization dynamics. This effect is empirically verified in Fig. 5b, where the validation performance on CIFAR100-LT becomes significantly more stable after incorporating samples from other subsets.

Algorithm 2 Alternate Group Training with Sampling "Others"

Input: Dataset D, teacher model T with backbone B and sub-classifiers $\{c_k\}_{k=1}^{n}$, total epochs E, batch size b, sampling ratio α

for *epoch* $\leftarrow 1$ **to** E **do**
  for each subset $g_i \in \{g_1, g_2, \cdots, g_n\}$ **do**
    Freeze all sub-classifiers $\{c_k\}_{k=1}^{n}$
    Activate sub-classifier c_i
    Sample $b \times (1-\alpha)$ instances from g_i
    Sample $b \times \alpha$ instances from $\{g_k\}_{k \neq i}$ using class-balanced sampling
    Construct mini-batch $\mathcal{I}$ from the combined samples
    Update backbone B and active sub-classifier c_i using $\mathcal{I}$
  end for
end for

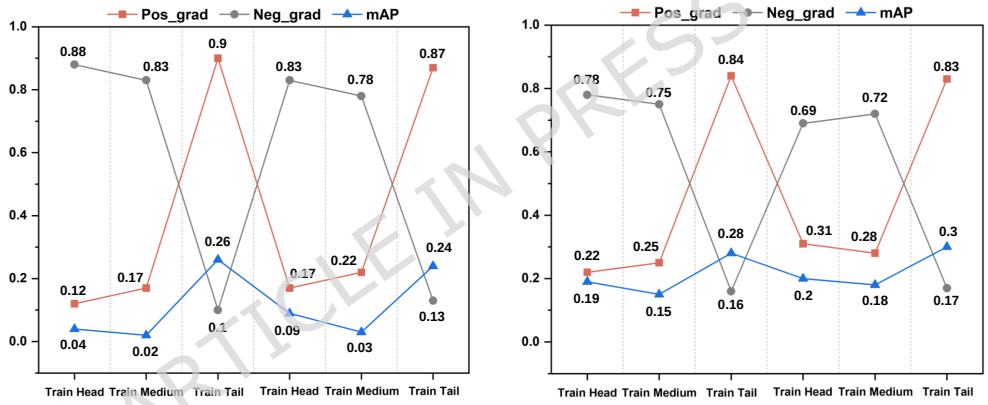
Output: Updated teacher model T

Algorithm 2 summarizes the AGT procedure augmented with sampling "others", which extends the base version in Algorithm 1 by modifying the mini-batch construction while keeping the update rule for sub-classifiers unchanged.

3.4.3 Gradient-based Re-weighting

While sampling "others" alleviates catastrophic forgetting at the group level, it does not fully resolve gradient imbalance at the class level, especially for tail categories with very limited positive samples. Following Equalization Loss (EQL) [40], we attribute this limitation to the imbalance between positive and negative gradients during optimization. However, unlike standard EQL, our mechanism specifically targets the sequential gradient imbalance inherent in alternate group training, effectively serving as a cross-group awareness constraint.

Under alternate group training, the optimization objective becomes inherently sequential. When training on a current subset g_i, samples belonging to g_i generate positive gradients for the active sub-classifier c_i, while simultaneously producing negative gradients for all inactive sub-classifiers. Conversely, samples from other subsets introduce competing gradients through the shared backbone. Although sampling "others" partially balances these effects across groups, it does not explicitly regulate the magnitude of gradients at the class level. As a result, for tail classes with scarce positive samples, negative gradients can still dominate, leading to suppression of their corresponding classifiers and reintroducing catastrophic forgetting.



(a) (b)

Fig. 6: Gradient dynamics of tail classes during alternate group training. (a) Gradient evolution without gradient-based re-weighting. (b) Gradient evolution with the proposed gradient-based re-weighting. The curves illustrate the positive gradients, negative gradients, and tail-class mAP across training on head, medium, and tail subsets.

As shown in Fig. 6a, for tail classes, the negative gradients dominate the learning signal, suppressing the classifier for these rare classes. This imbalance leads to catastrophic forgetting, especially for the tail categories. To address this, we introduce a class-wise gradient-based re-weighting mechanism, which adaptively amplifies under-represented positive gradients while suppressing dominant negative gradients, thereby balancing the gradient contributions across classes during training.

To explicitly control gradient magnitudes at the class level, we introduce a class-wise gradient-based re-weighting mechanism as shown in Algorithm 3. For each class j in a mini-batch $\mathcal{I}$, the positive and negative gradients with respect to the loss $\mathcal{L}$ are computed as:

$$\nabla_j^{pos}(\mathcal{L}) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y_j^i(p_j^i - 1), \quad \nabla_j^{neg}(\mathcal{L}) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}(1 - y_j^i)p_j^i \tag{1}$$

where p_j^i is the predicted probability of category j for the i-th instance, and y_j^i is the ground truth.

The imbalance ratio between positive and negative gradients is then defined as:

$$r_j = \frac{|\nabla_j^{pos}|}{|\nabla_j^{neg}| + \epsilon} \tag{2}$$

where ϵ is a small constant to prevent division by zero. For tail classes, r_j is typically very small because the cumulative magnitude of positive gradients is far outweighed by the overwhelming negative gradients from other categories.

To prevent these dominant negative gradients from suppressing the classifiers of rare categories, we define a class-wise negative weight w_j as:

$$w_j = \min(1, r_j^{\beta}) \tag{3}$$

where $\beta > 0$ is a hyperparameter that controls the intensity of suppression.

When r_j is small (severe imbalance), w_j becomes small, significantly reducing the contribution of negative gradients for class j. When r_j is large (mild imbalance), w_j approaches 1, leaving the class largely unaffected. Thus, the proposed mechanism selectively suppresses dominant negative gradients while preserving positive learning signals. The hyperparameter β determines how sensitively the suppression factor responds to gradient imbalance. A smaller β leads to a smoother and more moderate adjustment, while a larger β increases the curvature of the mapping and results in stronger suppression for severely imbalanced classes. In our experiments, we set $\beta = 2$ as a trade-off between responsiveness and training stability. Empirically, smaller values (*e.g.*, $\beta = 1$) were insufficient to correct strong imbalance, whereas larger values tended to over-suppress negative gradients and cause unstable oscillations during cyclic optimization.

To illustrate this behavior, we provide a numerical example. Consider a mini-batch of size 10.

For a tail class j with only 1 positive sample ($p = 0.6$) and 9 negative samples ($p = 0.2$): The average positive gradient magnitude is $|\nabla_j^{pos}| = \frac{1}{10} \times |0.6 - 1| = 0.04$. The average negative gradient magnitude is $|\nabla_j^{neg}| = \frac{1}{10} \times 9 \times 0.2 = 0.18$. The imbalance ratio is $r_j = 0.04/0.18 \approx 0.22$. With $\beta = 2$, the weight becomes $w_j = (0.22)^2 \approx 0.05$, which significantly suppresses the overwhelming negative gradients by 95%. In contrast, for a head class with 5 positive samples ($p = 0.7$) and 5 negative samples ($p = 0.1$): The magnitudes yield $|\nabla_j^{pos}| = \frac{1}{10} \times (5 \times |0.7 - 1|) = 0.15$ and $|\nabla_j^{neg}| = \frac{1}{10} \times (5 \times 0.1) = 0.05$. The ratio $r_j = 0.15/0.05 = 3.0$. Applying the clipping mechanism in Eq. (3), $w_j = \min(1, 3.0^2) = 1.0$.

This example demonstrates that the proposed mechanism adaptively safeguards tail classes from excessive suppression while ensuring that well-balanced head classes maintain their standard optimization gradients.

The weighted binary cross-entropy loss is then formulated as:

$$\mathcal{L} = -\sum_{i}^{N}\sum_{j}^{C}[y_j^i log p_j^i + w_j(1 - y_j^i)log(1 - p_j^i)] \tag{4}$$

This weighting strategy adaptively amplifies under-represented positive gradients while suppressing dominant negative gradients, thereby balancing gradient contributions across classes during cyclic optimization. The weighted binary cross-entropy loss is then used to update the active sub-classifier and the shared backbone, effectively mitigating catastrophic forgetting at the class level and complementing the group-level stabilization introduced by sampling "others". Fig. 6 further highlights the impact of this approach, showing that, without gradient re-weighting, the tail class gradients are heavily dominated by negative values, leading to a lower mAP. After applying the gradient re-weighting, the tail class gradients are more balanced, and the mAP improves significantly, reflecting reduced catastrophic forgetting and better representation of rare classes.

Algorithm 3 Loss Computation with Gradient-based Re-weighting

Input: Mini-batch $\mathcal{I}$, predicted probabilities $\{p_j^i\}$, labels $\{y_j^i\}$, hyperparameters β, ϵ
Output: Weighted loss $\mathcal{L}$

1: **for** each class $j = 1$ to C **do**
2: Compute ∇_j^{pos} and ∇_j^{neg} on $\mathcal{I}$ using Eq. (1)
3: Compute imbalance ratio r_j using Eq. (2)
4: Compute class-wise weight w_j using Eq. (3)
5: **end for**
6: Compute weighted binary cross-entropy loss $\mathcal{L}$ using Eq. (4)

Based on the above components, the complete teacher training strategy is formulated as a unified cyclic optimization procedure. Specifically, alternate subset updates, other-group sampling, and gradient-based re-weighting are jointly integrated within each mini-batch update. The overall training process is summarized in Algorithm 4.

Algorithm 4 Cyclic Alternate Group Training with Sampling and Gradient Re-weighting

Input: Long-tailed dataset D, teacher model T with backbone B and sub-classifiers $\{c_k\}_{k=1}^{n}$, total epochs E, batch size b, sampling ratio α
Output: Trained teacher model T

for *epoch* $\leftarrow 1$ **to** E **do**
  for each subset $g_i \in \{g_1, g_2, \cdots, g_n\}$ **do**
    Freeze all sub-classifiers $\{c_k\}_{k=1}^{n}$
    Activate sub-classifier c_i
    Sample $b \times (1-\alpha)$ instances from g_i
    Sample $b \times \alpha$ instances from $\{g_k\}_{k \neq i}$ using class-balanced sampling
    Construct mini-batch $\mathcal{I}$ from the combined samples
    for each class $j = 1$ to C **do**
      Compute ∇_j^{pos} and ∇_j^{neg} using Eq. (1)
      Compute imbalance ratio r_j using Eq. (2)
      Compute class-wise weight w_j using Eq. (3)
    end for
    Compute weighted binary cross-entropy loss using Eq. (4)
    Update backbone B and active sub-classifier c_i using the weighted loss
  end for
end for

3.5 Knowledge Distillation

After the teacher model is trained on generated subsets, we can distill it into a uniform student model. Specifically, for images from the i-th class, let z_i and $\hat{z}_i$ be the output logits of the teacher model and the student model, respectively, then we can get the soft targets q_i and $\hat{q}_i$ as:

$$q_i = \frac{exp(z_i/T)}{\sum_{0,1} exp(z_i/T)}, \quad \hat{q}_i = \frac{exp(\hat{z}_i/T)}{\sum_{0,1} exp(\hat{z}_i/T)} \tag{5}$$

where T is the distillation temperature and is usually set to be greater than 1 to increase the weight for smaller probabilities. And the distillation loss is defined as:

$$L_{KD_i} = KL(\hat{q}_i || q_i) = \hat{q}_i log \frac{\hat{q}_i}{q_i} \tag{6}$$

After that, we aim to integrate the knowledge distillation losses L_{KD_i} for each class into a unified loss. The straightforward approach is to sum them together; however, this may introduce issues related to label co-occurrence. Regarding a multi-label image x with labels $y_1 = 1$ and $y_2 = 1$, if y_1 and y_2 are assigned to different subsets, then x will be sampled twice to train the sub-classifiers. This means that multi-label images may be sampled multiple times, and since most multi-label images belong to the head classes, this exacerbates the class imbalance issue. Given our focus on the tail classes

while aiming to avoid negatively impacting the head classes, it would be unfair to distill all samples at the same rate and weight.

To solve this problem, we introduce a class-balanced distillation loss with respect to each sample. This allows us to assign appropriate weights to the samples during the distillation process, taking into account the class imbalance issue. By doing so, we aim to prioritize the tail classes while ensuring that the head classes are not adversely affected. This class-balanced distillation loss provides a more equitable and effective approach for knowledge transfer and model training. More specifically, for the class-balanced sampling strategy without label co-occurrence, all classes are assigned with an equal sampling probability $p_i = \frac{1}{C}$, where C is the number of classes. For each sample in the i-th class, they have the same probability $p_j = \frac{1}{N_i}$ to be sampled, where N_i is the number of i-th class samples. Therefore, the instance-level sampling probability of j-th image in i-th class is $p_i^j = \frac{1}{C}\frac{1}{N_i}$. When considering the multi-label sample, the instance-level sampling probability becomes the sum of each positive class i it contains, *i.e.*, $\hat{p_i}^j = \frac{1}{C}\sum_{y_k=1}\frac{1}{N_i}$. Then, we define a class-balanced weight for each sample denoted as $w_i^j = \frac{p_i^j}{\hat{p_i}^j}$. Finally, the class-balanced distillation loss is given as:

$$L_{KD} = \sum_{i}^{C}\sum_{j}^{N_i} KL(\hat{q}_i^j||q_i^j)\cdot w_i^j = \sum_{i}^{C}\sum_{j}^{N_i}\hat{q}_i^j log\frac{\hat{q}_i^j}{q_i^j}\cdot w_i^j \tag{7}$$

The final loss to train student model on the whole long-tailed dataset is as follows:

$$L = L_{BCE} + \gamma L_{KD} \tag{8}$$

where L_{BCE} is the binary cross entropy loss between the predicted labels of the student model and the ground truth, γ is a hyper-parameter to weight the class-balanced distillation loss and set to 1 in our experiments.

To provide an intuitive example, consider a three-class problem with $C = 3$, where the number of samples in each class is $N_1 = 1000$, $N_2 = 200$, and $N_3 = 50$. For a single-label sample from class 3, the instance-level sampling probability is computed as $p_3^j = \frac{1}{C}\times\frac{1}{N_3} = \frac{1}{3}\times\frac{1}{50} = 0.00667$ The weight for this single-label sample is 1, since there is no competition from other labels: $w_i^j = \frac{p_3^j}{p_3^j} = 1$. In contrast, for a multi-label sample containing both class 2 and class 3, the instance-level sampling probability is the sum of the probabilities for each class: $\hat{p}_i^j = \frac{1}{C}\left(\frac{1}{N_2}+\frac{1}{N_3}\right) = \frac{1}{3}\left(\frac{1}{200}+\frac{1}{50}\right) = 0.00833$. The corresponding weight for this multi-label sample is: $w_i^j = \frac{p_3^j}{\hat{p}_i^j} = \frac{0.00667}{0.00833} = 0.8 < 1$. Thus, multi-label samples are assigned a weight less than 1, down-weighting their influence during training. This ensures that multi-label samples, especially those dominated by head classes, do not overwhelm the distillation process.

3.6 Training Dynamics under Long-Tailed Multi-Label Learning

The proposed framework is motivated by the optimization characteristics of long-tailed multi-label learning. In this setting, challenges arise not only from label co-occurrence but also from severe class-frequency imbalance. In this subsection, we provide a brief analysis of these training dynamics to clarify the roles of relational grouping, alternate group training, and gradient-based reweighting.

3.6.1 Joint objective under multi-label learning

For multi-label classification, the overall loss can be written as:

$$L(\theta) = \sum_{j=1}^{C} L_j(\theta) \tag{9}$$

where L_j denotes the binary classification loss for class j, and all classes share the same backbone parameters θ. The gradient update is therefore as:

$$\nabla_\theta L(\theta) = \sum_{j=1}^{C} \nabla_\theta L_j(\theta) \tag{10}$$

When multiple labels co-occur in a sample, their gradients are accumulated through the shared representation:

$$\nabla_\theta L = \nabla_\theta L_{j_1} + \nabla_\theta L_{j_2} \tag{11}$$

Since the backbone is shared, gradients for different labels are generally coupled,*i.e.*,

$$\nabla_\theta L_{j_1} \cdot \nabla_\theta L_{j_2} \neq 0 \tag{12}$$

This coupling may lead to either reinforcement or interference depending on gradient alignment.

3.6.2 Effect of long-tailed distributions

Under long-tailed class distributions, head classes contain substantially more samples than tail classes. As a result, the gradients for head samples may be greater than those for tail samples:

$$\|\nabla_\theta L_{\text{head}}\| \gg \|\nabla_\theta L_{\text{tail}}\| \tag{13}$$

Even when tail classes appear in training, their effective contribution to parameter updates can be overwhelmed by gradients from head classes. In multi-label settings, head-class samples additionally generate negative gradients for tail classifiers, further amplifying this imbalance. This phenomenon is closely related to gradient starvation [53], where minority classes receive insufficient effective updates during optimization.

3.6.3 Sequential objective under alternate group training

In alternate group training, optimization is performed cyclically on relational subsets $\{g_1, \ldots, g_n\}$. At each stage, the effective objective becomes

$$L_{g_i}(\theta) = \sum_{j \in g_i} L_j(\theta) \tag{14}$$

which represents only a partial view of the full multi-label objective. Although the overall task remains unchanged, the optimization target shifts across stages. Parameters adapted to one subset may not remain optimal for others, introducing a sequentially changing objective during training.

Relational grouping mitigates this issue by clustering semantically related labels, which tends to align gradient directions within each subset. In contrast, frequency-based grouping considers only class counts and does not explicitly account for gradient compatibility. Classes with similar frequencies may still induce conflicting updates, whereas semantically correlated labels may produce more consistent gradients even if their frequencies differ.

3.6.4 Gradient-level re-balancing

To further address the imbalance induced by long-tailed distributions, we introduce a class-wise re-weighting scheme based on gradient statistics. Let ∇_j^{pos} and ∇_j^{neg} denote the positive and negative gradient components for class j. When $\|\nabla_j^{neg}\| \gg \|\nabla_j^{pos}$, the model updates for class j are dominated by negative signals, suppressing learning for rare categories. The weighting factor w_j adjusts negative gradients according to their gradient imbalance. By regulating gradient magnitudes rather than sample counts, the proposed mechanism stabilizes optimization under both long-tailed distributions and sequential group-wise training.

Overall, long-tailed multi-label learning can be understood as a coupled multi-objective optimization problem characterized by gradient interference, class-wise imbalance, and objective shifting. The proposed framework regulates these dynamics at both the grouping level and the gradient level, leading to more stable and balanced training.

3.7 Comparison with Previous Work in Methodology

This work is an extension of our previous work [22], which focuses on leveraging the relationships among retinal diseases through multi-task pre-training, region-based attention, and relation subsets division. In [22], a unified network is used for pre-training the backbone by combining segmentation and classification tasks on FGADR dataset [54]. Additionally, an automatic division of the original dataset into relational subsets is performed, followed by training multiple teacher models using a region-based mechanism to address label co-occurrence and class imbalance. Finally, the knowledge from these teacher models is distilled into a student model using a class-balanced distillation loss.

In this work, we adopt a shared backbone network and multiple sub-classifiers to design the teacher network. The teacher model is trained using an alternate training strategy on the entire dataset, whereas in [22], individual teacher models are trained on separate subsets without any interaction between them. This variation in training strategy significantly reduces the training cost of this method, particularly as the number of classes and subsets increases. Furthermore, our method allows for seamless integration with existing ensemble learning frameworks, overcoming the limitations faced by the training approach used in [22]. As a result, our approach achieves superior performance in comparison.

3.8 Implementation Details

To validate the effectiveness of the proposed method, we have conducted extensive experiments implemented with Pytorch and 4×RTX 3090 GPUs. The input color fundus images are resized to 512×512. We randomly crop 448×448 patches for training. For testing, center cropping of 448×448 patches is performed instead. The Adam optimizer is adopted with an initial learning rate of 0.001 and $\beta_1 = 0.9$, $\beta_2 = 0.99$. The mini-batch size is set to 64 and the Distillation temperature T is set to 5. We set α to 0.5 and β to 2 by default. Besides, we randomly split all samples to 80% for training and 20% for testing. All experiments are conducted with 5-fold cross-validation to produce more solid results. To avoid patient-level data leakage, all train/test splits are performed at the patient level rather than the image level. Unless otherwise specified, all methods are initialized with backbone networks pre-trained on the FGADR [54] dataset following the setting in RLKD [22].

4 Experiments

4.1 Datasets and Metrics

We have compared the results on the ODIR-5K [18] dataset and the RFMiD dataset [55] with other approaches. ODIR-5K consists of 7000 images with patient-level annotations and image-level diagnostic keywords. The ocular diseases include normal (N), diabetes (D), glaucoma (G), cataract (C), age-related macular degeneration (AMD), hypertension (H), myopia (M), and other diseases (O). We relabel the ODIR-5K to obtain image-level annotations according to the image-level diagnostic keywords, and all metrics are calculated based on the re-labeled image-level annotations. RFMiD is a retinal fundus dataset consisting of 3200 images annotated with 46 disease categories, exhibiting significantly more severe class imbalance and more frequent multi-label co-occurrence than ODIR-5K. Following common practice [56, 57], disease categories with fewer than 10 samples are merged into an "others" class, resulting in a total of 29 classes. The official data split is adopted, with 1920 images for training, 640 images for validation, and 640 images for testing. Since the test set annotations are not publicly available, all quantitative evaluations on RFMiD are conducted on the validation set. We use Cohen's kappa coefficient, F_1 score, and area under the receiver operating curve (AUC) as the metrics for the evaluation of

multi-label retinal disease recognition. These metrics are calculated as:

$$\begin{aligned} \text{kappa} &= \frac{p_o - p_e}{1 - p_e} \\ p_o &= \frac{\sum_{c=1}^{C} \text{TP}_c}{\sum_{c=1}^{C} (\text{TP}_c + \text{FN}_c)} \\ p_e &= \frac{\sum_{c=1}^{C} \text{TP}_c * (\text{TP}_c + \text{FN}_c)}{N^2} \end{aligned} \tag{15}$$

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \tag{16}$$

$$\begin{aligned} \text{AUC} &= \int_{x=0}^{1} \text{TPR}\left(\text{FPR}^{-1}(x)\right) \text{dx} \\ \text{TPR} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \end{aligned} \tag{17}$$

where C is the total number of classes, *i.e.*, 8 in this paper. TP, FP, TN, and FN are true positive predictions, false positive predictions, true negative predictions, and false negative predictions, respectively. TPR and FPR refer to true positive rate and false positive rate.

4.2 Comparison Results

The compared methods include exsiting methods which aim to recognize multi-label retinal diseases like CCT-Net [58], SCFKD [59], and other algorithms considering long-tailed classification such as Instance-Balanced Sampling, Class-Balanced Sampling, Focal Loss [37], RSKD [34], Distribution-Balanced Loss [38], and RIDE [42]. Our previous work RLKD [22] is also included in the comparison. For RSKD [34], since the original manual subset definitions are not publicly available, we follow the shot-based strategy and construct subsets based on class-frequency statistics, where categories are grouped into head, medium, and tail subsets according to their sample counts. To fairly compared with CCT-Net, we also take Dense-121 [60] as the backbone following the setting in CCT-Net.

From Table 1, with a ResNet-50 backbone, the proposed method achieves a Kappa score of 0.707 ± 0.014, outperforming DB Loss (0.673 ± 0.008), RSKD (0.660 ± 0.014), and RIDE (3 experts) (0.685 ± 0.010), and achieving performance comparable to RLKD (0.712 ± 0.011). Specifically, compared to Distribution-Balanced Loss (Kappa 0.673 ± 0.008), the proposed method improves the Kappa score by 3.4 percentage points (0.707 ± 0.014). Compared to Instance-Balanced Sampling (0.553 ± 0.010), the improvement reaches 15.4 percentage points. For example, RIDE (3 experts) achieves a Kappa score of 0.685 ± 0.010, which is 2.2 percentage points lower than the proposed method (0.707 ± 0.014) under the same ResNet-50 backbone. Under the DenseNet-121 backbone, the proposed method achieves a Kappa score of 0.735 ± 0.012. By integrating RIDE (3 experts), the Kappa score increases to 0.778 ± 0.008, demonstrating that the proposed framework can be effectively combined with multi-expert learning strategies. In addition to ODIR-5K, we further evaluate the proposed

Table 1: Long-tailed multi-label retinal diseases classification results on ODIR-5K [18]. The two best results are in red and blue.

Backbone	Methods	Kappa	F1	AUC
ResNet-50 [7]	IB Sampling	0.553 ± 0.010	0.874 ± 0.009	0.887 ± 0.012
	CB Sampling	0.601 ± 0.013	0.886 ± 0.011	0.920 ± 0.015
	Focal Loss [37]	0.625 ± 0.011	0.895 ± 0.009	0.930 ± 0.012
	DB Loss [38]	0.673 ± 0.008	0.930 ± 0.007	0.940 ± 0.009
	SCFKD [59]	0.635 ± 0.009	0.911 ± 0.007	0.927 ± 0.014
	RSKD [34]	0.660 ± 0.014	0.920 ± 0.013	0.935 ± 0.014
	RLKD [22]	0.712 ± 0.011	0.935 ± 0.013	0.944 ± 0.012
	RIDE (3 experts) [42]	0.685 ± 0.010	0.932 ± 0.007	0.941 ± 0.011
	Ours	0.707 ± 0.014	0.932 ± 0.014	0.941 ± 0.013
	Ours + RIDE (3 experts) [42]	0.722 ± 0.013	0.941 ± 0.010	0.953 ± 0.012
DenseNet-121 [60]	CCT-Net [58]	0.749 ± 0.007	0.952 ± 0.011	0.960 ± 0.013
	RIDE (3 experts) [42]	0.713 ± 0.010	0.943 ± 0.008	0.951 ± 0.011
	RLKD [22]	0.744 ± 0.010	0.960 ± 0.013	0.964 ± 0.010
	RLKD [22] + Focal Loss [37]	0.767 ± 0.012	0.965 ± 0.008	0.970 ± 0.014
	Ours	0.735 ± 0.012	0.954 ± 0.014	0.957 ± 0.011
	Ours + Focal Loss [37]	0.753 ± 0.008	0.957 ± 0.012	0.965 ± 0.015
	Ours + RIDE (3 experts) [42]	0.778 ± 0.008	0.973 ± 0.007	0.977 ± 0.009

Table 2: Long-tailed multi-label retinal diseases classification results on RFMiD dataset. The two best results are in red and blue.

Backbone	Methods	Kappa	F1	AUC
ResNet-50 [7]	IB Sampling	0.401	0.561	0.768
	CB Sampling	0.427	0.578	0.791
	Focal Loss [37]	0.449	0.592	0.812
	DB Loss [38]	0.486	0.634	0.836
	SCFKD [59]	0.458	0.612	0.802
	RSKD [34]	0.471	0.621	0.825
	RLKD [22]	0.534	0.661	0.858
	RIDE (3 experts) [42]	0.501	0.642	0.841
	Ours	0.519	0.648	0.846
	Ours + RIDE (3 experts) [42]	0.547	0.673	0.872
DenseNet-121 [60]	CCT-Net [58]	0.498	0.641	0.848
	RIDE (3 experts) [42]	0.516	0.653	0.856
	RLKD [22]	0.531	0.662	0.868
	RLKD [22] + Focal Loss [37]	0.549	0.674	0.881
	Ours	0.524	0.657	0.861
	Ours + Focal Loss [37]	0.539	0.666	0.872
	Ours + RIDE (3 experts) [42]	0.558	0.682	0.892

method on the RFMiD dataset, which is more challenging due to a larger number of disease categories and more severe class imbalance. As shown in Table 2, all methods exhibit an overall performance drop compared to ODIR-5K, which is expected given the increased difficulty of RFMiD.

Despite this performance degradation, the relative ranking of different methods remains largely consistent across the two datasets. On the RFMiD dataset with ResNet-50, the proposed method achieves a Kappa score of 0.519, outperforming DB Loss (0.486) and RIDE (3 experts) (0.501). When combined with RIDE, the performance further improves to 0.547, which is higher than RLKD (0.534) and RIDE (0.501) under the same backbone. With DenseNet-121, the proposed method combined with RIDE achieves a Kappa score of 0.558, outperforming RLKD + Focal Loss (0.549) and RIDE (0.516).

We also observe that extending the proposed framework with RIDE further improves performance on RFMiD, achieving the best results among all compared methods. This suggests that the proposed approach is complementary to multi-expert learning strategies and can benefit from increased model diversity.

Furthermore, a consistent performance gap between ResNet-50 and DenseNet-121 can be observed on both datasets. This difference is primarily attributed to the dense connections in DenseNet-121, which improve feature reuse and help capture more complex relationships between features. DenseNet-121 allows each layer to receive input from all previous layers, which facilitates more efficient gradient flow and enables the network to learn richer, more detailed feature representations. For retinal disease classification, where diseases can share complex visual relationships (such as between AMD and DR), DenseNet-121's ability to model these relationships more effectively leads to better generalization and higher performance compared to ResNet-50, which has more traditional residual connections that might not fully capture intricate relationships.

4.3 Ablation Studies

4.3.1 Effectiveness of Each Component

To investigate the contribution of each component, we design a series of ablation experiments. As shown in Table 3, the proposed framework mainly benefits from AGT and gradient-based re-weighting. Compared to the vanilla ResNet-50 baseline (Kappa 0.553 $\pm$ 0.010), incorporating AGT and KD improves performance to 0.645 $\pm$ 0.014. Further adding gradient-based re-weighting increases the Kappa score to 0.683 $\pm$ 0.012. Gradient-based re-weighting helps address the catastrophic forgetting in AGT. Finally, the class-balanced knowledge distillation can transfer knowledge to the student model in an effective way as well as avoid the class imbalance caused by multi-label samples.

4.3.2 Grouping Strategies

Ju *et al.* [34] divide the subsets using three different ways: shot-based, region-based, and feature-based. The latter two divisions require strong prior and clinical expertise

Table 3: Ablation study results on ODIR-5K [18]. The backbone is ResNet-50 [7]. AGT means alternate group training, KD is knowledge distillation, and RW is re-weighting. Pretraining indicates that the backbone is additionally pre-trained on the FGADR dataset, following the setting in RLKD [22].

Model	Kappa	F1	AUC
ResNet-50	0.553 ± 0.010	0.874 ± 0.009	0.887 ± 0.012
ResNet-50 + AGT + KD	0.645 ± 0.014	0.913 ± 0.012	0.933 ± 0.014
ResNet-50 + AGT + RW + KD	0.683 ± 0.012	0.925 ± 0.014	0.937 ± 0.013
ResNet-50 + AGT + RW + KD + Pretraining	$\mathbf{0.707 \pm 0.014}$	$\mathbf{0.932 \pm 0.014}$	$\mathbf{0.941 \pm 0.013}$

to complete. On the other hand, the shot-based division can be implemented automatically by counting the number of shots: many (head), medium, and few (tail). The comparison between shot-based division and our proposed relational subsets division can be seen in Table 4. The shot-based division still achieves good performance, which shows the effectiveness and superiority of the proposed AGT and gradient-based re-weighting. However, in the shot-based division, the relations among diseases can not be utilized well, leading to a decrease in performance.

Table 4: Different implementations and their effects of the key components. ResNet-50 [7] is used as the backbone.

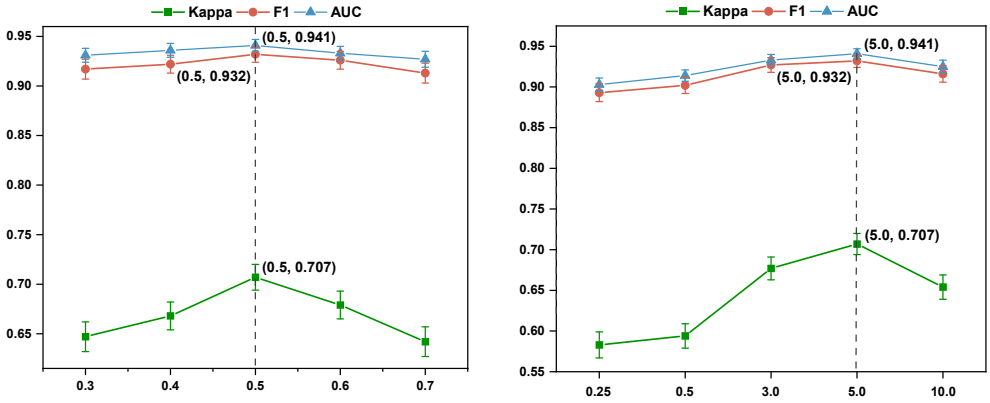
Module	Methods	Kappa	F1	AUC
Grouping Strategies	Shot-based	0.695 ± 0.010	0.921 ± 0.014	0.927 ± 0.015
	K-Means Clustering (Ours)	$\mathbf{0.707 \pm 0.014}$	$\mathbf{0.932 \pm 0.014}$	$\mathbf{0.941 \pm 0.013}$
Re-weighting	Count-based RW	0.672 ± 0.009	0.916 ± 0.010	0.924 ± 0.009
	Gradient-based RW (Ours)	$\mathbf{0.707 \pm 0.014}$	$\mathbf{0.932 \pm 0.014}$	$\mathbf{0.941 \pm 0.013}$
Knowledge Distillation	Unweighted	0.681 ± 0.014	0.922 ± 0.009	0.934 ± 0.012
	Class-balanced (Ours)	$\mathbf{0.707 \pm 0.014}$	$\mathbf{0.932 \pm 0.014}$	$\mathbf{0.941 \pm 0.013}$

Table 5: Comparison with our previous work RLKD [22] on ODIR-5K [18] regarding metrics. Performance can be further improved by combining RIDE [42]. ResNet-50 [7] is used as the backbone.

Model	Kappa	F1	AUC
Ours	0.707 ± 0.014	0.932 ± 0.014	0.941 ± 0.013
RLKD [22]	0.712 ± 0.011	0.935 ± 0.013	0.944 ± 0.012
Ours + RIDE (3 experts) [42]	$\mathbf{0.722 \pm 0.013}$	$\mathbf{0.941 \pm 0.010}$	$\mathbf{0.953 \pm 0.012}$

Table 6: Comparison with our previous work RLKD [22] on ODIR-5K [18] regrading computational cost during student training. Note that the parameter counts and flops are the sum of both teacher and student models.

Model	Training Time	GPU Memory	Params	FLOPs
Ours	**11h10min**	23.3GB	47.1M	2115.4G
RLKD [22]	18h40min	24.1GB	94.1M	4230.9G
Ours + RIDE (3 experts) [42]	13h20min	28.8GB	77.0M	2531.1G



(a) Effects of sampling ratio selection when sampling others in alternate group training. The values are an average of 5-fold results.

(b) Effects of temperature selection during knowledge distillation. The values are an average of 5-fold results.

Fig. 7: Ablation study of hyperparameter choices. (a) Sampling ratio selection when sampling others in alternate group training. (b) Temperature selection during knowledge distillation. Each value is averaged over 5 folds with error bars.

4.3.3 Sampling Ratio

The sampling ratio is an important hyper-parameter that significantly impacts the performance of alternate group training. By sampling "others," we can alleviate the suppression effect of the current sub-classifier on other sub-classifiers, but this comes at the cost of slowing down the optimization process. Setting the sampling ratio too large may prevent effective optimization of the current sub-classifier. Conversely, if the sampling ratio is too small, other sub-classifiers will be heavily suppressed, and catastrophic forgetting can not be effectively mitigated. The results of different sampling ratios are shown in Fig. 7, demonstrating that an intermediate sampling ratio yields better performance. The error bars reflect the variability of 5-fold cross-validation results, supporting that α=0.5 is the optimal choice. Therefore, we set the sampling ratio to 0.5 in the experiments by default. The observation is different from BAGS [41], which adopts a larger sampling ratio on the LVIS dataset [61]. We believe

that the difference in ratio depends on the data imbalance of the dataset. When dealing with severe data imbalance, a larger ratio may be more appropriate.

4.3.4 Re-weighting Approaches

To address the imbalanced gradients within each class, we propose gradient-based re-weighting to adjust the loss. Specifically, the weights of BCE loss are calculated according to the ratio of positive and negative gradients for each class. Besides calculating weights based on gradients, weights can also be calculated based on the ratio of positive and negative samples, called count-based re-weighting. With such a design, the ratio r_j in Eq. (2) is defined as $r_j = \frac{N_j^{pos}}{N_j^{neg}}$, where N_j^{pos} and N_j^{neg} represent the count of positive and negative samples for class j in current mini-batch. The comparison results are shown in Table 4. Obviously, gradient-based re-weighting is better than count-based re-weighting, which is similar to hard labels (count-based) and soft labels (gradient-based) in deep learning.

4.3.5 Knowledge Distillation

We employ the logit-level class-balanced weighted knowledge distillation to transfer knowledge from the teacher model to the student model. The comparison between weighted and unweighted distillation results is presented inTable 4. It can be observed that the unweighted distillation leads to a performance drop, indicating that the class imbalance issue caused by multi-label samples is not effectively addressed. Furthermore, we investigate the impact of temperature scaling on the performance. As shown in Fig. 7, we test five different temperature values: $T = 0.25, 0.5, 3, 5, 7$. We observe that $T = 5$ achieves the best performance, striking an optimal balance between regularizing the model and maintaining the sharpness of the teacher's soft labels. When T is too low (e.g., $T = 0.25$), the model becomes overly sensitive to the teacher's logits, resulting in poor generalization. On the other hand, when T is too high (e.g., $T = 7$), the model may become overly smoothed, leading to a loss of critical information that the teacher model is trying to convey. Therefore, $T = 5$ provides the best trade-off by retaining sufficient soft label information for effective distillation, which leads to improved performance in our method.

4.3.6 Comparison with RLKD

In contrast to RLKD [22], which utilizes three experts as the teacher models, our approach employs a single expert as the teacher model by default. The classifier of the teacher model is partitioned into multiple sub-classifiers, which are trained in an alternating manner on the relational subsets. From Table 5, we can see that our method achieves a significant reduction in training time compared to RLKD (only 60% of the training time required by RLKD), while maintaining a reasonable level of performance degradation. Furthermore, when extending the teacher models using RIDE [42], the performance can be further improved. It is worth noting that all experts in RIDE share the same backbone network, which ensures that the increase in training time is not substantial.

5 Discussion

5.1 Effectiveness of Gradient-based Re-weighting

The proposed gradient-based re-weighting mechanism is designed to explicitly regulate class-wise gradient imbalance introduced by AGT. Under AGT, classifiers corresponding to inactive subsets are repeatedly exposed to negative gradients, which may dominate the optimization process for tail classes and lead to catastrophic forgetting. While sampling "others" alleviates this issue at the group level, it does not directly control gradient magnitudes at the class level.

By estimating positive and negative gradients for each class and adaptively adjusting their contributions, the proposed re-weighting strategy stabilizes optimization for tail classes without modifying network architecture or requiring additional supervision. The gradient visualizations further confirm that re-weighting reduces the dominance of negative gradients for tail classes and enables more consistent updates across training cycles. This class-wise regulation complements the group-level stabilization of AGT, forming a unified mechanism to mitigate forgetting while preserving training efficiency.

5.2 Failure Cases and Limitations

Despite the overall performance improvements, several challenging scenarios remain. First, samples associated with multiple disease labels are particularly difficult to recognize. In our datasets, image-level samples annotated with three disease labels are extremely rare (only 8 samples). Moreover, multi-label evaluation at the sample level follows an exact-match criterion, where a prediction is considered correct only if all associated labels are simultaneously identified. As the number of labels per image increases, the combinatorial prediction space grows rapidly, making exact matching significantly more difficult than in single- or two-label cases. Even minor deviations in any individual label prediction may therefore lead to complete failure at the sample level. For instance, in one experimental split, only one three-label sample was present in the test set, and it was not correctly predicted under the exact-match criterion (sample-level accuracy = 0%). Although this observation is based on an extremely small sample size and should not be over-interpreted, it nonetheless reflects the intrinsic difficulty of high-cardinality multi-label prediction. This intrinsic combinatorial difficulty, coupled with severe data scarcity, makes such samples particularly prone to misclassification, even when gradient re-weighting is applied.

Second, extremely rare tail classes with very limited positive samples may still suffer from insufficient learning signals. Although gradient re-weighting alleviates the dominance of negative gradients, it cannot fully compensate for the lack of informative supervision. As a result, the learned representations for these classes may remain unstable, especially in the early stages of training.

Third, the proposed method assumes reasonably reliable annotations. In real-world clinical datasets, noisy or ambiguous labels are common, particularly for rare diseases. In such cases, gradient statistics may be biased by incorrect supervision, and re-weighting may amplify noisy gradients, leading to suboptimal generalization.

This limitation is inherent to gradient-driven balancing strategies and highlights the importance of integrating noise-robust learning techniques in future work.

5.3 Clinical and Practical Considerations

Although the proposed framework demonstrates consistent improvements on publicly available retinal fundus datasets, translating such models into real clinical environments involves additional considerations beyond retrospective benchmarking. In clinical practice, data acquisition conditions vary across devices, institutions, and patient populations. Therefore, prospective evaluation on independently collected cohorts would be necessary to assess model robustness under routine screening settings. Validation across multiple centers would further help determine whether the learned representations generalize beyond the distribution of curated research datasets.

Practical deployment also requires seamless integration into existing clinical workflows. In real-world settings, models must operate within hospital information systems and support efficient inference without disrupting routine diagnosis. Moreover, clinicians typically require interpretable outputs and calibrated confidence estimates to assist, rather than replace, human decision-making. Ensuring stability under distribution shifts and imperfect annotations—common in real clinical data—remains an important direction for further investigation.

Overall, while the proposed method provides a principled optimization framework for long-tailed multi-label retinal disease recognition, bridging the gap between algorithmic performance and clinical adoption will require careful validation, system-level integration, and continuous refinement in collaboration with medical practitioners.

6 Conclusion

In this paper, we propose a novel teacher–student framework for multi-label long-tailed retinal disease recognition. The proposed method leverages automatic relational subset division and alternate group training to efficiently train a single teacher model with balanced representations. To mitigate catastrophic forgetting during AGT, we introduce sampling "others" and a class-wise gradient-based re-weighting strategy that explicitly regulates positive and negative gradient contributions at the class level. The learned generalized representations and balanced classifiers are further distilled into a compact student model through knowledge distillation. Extensive experiments on two public datasets demonstrate that the proposed approach consistently outperforms existing methods for long-tailed multi-label recognition. Specifically, on the ODIR-5K dataset with ResNet-50, the method achieves a Kappa score of 0.707 ± 0.014. When combined with RIDE (3 experts), the Kappa score improves to 0.722 ± 0.013. Furthermore, when integrated with RIDE (3 experts) under DenseNet-121, the method achieves a Kappa score of 0.778 ± 0.008.

Several limitations and challenging scenarios, including extremely rare classes, multi-label coupling, and noisy annotations, are discussed in Sec. 5. Future work will explore more robust learning strategies and investigate how large-scale language or vision–language models [62] can be integrated as semantic prior providers within

our optimization-driven framework. Such models may offer complementary high-level representations, while the explicit gradient regulation mechanism introduced in this work can help stabilize their fine-tuning under long-tailed multi-label distributions.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 31771073.

Declarations

Conflict of interest

The authors declare no conflict of interest.

Data availability

The ODIR-5k dataset is available at https://odir2019.grand-challenge.org/dataset/. The RFMiD dataset is available at https://riadd.grand-challenge.org/download-all-classes/.

References

[1] Bourne, R., Steinmetz, J.D., Flaxman, S., Briant, P.S., Taylor, H.R., Resnikoff, S., Casson, R.J., Abdoli, A., Abu-Gharbieh, E., Afshin, A., *et al.*: Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the global burden of disease study. The Lancet Global Health **9**(2), 130–143 (2021)

[2] Mayya, V., S, S.K., Kulkarni, U., Surya, D.K., Acharya, U.R.: An empirical study of preprocessing techniques with convolutional neural networks for accurate detection of chronic ocular diseases using fundus images. Applied Intelligence **53**(2), 1548–1566 (2023)

[3] Li, Z., Wu, W., Wei, B., Li, H., Zhan, J., Deng, S., Wang, J.: Rice disease detection: Tli-yolo innovative approach for enhanced detection and mobile compatibility. Sensors **25**(8) (2025)

[4] Sungheetha, A., Sharma, R.: Design an early detection and classification for diabetic retinopathy by deep feature extraction based convolution neural network. Journal of Trends in Computer Science and Smart technology (TCSST) **3**(02), 81–94 (2021)

[5] Oh, K., Kang, H.M., Leem, D., Lee, H., Seo, K.Y., Yoon, S.: Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images. Scientific Reports **11**(1), 1–9 (2021)

[6] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer

[7] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[8] Sugeno, A., Ishikawa, Y., Ohshima, T., Muramatsu, R.: Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. Computers in Biology and Medicine **137**, 104795 (2021)

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

[10] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). IEEE

[11] Thomas, A., Harikrishnan, P., Krishna, A.K., Palanisamy, P., Gopi, V.P.: A novel multiscale convolutional neural network based age-related macular degeneration detection using oct images. Biomedical Signal Processing and Control **67**, 102538 (2021)

[12] Kadry, S., Rajinikanth, V., González Crespo, R., Verdú, E.: Automated detection of age-related macular degeneration using a pre-trained deep-learning scheme. The Journal of Supercomputing **78**(5), 7321–7340 (2022)

[13] He, T., Zhou, Q., Zou, Y.: Automatic detection of age-related macular degeneration based on deep learning and local outlier factor algorithm. Diagnostics **12**(2), 532 (2022)

[14] Li, Y., Shen, J., Mao, Z.: Coocnet: a novel approach to multi-label text classification with improved label co-occurrence modeling. Applied Intelligence **54**(17), 8702–8718 (2024)

[15] Qin, S., Wu, H., Zhou, L., Zhao, Y., Zhang, L.: Tae: Topic-aware encoder for large-scale multi-label text classification. Applied Intelligence **54**(8), 6269–6284 (2024)

[16] Sheng, B., Pan, D., Li, X.: Dynamic dual mining framework for long-tailed out-of-distribution detection. Applied Intelligence **55**(11), 783 (2025)

[17] Hao, L., Yang, J., Zhang, Y.: Balanced loss function for long-tailed semi-supervised ship detection. Applied Intelligence **55**(13), 942 (2025)

[18] Peking University: International Competition on Ocular Disease Intelligent Recognition (ODIR-2019). Accessed: 2025-08-04 (2019). https://odir2019.grand-challenge.org

[19] Fu, P., Ruhaiyem, N.I.R., Wang, J.: Reweighting balanced representation learning for long tailed image recognition in multiple domains. Scientific Reports **15**(1), 23948 (2025)

[20] Wang, Y., Chang, Y., Qin, Y., Zhao, Y., Wei, S.: Unbiased sample selection and label improvement for mitigating noisy labels in class-imbalanced datasets. IEEE Transactions on Circuits and Systems for Video Technology (2025)

[21] Zhang, S., Chen, C., Hu, X., Peng, S.: Balanced knowledge distillation for long-tailed learning. Neurocomputing **527**, 36–46 (2023)

[22] Zhou, Q., Zou, H., Wang, Z.: Long-tailed multi-label retinal diseases recognition via relational learning and knowledge distillation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 709–718 (2022). Springer

[23] Chous, A.: Do diabetes, diabetic retinal disease contribute to macular degeneration? Optometry Times Journal **15** (2023)

[24] Yongpeng, Z., Yaxing, W., Jinqiong, Z., Qian, W., Yanni, Y., Xuan, Y., Jingyan, Y., Wenjia, Z., Ping, W., Chang, S., *et al.*: The association between diabetic retinopathy and the prevalence of age-related macular degeneration—the kailuan eye study. Frontiers in Public Health **10**, 922289 (2022)

[25] Chen, Y.W., Wu, T.Y., Wong, W.H., Lee, C.Y.: Diabetic retinopathy detection based on deep convolutional neural networks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1030–1034 (2018). IEEE

[26] Zhou, T.H., Liu, Y.M., Xie, W.C., Li, H.N., Wang, L.: Automatic identification and classification method for diabetic retinopathy ffa image processing. In: Advances in Intelligent Information Hiding and Multimedia Signal Processing: Proceeding of the 16th International Conference on IIHMSP in Conjunction with the 13th International Conference on FITAT, November 5-7, 2020, Ho Chi Minh City, Vietnam, Volume 1, pp. 204–210 (2021). Springer

[27] Eladawi, N., Elmogy, M., Khalifa, F., Ghazal, M., Ghazi, N., Aboelfetouh, A., Riad, A., Sandhu, H., Schaal, S., El-Baz, A.: Early diabetic retinopathy diagnosis based on local retinal blood vessel analysis in optical coherence tomography angiography (octa) images. Medical Physics **45**(10), 4582–4599 (2018)

[28] Nair, A.T., Muthuvel, K.: Blood vessel segmentation and diabetic retinopathy recognition: an intelligent approach. Computer Methods in Biomechanics and

Biomedical Engineering: Imaging & Visualization **8**(2), 169–181 (2020)

[29] Yang, Y., Li, T., Li, W., Wu, H., Fan, W., Zhang, W.: Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 533–540 (2017). Springer

[30] Wu, W., Chen, Z., Ma, X., Zhang, W., Qiu, X., Song, S., Huang, X., Ma, F., Xiao, J.: Contrastive prompt clustering for weakly supervised semantic segmentation (2025). https://arxiv.org/abs/2508.17009

[31] He, J., Li, C., Ye, J., Qiao, Y., Gu, L.: Multi-label ocular disease classification with a dense correlation deep neural network. Biomedical Signal Processing and Control **63**, 102167 (2021)

[32] Ju, L., Wang, X., Yu, Z., Wang, L., Zhao, X., Ge, Z.: Long-tailed multi-label retinal diseases recognition using hierarchical information and hybrid knowledge distillation. arXiv preprint arXiv:2111.08913 (2021)

[33] Cheng, Y., Ma, M., Li, X., Zhou, Y.: Multi-label classification of fundus images based on graph convolutional network. BMC Medical Informatics and Decision Making **21**(2), 1–9 (2021)

[34] Ju, L., Wang, X., Wang, L., Liu, T., Zhao, X., Drummond, T., Mahapatra, D., Ge, Z.: Relational subsets knowledge distillation for long-tailed retinal diseases recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 3–12 (2021). Springer

[35] Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks **106**, 249–259 (2018)

[36] Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., *et al.*: Balanced meta-softmax for long-tailed visual recognition. Advances in Neural Information Processing Systems **33**, 4175–4186 (2020)

[37] 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)

[38] Wu, T., Huang, Q., Liu, Z., Wang, Y., Lin, D.: Distribution-balanced loss for multi-label classification in long-tailed datasets. In: European Conference on Computer Vision, pp. 162–178 (2020). Springer

[39] Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., Yan, J.: Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11662–11671 (2020)

[40] Tan, J., Lu, X., Zhang, G., Yin, C., Li, Q.: Equalization loss v2: A new gradient balance approach for long-tailed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1685–1694 (2021)

[41] Li, Y., Wang, T., Kang, B., Tang, S., Wang, C., Li, J., Feng, J.: Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10991–11000 (2020)

[42] Wang, X., Lian, L., Miao, Z., Liu, Z., Yu, S.: Long-tailed recognition by routing diverse distribution-aware experts. In: International Conference on Learning Representations (2021)

[43] Zhang, Y., Hooi, B., Hong, L., Feng, J.: Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. Advances in Neural Information Processing Systems **3** (2022)

[44] Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2022, p. 3876 (2022)

[45] Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2** (2024)

[46] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)

[47] Zhou, Q., Gao, Y., Zou, H., Luo, F., Bai, X.: Anatomy-aware adaptation of pre-trained models for medical difference visual question answering. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2283–3388 (2025). IEEE

[48] Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H), pp. 353–367 (2023). PMLR

[49] Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1), 100–108 (1979)

[50] Li, B., Han, Z., Li, H., Fu, H., Zhang, C.: Trustworthy long-tailed classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6970–6979 (2022)

[51] Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: International Conference on Learning Representations (2019)

[52] Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Advances in Neural Information Processing Systems **32** (2019)

[53] Pezeshki, M., Kaba, O., Bengio, Y., Courville, A.C., Precup, D., Lajoie, G.: Gradient starvation: A learning proclivity in neural networks. Advances in Neural Information Processing Systems **34**, 1256–1272 (2021)

[54] Zhou, Y., Wang, B., Huang, L., Cui, S., Shao, L.: A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability. IEEE Transactions on Medical Imaging **40**(3), 818–828 (2020)

[55] Pachade, S., Porwal, P., Thulkar, D., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Giancardo, L., Quellec, G., Mériaudeau, F.: Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research. Data **6**(2), 14 (2021)

[56] Ju, L., Yu, Z., Wang, L., Zhao, X., Wang, X., Bonnington, P., Ge, Z.: Hierarchical knowledge guided learning for real-world retinal disease recognition. IEEE Transactions on Medical Imaging **43**(1), 335–350 (2023)

[57] Li, T., Sheng, B.: Msce-lt: Multi-label supervised contrastive enhancement for long-tailed retinal diseases recognition. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2128–2133 (2024). IEEE

[58] Zhou, Y., Huang, L., Zhou, T., Shao, L.: Cct-net: Category-invariant cross-domain transfer for medical single-to-multiple disease diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8260–8270 (2021)

[59] He, J., Li, C., Ye, J., Qiao, Y., Gu, L.: Self-speculation of clinical features based on knowledge distillation for accurate ocular disease classification. Biomedical Signal Processing and Control **67**, 102491 (2021)

[60] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

[61] Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance

segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5356–5364 (2019)

[62] Wu, W., Chen, X., Chen, Z., Jiang, J.-E., Tsang, K.-F., Huang, X., Ma, F., Xiao, J.: Tag-enriched multi-attention with large language models for cross-domain sequential recommendation. IEEE Transactions on Consumer Electronics (2025)