ARConvL: Adaptive Region-Based Convolutional Learning for Multi-class Imbalance Classification

Shuxian Li[1,2,3], Liyan Song(✉)[1,2], Xiaoyu Wu[4], Zheng Hu[4], Yiu-ming Cheung[3], and Xin Yao(✉)[1,2]

[1] Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology (SUSTech), Shenzhen, China.
[2] Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation, Department of Computer Science and Engineering, Southern University of Science and Technology (SUSTech), Shenzhen, China.
[3] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China.
[4] RAMS Reliability Technology Lab, Huawei Technology Co., Ltd., Shenzhen, China.

Abstract. Real-world image classification usually suffers from the multi-class imbalance issue, probably causing unsatisfactory performance, especially on minority classes. A typical way to address such problem is to adjust the loss function of deep networks by making use of class imbalance ratios. However, such static between-class imbalance ratios cannot monitor the changing latent feature distributions that are continuously learned by the deep network throughout training epochs, potentially failing in helping the loss function adapt to the latest class imbalance status of the current training epoch. To address this issue, we propose an adaptive loss to monitor the evolving learning of latent feature distributions. Specifically, the class-wise feature distribution is derived based on the region loss with the objective of accommodating feature points of this class. The multi-class imbalance issue can then be addressed based on the derived class regions from two perspectives: first, an adaptive distribution loss is proposed to optimize class-wise latent feature distributions where different classes would converge within the regions of a similar size, directly tackling the multi-class imbalance problem; second, an adaptive margin is proposed to incorporate with the cross-entropy loss to enlarge the between-class discrimination, further alleviating the class imbalance issue. An adaptive region-based convolutional learning method is ultimately produced based on the adaptive distribution loss and the adaptive margin cross-entropy loss. Experimental results based on public image sets demonstrate the effectiveness and robustness of our approach in dealing with varying levels of multi-class imbalance issues.

Keywords: Multi-class imbalance classification · Deep learning · Adaptive loss · Feature engineering · Margin.

1 Introduction

In real-world applications of image classification such as human behavior recognition [24], video classification [27], and medical decision making [19], image

classes usually exhibit the multi-class imbalance issue, for which some classes are under-represented as minorities while others are over-represented as majorities. Catering for this multi-class imbalance issue is important to retain good predictive performance, especially for those minority classes. Taking image classifications in the medical domain as an example, the number of images related to rare diseases is usually much less than those related to common diseases and healthy cases. Neglecting this issue would probably result in poor predictive performance on minority classes, which poses severe threats to patients afflicted with rare diseases and even undermines the public health service system [19].

Existing approaches for multi-class imbalance learning can be grouped into three categories [13,36,15,41] that are data-level [25,37], model-level [42,38], and cost-sensitive approaches [20,6,23]. Cost-sensitive approach is the most popular and efficient approach in mitigating the image multi-class imbalance issue, which designates different weights to training samples of different classes to adjust the loss function [41]. The weighting is typically designed based on class imbalance ratios [20,6,3,23]. However, imbalance ratios remain static and cannot monitor the changing latent feature distributions that are continuously learned by the deep network throughout training epochs, potentially failing in helping the loss functions to adapt to the latest class imbalance status. Latent feature distributions have shown to be beneficial to the multi-class imbalance learning [11,21], and thus is especially taken into account in this paper.

To the best of our knowledge, there have been only a few studies employing derived latent feature distributions to facilitate multi-class imbalance learning [11,21]. However, they all rely on strict assumptions of the latent feature distribution such as the Gaussian distribution, and cannot adaptively learn the latent feature distribution of entire training samples [11,21]. Our approach enables practical learning of the latent feature space by defining a class-wise *region*, within which most feature points of the same class can be enclosed. Concretely, each region in the latent feature space corresponds to a single class, which is outlined by a *class center* depicting the geometric location of the feature points of that class and a *radius* depicting the spread of the feature points around the class center. In this way, we do not need to rely on any strict assumption on the latent feature distribution, which will be adaptively learned throughout training epochs. Our region learning module utilizes the region loss to derive the class-wise region over time during the training process of the deep network.

The class imbalance problem can then be addressed based on the derived class regions in the latent feature space from two perspectives. First, we propose an adaptive distribution loss to guide the learning process of the class-wise latent feature distribution, so that all class regions can be gradually enclosed within a benchmark radius. As a result, the decision boundary constructed based on class regions in the latent feature space would be unbiased towards any class, dealing with class imbalance directly. Second, we propose an adaptive margin as a mediator to upgrade the original cross-entropy so that between-class discrimination can be enlarged to eliminate possible overlaps of feature points of different classes, further alleviating the class imbalance issue. Ultimately, we

construct the convolutional networks by optimizing the adaptive distribution loss and the adaptive margin cross-entropy loss simultaneously, producing our **A**daptive **R**egion-Based **Conv**olutional **L**earning (ARConvL). In summary, our main contributions are:

- We propose a region learning module based on the region loss to derive class-wise regions, each of which consists of a center and a radius, continuously monitoring the class distribution without posing any strict assumption on the latent feature space;
- Based on the derived class regions, we propose an adaptive distribution loss to optimize the class-wise latent feature distribution, so that feature points of different classes are optimized to be enclosed within a benchmark radius, addressing the class imbalance problem directly;
- Based on the derived class regions, we propose an adaptive margin as a mediator to upgrade the loss function, producing our adaptive margin cross-entropy loss, so that the between-class discrimination can be improved, further alleviating multi-class imbalance learning;
- We experimentally investigate the effectiveness and robustness of our proposed ARConvL in dealing with different levels of class imbalance.

The remainder of this paper is organized as follows. Section 2 presents related work. Section 3 proposes ARConvL. Experimental setup and results are discussed in Section 4. The paper is concluded in Section 5.

2 Related Work

2.1 Multi-class Imbalance Learning

Existing approaches of multi-class imbalance learning can be generally grouped into three categories: data-level approaches, model-level approaches, and cost-sensitive approaches [13,36,15,41].

Data sampling is a representative of data-level approaches, which synthetically balance the training set by under-sampling the majorities or (and) over-sampling the minorities in the image space [13]. Traditional sampling methods such as RUS (Random Under-Sampling) [13], ROS (Random Over-Sampling) [13], SMOTE (Synthetic Minority Over-sampling Technique) [4], and ADASYN (Adaptive Synthetic Sampling Approach) [12] are typically used for class imbalance learning with the numerical features. Several studies extend ROS and RUS to the image data [18,2]. Due to the popularity of deep learning, generative models are also widely used as over-sampling techniques to tackle the multi-class imbalance problem for the image data [25,37].

Ensemble learning is a representative of model-level approaches that has been popularly used for multi-class imbalance learning. Good examples are AdaBoost [10], AdaBoost.NC [35,36], SMOTEBoost [5], and RUSBoost [29]. Methods of this category need to train multiple classifiers, and when it comes to deep learning, the process would often be time-consuming [32,42,38].

Cost-sensitive methods deal with the class imbalance issue by designating different weights to training samples or (and) classes to distinguish the losses posed to the majority vs the minority classes. [9,13,1]. The weights are usually incorporated with the loss function of the deep network to deal with the multi-class imbalance problem for image data [20,6]. Such methods that make alterations to the loss function are also known as the loss modification based methods, for which we present into more details in the subsequent section.

2.2 Loss Modification Based Methods

Methods of this type deal with multi-class imbalance problem via the modification of the loss function in the deep networks. Re-weighting and logit adjustment are two common approaches for loss modification [41].

For re-weighting approaches, the sample (class) weights are usually encoded into the cross-entropy loss or softmax, rephrasing the loss functions [41]. The main challenge of the re-weighting approaches is how to set proper weights. Lin et al. design sample weights based on their classification difficulties and class imbalance ratios, which are then incorporated into the cross-entropy loss, contributing to the focal loss [20]. Cui et al. design class weights based on a novel effective number, contributing to the class-balanced loss [6]. Besides encoding sample (class) weights into the loss, more related strategies include setting weights directly on the softmax function [28,33].

Logit adjustment approaches tune the logit value of the softmax function to tackle the multi-class imbalance problem [41]. Margins between classes can be produced based on class imbalance ratios to adjust the logit [3]. In 2020, Liu et al. encode the information of feature distribution based on the Gaussian distribution assumption into the logit, enlarging the margin between the minority classes and the majority classes [21]. More recently, Menon et al. adjust the logit based on label frequency to distinguish margins of different classes, contributing to the logit adjustment loss [23].

This paper aims for proposing an adaptive region-based learning method to derive feature distributions adaptively across training epochs without posing any strict assumption to the distribution. With this in mind, we can upgrade the loss function, dealing with the class imbalance issue.

2.3 Convolutional Prototype Learning

Compared to the traditional CNN, Convolutional Prototype Learning (CPL) utilizes L_2-norm, rather than the cosine similarity, to compute the distance (similarity) between the feature point and its connection weight vector [40]. Thus, feature distributions learned with L_2-norm can present hyper-sphere distributions in the latent feature space. The connection weights between the feature layer and the distance output layer can be used as centers of classes. In 2018, Yang et al. point out that this framework would learn more robust features especially after adding an extra regularization term [40]. Later in 2019, Hayat et al. further propose the affinity loss based on L_2-norm, where a hyper-parameter needs to be predefined

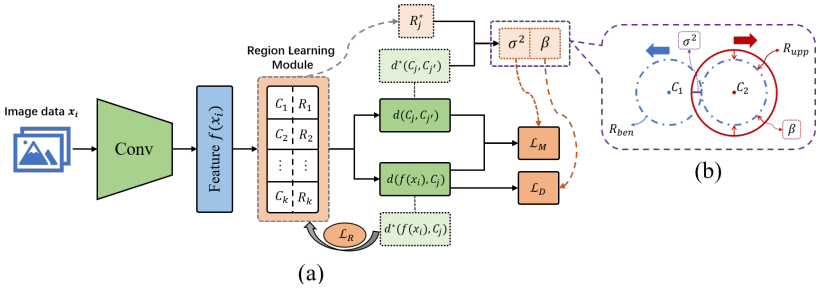


Fig. 1. Overview learning process of our proposed ARConvL on each training batch. Fig.1(b) further illustrates the benchmark radius, upper-bound radius, and corresponding margin σ^2 that measures the potential overlap between class regions.

and encoded to the loss to manipulate the margin manually [11]. By considering the distribution of class centers, Hayat et al. also propose a loss to force class centers evenly distributed, so that the discrimination between classes would be similar, alleviating the class imbalance problem [11].

CPL has the advantage of describing geometric characteristics of decision boundaries in a straightforward way such as hyper-sphere. Thus, we opt for CPL as our base framework to learn feature distributions of image data.

3 ARConvL

This section proposes **A**daptive **R**egion-Based **Conv**olutional **L**earning (ARConvL) for multi-class imbalance learning. Section 3.1 outlines the learning framework, followed by Section 3.2 presenting the way to adaptively learn class-wise regions across the training process. Sections 3.3 and 3.4 adaptively optimize the class-wise latent feature distribution and adaptively produce the between-class margin cross-entropy loss, respectively, for multi-class imbalance learning.

3.1 Overview of ARConvL

Figure 1 shows the learning process of ARConvL on each training batch $\{x_i, i = 1, \cdots, n\}$, based on which feature points $\{f(x_i)\}$ in the latent feature space are trained via convolutional layers. Such latent features are then connected with the region learning module to derive class regions, each of which consists of a class center C and a class radius R learned from the region loss $\mathcal{L}_R$. Based on the class regions, two loss functions are proposed from two perspectives to cater for class imbalance learning: the first perspective aims to optimize the class-wise latent feature distribution to be enclosed within a region with a benchmark radius based on the distribution loss $\mathcal{L}_D$; the second perspective aims to enlarge the distance between class regions via the proposed margin as a mediator to produce the margin loss $\mathcal{L}_M$. In formulation, the loss of ARConvL is

$$\mathcal{L} = \mathcal{L}_R + \mathcal{L}_D + \mathcal{L}_M \ , \tag{1}$$

where $\mathcal{L}_R$, $\mathcal{L}_D$, and $\mathcal{L}_M$ are the region, distribution, and margin cross-entropy loss functions, respectively. We will present the design of the three loss functions in detail in the subsequent sections.

As explained in Section 2.3, we adopt CPL as the base framework to boost geometric characteristics of the latent feature space [40] and thus L_2-norm rather than the cosine similarity is adopted as the distance metric in this paper. Accordingly, the distance of a sample from a region is $d(f(x_i), C_j) = \|f(x_i) - C_j\|_2$ and the distance between two regions is $d(C_j, C_{j'}) = \|C_j - C_{j'}\|_2$, where $j, j' \in \{1, \cdots, k\}$, $i \in \{1, \cdots, n\}$, k is class number, and n is training batch size.

3.2 Region Learning Module

The region learning module aims to compute class-wise regions based on feature points adaptively across training epochs. In particular, class centers $\{C_j\}$ are obtained as the network weights connecting the learned latent features. We then propose the region loss $\mathcal{L}_R$ to learn the radius R_j for each class j as

$$\mathcal{L}_R = \frac{1}{n} \sum_{j=1}^{k} \alpha_j [\sum_{\substack{i=1, \\ x_i \in j}}^{n} (\max\{0,\ d^*(f(x_i), C_j) - R_j\}^2 + \gamma \cdot R_j^2)], \tag{2}$$

where n is the training batch size, $\alpha_j = \max_{j'}(N_{j'})/N_j$ quantifies the emphasis to be placed on class j versus other classes, and N_j denotes the total number of training samples of class j. We are employing $\gamma \cdot R_j^2$ as the regularization term for $\gamma \in [0, 1]$. We set $\gamma = 0.05$ in this paper based on our preliminary experiments. Figure 2 illustrates the mechanism of the region loss: if a feature point falls within its corresponding class region, it causes zero penalty to $\mathcal{L}_R$; otherwise, this feature point contributes the penalty of $[d(x_i, C_{y_i}) - R_{y_i}]^2$ to $\mathcal{L}_R$.

We presume that class radii $\{R_j\}$ are learnable variables, and other variables such as variables in $d(f(x_i), C_j)$ are frozen as non-learnable[5]. We attach the superscript "*" to variables to indicate that they are non-learnable variables. For example, the non-learnable distance between feature point and class center is denoted as $d^*(f(x), C)$ in Fig. 1.

Figure 3 illustrates class regions learned by the region learning module of ARConvL. We deliberately set the two-dimensional latent feature space to facilitate visualization. In our experimental studies, the latent feature dimension is set to 64 to attain good performance. We can see that the majority classes often learn regions of larger radii compared to those of minorities.

3.3 Optimizing Class-Wise Latent Feature Distribution

Based on the derived class regions, this section aims to optimize the latent feature distribution of each class to be enclosed within a region surrounding the class center with a benchmark radius that is universal to all classes and

[5] In Tensorflow, we can annotate and freeze non-learnable variables using the command `get_static_value()`.

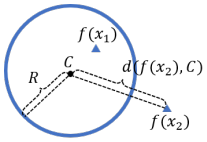


Fig. 2. Illustration of the region loss $\mathcal{L}_R$ in Eq. (2). Given a class region with center C and radius R, $f(x_1)$ locates in the region and thus contributes null penalty to $\mathcal{L}_R$; whereas, $f(x_2)$ contributes the penalty of $[d(f(x_2), C) - R]^2$ to $\mathcal{L}_R$.

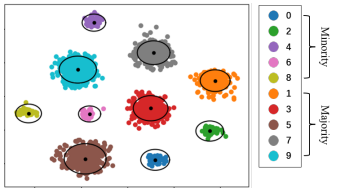


Fig. 3. Illustration of the class regions in the latent feature space that are learned by the region learning module of ARConvL in MNIST.

can accommodate most feature points of this class. In this way, the decision boundary constructed based on class regions in the latent feature space would be unbiased towards any class, dealing with the class imbalance issue directly. Thereby, one can rely on the class region to optimize the class-wise latent feature distribution which does not need to pose any assumption (e.g., Gaussian) on it.

The *benchmark radius* is defined for all classes based on the derived class regions as

$$R_{ben} = \min_{j \in \{1, \cdots k\}} d_{\min}(C_j)/2, \tag{3}$$

where $d_{\min}(C_j) = \min_{j' \neq j} d^*(C_j, C_{j'})$ is the minimum distance of two different class centers and $d^*(C_j, C_{j'})$ is a non-learnable variable as explained in Sec. 3.1. The *upper-bound radius* of all classes is formulated as

$$R_{upp} = \max_{j=1, \cdots, k} (R_j^*, \ \ d_{\min}(C_j)/2 \ ). \tag{4}$$

where R_j^* is the radius of class j and is non-learnable in the learning process of distribution loss. The upper-bound radius indicates the possible largest value that feature points of each class would spread surrounding the class center. We have $R_{ben} \leq R_{upp}$.

The idea is to move from the upper-bound radius towards the benchmark radius downside, as shown in Fig.1(b), enforcing the class-wise latent feature distribution to be enclosed within similar-sized regions. Thereby, the decision boundary would be unbiased towards/against any class. We formulate the ideal scenario of the class regions as $R_{ben} = R_{upp}$. However, in the practical learning process of latent feature distribution, $R_{upp} > R_{ben}$ frequently happens, so that the classes having larger radii than the benchmark value should be penalized. The idea is formulated as

$$\beta = \min(1, \ \ (R_{upp}/R_{ben})^2 - 1), \tag{5}$$

quantifying the emphasis on $\mathcal{L}_D$ versus other loss functions of $\mathcal{L}$ in Eq. (1)

Overall, we integrate the *adaptive distribution loss* $\mathcal{L}_D$ to optimize the latent feature distribution of each class as

$$\mathcal{L}_D = \frac{\beta}{n} \sum_{j=1}^{k} \alpha_j \sum_{\substack{i=1, \\ x_i \in j}}^{n} d^2(f(x_i), C_j), \tag{6}$$

where $\sum_{x_i \in j} d(f(x_i), C_j)^2$ accumulates the distance of feature points of class j from their class center C_j and α_j and β are with the same meaning as in Eqs. (2) and (5), respectively. In particular, when the upper-bound R_{upp} approaches the benchmark R_{ben} downside, $\beta \to 0$, leading to zero penalty to $\mathcal{L}_D$.

By introducing an adaptive penalization β, $\mathcal{L}_D$ can also be viewed as a regularization term of the overall loss $\mathcal{L}$ in Eq. (1). Our experiments show that $\mathcal{L}_D$ has significant benefit to the prediction performance, being consistent with the study of Yang et al. [40].

3.4 Enlarging Margin Between Classes

We define the *adaptive margin* as

$$\sigma^2 = \min(R_{ben}^2, \quad R_{upp}^2 - R_{ben}^2), \tag{7}$$

where R_{ben} and R_{upp} are the benchmark and upper-bound radii in Eqs. (3) and (4), respectively. As shown in Fig.1(b), margin σ^2 measures the overlap between the derived class regions in the latent feature space, for which the learning algorithm should have pushed the class regions away from each other to improve their discrimination.

To incorporate the margin into the cross-entropy loss, we rephrase the softmax function for a given feature point x_i as

$$p(x_i \in j) = \frac{\eta_j \cdot e^{-d^2(f(x_i), C_j)}}{\eta_j \cdot e^{-d^2(f(x_i), C_j)} + \sum_{j' \neq j} \eta_{j'} \cdot e^{-d^2(f(x_i), C_{j'}) + \sigma^2}},$$

where $\eta_j = N_j / (\sum_{j'=1}^{k} N_{j'})$ is the imbalance ratio that was shown to be beneficial for class imbalance learning when embedded into the softmax function [23] and N_j denotes the total number of samples of class j. To push the class centers being evenly distributed in the latent feature space, we formulate the softmax function of class center C_j based on the margin σ^2 as

$$p(C_j) = \frac{e^{-d^2(C_j, C_j)}}{e^{-d^2(C_j, C_j)} + \sum_{j' \neq j} e^{-d^2(C_j, C_{j'}) + \sigma^2}} = \frac{1}{1 + \sum_{j' \neq j} e^{-d^2(C_j, C_{j'}) + \sigma^2}}.$$

Overall, we integrate the *adaptive margin cross-entropy loss* $\mathcal{L}_M$ for this training batch across all classes as

$$\mathcal{L}_M = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} -\log(p(x_i \in j)) + \frac{1}{k} \sum_{j=1}^{k} -\log(p(C_j)). \tag{8}$$

Optimizing $\mathcal{L}_M$ can simultaneously enlarge margins between classes at both the feature point level and the class center level, alleviating the class imbalance issue.

4 Experimental Studies

This section aims to investigate the effectiveness of the proposed ARConvL through a series of experiments. The code is available online [6].

[6] Code and supplementary material: https://github.com/shuxian-li/ARConvL

4.1 Experimental Setup

Following previous studies [11,25,37], our experiments are conducted based on 4 public image repositories: MNIST [8], Fashion MNIST [39], SVHN [26], and Cifar10 [17], each of which contains 10 classes labeled from class 0 to 9. To emulate varying class imbalance levels, we randomly sample $1/q$ of the training data from even classes (i.e., 0, 2, 4, 6, and 8) as the minority classes for which $1/q \in \{1, 1/10, 1/20, 1/50, 1/100\}$, following previous studies [16,11]. All training samples of odd classes are retained as the majority classes. For instance, 5 MNIST-related datasets are produced as MNIST-1, MNIST-10, MNIST-20, MNIST-50, and MNIST-100. Table 1 of the supplementary material provides description of the datasets used in the study.

For image sets produced from MNIST and Fashion MNIST, the input size is (28, 28, 1) and we employ a simple backbone consisting of two sets of double convolutional layers connected with one max-pooling layer, one flatten layer, one dense layer, and a batch normalization layer in sequence. For image sets produced based on SVHN and Cifar10, the input size is (32, 32, 3) and we employ ResNet [30] with depth 44 as the backbone. All methods are set under the same CPL framework. The latent space dimension is 64 and the training batch size is 128 in our experiments. Stochastic Gradient Descent (SGD) is used as the optimizer with the momentum 0.9, and the initial learning rate is set to 0.1 for all datasets. For MNIST, Fashion MNIST, and SVHN, the total number of training epochs is set to 50, and we decay the learning rate by 0.1 at the 26-th and 41-th epochs. For Cifar10, the total number of training epochs is set to 100, and we decay the learning rate by 0.1 at the 51-th and 81-th epochs. For the proposed ARConvL, the learning rate for the radius variables is set to 0.001 at the first three epochs and 0.01 at the remaining training epochs.

We randomly select 90% of the training samples for model training and the remaining 10% are reserved for validation, so that we can decide the best model in terms of G-mean [31] out of the models created across training epochs as our learned deep model. This is to alleviate the over-fitting issue which may particularly impact the deep learning process. We evaluate predictive performance of deep models on spare test sets.

ARConvL are compared against 2 baseline methods, namely CPL [40] and GCPL [40], and 5 state-of-the-art methods, namely Focal Loss ("Focal") [20], Class Balanced Loss ("CB") [6], Class Balanced Focal Loss ("CB Focal") [6], Affinity Loss ("Affinity") [11], and Logit Adjustment Loss ("LA") [23]. Table 2 of the supplementary material reports the parameter settings for those methods.

G-mean [31] and class-wise accuracy are used to evaluate performance for being popularly used and shown to be robust in class imbalance learning [16,11]. Experiments are repeated 10 times, and the average performance (mean) $\pm$ standard deviation (std) are reported. Friedman tests or Wilcoxon-signed rank tests are used to detect statistically significant difference between more than two or two methods across datasets [7]. Given rejection of H0, Holm-Bonferroni correction [14] is conduced as the post-hoc test.

Table 1. G-means (%) of the investigated methods. Each entry is the mean±std of 10 times. The last column corresponds to our ARConvL. The best model on each dataset is highlighted in bold. The last row lists the average ranks (avgRank) of each model across datasets. Significant difference against ARConvL is highlighted in yellow.

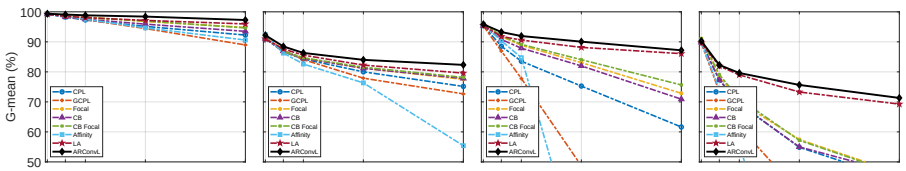
Data	CPL	GCPL	Focal	CB	CB Focal	Affinity	LA	ARConvL
Mnist-1	99.2±0.1	99.4±0.0	99.4±0.1	99.2±0.1	99.4±0.1	**99.5±0.1**	99.2±0.1	99.4±0.1
Mnist-10	98.2±0.2	98.5±0.1	98.8±0.2	98.3±0.1	98.6±0.3	98.7±0.2	98.5±0.1	**99.1±0.0**
Mnist-20	97.3±0.2	97.4±0.3	98.3±0.3	97.5±0.3	98.1±0.3	97.5±0.4	98.1±0.3	**98.8±0.2**
Mnist-50	95.1±0.3	94.4±0.4	96.8±0.5	95.9±0.3	97.0±0.5	94.7±1.4	97.1±0.4	**98.4±0.3**
Mnist-100	92.3±0.8	89.0±1.3	94.8±0.9	93.5±0.9	94.6±0.8	90.6±1.5	96.0±0.5	**97.3±0.6**
Fashion-1	91.0±0.2	92.0±0.2	91.4±0.4	91.1±0.2	91.4±0.4	**92.4±0.2**	91.0±0.2	92.2±0.2
Fashion-10	86.6±0.6	87.1±0.4	86.8±0.5	86.9±0.5	86.9±0.4	86.2±0.3	87.6±0.3	**88.5±0.5**
Fashion-20	84.3±0.6	84.1±0.8	84.6±0.8	84.5±0.7	84.7±0.6	82.6±0.8	85.6±0.5	**86.3±0.7**
Fashion-50	80.0±1.0	77.9±1.4	81.2±1.0	81.2±1.1	81.6±1.2	76.3±1.8	82.2±1.9	**84.0±1.0**
Fashion-100	75.1±2.7	72.7±2.1	77.5±2.2	77.8±1.5	78.2±1.3	55.4±20.2	79.6±2.5	**82.3±0.8**
SVHN-1	95.4±0.1	95.3±0.2	96.0±0.2	95.4±0.1	**96.1±0.1**	95.8±0.1	95.4±0.1	95.9±0.2
SVHN-10	88.5±0.8	86.9±0.9	91.7±0.7	90.8±0.3	92.0±0.2	90.4±0.4	91.8±0.4	**93.3±0.2**
SVHN-20	83.5±1.5	77.8±2.8	88.8±0.6	87.9±0.4	89.2±0.7	84.7±0.9	90.6±0.4	**91.9±0.5**
SVHN-50	75.3±0.6	48.7±8.0	83.1±0.6	82.0±0.7	84.1±0.7	15.4±14.8	88.2±1.4	**90.1±1.0**
SVHN-100	61.6±2.5	0.0±0.0	72.9±2.6	70.9±2.3	75.6±1.7	0.0±0.0	86.1±0.9	**87.2±1.1**
Cifar10-1	89.7±0.2	89.6±0.2	91.1±0.2	89.8±0.3	**91.1±0.2**	89.9±0.3	89.6±0.3	90.3±0.4
Cifar10-10	77.3±0.6	73.4±1.5	78.6±0.8	77.3±0.9	79.1±0.4	74.1±1.1	81.9±0.4	**82.3±0.6**
Cifar10-20	69.5±1.1	61.6±1.5	69.9±1.2	69.3±1.8	71.0±1.0	54.9±3.9	79.0±0.6	**79.6±0.6**
Cifar10-50	54.9±3.0	39.9±4.1	57.5±2.6	55.0±3.3	57.2±3.2	0.0±0.0	73.3±1.6	**75.6±0.6**
Cifar10-100	43.9±1.2	5.0±7.8	46.2±2.2	46.0±2.5	45.8±3.4	0.0±0.0	69.3±2.3	**71.3±1.2**
avgRank	6.45	6.825	3.5	5.2	3.25	6.175	3.2	1.4

4.2 Performance Comparison

This section discusses performance comparisons between our ARConvL against its competitors for multi-class imbalance learning. Comparisons in terms of G-mean are reported in Table 1; comparisons in terms of class-wise accuracy present the same conclusions and can be found in Section 3.1 of the supplementary material for space reason. The last column corresponds to ARConvL

Table 1 shows that our ARConvL achieves the best G-means in 16 out of 20 datasets, showing the effectiveness of our approach in dealing with varying levels of class imbalance. Friedman tests at the significance level 0.05 reject H0 with the p-value 0, meaning that there is significant difference between methods. The average rank ("avgRank") at the last row provides a reasonable idea of how well each method performs compared to others. The average rank of ARConvL is 1.4, being the best (lowest value) among all competing methods. This indicates that our method generally performs the best across datasets with different levels of class imbalance. ARConvL is then chosen as the control method to conduct post-hoc tests for performing the best among all classifiers. Post-hoc tests show that the proposed ARConvL significantly outperforms all competitors.

Note that GCPL obtains zero G-mean in SVHN-100; Affinity obtains zero G-means individually in SVHN-100, Cifar10-50, and Cifar10-100. Further exploitation finds that the corresponding method got zero recalls in certain minority classes, thereby resulting in zero G-means. Such poor recalls usually occur in severely imbalanced scenarios.



(a) MNIST (b) Fashion MNIST (c) SVHN (d) Cifar10

Fig. 4. Performance deterioration in terms of G-mean (%) with the increase of class imbalance levels. The x-axis represents different class imbalance levels, and the y-axis represents G-means. We show G-mean between 50 and 100 to facilitate visualization.

4.3 Performance Deterioration with Increasing Imbalance Levels

This section investigates the relation between the class imbalance levels and predictive performance of all investigated methods on each image repository. Figure 4 shows experimental results in terms of G-means. We can see that all methods achieve similar G-means in the original image repository for the case $q = 1$. With the increase of class imbalance levels with larger q, performance of all methods declines. The proposed ARConvL usually achieves better G-means than its competitors when datasets become more imbalanced, demonstrating better robustness of ARConvL against different levels of class imbalance. Experimental results in terms of class-wise accuracy show the same pattern and are reported in Figure 1 and Section 3.2 of the supplementary material.

4.4 Effect of Each Adaptive Component of ARConvL

This section investigates the effect of each adaptive component of the overall loss in Eq. (1). Particularly, the region loss $\mathcal{L}_R$ is indispensable to derive the class-wise regions and thus should not be eliminated; the adaptive margin cross-entropy loss $\mathcal{L}_M$ contains two adaptive components, namely the adaptive margin σ^2 and the loss for class centers $\frac{1}{k}\sum_{j=1}^{k} -\log(p(C_j))$. Therefore, effects of the adaptive distribution loss $\mathcal{L}_D$ in Eq. (6), the adaptive margin σ^2 (of $\mathcal{L}_M$ in Eq. (8)), and the penalty on class centers (of $\mathcal{L}_M$) are investigated individually.

For the space reason, we only report experimental results in terms of G-means in this section. Experimental results in terms of class-wise accuracy lead to the same conclusions and are provided in Section 3.3 of the supplementary material.

Effect of Adaptive Distribution Loss To conduct this investigation, the adaptive parameter β of ARConvL is fixed and chosen from $\{0, 0.5, 1\}$. In particular, ARConvL without the adaptive distribution loss is equivalent to the case $\beta = 0$. Pair-wise comparisons in terms of G-means between ARConvL in Table 1 and the degraded ARConvL with non-adaptive β in Table 2(a) show the performance deterioration in most cases.

Given fixed $\beta = 0$ and $\beta = 1$, Wilcoxon signed rank tests reject H0 with p-values 0.0017 and 0.04, respectively, showing significant difference in predictive performance between ARConvL and the degraded versions. Average ranks

Table 2. G-means (%) of the degraded ARConvL with non-adaptive β. Each entry is the mean±std of 10 times. Better pair-wise performance compared to ARConvL in Table 1 is highlighted in bold. The last row lists average ranks (avgRank) of ARConvL vs the degraded version across datasets. Significant difference is highlighted in yellow.

	(a) Non-adaptive β			(b) Non-adaptive σ^2			(c) ARC-C
Data	$\beta = 0$	$\beta = 0.5$	$\beta = 1$	$\sigma^2 = 0$	$\sigma^2 = 0.5$	$\sigma^2 = 1$	ARC-C
Mnist-1	99.3±0.0	99.3±0.1	99.3±0.1	**99.4±0.0**	99.4±0.0	**99.4±0.1**	99.3±0.0
Mnist-10	98.7±0.1	99.1±0.0	99.0±0.1	99.0±0.1	99.0±0.1	99.1±0.1	99.1±0.1
Mnist-20	98.1±0.3	**98.9±0.1**	**98.9±0.1**	**98.9±0.1**	98.8±0.2	**98.9±0.2**	**98.9±0.1**
Mnist-50	97.1±0.4	**98.5±0.2**	**98.6±0.2**	98.4±0.2	**98.4±0.3**	**98.5±0.2**	98.4±0.2
Mnist-100	95.6±0.6	**97.6±0.3**	**97.8±0.3**	**97.4±0.4**	**97.6±0.4**	97.3±0.6	**97.4±0.5**
Fashion-1	91.4±0.2	92.2±0.2	92.0±0.2	91.8±0.1	92.1±0.2	92.1±0.2	91.7±0.2
Fashion-10	87.4±0.2	**88.7±0.3**	88.3±0.5	88.3±0.4	**88.5±0.4**	88.4±0.4	85.2±1.2
Fashion-20	84.8±1.1	**86.6±0.9**	**86.4±0.9**	**86.5±1.0**	86.3±1.3	86.3±0.8	83.4±1.8
Fashion-50	81.7±1.4	82.8±2.8	**84.3±0.8**	**84.6±0.6**	**84.6±0.5**	83.8±1.3	81.8±1.4
Fashion-100	79.6±2.2	81.3±1.0	81.7±1.4	82.2±1.5	81.9±1.7	81.9±1.5	80.7±1.7
SVHN-1	**96.3±0.1**	95.7±0.3	94.9±0.7	95.3±0.2	95.6±0.2	**95.9±0.2**	12.1±1.5
SVHN-10	93.0±0.5	93.2±0.4	92.0±1.6	92.0±0.4	92.3±0.6	92.8±0.3	54.1±34.8
SVHN-20	91.3±0.5	**92.1±0.4**	91.8±0.7	90.2±1.1	90.1±1.8	91.2±1.5	75.3±21.4
SVHN-50	88.0±1.3	89.3±1.4	89.7±1.7	88.0±1.1	88.8±1.3	89.4±0.5	79.6±2.1
SVHN-100	83.0±3.5	86.7±0.8	85.2±5.7	84.9±2.7	85.0±2.2	85.7±2.8	76.9±4.6
Cifar10-1	**92.0±0.2**	90.2±0.5	89.6±0.5	89.3±0.5	90.0±0.4	**90.3±0.5**	68.2±8.5
Cifar10-10	**82.5±0.6**	**82.9±0.6**	81.8±1.0	80.1±1.1	81.6±0.7	81.9±0.7	64.1±1.4
Cifar10-20	78.2±0.7	**80.0±0.7**	79.4±1.1	77.0±1.4	78.7±0.6	79.4±0.8	62.5±1.6
Cifar10-50	70.9±1.9	75.3±0.9	75.6±0.8	73.8±1.4	74.8±1.0	74.6±1.7	61.4±1.6
Cifar10-100	62.7±2.8	68.2±3.5	**71.6±0.7**	69.1±3.0	70.1±2.6	69.3±3.1	59.6±3.5
avgRank	1.15/1.85	1.4/1.6	1.3/1.7	1.25/1.75	1.2/1.8	1.25/1.75	1.1/1.9

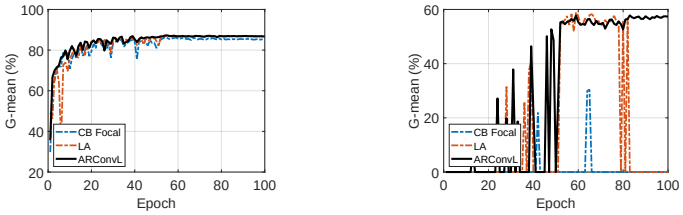
are 1.15 and 1.3 for ARConvL vs 1.85 and 1.7 for the degraded versions, respectively. This means that adaptively learning β throughout the training epochs has significantly beneficial effect on predictive performance.

Given fixed $\beta = 0.5$, Wilcoxon signed rank test does not find significant difference between ARConvL and the degraded version with p-value 0.39. Further analyses found that on the datasets that the degraded version outperforms, performance deterioration of ARConvL is at most 0.79% in Cifar10-10; whereas on the datasets that ARConvL outperforms, performance superiority can be as high as 4.59% in Cifar10-100, with the average improvement at 0.80%. This indicates that the degraded ARConvL may cause relatively large performance decline compared to the small performance improvement it may have.

Overall, the experimental investigation shows the effectiveness of the adaptive distribution loss, in view of the adaptive β, on retaining good performance in multi-class imbalance learning.

Effect of Adaptive Margin To conduct this investigation, the adaptive margin σ^2 in $\mathcal{L}_M$ of ARConvL is fixed and chosen from $\{0, 0.5, 1\}$. In particular, ARConvL without the adaptive margin is equivalent to the case $\sigma^2 = 0$. Pairwise comparisons in terms of G-means between ARConvL in Table 1 and the degraded ARConvL with non-adaptive σ^2 in Table 2(b) show the performance deterioration in the vast majority of cases.

Given σ^2 with those fixed values, Wilcoxon signed rank tests reject H0 with p-values 0.0045, 0.0036, and 0.0057, respectively, showing significant difference



(a) CelebA (b) iNaturalist 2018

Fig. 5. Training curves of ARConvL, LA, and CB Focal on CelebA (left) and iNaturalist 2018 (right).

in predictive performance between ARConvL and the degraded versions with non-adaptive σ^2. Performance comparisons in terms of average ranks further show the significance of such performance deterioration of the degraded ARConvL. This means that adaptively learning σ^2 throughout the training epochs has significantly beneficial effect on predictive performance, demonstrating the effectiveness of the adaptive margin on retaining good performance in multi-class imbalance learning.

Effect of Loss for Class Centers To conduct this investigation, we produce the degraded version of ARConvL (denoted as "ARC-C") by eliminating the loss for class centers $\frac{1}{k}\sum_{j=1}^{k} -\log(p(C_j)$ from $\mathcal{L}_M$ in Eq. (8). The loss of ARC-C in accordance with $\mathcal{L}_M$ is degraded as $\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n} -\log(p(x_i \in j)$. Performance comparisons in terms of G-means between ARConvL in Table 1 and the degraded ARC-C in Table 2(c) show the performance deterioration in almost all cases.

Wilcoxon signed rank test rejects H0 with p-value $3.38 \cdot 10^{-4}$, showing significant difference in predictive performance between ARConvL and the degraded ARC-C. Performance comparisons in terms of average ranks further show the significance of such performance deterioration eliminating the loss for class centers, demonstrating the effectiveness of the loss for class centers in multi-class imbalance learning.

4.5 Utility in Large-Scale Datasets

To demonstrate the proposed ARConvL can be utilized on large-scale real-world datasets, we present training curves of ARConvL and the two most competitive methods CB Focal [6] and LA [23] on two additional large-scale datasets, namely CelebA [22] and iNaturalist 2018 [34]. For CelebA, only five non-overlapping classes (blonde, black, bald, brown, and gray) are kept following previous work [37]. Details of these datasets are shown in Section 1 of the supplementary material. The input size is (64, 64, 3) for CelebA and (224, 224, 3) for iNaturalist 2018. We employ ResNet [30] with depth 56 as the backbone in this extra study. The training batch size is set to 64; the total number of training epochs is set to 100. We decay the learning rate by 0.1 at the 51-th and 81-th epochs.

Training curves on those large-scale datasets are shown in Fig. 5. Fig. 5(a) shows that ARConvL outperforms CB Focal across all training epochs; ARConvL yields better or similar performance compared to LA and it can converge faster than LA within 52 epochs. Fig. 5(b) shows similar experimental results: ARConvL achieves better G-means at most training epochs and possesses better convergence than its competitors. In particular, between the training epoch 52 and 78, LA and ARConvL achieve similar performance, and after the training epoch 82, ARConvL outperforms LA. All methods confront with zero G-means at some training epochs, meaning that they fail in detecting any example of some class(es). Performance in terms of class-wise accuracy shows the same experimental results and can be found in Section 3.4 of the supplementary material. Therefore, experimental results on two large-scale datasets show the utility of the proposed ARConvL over its competitors.

5 Conclusion

This paper proposes ARConvL for multi-class imbalance learning, which derives class-wise regions in the latent feature space adaptively throughout training epochs. Latent feature distributions can then be well depicted by class regions without relying on any strict assumption. Based on the derived class regions, we address the multi-class imbalance issue from two perspectives. First, an adaptive distribution loss is proposed to optimize the class-wise latent feature distribution, by pushing down the upper-bound of the radii to approach the benchmark radius, directly tackling the multi-class imbalance problem. Second, an adaptive margin cross-entropy loss is proposed by employing the defined margin as a mediator to improve the discrimination between classes, further alleviating the class imbalance problem.

Experimental results based on plenty of real-world image sets demonstrated the superiority of our ARConvL to SOTA methods. Investigations on the performance deterioration with respect to different imbalance ratios showed the robustness of the proposed method. Ablation studies demonstrated the effectiveness of the adaptive distribution loss and the adaptive margin cross-entropy loss in the learning process. Experiments on two large-scale real-world image sets showed the utility of ARConvL on large-scale datasets.

Future work includes additional experimental investigations to better understand how data noise and missing data affect the performance of our proposed method and the extension of ARConvL by having multiple regions assigned to each class (instead of only one).

Acknowledgements This work was supported by National Natural Science Foundation of China (NSFC) under Grant No. 62002148 and Grant No. 62250710682, Guangdong Provincial Key Laboratory under Grant No. 2020B121201001, the Program for Guangdong Introducing Innovative and Enterpreneurial Teams under Grant No. 2017ZT07X386, and Research Institute of Trustworthy Autonomous Systems (RITAS).

References

1. Alejo, R., Sotoca, J.M., Valdovinos, R.M., Casañ, G.A.: The multi-class imbalance problem: Cost functions with modular and non-modular neural networks. In: International Symposium on Neural Networks. pp. 421–431. Springer (2009)
2. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural networks **106**, 249–259 (2018)
3. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Proceedings of the 33rd International Conference on Neural Information Processing Systems pp. 1567–1578 (2019)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**(1), 321–357 (2002)
5. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: SMOTEBoost: Improving prediction of the minority class in boosting. In: European conference on principles of data mining and knowledge discovery. pp. 107–119. Springer (2003)
6. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
7. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7**, 1–30 (2006)
8. Deng, L.: The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine **29**(6), 141–142 (2012)
9. Elkan, C.: The foundations of cost-sensitive learning. In: International joint conference on artificial intelligence. vol. 17, pp. 973–978 (2001)
10. Freund, Y., Schapire, R.E.: A desicion-theoretic generalization of on-line learning and an application to boosting. In: Computational Learning Theory. pp. 23–37. Springer Berlin Heidelberg (1995)
11. Hayat, M., Khan, S., Zamir, S.W., Shen, J., Shao, L.: Gaussian affinity for max-margin class imbalanced learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6469–6479 (2019)
12. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). pp. 1322–1328. IEEE (2008)
13. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263–1284 (2009)
14. Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics **6**(2), 65–70 (1979)
15. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of Big Data **6**(1), 1–54 (2019)
16. Khan, S.H., Hayat, M., Bennamoun, M., Sohel, F.A., Togneri, R.: Cost-sensitive learning of deep feature representations from imbalanced data. IEEE Transactions on Neural Networks and Learning Systems **29**(8), 3573–3587 (2018)
17. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto, Toronto, Ontario (2009)
18. Lee, H., Park, M., Kim, J.: Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. In: 2016 IEEE international conference on image processing (ICIP). pp. 3713–3717. IEEE (2016)

19. Liang, L., Jin, T., Huo, M.: Feature identification from imbalanced data sets for diagnosis of cardiac arrhythmia. In: International Symposium on Computational Intelligence and Design. vol. 02, pp. 52–55. IEEE (2018)
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
21. Liu, J., Sun, Y., Han, C., Dou, Z., Li, W.: Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2970–2979 (2020)
22. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
23. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. In: International Conference on Learning Representations (2021)
24. M'hamed, B.A., Fergani, B.: A new multi-class WSVM classification to imbalanced human activity dataset. Journal of Computers **9**(7), 1560–1565 (2014)
25. Mullick, S.S., Datta, S., Das, S.: Generative adversarial minority oversampling. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1695–1704 (2019)
26. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011)
27. Pouyanfar, S., Chen, S.C., Shyu, M.L.: Deep spatio-temporal representation learning for multi-class imbalanced data classification. In: International Conference on Information Reuse and Integration. pp. 386–393. IEEE (2018)
28. Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., Li, h.: Balanced meta-softmax for long-tailed visual recognition. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 4175–4186. Curran Associates, Inc. (2020)
29. Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans **40**(1), 185–197 (2010)
30. Shah, A., Kadam, E., Shah, H., Shinde, S., Shingade, S.: Deep residual networks with exponential linear unit. In: Proceedings of the third international symposium on computer vision and the internet. pp. 59–65 (2016)
31. Sun, Y., Kamel, M.S., Wang, Y.: Boosting for learning multiple classes with imbalanced class distribution. In: International Conference on Data Mining. pp. 592–602. IEEE (2006)
32. Taherkhani, A., Cosma, G., McGinnity, T.M.: AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. Neurocomputing **404**, 351–366 (2020)
33. Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., Yan, J.: Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11662–11671 (2020)
34. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)

35. Wang, S., Chen, H., Yao, X.: Negative correlation learning for classification ensembles. In: International Joint Conference on Neural Networks. pp. 1–8. IEEE (2010)
36. Wang, S., Yao, X.: Multiclass imbalance problems: Analysis and potential solutions. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **42**(4), 1119–1130 (2012)
37. Wang, X., Lyu, Y., Jing, L.: Deep generative model for robust imbalance classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14124–14133 (2020)
38. Xiang, L., Ding, G., Han, J.: Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In: European Conference on Computer Vision. pp. 247–263. Springer (2020)
39. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
40. Yang, H.M., Zhang, X.Y., Yin, F., Liu, C.L.: Robust classification with convolutional prototype learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3474–3482 (2018)
41. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. arXiv preprint arXiv:2110.04596 (2021)
42. Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M.: 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. pp. 9719–9728 (2020)