Long-Tailed Multi-Label Retinal Diseases Recognition Using Hierarchical Information and Hybrid Knowledge Distillation

Lie Ju\textsuperscript{1,2} Xin Wang\textsuperscript{1} Zhen Yu\textsuperscript{2} Lin Wang\textsuperscript{1,2,3} Xin Zhao\textsuperscript{1} Zongyuan Ge\textsuperscript{1,2}
\textsuperscript{1} Airdoc LLC, China  
\textsuperscript{2} Monash Medical AI Group, Monash University, Australia  
\textsuperscript{3} Harbin Engineering University, China

julie334600@gmail.com; Zongyuan.ge@monash.edu
https://www.monash.edu/mmai-group

Abstract

In the real world, medical datasets often exhibit a long-tailed data distribution (i.e., a few classes occupy most of the data, while most classes have rarely few samples), which results in a challenging imbalance learning scenario. For example, there are estimated more than 40 different kinds of retinal diseases with variable morbidity, however with more than 30+ conditions are very rare from the global patient cohorts, which results in a typical long-tailed learning problem for deep learning-based screening models. Moreover, there may exist more than one kind of disease on the retina, which results in a multi-label scenario and bring label co-occurrence issue for re-sampling strategy. In this work, we propose 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. Then, an instance-wise class-balanced sampling strategy and hybrid knowledge distillation manner are firstly introduced to learn from the long-tailed multi-label distribution. Our experiments training on the retinal dataset of more than one million samples demonstrate the superiority of our proposed methods which outperform all competitors and significantly improve the recognition accuracy of most diseases especially those rare diseases.

1. Introduction

Retinal diseases such as diabetic retinopathy and glaucoma are the leading causes of blindness [23, 15]. Some early signs are usually ignored, but the vision loss or even blindness caused by no timely intervention is usually irreversible [24]. Fundus screening methods such as fundus camera and OCT are considered to be effective for retinal diseases recognition. However, manual fundus screening is time-consuming, which requires professional ophthalmologists to screen fundus images and present reports. The automated screening for retinal diseases has long been recognized and attracted great attention [9, 25]. Recent studies have demonstrated successful applications of deep learning-based models for retinal disease screening such as diabetic retinopathy (DR) and glaucoma [15, 11]. Gargeya et al. [12] used features extracted by CNN and metadata information fed into a decision tree model for binary classification. [1] proposed a Multiple Instance Learning and uncertainty-based framework for DR grading. For glaucoma detection, the OD/OC area is always considered. Dos et al. [8] designed a phylogenetic diversity indexes module for semantic features extraction for the diagnosis of glaucoma. Fu et al. [11] proposed a novel disc-aware ensemble network, which integrates the deep hierarchical context of the global fundus image and the local optic disc region.

Many previous works have been proved to be effective for some specific diseases and lesions under a controlled experimental environment. However, the pathological changes of the fundus are extremely complex. For instance, there are estimated more than 40 different kinds of...
retinal diseases. Those diseases and lesions that occurred less frequently in the training set may not perform well in real clinical test settings due to the algorithm failing to generalize those pathologies. The diagnosis of multiple diseases is very significant in clinical practice, especially for those rare diseases. For human expertise, common diseases are easy to be diagnosed while rare diseases are easy to be missed. Improving the accuracy of rare diseases recognition can become more useful for clinicians within the CAD system. Recently, Wang et al. [32] designed a three-stream framework that goals to detect the 36 kinds of retinal lesions/diseases. Quellec et al. [29] used few-shot learning to perform rare pathologies detection on the OPHDIAT dataset [28], which consists of 763,848 fundus images with 41 conditions. A recently released dataset [20] collected 53 different kinds of fundus diseases that may appear in clinical screening. As Fig. 1 shows, this dataset shows a typical long-tailed distribution attribute, whose ratio of head class to tail class exceeds 100, indicating a serious class-imbalance issue, which is not uncommon in many medical datasets. Besides, some samples may contain more than a single retinal disease label, leading to a label co-occurrence issue for re-balance of the data such as a re-sampling strategy.

In this work, we propose a novel framework for learning from long-tailed multi-label retinal diseases. The success of our proposed framework is affected by two key observations in most existing long-tailed classification works. First, incorporating the hierarchy in the model would help improve generalization on classes especially for those minority classes, by leveraging shared features among hierarchically-related classes [7]. Second, the model learning from the original distribution obtains a better representation, i.e., convolutional layers and the model learning from a re-balanced distribution shows a fairer and more unbiased classification for all categories [40, 22]. Following these two observations, we leverage the prior knowledge in retinal diseases, e.g., the hierarchical information for all categories from coarse to fine, to help the model train a better representation under a hierarchy-sensible manner. Then, we use the vanilla instance-balanced sampling and a novel instance-wise class-balanced sampling to train two teacher models. Then, we distill the knowledge of two teacher models into a unified student model.

The main contributions can be summarized as follows:

1. We first train from the database with more than one million fundus images and develop a novel framework to recognize more than 50 kinds of retinal diseases.

2. We inject hierarchy knowledge in retinal diseases into the training for a well-generalized representation with more semantic information, which can be naturally shared across data between coarse or fine categories.

3. We exploit a hybrid knowledge distillation manner to help the model simultaneously learn from the teacher models with better representation or fairer classifier.

4. Extensive experiments on two datasets with the different number of samples and imbalanced ratio are conducted. The results demonstrate our proposed methods can well improve the recognition accuracy for those rare diseases without losing the performance on common diseases.

2. Related Work

2.1. Retinal Diseases Recognition

The CAD methods for retinal diseases such as DR, Glaucoma and AMD have been long recognized. [15, 12] proposed to use a CNN for binary classification of with/without DR. [38] presented an ensemble strategy to perform two-class and four-class classifications. Some methods [36, 13, 10] proposed to combine the segmentation results and grading results to provide the interpretability with the internal correlation between DR levels and lesions. Besides the OC/OD area-based method [11] mentioned above, there are some previous works that generate evidence maps for glaucoma diagnosis. [39] proposed a two-stage cascaded approach which obtains unsupervised feature representation of fundus image with a CNN and CDR value regression by random forest regressor. The automated diagnosis of AMD has also been studied [14, 3]. [21] leverage the common features between DR and AMD, then improve both accuracy under a knowledge distillation and multi-task manner.

Although some specific retinal diseases are fully studied, the condition of the fundus is complex. Most of the existing methods do not promise the robustness and ability of diagnosis for other diseases especially those rare diseases. Kaggle EyePACS dataset [4] consists of 35,126 training images graded into five DR stages. However, it is found that there are at least more than 30 kinds of retinal diseases that are mislabeled by the original annotators [20]. Ignoring those out-of-distribution categories leads to a huge risk. [32] firstly used a multi-task framework to detect 36 kinds of retinal diseases respectively. However, it requires extra annotations for the location of the optic disc and macula area. [29] also take the features among diseases into consideration and leverage the few-shot learning technique to perform rare pathologies recognition. The existing methods are inspiring but show some limitations: those methods consider the prior knowledge of retinal diseases such as the common features (lesions) and locations shared by some diseases, but still lack direct insights for leveraging some prior knowledge or cognitive laws.
2.2. Long-tailed Classification

Re-sampling methods try to balance the distribution by over-sampling the minority-class samples or under-sampling the majority class samples in some manners. [22] proposed a two-stage training strategy to train the feature extractor and classifier from a uniform and re-sampling distribution respectively. [40] proposed to design a two-branch network to achieve one-stage training with the same insight. [33] presented a curriculum learning-based sampling scheduler. Most of these works focus on single-label datasets. Since the over-sampling on minority class will sometimes sample the majority class at the same time due to the label co-occurrence and leads to a new imbalanced condition. To handle this challenge, [34] extended the re-sampling strategy into a multi-label scenario and proposed a regularization term to overcome the over-suppression from negative samples.

Re-weighting methods aim to design loss functions for more robust learning from imbalanced data. Focal loss [26] learned to weigh the samples and achieve Hard Samples Mining according to the output probability. The increase of the sample brings about a diminishing return on performance for those majority classes. [6] proposed to assign a new variable to increase the benefit from those effective samples. [5] proposed an effective training strategy, which allows the model to learn an initial representation while avoiding some of the complications associated with re-weighting or re-sampling.

Transfer Learning methods aim to learn the general knowledge from the majority classes and then transfer it to the minority classes. OLTR [27] learns a set of dynamic meta embedding to transfer the visual information knowledge of the head to the tail category. Also, they design a memory set that allows tail categories to utilize relevant head information through calculating the similarity. [35] found that learning from a less-imbalanced subset suffers from little performance loss and proposed to train multiple teacher models from shot-based subsets, then guide the training of a unified student model.

3. Datasets

To address the long-tailed retinal diseases recognition challenge, we evaluate the proposed methods on three datasets with different scales including the number of samples, the number of classes, the imbalance ratio ($\rho$ in Table 1) and the label co-occurrence (e.g., label cardinality). Before introducing the three datasets, we first give the quantitative metrics for the main natures of long-tailed retinal diseases recognition, the first is how imbalanced the original distribution is. Formally, for the original distribution $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where each $x$ indicates a feature vector and each $y$ is an associated one-hot label. The indexes of $k$ kinds of categories are sorted from most to least with $N_{c_1} > N_{c_2} > \ldots > N_{c_k}$ where $N_{c_i}$ denotes the number of the $i_{th}$ category. Thus, the imbalance ratio can be simply described as $\rho = \frac{N_{c_1}}{N_{c_k}}$.

Besides, as we claimed that there is label co-occurrence among the original long-tailed distribution and most resampling strategies will lead to a new inner-class imbalance [34] (we have a detailed analysis on Sec.). Here, we introduce several useful indicators for the measurement of the label-occurrence [30, 37]. The first is label cardinality which indicates the average number of labels per example, then the $L_{Card}$ can be normalized as label density by the number of possible labels in the label space: $L_{Card}(S) = \frac{1}{n}\sum_{i=1}^{n}|y_i|$, $L_{Den}(S) = \frac{1}{n^{2}} \cdot L_{Card}(S)$.

Our collected datasets are acquired from private hospitals with a time span of 10 years, and each image was labeled by 3 - 8 senior ophthalmologists. A sample is retained only if more than half of the ophthalmologists are in agreement with the disease label, or the sample will be assigned into re-labeled processing with the discussion of all ophthalmologists. There are representiative 70 kinds of categories\(^1\).

In this study, more than one million samples are selected to form the two datasets, which consist of more than 50 kinds of retinal diseases, also, this is the first study that attempts to train the DL-based model for retinal diseases recognition from a database with more than one million samples. Both two datasets exhibit an extreme long-tailed distribution with

\(^1\)Please refer to our supplementary files for more details
Also, we have class-balanced sampling and instance-balanced sampling. Especially, to better utilize the proposed methods, all diseases are mapped into a semantic hierarchy tree. Specifically, we regard the 50+ kinds of diseases as the base category, and we divide them into several subsets according to their characteristics. There are three hierarchies for each base category in total, as shown in Fig. 2. For instance, a sample is labeled as big drusen, and it also belongs to drusen and AMD. There are also three base categories in the drusen subset: small drusen, medium drusen and non-macular drusen. We give a detailed description of the hierarchy form in our supplementary files.

4. Methodology

In this section, we first define some basic notations for long-tailed multi-label classification in retinal diseases recognition. An analysis is given on how the traditional re-sampling strategies do not work well on a multi-label setting for long-tailed classification. Then we present an instance-wise class-balanced sampling technique to handle the discussed issue. Finally, we propose to use hybrid multiple knowledge distillation to bridge the gap for feature and classifier bias proposed in two-stage training works [40, 22].

4.1. Problem Definition

Suppose the original distribution \( S = \{(X,Y)\} \), where \( X = \{x_1, x_2, ..., x_n\} \) are the \( N \) training examples, and \( Y = \{y_1, y_2, ..., y_n\} \) denotes the associated labels. Specially, for a multi-label setting, each \( y_i = [y_{i1}, y_{i2}, ..., y_{ik}] \), where \( k \) denotes the total number of possible categories and \( y_i \in \{0, 1\}^k \), i.e., \( y_{ij} = 1 \) indicates the presence of the category \( j \) in image \( i \) and \( y_{ij} = 0 \) otherwise. Specifically, all indexes of category are sorted into \( \{c_1, c_2, ..., c_k\} \) from most to least. Also, we have \( N_{c_1} >> N_{c_k} \). Our goal is to train a deep neural network model (feature extractor \( f(\cdot) \) and classifier \( g(\cdot) \)) which can accurately recognize all classes.

4.2. Sampling

In this section, we will introduce two commonly-used sampling strategies in the long-tailed multi-class classification: instance-balanced sampling and class-balanced sampling. Then we give an experimental analysis on why these two sampling ways do not work well in a multi-label setting. Finally, we propose to use an instance-wise class-balanced sampling to handle this scenario.

4.2.1 Instance-balanced Sampling

In this multi-label classification, we always train a deep neural network by minimizing a binary cross-entropy (BCE) loss, which can be formulated as follows:

\[
L_{\text{BCE}} = \frac{1}{kN} \sum_{i=1}^{N} \sum_{j=1}^{k} [y_{ij} \log(\hat{y}_{ij}) + (1 - y_{ij}) \log(1 - \hat{y}_{ij})],
\]

where \( \hat{y}_{ij} \) denotes the predicted results for the \( i^{th} \) sample in the index of \( j^{th} \). It can be seen that each example \( x_i \in X \) obtains the same sampling probability in a mini-batch during the training. However, the head classes will be sampled more frequently than the tail classes in a long-tailed distribution dataset and the model tends to under-fit those classes with less samples, which results in a prediction bias.

4.2.2 Class-balanced Sampling

Given the sampling probability \( p_i \) for each sample in a mini-batch, and we have \( p_i = \frac{N_{c_i}}{N} \) in an instance-based sampling. Instead, class-balanced sampling aims to assign an equal sampling probability \( p_{ck} = \frac{1}{K} \) for each category in a mini-batch, as Fig. 4-(b) shows. Class-balanced sampling is considered to be a simple but effective trick for training imbalanced data and necessary component in many state-of-the-art works [40, 22]. However, there are two disadvantages especially for a multi-label setting. First, under an idea condition, we have \( K \) examples from \( K \) kinds of categories in a mini-batch with batch-size of \( K \). In this case, the instance-level sampling probability becomes \( p_i^* = \frac{p_i}{N_j} \). Since \( N_{c_1} >> N_{c_k} \), those examples from the head classes will be less utilized. Not all samples are ‘seen’ by the model, and the feature space learned is incomplete. Second, due to the label co-occurrence, over-sampling a sample in tail class will lead to a re-sampling the head class at the same time when there both exists categories from both in this sample, as Fig. 4-(c) shows.

4.2.3 Instance-wise Class-balanced Sampling

From Fig. 4-(c) shows, although wrongly sampling those samples with label co-occurrence can release new relative imbalance, we hope to ‘ignore’ those samples covering too many categories, so we can try to avoid sampling such samples as much as possible. Here, we present the instance-wise class-balanced sampling (ICS) strategy as [34, 17]. Since we have the expected ideal sampling probability for one sample from category \( j \): \( p_i^A = \frac{N_{c_j}}{N} \) and the actual sampling probability: \( p_i^A = \sum_{j=1}^{k} 1_j \cdot p_i^j \). Then, a sampling factor \( \delta \) is used to re-balanced the sampling probability \( \delta_i^j = \frac{p_i^j}{p_i^A} \), then the BCE loss becomes:

\[
L_{\text{BCE}} = \frac{1}{kN} \sum_{i=1}^{N} \sum_{j=1}^{k} [y_{ij} \log(\hat{y}_{ij}) + (1 - y_{ij}) \log(1 - \hat{y}_{ij})],
\]

(BCE) loss, which can be formulated as follows:
Figure 3. The overall framework of our proposed methods. The framework consists of three key components. First, we use the hierarchical information of retinal diseases as the prior knowledge for the pre-training of the model $g_1(f_1(\cdot))$. Then, a instance-wise class-balanced loss is introduced to help train the model $g_2(f_2(\cdot))$ which tends to pay more attention to those rare diseases in re-sampling distribution. Finally, the model $g_1(f_1(\cdot))$ with better representation and $g_2(f_2(\cdot))$ with fairer weights of classification are both distilled into a unified student model under a hybrid knowledge distillation manner.

Figure 4. The illustration of sampling probability under different sampling strategies: (a) instance-balanced sampling; (b) expected sampling; (c) class-balanced sampling; (d) instance-wise class-balanced sampling.

However, we find that some mild diseases which are also in head class such as tessellated fundus occurs in more than half of the samples from other categories, which makes the sampling factor $\delta_{ij}^m$ to zero and makes it difficult to make the optimization. Unlike [34] with two hyper-parameters to map the sampling factor to getting close to 1, we directly square it, which can increase $\delta_{ij}^m$ rapidly from near zero, and the sampling probability in those samples with different co-occurrence levels will not be too close. For the simplification, we will use $L_{BCE}$ to refer to the ICS.

4.3. Pre-Training with Hierarchical Information

Previous works on retinal diseases classification aim to learn a model which has a good ability to classify some specific retinal diseases such as DR or Glaucoma [11, 15]. As we claimed before, only detecting a single kind of disease lacks the global monitor for the conditions of patients. There are some complications that denote an implicit relationship between categories. Also, Given a sample from unknown classes, a trained model may have no ability to handle the challenge, e.g., giving a hypertension fundus to a DR classification model. Such approaches only assume mutually exclusive, unstructured labels [7].

We have the semantic-related subsets division in Sec. III-Datasets. Different from [19], which present 3 kinds of 2-level relational subsets generation: shot-based, region-based and feature-based. In this study, after the discussion with more than 10 senior ophthalmologists, we refine the hierarchical mapping by taking both region-information and feature information into consideration under a reasonable medical motivation. We have included a detailed description of hierarchical mapping in our supplementary files. In the following parts, we explore how to pass the designed hierarchical information to deep neural networks training. In contrast to previous works for hierarchical images classification [31, 7] which focus on the classification performance on each level, we only care about the lowest level (defined as base classes in this study) results and leverage the higher-level hierarchical information as auxiliary information for improving training base classes.

4.3.1 ML Marginalization classifier (MC)

MC is firstly introduced in [7] to make the model aware of the relationship between lower levels and higher levels in the multi-class model. In this study, we extend it to the form of multi-label classification (MLMC). Here, a single classi-
fier \( g(\cdot) \) outputs a probability distribution over the final level (base classes). Instead of having classifiers for the remaining higher \( M - 1 \) levels, we compute the probability distribution over each one of these by summing the probability of their corresponding lower-hierarchical classes. Given the category \( \hat{c}_k \) in a upper level and its corresponding lower categories \( \{c_1, c_2, \ldots, c_o\} \), we have its predicted probabilities:

\[
\hat{Y}_{c_{ik}} = D(\sum_{j=1}^{o} Y_{c_{ij}}^{l_{i-1}});
\]

\[
L_{BCE}^{l_i} = L_{BCE}(\hat{Y}_{c_{ik}}^{l_i}, Y^{l_i}),
\]

where \( D(\cdot) \) denotes the Sigmoid activation function. It should be noted that, for the probabilities over the upper layer of the final layer, we directly use the logits from the output logits \( g(f(X)) \):

\[
L_{BCE} = L_{BCE}(D(g(f(X))), Y^{L_i}).
\]

Obviously, ICS factor can be added into MLMC to adjust the fitting weights for each level.

### 4.4. Hybrid Multiple Knowledge Distillation

[35, 19] have explored knowledge distillation in long-tailed classification. The original long-tailed distribution is divided into several subsets which are in a relatively balanced status. Those subsets are used to train several teacher models which are distilled into a unified student model. However, we find that (1) the divided subsets result in an incomplete distribution and limited features representation learned by teacher model makes it have upper bound of performance; (2) most performance improvements benefit from the knowledge distillation instead of learning from a relatively balanced subset; (3) some rare diseases can not directly divided by independent feature-based or region-based rules proposed by [19]. In this study, we use the entire datasets for the teacher models training and propose a hybrid multiple knowledge distillation method.

Firstly, we trained two teacher models \( F_{t_1} = g_{t_1}(f_{t_1}(\cdot)) \) and \( F_{t_2} = g_{t_2}(f_{t_2}(\cdot)) \) using general classification loss (e.g., MLMC) and re-sampling classification loss (e.g., ICS-MLMC). Inspired by two-stage training [40, 22, 16] that model learns a good representation for feature extractor under an original distribution, we first leverage the feature distillation from \( F_{t_1} \) to the student model. Given the student model \( F_s = g_{s}(f_{s}(\cdot)) \), the feature-level knowledge distillation can be formulated as follows:

\[
L_{F-KD} = l_{sim}(v_{t_1}, v_{s}),
\]

where the extracted feature \( v = f(x) \), and \( l_{sim}() = 1 - \cos(\cdot) \) to calculate the cosine distance (similarity) between output features from teacher and student models. Then, the standard KD is used to distill the knowledge from \( F_{t_2} \) to the student model, which can be formulated as:

\[
L_{C-KD} = l_{KD}(g_{s}(f_{s}(X)/T)|g_{t_2}(f_{t_2}(X)/T)),
\]

where \( T \) is the hyper-parameter for temperature scaling and \( l_{KD} \) is the Kullback-Leibler divergence loss:

\[
l_{KD}(p_1|p_2) = p_1 \cdot \ln \frac{p_1}{p_2}.
\]

Hence, we have the total loss:

\[
L_{total} = (1 - \alpha - \beta)L_{BCE} + \alpha(\gamma L_{F-KD}) + \beta L_{C-KD},
\]

where \( \alpha \) and \( \beta \) are used to control the KD loss weights while \( \gamma \) for prevent \( L_{F-KD} \) from being close to zero. For the simplification of hyper-parameters search, we keep \( \alpha = \beta \).

### 5. Experiments

#### 5.1. Implementation Details

We use ResNet-50 [18] with pre-trained weights from ImageNet as our backbone network. The input size is 512 × 512. We apply Adam to optimize the model. The learning rate starts at \( 1 \times 10^{-3} \) and reduces ten-fold when there is no drop in validation loss till \( 1 \times 10^{-7} \) with the patience of 5. The \( \alpha \) and \( \beta \) are set as \( 0.2 \). The \( \gamma \) is set as \( 10 \). We apply some data-augmentation transformations during the training phase, such as random crop and flip. All the results are calculated after 5-time running with the batch-size of 128. In all presented results, we report the 5-trial average performance. All experiments are implemented with the Pytorch platform and 8 × GTX 3090 GPUs.

#### 5.2. Metrics

Following most long-tailed multi-label works [34, 16], we use the mean average precision (mAP) as the evaluation metric. The sorted classes are divided into many, medium and few groups. Then, the macro mAPs are calculated on those categories belonging to the same group. For the global evaluation, we report the average performance of three groups (denoted by average) and all categories in a macro manner (denoted by global).

#### 5.3. Comparison Study

In this section, we will give a comparison study with various baselines including some state-of-the-art works for long-tailed classification: (1) Empirical Risk Minimization (ERM); (2) vanilla Re-sampling (RS); (3) vanilla Re-weighting (RW); (4) Focal Loss [26]; (5) OLTR [27]; (6) LDAM [5]; (7) CB Loss [6]; (8) DBLoss [34]; (9) ASL [2].

The overall results are shown in Table 2. We first present the performance of many, medium and few to monitor how different approaches affect those groups with the
Table 2. mAP performance of the proposed method and comparison methods on two datasets Retina-100K and Retina-1M.

| Dataset | Retina-100K | Retina-1M |
|---------|-------------|-----------|
| Methods | many | medium | few | average | global | many | medium | few | average | global |
| ERM | 70.89 | 71.93 | 35.90 | 59.57 | 49.40 | 84.70 | 61.57 | 41.83 | 62.70 | 46.91 |
| RS | 65.72 | 67.17 | 36.99 | 56.63 | 48.24 | 80.66 | 58.45 | 42.72 | 60.61 | 44.32 |
| RW | 71.32 | 73.11 | 39.03 | 61.15 | 51.72 | 84.02 | 63.98 | 43.69 | 63.90 | 47.36 |
| Focal Loss [26] | 72.84 | 73.37 | 39.13 | 61.78 | 52.00 | 85.74 | 62.93 | 43.99 | 63.63 | 47.20 |
| OLTR [27] | 70.22 | 72.08 | 39.00 | 60.43 | 50.77 | 80.10 | 62.31 | 42.01 | 61.47 | 45.93 |
| LDAM [5] | 71.24 | 73.75 | 39.12 | 61.37 | 51.88 | 85.62 | 63.79 | 43.45 | 63.65 | 47.26 |
| CBLoss-Focal [6] | 51.93 | 50.66 | 20.79 | 41.13 | 32.18 | 77.14 | 50.69 | 20.66 | 49.50 | 40.11 |
| DBLoss-Focal [34] | 72.61 | 72.39 | 38.59 | 61.27 | 51.37 | 85.99 | 62.15 | 43.22 | 63.79 | 47.30 |
| ASL [2] | 72.94 | 73.67 | 39.21 | 61.94 | 52.15 | 85.10 | 63.79 | 43.45 | 64.11 | 47.55 |
| baseline-original | 71.18 | 72.33 | 37.39 | 60.29 | 50.47 | 85.02 | 62.27 | 42.17 | 63.15 | 47.00 |
| baseline-ICS | 70.67 | 73.40 | 40.95 | 61.67 | 52.94 | 84.86 | 62.39 | 43.44 | 63.56 | 47.12 |
| Ours | 73.86 | 74.75 | 43.82 | 64.14 | 55.41 | 85.79 | 64.00 | 44.28 | 64.69 | 48.23 |

different number of samples. We also calculate the average results of three groups and the global average mAP (macro). Among all baselines, ASL achieves the best results - 61.94% mAP of average results over three groups and 52.15% over all classes on Retina-100K, which is followed by Focal Loss - 61.78% mAP. We also find although RS increases the performance on few-shot classes, but also brings catastrophic destroy to many-shot classes (70.89% → 65.72%) and medium-shot classes (71.93% → 67.17%), resulting from the label co-occurrence under a multi-label setting. We fail to obtain a satisfactory performance in term of CBLoss which is designed for single-label long-tailed classification.

In our framework, we first train two baseline models with different sampling strategies as teacher models, then we will distill the knowledge from two baseline teacher models into a unified student model under a hybrid distillation manner. Those two models are denoted by "baseline-original" and "baseline-ICS" respectively. Take Retina-100K for instance, the results of "baseline-original" show that the hierarchical pre-training can benefit the ERM baseline model (from 59.57% mAP to 60.29% mAP) without any sampling strategy. Then, the ICS further improves the overall performance (from 60.29% mAP to 61.67%) with a little loss on the many-shot group (from 71.18% mAP to 70.67% mAP). The last row shows the performance of our proposed methods, which outperform all competitors on both two-scale datasets Retina-100K and Retina-1M and the superiority holds for all metrics.

5.4. Ablation Study

5.4.1 Components Analysis

To figure out what components make our methods performant, we conduct an ablation study and the results are shown in Table. 3. As we present in Table. 2, we first test our proposed MLMC to have the hierarchical information embedded into the model pre-training. It can be seen that the overall mAP increases from 59.57% to 60.29%. The adoption of the ICS strategy can also bring performance gain to the baseline ERM model, with 0.87% improvement. The result of ICS denotes that this sampling strategy can reduce the risk of oversampling on those samples with more label co-occurrence. Then we test the effectiveness of combining the MLMC pre-training and ICS strategy, the overall mAP reaches 61.67%. The "cRT" denotes that if we train the "baseline-ICS" model in a two-stage manner (e.g., cRT [22]). We first train the baseline model with MLMC but use a regular sampling strategy (e.g., class-balanced sampling). Then we fine-tune the model with ICS strategy and the overall mAP is 62.28%. Overall, the hybrid KD benefits the model most with a 1.86% improvement in terms of the average mAP. The same findings can be found in Retina − 1M too. However, ICS seems to less improve the performance (only 0.18%) since more imbalanced and serious label co-occurrence issue it has.

5.4.2 Distillation

In this section, we evaluate the different knowledge distillation techniques. Our proposed hybrid KD manner consists of two basic KD ways with different intermediate outputs from the DL-based model, e.g., the features from the last convolutional layer (Eq. 6) and logits from the last FC layer (Eq. 7). We show the overall results on Retina-100K in
Fig. 5 and denote the two basic KD techniques as ‘Feature-level’ and ‘Logits-level’ respectively. It should be noted that since we keep $\alpha = \beta$ in Eq. 8, there is different range of the hyper-parameter selection for $\alpha$ or $\beta$. From the results, we can see that both single KD techniques can benefit the baseline model except when $\alpha = 0.5$ for the Feature-level KD, with the 0.95% loss.

Then we investigate how the temperature scaling affects the performance. Three values are considered and $T = 3$ outperform all competitors when $\alpha = [0, 0.1, 0.25]$. The best results are achieved when $T = 3$ and $\alpha = 0.2$. However, when $T = 10$, there is no any result exceeds the baseline model for all $\alpha$ values except $\alpha = 0.2$. It can be concluded that a smaller T value can benefit the KD phase or it will do harm to the performance.

5.5. Group-wise Analysis

5.5.1 In Groups of Shots

In Fig. 6, we visualize the increments of different classes in groups of shots to better view that how different approaches affect the performance. Classic re-sampling strategy wrongly samples the samples with co-occurrence and brings new relatively imbalanced status. Although several categories in tail classes obtain improvements on mAP, the performance of almost all head and middle classes decreased, leading to an overall performance drop. Then, we find that RW is agnostic to the distribution of the data, it will assign fair weights to the classifier when giving the predictions but the improvements are still marginal. ASL and our proposed methods can both help the model learn an overall well-generalized representation and unbiased classifier, which contribute to the global performance gain.

5.5.2 In Groups of Diseases

As we claimed that hierarchy knowledge help train the model with a well-generalized representation since more semantic information is added, this information can be naturally shared across data between coarse or fine categories. In Table 4, we show how the proposed methods improve the performance in higher/coarse levels. Here, we select five representative coarse diseases, which are denoted by mild, macula, vessels, Optic Disc (OD) and Rare diseases. Another widely-used training loss in hierarchical classification Hierarchical Loss [7] is evaluated. Although hierarchical loss can help improve most categories at a coarse level. However, it lacks the ability to tackle the imbalance issue. Our proposed methods can well combine the hierarchical learning and instance-wise re-sampling strategy, and the performance of recognition of those rare diseases has also been significantly improved.

6. Conclusion

In this paper, we discussed the necessity and challenges of training a multi-disease fundus recognition model. Then, we proposed a novel framework that leverages the hierarchical information as prior knowledge for the pre-training. Also, an instance-wise class-balanced sampling strategy and hybrid knowledge distillation were introduced to tackle the multi-label long-tailed issue. We first trained from one million fundus images and the experiment results demonstrated that the effectiveness of our proposed methods outperformed all competitors over all metrics.

Table 4. The comparison of performance on 5 selected coarse categories.

|         | Mild | Macula | Vessels | OD   | Rare |
|---------|------|--------|---------|------|------|
| Original| 66.86| 55.02  | 41.37   | 32.36| 18.28|
| HL [7]  | 67.44| 57.23  | 45.06   | 33.04| 17.95|
| Proposed| 71.33| 58.22  | 49.32   | 38.63| 27.58|
References

[1] T. Araújo, G. Aresta, L. Mendonça, S. Penas, C. Maia, Á. Carneiro, A. M. Mendonça, and A. Camplinho. Dr—graduate: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. Medical Image Analysis, 63:101715, 2020.

[2] E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Proter, and L. Zelnik-Manor. Asymmetric loss for multi-label classification. arXiv preprint arXiv:2009.14119, 2020.

[3] P. M. Burlina, N. Joshi, K. D. Pacheco, D. E. Freund, J. Kong, and N. M. Bressler. Use of deep learning for detailed severity characterization and estimation of 5-year risk among patients with age-related macular degeneration. JAMA ophthalmology, 136(12):1359–1366, 2018.

[4] E. California Healthcare Foundation. Diabetic retinopathy detection. https://www.kaggle.com/c/diabetic-retinopathy-detection.

[5] K. Cao, C. Wei, A. Gaidon, N. Areachiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. arXiv preprint arXiv:1906.07413, 2019.

[6] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9268–9277, 2019.

[7] A. Dhall, A. Makarova, O. Ganea, D. Pavllo, M. Greeff, and A. Krause. Hierarchical image classification using entailment cone embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 836–837, 2020.

[8] M. V. dos Santos Ferreira, A. O. de Carvalho Filho, A. D. de Sousa, A. C. Silva, and M. Gattass. Convolutional neural network and texture descriptor-based automatic detection and diagnosis of glaucoma. Expert Systems with Applications, 110:250–263, 2018.

[9] O. Faust, A. U. Rajendra, E. Y. K. Ng, K. H. Ng, and J. S. Suri. Algorithms for the automated detection of diabetic retinopathy using digital fundus images: A review. Journal of Medical Systems, 36(1):145–157, 2012.

[10] A. Foo, W. Hsu, M. L. Lee, G. Lim, and T. Y. Wong. Multi-task learning for diabetic retinopathy grading and lesion segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 13267–13272, 2020.

[11] H. Fu, J. Cheng, Y. Xu, C. Zhang, D. W. K. Wong, J. Liu, and X. Cao. Disc-aware ensemble network for glaucoma screening from fundus image. IEEE transactions on medical imaging, 37(11):2493–2501, 2018.

[12] R. Gargeya and T. Leng. Automated identification of diabetic retinopathy using deep learning. Ophthalmology, 124(7):962–969, 2017.

[13] W. M. Gondal, J. M. Köhler, R. Grzeszick, G. A. Fink, and M. Hirsch. Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. In 2017 IEEE international conference on image processing (ICIP), pages 2069–2073. IEEE, 2017.

[14] A. Govindaitaiah, M. A. Hussain, R. T. Smith, and A. Bhuiyan. Deep convolutional neural network based screening and assessment of age-related macular degeneration from fundus images. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pages 1525–1528. IEEE, 2018.

[15] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22):2402–2410, 2016.

[16] H. Guo and S. Wang. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15089–15098, 2021.

[17] L.-Z. Guo, Z. Zhou, J.-J. Shao, Q. Zhang, F. Kuang, G.-L. Li, Z.-X. Liu, G.-B. Wu, N. Ma, Q. Li, et al. Learning from imbalanced and incomplete supervision with its application to ride-sharing liability judgment. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[19] L. Ju, X. Wang, L. Wang, T. Liu, X. Zhao, T. Drummond, D. Mahapatra, and Z. Ge. Relational subsets knowledge distillation for long-tailed retinal diseases recognition. arXiv preprint arXiv:2104.11057, 2021.

[20] L. Ju, X. Wang, L. Wang, D. Mahapatra, X. Zhao, M. Harandi, T. Drummond, T. Liu, and Z. Ge. Improving medical image classification with label noise using dual-uncertainty estimation. arXiv preprint arXiv:2103.00528, 2021.

[21] L. Ju, X. Wang, X. Zhao, H. Lu, D. Mahapatra, P. Bonnington, and Z. Ge. Synergic adversarial label learning for grading retinal diseases via knowledge distillation and multi-task learning. IEEE Journal of Biomedical and Health Informatics, 2021.

[22] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis. Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217, 2019.

[23] B. Klein. Overview of epidemiologic studies of diabetic retinopathy. Ophthalmic Epidemiology, 14(4):179–183, 2007.

[24] I. Kocur and S. Resnikoff. Visual impairment and blindness in europe and their prevention. British Journal of Ophthalmology, 86(7):716–722, 2002.

[25] T. Li, W. Bo, C. Hu, H. Kang, H. Liu, K. Wang, and H. Fu. Applications of deep learning in fundus images: A review. Medical Image Analysis, page 101971, 2021.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017.
[27] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. Large-scale long-tailed recognition in an open world. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2537–2546, 2019. 3, 6, 7

[28] P. Massin, A. Chabouis, A. Erginay, C. Viens-Bitker, A. Leclaire-Collet, T. Meas, P.-J. Guillausseau, G. Choupot, B. André, and P. Denormaldie. Ophdiatto©: A telemedical network screening system for diabetic retinopathy in ile-de-france. Diabetes & metabolism, 34(3):227–234, 2008. 2

[29] G. Quellec, M. Lamard, P.-H. Conze, P. Massin, and B. Cochen. Automatic detection of rare pathologies in ophthalmographs using few-shot learning. Medical image analysis, 61:101660, 2020. 2

[30] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. Machine learning, 85(3):333–359, 2011. 3

[31] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor. Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972, 2021. 5

[32] X. Wang, L. Ju, X. Zhao, and Z. Ge. Retinal abnormalities recognition using regional multitask learning. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 30–38. Springer, 2019. 2

[33] Y. Wang, W. Gan, J. Yang, W. Wu, and J. Yan. Dynamic curriculum learning for imbalanced data classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5017–5026, 2019. 3

[34] T. Wu, Q. Huang, Z. Liu, Y. Wang, and D. Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In European Conference on Computer Vision, pages 162–178. Springer, 2020. 3, 4, 5, 6, 7

[35] L. Xiang, G. Ding, and J. Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In European Conference on Computer Vision, pages 247–263. Springer, 2020. 3, 6

[36] Y. Yang, T. Li, W. Li, H. Wu, W. Fan, and W. Zhang. 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, pages 533–540. Springer, 2017. 2

[37] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering, 26(8):1819–1837, 2013. 3

[38] W. Zhang, J. Zhong, S. Yang, Z. Gao, J. Hu, Y. Chen, and Z. Yi. Automated identification and grading system of diabetic retinopathy using deep neural networks. Knowledge-Based Systems, 175:12–25, 2019. 2

[39] R. Zhao, X. Chen, X. Liu, Z. Chen, F. Guo, and S. Li. Direct cup-to-disc ratio estimation for glaucoma screening via semi-supervised learning. IEEE journal of biomedical and health informatics, 24(4):1104–1113, 2019. 2

[40] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9719–9728, 2020. 2, 3, 4, 6