IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 May 2023, accepted 30 May 2023, date of publication 2 June 2023, date of current version 7 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3282455

RESEARCH ARTICLE

Long-Tailed Recognition by Hierarchical Rebalancing Dual-Classifier

JUNSONG ZHANG[1,2], LINSHENG GAO[2], HAO LI[2], AND HAO ZHOU[3]
[1]School of Information and Communication Engineering, Communication University of China, Beijing 100024, China
[2]North China Institute of Science and Technology, Beijing 101601, China
[3]Department of Operational Research and Planning, Naval University of Engineering, Wuhan 430033, China

Corresponding author: Hao Zhou (zhouhao930324@163.com)

This work was supported by the Hubei Provincial Natural Science Foundation of China under Grant 2022CFC049.

ABSTRACT Image classification techniques have succeeded greatly on various large-scale visual datasets using deep convolution neural networks. However, previous deep models usually suffer severe performance degradation in highly skewed datasets, which restricts their practical application. In this paper, we propose a novel Hierarchical Rebalancing Dual-Classifier model for long-tailed recognition. To better identify the tail samples and maintain the performance of head classes, we propose a dual-classifier framework with a uniform sampler for performing their duties. For balancing the learning of feature representation and classifiers, a dynamic weight is introduced to adjust the model's attention. To alleviate the feature deviation between training data and testing data, a hierarchical rebalancing loss is designed for the re-weighting branch, which adjusts the decision values in predicted logits to facilitate the model actively compensating for tail categories. Finally, we conduct extensive experiments on standard long-tailed benchmarks Cifar10-LT, Cifar100-LT, ImageNet-LT, and iNaturalist2018, demonstrating the effectiveness and superiority of our HRDC.

INDEX TERMS Image classification, long-tailed distribution, imbalance learning, dual-classifier framework, hierarchical rebalancing loss, dynamic weight.

I. INTRODUCTION

With the development of deep convolution neural networks (DCNNs), the performance of image classification has achieved great success with high-quality and large-scale datasets [1], [2], [3], [4], such as ImageNet2012 [5], MS-COCO [6], and so on. Unlike the carefully selected images with uniform distributions of labels in these datasets, there are more significant challenges for real-world data, in which they are imbalanced and long-tailed [7], [8], [9]. As shown in Figure 1, the label distribution is highly skewed, where a few categories occupy most of the samples and most categories only have rarely a few data. When tackling such long-tailed data, current deep models are difficult to achieve outstanding performance [10], [11], [12], because they tend to fit the head classes and cause the tail category features to be under-expressed.

In the literature, class rebalancing is the common way to deal with long-tailed distribution, which includes re-sampling [13], [14] and re-weighting [15], [16], [17]. Re-sampling enables the models to be trained with relatively balanced samples through different sampling strategies, which can alleviate the extreme imbalance between different categories [18], [19]. The sampling strategies contain class rebalancing sampling [20], [21], square-root sampling [22], progressively-balanced sampling [16], etc. Re-weighting mainly weights the losses of different categories so that the model can treat different samples more equally [23], [24], [25], and it can alleviate the prejudice of classifiers against these categories. The classical re-weighting strategies are Focal loss [24], Class-balanced loss [20], Equalization loss [23], and so on. Besides, metric learning [26], [27], meta-learning [8], and transfer learning [28], [29], [30] also can

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

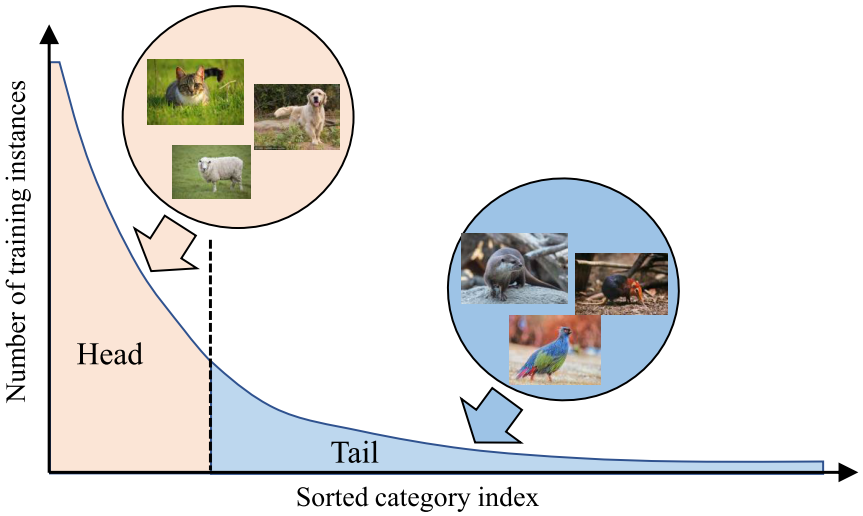

FIGURE 1. The long-tailed data distribution in real-world.

be used to relieve the troubles of long-tailed recognition. However, although the existing methods improve the accuracy of tail classes, it inevitably sacrifices the performance of the head classes, which presents a typical ''seesaw'' [31], [32]. For example, the re-sampling may overfit the tail classes by oversampling or underfit the head classes by undersampling. Zhou et al. [32] proposed that the re-sampling strategies might unexpectedly damage the feature representative ability of the deep models. The re-weighting methods balance the weight of classifiers among different categories, but distort the original distributions [7], [16]. It is difficult for DCNNs to optimize with the re-weighting, which also severely impairs the generalization ability of deep models.

In this paper, we propose a novel Hierarchical Rebalancing Dual-Classifier (HRDC) model for long-tailed visual recognition. HRDC mainly consists of a dual-classifier framework with a unique uniform sampler and a hierarchical rebalancing loss function to guide the training of the HRDC. It aims to improve the performance of tail classes and alleviate the degradation of head performance. Inspired by the BBN model [32], the features learned by cross entropy can be better represented. We design a dual-classifier framework that includes a uniform sampler and two classifiers. A plain classifier equipped with the cross entropy loss is used for learning universal patterns, and another re-weighting classifier is for rebalancing training. In the training process, the focus of HRDC will be gradually shifted from the plain classifier to the re-weighting classifier for both the learning of the feature representation and classifiers. Different from ensemble methods [31], [33], [34], which usually take different sampling strategies for several branches, HRDC adopts the unique uniform sampler for these two classifiers without any dataset division for concise and convenient model training. Furthermore, to alleviate the side effect of re-weighting, we propose a hierarchical rebalancing loss function for the rebalancing branch, which tries to improve the performance of tail classes and maintain the feature representation of head classes. The improvements of Top-1 accuracy on four mainstream long-tailed datasets: Cifar10-LT, Cifar100-LT, ImageNet-LT, and iNaturalist 2018, show the effectiveness of our HRDC than other state-of-the-art methods with re-weighting strategies. The main contributions of our paper are summarized as follows.

- We propose a novel dual-classifier framework with a uniform sampler for long-tailed recognition, in which a plain classifier maintains the performance of head classes and a re-weighting classifier effectively identifies the tail samples.
- We introduce a dynamic weight to adjust the model's attention to different classifiers in training, which can protect the ability of feature representation and improve the classifier learning.
- We design a hierarchical rebalancing loss to guide the training of the re-weighting branch. By adjusting the decision values in predicted logits, the model can effectively alleviate the feature deviation between training data and testing data.

II. RELATED WORKS

A. DATA DISTRIBUTION REBALANCING

Re-sampling is one of the most widely-used schemes for long-tailed recognition to pursue the rebalancing of training samples [13], [16]. It can over-sample the rare instances [18], [35] or under-sample the frequent instances [19], [36] to balance the data distribution. The common ways of re-sampling are random over-sampling and random under-sampling. In recent years, most researches have focused on class rebalancing for long-tailed distribution, including class rebalancing sampling [20], square-root sampling [22], progressively-balanced sampling [16], etc. Different from instance-balanced sampling, each class can be selected with equal probabilities in class-balanced sampling. In square-root

sampling [22], the possibilities of being selected for each class are related to the square root of the frequency of the corresponding category. The progressively-balanced sampling [16] computes the sampling probability by interpolating progressively between instance-balanced and class-balanced sampling. In addition, Dynamic curriculum learning (DCL) [37] proposed a dynamic sampling strategy. As training goes by, the sampling probability will be reduced if more instances from one class are sampled.

B. COST-SENSITIVE LEARNING

Cost-sensitive learning tries to adjust the loss value of different categories to re-balance data distribution [17], [38]. Re-weighting is one of the classical methods [23], [24], [25], [28], [39], which assigns weights to different categories according to the labeling frequency of training samples, namely weighted softmax loss. Park et al. [40] improved the loss by adjusting the influence of label frequencies on loss weights based on sample influence. Ren et al. [20] proposed the balanced softmax loss to adjust model predictions through the label frequencies in the training process, which can soften the biases from long-tailed distribution by the prior knowledge. Without the label frequencies, Cui et al. [41] introduced the ''effective number'' to approximate the number of expected samples of different categories. Lin et al. [24] proposed the Focal loss and explored the re-weighting based on the difficulty of predictions. In equalization loss [23], they directly ignore the loss values of tail-class samples for the negative pairs of head classes to balance the contributions of different categories.

C. ENSEMBLE LEARNING

The ensemble models mainly generate and combine multiple network modules, such as multiple experts and branches, to deal with the problem of long-tail recognition [31], [32], [33], [34], [42], [43]. The existing ensemble methods usually train multi-experts with different dataset divisions [31] or with varying strategies of sampling [32]. Zhou et al. [32] proposed the BBN model, which includes a conventional learning branch with uniform sampler and a rebalancing branch with reversed sampler. Similar to BBN, Wang et al. [42] proposed a dual classification head scheme for long-tail instance segmentation. Xiang et al. [34] divide the long-tailed dataset into several balancing subsets, and each subset is used to train an expert. Cai et al. [31] divide the dataset into multiple skill-diverse and overlap subsets for training experts with different domains.

III. PROPOSED METHODOLOGY

In this section, we discuss the methodology and implementation details of HRDC model. We show the problem formalization and the framework of our HRDC in section III-A. The HRDC can be divided into the following three parts: (1) With the uniform sampler, image features are learned and extracted through the backbone framework in Section III-B; (2) In Section III-C, we propose the dual-classifier framework for long-tailed recognition and design a dynamic-adaptive weight for model training, which shifts the focus of our model from the general learning to imbalancing learning; (3) The hierarchical rebalancing loss has been introduced for re-weighting classifier to improve the tail recognition and maintain the performance of head classes in Section III-D.

A. MODEL OVERVIEW

1) PROBLEM FORMALIZATION

Deep long-tailed visual recognition is to learn a DCNN model from a highly skewed image dataset, where a few head classes have massive samples and lots of tail classes are only a few samples. Let $D_s = \{x_i, y_i\}_{i=1}^{n_T}$ denotes the training set in long-tailed recognition, where the class label of the sample x_i is y_i. The total number of training samples n_T in C categories can be denoted as $n_T = \sum_{k=1}^{C} n_k$, where n_k is the sample number of K-th categories. Without loss of generality, we sort the categories according to the cardinality in decreasing order of sample number. If $i_1 < i_2$, then $n_{i_1} \geq n_{i_2}$ and $n_1 \gg n_K$. The imbalance ratio can be defined as n_1/n_K.

The learning of long-tailed distribution faces two challenges. On the one hand, the imbalanced training data makes the predictions biased toward the head classes. On the other hand, the tail classes are usually under-represented, leading to the poor identification of tail samples. The existing methods seek to improve the accuracy of tail classes by rebalancing and over-sampling, but it inevitably hurts the head classes [16], [32]. In this paper, we design a dual-classifier framework with dynamic-adaptive learning, and propose a hierarchical rebalancing loss for the re-weighting classifier, which can maintain the feature representation of the model and improve the classifier ability for tail classes.

2) THE FRAMEWORK OF HRDC

As shown in Figure 2, we briefly summarize the framework of our Hierarchical Rebalancing Dual-Classifier (HRDC) model. It consists of three modules: (1) Sampler and feature extraction module is responsible for sampling images and extracting visual features through the backbone network. We take the uniform sampler for instance sampling in HRDC. (2) Dual-classifier module includes two classifiers. One of them is used for learning universal patterns from the original data distribution. Equipped with the plain classifier, HRDC can conduct representation learning for better features. Another is used for rebalancing learning by re-weighting, which can pay more attention to the tail samples. The image features are inputted to the dual-classifier module, and then HDRC generates two logits by these classifiers, respectively. The predicted outputs are aggregated and fused by the dynamic-adaptive weight during training. (3) Rebalancing training module contains two kinds of loss functions to guide the model training. The conventional cross-entropy

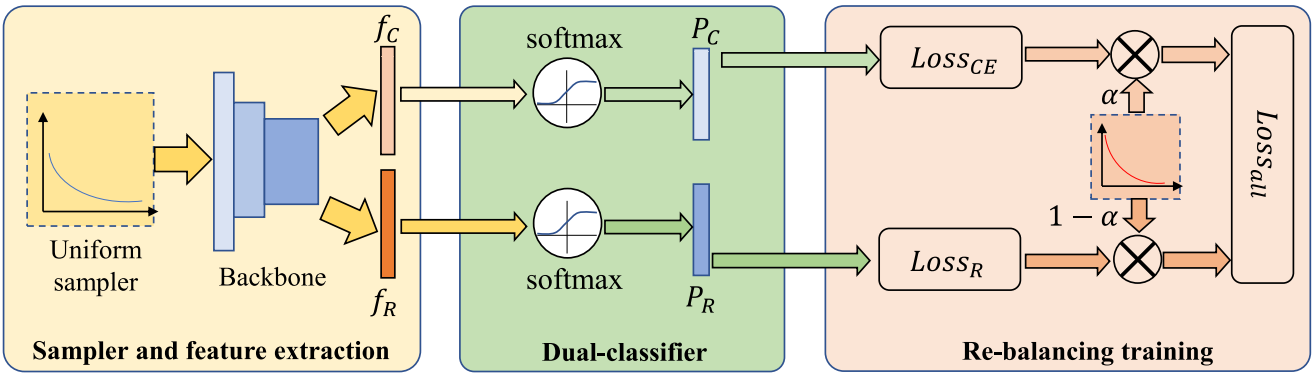

FIGURE 2. The briefly framework of our Hierarchical Rebalancing Dual-Classifier (HRDC) model.

loss is used for plain classifier training. For the rebalancing classifier, we propose a novel hierarchical rebalancing loss to distinguish different frequency categories. Finally, the model will gradually transfer the focus from universal patterns to rebalance learning, to improve the discrimination ability of tail samples and alleviate the decline of the performance of head classes.

B. SAMPLER AND FEATURE EXTRACTION MODULE

The sampler and feature extraction module is the basic network of our model. Different from other multi-branches models, we use the uniform sampler for dual classifier learning without any rebalancing sampling. In instance-balanced sampling, each sample has an equal probability of being selected. We take the ResNet [44] or ResNeXt [45] network as the backbone to learn and extract the image features.

C. THE DESIGN AND FUSION OF THE DUAL-CLASSIFIER FRAMEWORK

In this section, we will elaborate on the details of dual-classifier framework. After the instance sampling, the training samples (x_i, y_i) are inputted into the backbone network for feature extraction, and the model generates the feature vectors $f_i \in \mathbb{R}^D$ via global average pooling. As mentioned above, the dual classifier structure includes a plain classifier for feature learning and a re-weighting classifier for rebalancing learning. The re-weighting classifier is mainly composed of linear classifier and rebalancing loss function. Let ω_R^T, b_R, Φ, and p_R denotes the model classifier, bias, softmax function, and prediction probabilities, respectively. And the model prediction probability p_R^i of the i-th sample in re-weighting branch can be expressed as:

$$p_R^i = \Phi(\omega_R^T f_i + b_R). \tag{1}$$

During training, the model prediction probability p_R^i will be sent into the rebalancing loss to calculate the loss value ℓ_R:

$$\ell_R^i = E_R(p_R^i, y_i), \tag{2}$$

where $E_R(,)$ is the hierarchical rebalancing loss function for re-weighting, which will be elaborated in Section III-D.

Similarly, the plain classifier consists of the linear classifier and conventional cross-entropy loss function. Let ω_C^T, b_C, Φ, and p_C denotes the model classifier, bias, softmax function, and prediction probabilities in the traditional branch, respectively. And the model prediction probability p_C^i of the i-th sample can be computed as:

$$p_C^i = \Phi(\omega_C^T f_i + b_C). \tag{3}$$

And the corresponding loss value ℓ_C is:

$$\ell_C^i = E_C(p_C^i, y_i), \tag{4}$$

where $E_C(,)$ denotes the conventional cross-entropy loss function.

Furthermore, we design a dynamic learning strategy in model training, which enables the attention of the model gradually changes from the traditional branch to re-weighting branch. Previous studies have shown that the model training with cross entropy loss tends to learn better features for superior classification results than other rebalancing ways. Therefore, the model should first focus on the outputs from the plain classifier to strengthen the ability of feature representation. As the training goes by, the model gradually shifts its focus from feature learning to classifier learning. It should pay more attention to the predictions of the re-weighting classifier at the end of the training to improve the contributions and recognition of tail samples. With the dynamic learning strategy, two classifiers perform their duty well for both feature representation learning and tail sample predictions. And our HRDC can avoid damaging the frequent categories when emphasizing the rare samples. Specifically, a dynamic-adaptive weight α is designed to adjust the focus between the loss ℓ_C^i of plain classifier and the loss ℓ_R^i of re-weighting classifier in HRDC. The total loss fused two branches can be computed as:

$$\ell^i = \alpha \ell_C^i + (1-\alpha)\ell_R^i. \tag{5}$$

In the training phase, the loss ℓ_C^i of plain classifier will be multiplied by α, and the loss ℓ_R^i of re-weighting classifier will

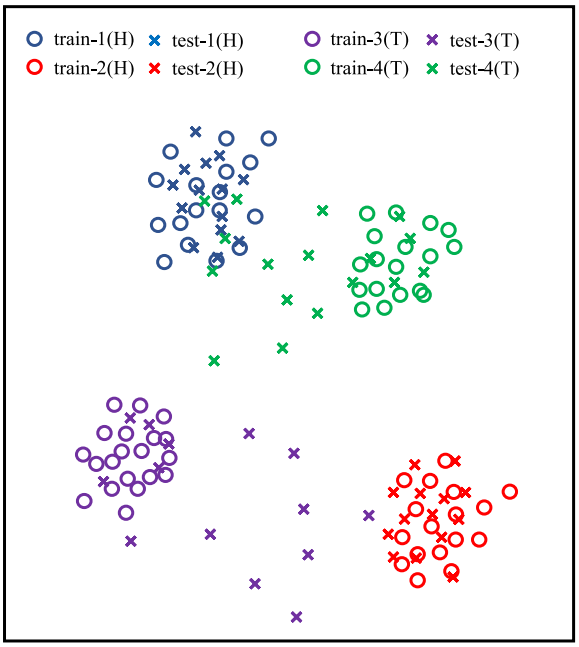

FIGURE 3. The feature deviation between training and test data for tail classes. H: Head classes; T: Tail classes.

be multiplied by $(1-\alpha)$. The dynamic-adaptive weight α is automatically adjusted according to the training epochs. With the progress of training, the weight α will gradually decrease to transfer the model attention from the plain classifier to the re-weighting classifier. Concretely, the weight α can be computed as:

$$\alpha = (\frac{T - T_{\max}}{T_{\max}})^2, \tag{6}$$

where T denotes the current epoch, and $T_{\max}$ denotes the number of total training epochs.

D. HIERARCHICAL REBALANCING LOSS FOR RE-WEIGHTING CLASSIFIER

In this section, we mainly introduce the hierarchical rebalancing loss for the re-weighting classifier. As shown in Figure 3, there is a feature deviation between training and test data for rare categories in long-tailed distribution. And the fewer the training samples of a class are, the larger the deviation is, which may cause the poor performance of tail classes. Therefore, to alleviate the feature deviation, we propose a novel hierarchical rebalancing loss that introduces the hierarchy into the class-dependent temperatures loss (CDT) [46]. Equipped with the hierarchical rebalancing loss, the re-weighting classifier should force the DCNNs to focus on the tail samples in training. And it can enlarge the decision values for tail classes to offset the feature deviation between training and test data.

Inspired by the CDT, we introduce a factor α to re-adjust the logits predicted by the re-weighting classifier. For the i-th item in logits, the decision values l_i can be adjusted as:

$$l_i = \frac{\omega_i^T f(x_n)}{a_i}, \tag{7}$$

where ω_i^T is the i-th weight in re-weighting classifier and $f(x_n)$ is the feature inputted to the classifier. In CDT, the factor $a_i = (\frac{N_{\max}}{N_i})^\gamma$, where γ is the hyperparameter, and $N_{\max}$ is the number of samples for the most frequent category. In general, the factors of tail classes are larger than head classes. Thus, the classifier needs to generate larger logits value for tail samples and pays more attention to the rare categories. However, the model tends to classify samples into header classes in long-tailed distribution. Figure 4 shows the 2-norm of classifier weights for each class, and we can find that the weights of tail classes in the classifier are slight. For the tail classes, the corresponding decision values could be too small to be adjusted by DCNNs, because their classifier weights and the related factors severely reduce the original logits. Therefore, we propose a hierarchical rebalancing loss to adjust the decision values and alleviate the feature deviation, in which the model will adjust the factors according to the different hierarchies.

We divide all categories into three levels according to their number of training samples: head, medium, and tail. The average factors $\bar{a}_h$, $\bar{a}_m$, and $\bar{a}_t$ are calculated by the factors a_i of different subsets for each level, respectively. The new factor a_i^h of head classes can be updated as:

$$a_i^h = \begin{cases} (\frac{N_{\max}}{N_i})^\gamma, & (\frac{N_{\max}}{N_i})^\gamma < \bar{a}_h, \\ \bar{a}_h, & (\frac{N_{\max}}{N_i})^\gamma \geq \bar{a}_h. \end{cases} \tag{8}$$

For the i-th item in the logits belonging to the head level, its factor will be replaced by the average factor $\bar{a}_h$ when the original $(\frac{N_{\max}}{N_i})^\gamma$ is large than $\bar{a}_h$, and be kept with the $(\frac{N_{\max}}{N_i})^\gamma$ otherwise. Similarly, we can calculate the new factors a_i^m of medium classes and a_i^t of tail classes. By replacing the factors greater than the hierarchical average with the average, the classifier can effectively smooth the weights between different categories. In particular, the largest factors of the tail classes will be replaced by the average to prevent their decision values from exceeding the adjustable range. The hierarchical factors can be conducive to distinguishing the boundaries between the head, medium, and tail levels. It also makes the tail samples not be ignored by the model, to promote the model better to compensate for the feature deviation between training data and testing data. In conclusion, the hierarchical rebalancing loss can be expressed as:

$$\ell_R = -\log\left(\frac{\exp(\frac{\omega_{y_n}^T f(x_n)}{a_{y_n}})}{\sum_i \exp(\frac{\omega_i^T f(x_n)}{a_i})}\right). \tag{9}$$

In plain classifier, the probabilities p_C^i predicted by the classifier will be input into the cross entropy function to

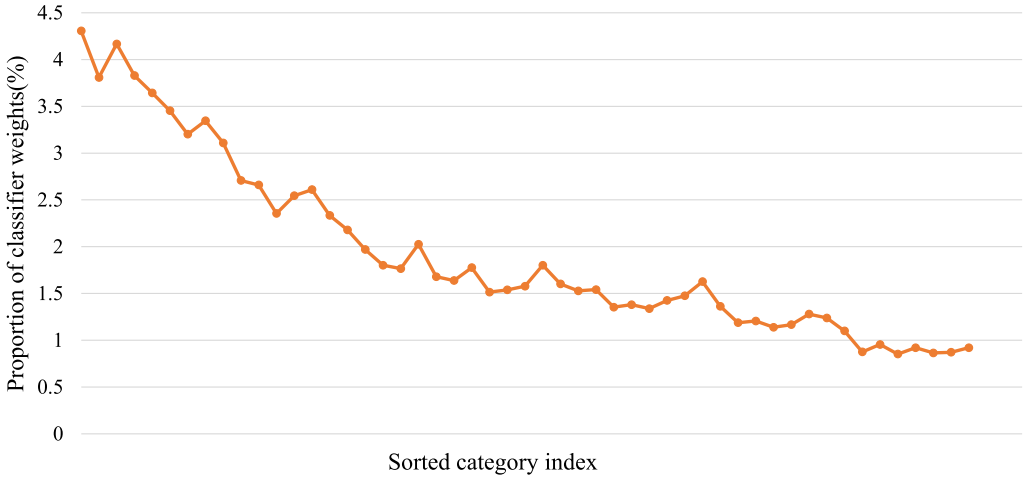

FIGURE 4. The proportion of classifier weights $\|w_i\|$ for each class.

TABLE 1. The details of four imbalanced datasets.

Dataset	Num of classes	Num of sample	Imablance ratio
CIFAR10-LT [15]	10	50K	10,50,100
CIFAR100-LT [15]	100	50K	10,50,100
ImageNet-LT [8]	1K	186K	256
iNaturalist2018 [47]	8K	437K	500

calculate the loss value ℓ_C^i:

$$\ell_C^i = -\sum_{k=1}^{K} y_{i_k} \log(p_C^{i_k}), \tag{10}$$

where $k \in [1, 2, 3, \ldots, K]$ and K is the number of categories.

IV. EXPERIMENTS

A. DATASETS

We follow the standard evaluation protocol [7], [8], [15] in long-tailed classification, in which the models are trained on long-tailed datasets and evaluated on the test set with uniform distribution. We evaluate our model on four long-tailed datasets, including Cifar10-LT [15], Cifar100-LT [15], ImageNet-LT [8], and iNaturalist 2018 [47]. The details of four datasets are shown in Table.1.

Cifar10-LT and Cifar100-LT are derived from the balanced Cifar dataset and contain 10 and 100 categories, respectively. The imbalance factor β is used to adjust the inclination of the training set, where $\beta = \frac{N_{max}}{N_{min}}$ denotes the ratio between the most frequent class to the least frequent class. We set the imbalance factor as 10, 50, and 100, respectively.

ImageNet-LT dataset, proposed by Liu et al. [8], is the long-tailed version of large-scale ImageNet dataset [48]. It contains 186K images and 1000 categories. The number of different categories ranges from 5 to 1280 samples, and its imbalance factor is 256.

iNaturalist 2018 is a real-world large-scale dataset for species classification, which contains 437K images and 8000 categories. It suffers from an extreme distribution imbalance, and the imbalance factor is 500.

B. IMPLEMENTATION DETAILS

For the Cifar10-LT and Cifar100-LT datasets, we follow [15] and [32] to pre-process and enhance the training data. We randomly crop a 32 × 32 patch from the original image or its horizontal flip. Four pixels are padded on each side of images. To maintain consistency with the previous models, we adopt ResNet-32 [44] as the backbone network of our model. The model is trained by standard stochastic gradient descent (SGD) with a momentum of 0.9. The model is trained 200 epochs with batchsize of 128 on an NVIDIA 3090TI GPU. The initial learning rate is 0.1, and the first five epochs are optimized by linear warm-up learning strategy.

For a fair comparison, we conducted experiments on ImageNet-LT and iNaturalist 2018 datasets with the same settings as [32]. We use ResNeXt-50 [45] and ResNet-50 [44] as the backbone networks on ImageNet-LT and iNaturalist 2018 datasets, respectively. For all experiments, the models are trained by standard stochastic gradient descent (SGD) with a momentum of 0.9 on four NVIDIA 2080TI GPUs. The images are cropped by 224 × 224. We also decrease the learning rate from 0.1 with a cosine schedule.

C. COMPARISONS WITH STATE-OF-THE-ART MODELS

1) THE RESULTS ON LONG-TAILED CIFAR DATASETS

We have conducted lots of experiments on CIFAR10-LT and CIFAR100-LT datasets with imbalance ratios 10, 50, and 100. And we compare our results with the re-weighting state-of-the-art models and some models with other re-balancing strategies, including baseline (Cross entropy), Focal loss [24], Mixup [49], CE-DRW [15], CB-Focal [41], LDAM-DRW [15], BBN [32], TDE [7], CDT [46], and so on. The results are shown in Table 2, and the best results are highlighted in bold face.

From Table 2, we can find that our HRDC model achieves performance improvements on all settings and gets optimal results. Specifically, compared with the baseline, the top-1 error rate on CiFar10-LT dataset with imbalance rate

TABLE 2. Top-1 error rates on CIFAR10-LT and CIFAR100-LT datasets with IFs = 10, 50, 100.

Methods	CIFAR10-LT			CIFAR100-LT		
	100	50	10	100	50	10
Baseline	29.64	25.19	13.61	61.68	56.15	44.29
Focal loss [24]	29.62	23.28	13.34	61.59	55.68	44.22
Mixup [49]	26.94	22.18	12.9	60.46	55.01	41.98
CE-DRW [15]	23.66	20.03	12.44	58.49	54.71	41.88
CB-Focal [41]	25.43	20.73	12.9	60.4	54.83	42.01
LDAM-DRW [15]	22.97	18.97	11.84	57.96	53.38	41.29
BBN [32]	20.18	17.82	11.68	57.44	52.98	40.88
TDE [7]	19.40	16.40	11.50	55.90	49.70	40.40
CDT [46]	20.60	-	10.60	55.70	-	41.10
Ours	19.40	16.31	10.45	54.85	49.60	40.16

TABLE 3. Top-1 error rates on ImageNet-LT dataset.

Methods	ResNeXt50
Baseline	52.4
Mixup [49]	53.7
Focal loss [24]	65.4
Range loss [27]	64.9
Lifted [50]	64.8
TDE [7]	48.1
OLTR [8]	62.3
LWS [16]	48.5
ResLT [33]	47.1
Ours	46.8

100 decreases from 29.64% to 19.40%. The HRDC gets 1.20% accuracy improvements than CDT model, which proves that the dual-classifier framework proposed in our HRDC can effectively promote the learning of features and classifiers than those single branch models. Under the extreme imbalance, the top-1 error rate of our HRDC is reduced by 2.59% on CiFar100-LT dataset with imbalance rate 100 than BBN with two sampling branches. Compared with other SOTA LDAM-DRW, TDE, and CDT models, the HRDC achieves 3.11%, 1.05%, and 0.85% accuracy improvements, respectively.

2) THE RESULTS ON LONG-TAILED ImageNet DATASET

Table 3 shows the comparison of our model and other latest models on ImageNet-LT dataset, including baseline (Cross entropy), Mixup [49], Focal loss [24], Range loss [27], Lifted [50], TDE [7], OLTR [8], LWS [16], and ResLT [33]. All models are trained on ResNeXt50 as the backbone network. From Table 3, the HRDC has achieved 46.8% top-1 error rate, with a 5.6% reduction compared with the baseline, which is the best performance among all models. Compared with the TDE, LWS, and ResLT, the accuracy of our HRDC gets 1.3%, 1.7%, and 0.3% gains, respectively, which proves its superiority. Following Cui et al. [33], we also report the results on three divisions: many-shot (more than 100 images), medium-shot (20 100 images), and few-shot (less than 20 images). Figure 5 shows the results of three divisions for different methods. And we achieve the best outstanding performance in head, medium, and tail classes, demonstrating

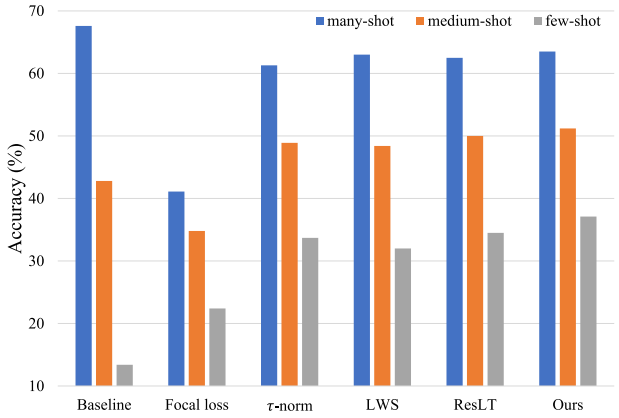

FIGURE 5. The accuracy of many-shot, medium-shot, and few-shot for different methods on ImageNet-LT dataset.

TABLE 4. Top-1 error rates on iNaturalist2018 dataset.

Methods	ResNet50
Baseline	42.8
CE-DRW [15]	36.3
CB-Focal [41]	38.9
LDAM [15]	35.4
LDAM-DRW [15]	32.0
BBN [32]	30.4
LWS [16]	30.5
τ-norm [16]	30.7
TDE [7]	35.6
LADE [51]	30.0
Balanced softmax [20]	30.2
CDT [46]	30.9
Ours	29.8

that our HRDC can maintain the feature representation of head classes and improve the recognition for tail samples.

3) THE RESULTS ON iNaturalist 2018 DATASET

Table 4 shows the results on iNaturalist 2018 dataset. The comparative models include baseline (Cross entropy), CE-DRW [15], CB-Focal [41], LDAM [15], LDAM-DRW [15], BBN [32], LWS [16], τ-norm [16], TDE [7], LADE [51], Balanced softmax [20], and CDT [46]. For iNaturalist 2018, we report our performance with ResNet50 backbone. From Table 4, the Top-1 error rates of the baseline and CDT models are 42.8% and 30.9%. The performance improvement proves that it is effective to adjust the decision value in logits. By designing the dual-classifier framework, the Top-1 error rate of the HRDC decreases by 13.0% and 1.1% respectively than baseline and CDT models and achieves the best performance among all models. Compared with the BBN, Balanced softmax, and LADE models, the performance of our model is improved by 0.6%, 0.4%, and 0.2%, respectively. The improvements demonstrate that the HRDC can reasonably learn the features representations and guide the classifier learning facing the large-scale imbalanced datasets, to better distinguish the head classes from the tail classes and offset feature deviation.

TABLE 5. The results of different loss functions in re-weighting branch.

Classifer1	Classifer2	ACC rate
CE	CE	76.75
	HCEloss	78.72
	CBCE	79.44
	CDT	81.89
	LabelSmooth	77.44
	LabelAwareSmooth	79.02
	CSCE	80.11
	Focal loss	78.21
	SEQL	79.07
	Ours	83.69

TABLE 6. The results of different fusion ways and learning strategies for our HRDC.

Fusion ways \ Learning strategies	Average fusion	Head fusion	Tail fusion
Equal weight	79.6	80.82	80.1
Segmented weight	80.99	82	80.37
Cosine decay	81.5	81.74	82.15
Ours	82.1	83.2	83.69

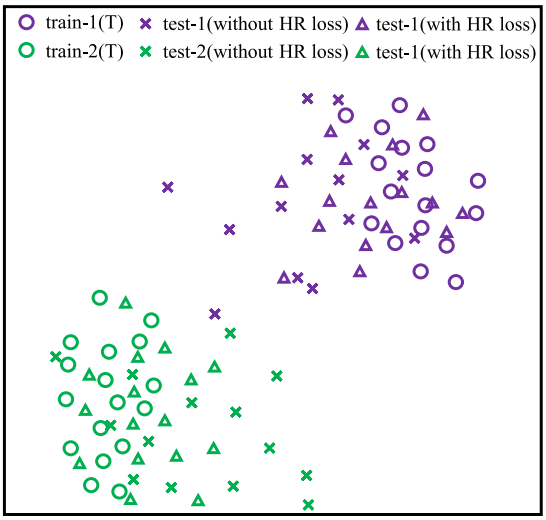

FIGURE 6. The feature deviation in tail classes with and without the hierarchical rebalancing loss (HR loss).

D. ABLATION STUDY

To better verify the effectiveness of the dual-classifier framework and the hierarchical rebalancing loss in HRDC, we conduct two ablation studies.

1) DIFFERENT LOSS FUNCTIONS in RE-WEIGHTING BRANCH

To verify the effectiveness of hierarchical rebalancing loss, we use the classifiers equipped with different loss functions in the re-weighting branch to carry out ablation experiments. Taking the cross entropy function as a benchmark in re-weighting branch, we compare our hierarchical rebalancing loss with other re-weight loss functions, such as HCEloss, CBCE, CDT, LabelSmooth, LabelAwareSmooth, CSCE, Focal loss, and SEQL. Table 5 shows the results of different loss functions. From Table 5, we can find that these cost-sensitive rebalancing losses in the re-weighting branch can effectively improve the recognition performance for tail samples. For example, equipped with the Focal loss, SEQL, and CECS, the performance of the model is improved by 5.48%, 4.62%, and 3.58%, respectively. More importantly, our hierarchical rebalancing loss achieves the best performance, and its accuracy is 83.69%, which is 1.80% higher than CDT loss. Figure 6 shows the feature deviation in tail classes with and without the hierarchical rebalancing loss. And we can easily find that the spatial distributions of triangles are more similar to that of circles, which proves that the hierarchical rebalancing loss alleviates feature deviation for tail classes.

2) DIFFERENT FUSION WAYS AND LEARNING STRATEGIES FOR DUAL-CLASSIFIER FRAMEWORK

To better understand the dual classifiers framework, we explore different dynamic learning strategies between two branches during training and different fusion weights for dual classifiers in inference. To transfer the model attention, we design three dynamic learning strategies, including equal weight, segmented weight, and cosine decay. For the fusion of two classifiers, we design three different fusion methods, including average fusion, head fusion, and tail fusion.

The results are shown in Table 6. We can find that the performance of equal weight is worse than other learning strategies, which shows that dynamically adjusting the weights between different branches is conducive to the learning of features and classifiers at different stages. Our model gets the best 83.69% accuracy, with 3.59%, 3.32%, and 1.54% improvements than equal weight, segmented weight, and cosine decay. In addition, the performance of tail fusion is better than average fusion and head fusion. Its performance achieves 1.59% and 0.49% improvements, which is adopted in our model.

V. CONCLUSION

In this paper, we propose a novel Hierarchical Rebalancing Dual-Classifier model (HRDC) for long-tailed visual recognition, in which each classifier performs its duties for the learning of feature representation and classifiers. To balance different branches, a dynamic weight is introduced to our dual-classifier framework for shifting the model focus from the plain classifier to the re-weighting classifier during training. To alleviate the feature deviation, we design a hierarchical rebalancing loss for re-weighting branch. By altering the decision values in predicted logits, our model will try to compensate the tail samples actively. Finally, by conducting extensive experiments on four long-tailed datasets, we proved that our HRDC could achieve the best results on imbalanced benchmarks. And the ablation studies further verify the effectiveness of all modules.

REFERENCES

[1] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6450–6458.

[2] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, G. Izacard, A. Joulin, G. Synnaeve, J. Verbeek, and H. Jégou, "ResMLP: Feedforward networks for image classification with data-efficient training," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 5314–5321, Apr. 2023.

[3] D. Xue, X. Zhou, C. Li, Y. Yao, M. M. Rahaman, J. Zhang, H. Chen, J. Zhang, S. Qi, and H. Sun, "An application of transfer learning and ensemble learning techniques for cervical histopathology image classification," *IEEE Access*, vol. 8, pp. 104603–104618, 2020.

[4] G. Liang and H. Wang, "I-CNet: Leveraging involution and convolution for image classification," *IEEE Access*, vol. 10, pp. 2077–2082, 2021.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[7] K. Tang, J. Huang, and H. Zhang, "Long-tailed classification by keeping the good and removing the bad momentum causal effect," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1513–1524.

[8] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2532–2541.

[9] H. Zhou, J. Zhang, T. Luo, Y. Yang, and J. Lei, "Debiased scene graph generation for dual imbalance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4274–4288, Apr. 2023.

[10] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[11] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[13] T. Wang, Y. Li, B. Kang, J. Li, J. Liew, S. Tang, S. Hoi, and J. Feng, "The devil is in classification: A simple framework for long-tail instance segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 728–744.

[14] Z. Zhang and T. Pfister, "Learning fast sample re-weighting without reward data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 705–714.

[15] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[16] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[17] P. Zhao, Y. Zhang, M. Wu, S. C. H. Hoi, M. Tan, and J. Huang, "Adaptive cost-sensitive online classification," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 2, pp. 214–228, Feb. 2019.

[18] J. Byrd and Z. Lipton, "What is the effect of importance weighting in deep learning?" in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 872–881.

[19] H. Guo and S. Wang, "Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15084–15093.

[20] J. Ren, C. Yu, X. Ma, H. Zhao, and S. Yi, "Balanced meta-softmax for long-tailed visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4175–4186.

[21] S. Guo, R. Liu, M. Wang, M. Zhang, S. Nie, S. Lina, and N. Abe, "Exploiting the tail data for long-tailed face recognition," *IEEE Access*, vol. 10, pp. 97945–97953, 2022.

[22] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 181–196.

[23] J. Tan, C. Wang, B. Li, Q. Li, W. Ouyang, C. Yin, and J. Yan, "Equalization loss for long-tailed object recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11659–11668.

[24] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[25] M. A. Jamal, M. Brown, M. Yang, L. Wang, and B. Gong, "Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7607–7616.

[26] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5375–5384.

[27] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5419–5428.

[28] Y.-X. Wang, D. Ramanan, and M. Hebert, "Learning to model the tail," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[29] Y. Zhong, W. Deng, M. Wang, J. Hu, J. Peng, X. Tao, and Y. Huang, "Unequal-training for deep face recognition with long-tailed noisy data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7804–7813.

[30] F. Zhang, H. Fan, K. Wang, Y. Zhao, X. Zhang, and Y. Ma, "Research on intelligent target recognition integrated with knowledge," *IEEE Access*, vol. 9, pp. 137107–137115, 2021.

[31] J. Cai, Y. Wang, and J. Hwang, "ACE: Ally complementary experts for solving long-tailed recognition in one-shot," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 112–121.

[32] B. Zhou, Q. Cui, X. Wei, and Z. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9716–9725.

[33] J. Cui, S. Liu, Z. Tian, Z. Zhong, and J. Jia, "ResLT: Residual learning for long-tailed recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3695–3706, Mar. 2023.

[34] L. Xiang, G. Ding, and J. Han, "Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 247–263.

[35] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput.* Cham, Switzerland: Springer, 2005, pp. 878–887.

[36] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 39, no. 2, pp. 539–550, Apr. 2009.

[37] Y. Wang, W. Gan, J. Yang, W. Wu, and J. Yan, "Dynamic curriculum learning for imbalanced data classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5016–5025.

[38] Y. Zhang, P. Zhao, S. Niu, Q. Wu, J. Cao, J. Huang, and M. Tan, "Online adaptive asymmetric active learning with limited budgets," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2680–2692, Jun. 2021.

[39] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Deep imbalanced learning for face recognition and attribute prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2781–2794, Nov. 2020.

[40] S. Park, J. Lim, Y. Jeon, and J. Y. Choi, "Influence-balanced loss for imbalanced visual classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 715–724.

[41] Y. Cui, M. Jia, T. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9260–9269.

[42] X. Wang, L. Lian, Z. Miao, Z. Liu, and S. X. Yu, "Long-tailed recognition by routing diverse distribution-aware experts," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[43] X. Dong and J. Shen, "Triplet loss in Siamese network for object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 459–474.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[45] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[46] H.-J. Ye, H.-Y. Chen, D.-C. Zhan, and W.-L. Chao, "Identifying and compensating for feature deviation in imbalanced deep learning," 2020, *arXiv:2001.01385*.

[47] G. Van Horn, O. M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The iNaturalist species classification and detection dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8769–8778.
[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[49] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018.
[50] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4004–4012.
[51] Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, and B. Chang, "Disentangling label distribution for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6622–6632.

JUNSONG ZHANG received the bachelor's degree in electronic and information engineering from the North China Institute of Aerospace Engineering, in 2009, and the master's degree in communication and information system from the Communication University of China, in 2011, where he is currently pursuing the Ph.D. degree in communication and information system. He is also a Lecturer with the North China Institute of Science and Technology. He has authored more than ten articles in journals and conferences. His research interests include image recognition and artificial intelligence.

LINSHENG GAO received the bachelor's and master's degrees in mining engineering from the China University of Mining and Technology, in 2010 and 2013, respectively. He is currently a Lecturer with the North China Institute of Science and Technology. He has authored more than 30 articles in journals and conferences. His research interests include underground coal mining and mine disaster prevention.

HAO LI received the bachelor's degree in safety engineering from the China University of Mining and Technology, in 2005, and the master's degree in mining engineering from the Xi'an University of Science and Technology, in 2010. He is currently a Lecturer with the North China Institute of Science and Technology. He has authored more than ten articles in journals and conferences. His research interest includes application of optical fiber sensing technology in mine.

HAO ZHOU received the B.S. degree in information system engineering and the M.S. and Ph.D. degrees in control science and engineering from the National University of Defense Technology, Changsha, China, in 2014, 2016, and 2021, respectively. He is currently a Research Assistant with the Naval University of Engineering. He has authored more than 20 articles in journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Public Relations Journal*, and ICME. His current research interests include causal inference, scene graph generation, and image understanding. He was a Reviewer of several conferences, including CVPR and ICCV.