Parameter-Free Loss for Class-Imbalanced Deep Learning in Image Classification

Jie Du, Yanhong Zhou, Peng Liu, Chi-Man Vong, *Senior Member, IEEE*, and Tianfu Wang

***Abstract*— Current state-of-the-art class-imbalanced loss functions for deep models require exhaustive tuning on hyperparameters for high model performance, resulting in low training efficiency and impracticality for nonexpert users. To tackle this issue, a parameter-free loss (PF-loss) function is proposed, which works for both binary and multiclass-imbalanced deep learning for image classification tasks. PF-loss provides three advantages: 1) training time is significantly reduced due to NO tuning on hyperparameter(s); 2) it dynamically pays more attention on minority classes (rather than outliers compared to the existing loss functions) with NO hyperparameters in the loss function; and 3) higher accuracy can be achieved since it adapts to the changes of data distribution in each mini-batch instead of the fixed hyperparameters in the existing methods during training, especially when the data are highly skewed. Experimental results on some classical image datasets with different imbalance ratios (IR, up to 200) show that PF-loss reduces the training time down to 1/148 of that spent by compared state-of-the-art losses and simultaneously achieves comparable or even higher accuracy in terms of both G-mean and area under receiver operating characteristic (ROC) curve (AUC) metrics, especially when the data are highly skewed.**

***Index Terms*— Class-imbalanced deep learning, dynamic changes, hyperparameters tuning, loss function, parameter-free.**

I. Introduction

Learning deep model from class-imbalanced data is a significant challenge for most artificial intelligent classification applications [1]–[3], in which the critical and highly interested classes (called minority) have much fewer samples than other classes (called majority) [4]–[6]. For instance, in disease detection [7], [8], most training data are collected from healthy persons, but very few data are available from patients; in video surveillance [9], only very few scenes capture improper behaviors and most are normal scenes. If traditional algorithms for balanced data are used to classify such data, the patients (improper scenes) are easily misclassified to healthy persons (normal scenes), which causes irretrievable loss [7], [9].

In general, there are two common strategies to resolve the class-imbalanced problem: resampling and reweighting [10]. In resampling, oversampling the minority samples or undersampling the majority ones or both are used to achieve a balanced dataset in quantity. In reweighting, the loss function is redesigned to force the model to focus more attention on the easily misclassified minority samples. However, there are some disadvantages of applying resampling on deep learning [11]. In detail, oversampling easily causes the model suffering from overfitting and also lower training efficiency due to the increased amount of redundant samples, while undersampling causes information loss and gets poor performance on majority classes. Hence, present works aim to design a better class-imbalanced loss [11].

In the literature, several popular class-imbalanced loss functions [12]–[18] have been proposed. Weighted cross-entropy (WCE) loss is a commonly used loss function for class-imbalanced deep learning, which associates a relatively large class weight with the loss of minority samples and a small one with the loss of majority samples. Under this way, the WCE-loss focuses on the loss of minority samples and aims to accurately classify them. However, there is a significant issue: every time, an exhaustive tuning on the hyperparameter (i.e., class weight) is necessary to fit for data [11], which is impractical for nonexpert users and burdens the training time [19].

Recently, focal loss [13], [14] is proposed for resolving the class-imbalanced problem, which associates large weights based on estimated class probabilities with the losses of easily misclassified (i.e., hard) samples. However, harmful samples (e.g., outliers or mislabeled data) are also hard samples [11], [20], and hence, the focal loss is highly sensitive to outliers. Even if there are no harmful samples, focusing on hard samples does not fully address class-imbalanced problem. In detail, it is very common to have more majority hard samples than minority hard ones. In fact, focusing on hard samples is similar to concentrating on support vectors in support vector machine (SVM), and SVM suffers from class-imbalanced problem as well [21], [22]. In order to alleviate this issue, in the focal loss, the class weight is also assigned to each sample loss. Consequently, the focal loss has two hyperparameters to determine the weights associated with hard samples and hard classes. With multiple hyperparameters, the training time would be increased exponentially because every combination of these hyperparameters should be tested.

More recently, other loss functions (including average-precision loss (AP-loss) [17], distribution ranking loss (DR-loss) [15], gradient harmonizing mechanism loss (GHM-loss) [16], and label-distribution-aware margin loss (LDAM) [18]) for class-imbalanced deep learning are also proposed, but all of them have hyperparameters to be exhaustively tuned (detailed in Section II-D). With hyperparameters, class-imbalanced deep model may suffer from the following limitations.

1) Huge amount of time and effort are required for exhaustive tuning on the hyperparameters for satisfactory model performance, especially under big data.
2) In class-imbalanced data, the data distribution may be dynamically changed (nonstationary environment) in each mini-batch [16], [23], [24], while the hyperparameters are usually fixed during training. Hence, the deep model with hyperparameters is not able to adapt to the dynamic changes in data distribution.

Manuscript received October 9, 2020; revised February 24, 2021 and July 13, 2021; accepted September 3, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 62006160; in part by the Educational Commission of Guangdong Province under Grant 2020KQNCX062; in part by the Shenzhen Fundamental Research Program under Grant 20200813102946001; in part by the Science and Technology Development Fund, Macau, under Grant 0112/2020/A and Grant 004/2019/AFJ; and in part by the National Natural Science Foundation of China under Grant 81771922, Grant 62071309, Grant 61801305, Grant 81971585, and Grant 61871274. *(Corresponding author: Chi-Man Vong.)*

Jie Du, Yanhong Zhou, and Tianfu Wang are with the Health Science Center, School of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China, also with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Shenzhen University, Shenzhen 518060, China, and also with the Marshall Laboratory of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: dujie@szu.edu.cn; meiszyh@163.com; tfwang@szu.edu.cn).

Peng Liu and Chi-Man Vong are with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yb77427@um.edu.mo; cmvong@um.edu.mo).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TNNLS.2021.3110885.

Digital Object Identifier 10.1109/TNNLS.2021.3110885

In this work, a parameter-free loss (PF-loss) for tackling both class-imbalanced and nonstationary environment problems in deep learning-based image classification tasks is proposed, which dynamically pays more attention to minority classes without using hyperparameters. It is known that traditional cross-entropy (CE) loss treats each sample equally regardless of its class. For majority classes having high data quantity, CE-loss pays more attention to them in order to get a low loss value, while in the proposed PF-loss, the preference of data quantity in majority classes is eliminated and each class (rather than each sample) is treated equally regardless of its data quantity. Hence, for the easily misclassified minority classes, PF-loss dynamically focuses more attention on them to get a low loss value. Further discussion is in Section III. The main contributions of this work are summarized as follows.

1) Our PF-loss function does not have hyperparameters so that exhaustive time and effort for their tuning is eliminated.
2) Our PF-loss dynamically focuses more attention on the minority classes in both binary and multiclass-imbalanced learning.
3) Compared to WCE-loss and focal loss whose performances are negatively influenced by the nonstationary environment because of fixed class weights, our PF-loss is adaptive to the dynamically changed class-imbalanced ratio within each mini-batch because of NO hyperparameters, which produces higher model accuracy, especially when the data are highly imbalanced.
4) The proposed PF-loss more effectively addresses the class-imbalanced problem by focusing on minority samples rather than the outliers (e.g., in WCE-loss and focal loss).

Section II provides a short review on CE-loss, WCE-loss, and focal loss, followed by other popular class-imbalanced losses. Section III details our method: PF-loss. Section IV shows the experimental results with analysis and discussion. Finally, a conclusion is drawn in Section V.

II. Related Work

In this section, we will give a brief introduction of CE-loss, WCE-loss, focal loss, and other popular class-imbalanced loss functions (including AP-loss, DR-loss, GHM-loss, and LDAM) and show their differences using a simple but classic example.

A. CE Loss

CE-loss is the most popular loss function for deep learning [25], [26]; however, it is designed based on the assumption that the data are balanced. The formula of CE-loss is

$$\mathrm{CE} = -\sum_{i=1}^{N} \log(p_i) \tag{1}$$

and

$$p_i = \sum_{j=1}^{M} y_{i,j}\, p_{i,j} \tag{2}$$

$$y_{i,j} = \begin{cases} 1, & \text{if } i\text{th sample } \in \text{class } j \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where N is the number of samples, M is the number of classes, $p_{i,j}$ is the estimated probability that the ith sample is of class j, and hence, p_i is the correctly classified probability of ith sample.

Example: Assume that there are one minority sample and three majority ones, and their estimated class probabilities [i.e., p_i in (1)] are 0.4, 0.8, 0.7, and 0.4 at the kth iteration. Hence, $\mathrm{CE}^k = -\log(0.4 \times 0.8 \times 0.7 \times 0.4) = 1.05$.

In one case, if the probability of the last *majority* sample is increased from 0.4 to 0.9 at the $(k+1)$th iteration, $\mathrm{CE}_{\mathrm{ma}}^{k+1} = -\log(0.4 \times 0.8 \times 0.7 \times 0.9) = 0.70$.

In another case, if the probability of the *minority* sample is increased from 0.4 to 0.9 at the $(k+1)$th iteration, $\mathrm{CE}_{\mathrm{mi}}^{k+1} = -\log(0.9 \times 0.8 \times 0.7 \times 0.4) = 0.70$.

This example verifies that CE-loss treats *each sample* equally regardless of its class, and hence, CE-loss cannot resolve the class-imbalanced problem [13].

B. WCE Loss

In order to tackle class-imbalanced problem, WCE-loss is designed, whose formula is

$$\mathrm{WCE} = -\sum_{i=1}^{N} \alpha_i \log(p_i) \tag{4}$$

where α_i is the class weight associated with the ith sample. α_i has a large value if the ith sample is a minority one; otherwise, it has a small value.

In most studies [12], [27], [28], $\alpha_i = 1/N_j$ through lots of parameters tuning, where N_j is the number of samples in class j, if the ith sample belongs to class j.

Example: The example in Section II-A is used again to show the difference between CE-loss and WCE-loss.

Initially, at the kth iteration, $\mathrm{WCE}^k = -\log(0.4) - (1/3)\log(0.8) - (1/3)\log(0.7) - (1/3)\log(0.4) = 0.615$.

When *majority* probability increases, $\mathrm{WCE}_{\mathrm{ma}}^{k+1} = -\log(0.4) - (1/3)\log(0.8) - (1/3)\log(0.7) - (1/3)\log(0.9) = 0.497$.

When *minority* probability increases, $\mathrm{WCE}_{\mathrm{mi}}^{k+1} = -\log(0.9) - (1/3)\log(0.8) - (1/3)\log(0.7) - (1/3)\log(0.4) = 0.262$.

We can clearly observe that $|0.615 - 0.497| < |0.615 - 0.262|$, i.e., with the same increases of majority and minority probabilities, a large decrease of WCE-loss occurs only when minority probability increases. Hence, WCE-loss focuses on minority classes and resolves the class-imbalanced problem.

C. Focal Loss

In the focal loss, for binary classification, a modulating factor $(1-p_t)^\gamma$ is multiplied to the CE-loss of tth sample as follows:

$$\mathrm{FL} = -(1-p_t)^\gamma \log(p_t) \tag{5}$$

where

$$p_t = \begin{cases} p, & \text{if } t\text{th sample } \in \text{minority} \\ 1-p, & \text{if } t\text{th sample } \in \text{majority} \end{cases} \tag{6}$$

where p is the estimated probability of tth sample for the class with minority label.

Hence, the focal loss focuses on the samples with low probabilities (i.e., hard samples).

Similarly, for multiclass classification with N samples

$$\mathrm{FL} = -\sum_{i=1}^{N} (1-p_i)^\gamma \log(p_i). \tag{7}$$

Equation (2) is used to calculate p_i as well.

Example: For the example in Section II-A, γ is set to 2 (default setting in focal loss), and $\mathrm{FL}^k = -0.6^2\log(0.4) - 0.2^2\log(0.8) - 0.3^2\log(0.7) - 0.6^2\log(0.4) = 0.304$.

When *majority* probability increases, $\mathrm{FL}_{\mathrm{ma}}^{k+1} = -0.6^2\log(0.4) - 0.2^2\log(0.8) - 0.3^2\log(0.7) - 0.1^2\log(0.9) = 0.162$.

When *minority* probability increases, $\mathrm{FL}_{\mathrm{mi}}^{k+1} = -0.1^2\log(0.9) - 0.2^2\log(0.8) - 0.3^2\log(0.7) - 0.6^2\log(0.4) = 0.162$.

Hence, the focal loss focuses on hard samples regardless of its class. If there are more majority hard samples than minority hard ones, Focal loss will not expect to focus on majority classes indirectly.

In order to alleviate this issue, the class weight α_i in WCE-loss is also associated with each sample in the focal loss

$$\mathrm{FL}(\alpha) = -\sum_{i=1}^{N} \alpha_i (1-p_i)^{\gamma} \log(p_i). \tag{8}$$

Example: Under the same example, $\mathrm{FL}(\alpha)^k = 0.197$, $\mathrm{FL}(\alpha)^{k+1}_{\mathrm{ma}} = 0.149$, and $\mathrm{FL}(\alpha)^{k+1}_{\mathrm{mi}} = 0.054$. Obviously, $|0.197-0.149| < |0.197-0.054|$, i.e., focal loss with class weight [$\mathrm{FL}(\alpha)$] decreases largely when minority probability increases, and thus, $\mathrm{FL}(\alpha)$ pays more attention to hard minority samples than hard majority ones.

However, $\mathrm{FL}(\alpha)$ has two hyperparameters (α and γ) to be tuned, which requires lots of effort.

D. Other Class-Imbalanced Loss Functions

More recently, several loss functions for class-imbalanced deep learning are proposed, including DR-loss, AP-loss, GHM-loss, and LDAM.

DR-loss ranks the distribution or estimated probability of minority above that of majority. Similarly, AP-loss replaces classification task with ranking task and adopts average precision as their target loss. GHM-loss is similar to WCE-loss except that the class weight α is determined by gradient density. LDAM is designed to encourage larger margins for minority classes. However, all of them have hyperparameters to be tuned.

In detail, DR-loss has a hyperparameter of smoothing term L to control the smoothness of the loss function [15]. AP-loss has a hyperparameter δ to determine the threshold in the piecewise step function [17]. In addition, GHM-loss has a hyperparameter M (the number of unit regions) to determine the gradient density of samples [16]. LDAM has a hyperparameter C to be tuned to determine the margins of classes.

In summary, state-of-the-art studies usually require exhaustive hyperparameters tuning to obtain an accurate classifier for class-imbalanced deep learning. In contrast, in this work, we proposed a PF-loss that can dynamically pay more attention to minority classes without any hyperparameters.

III. Proposed Method

A. Parameter-Free Loss

Learning from CE-loss, the main reason that causes misclassification on minority samples is the data scarcity of minority classes. In other words, the majority classes have relatively more samples than minority classes and, hence, attracts more attention from the loss function. In this work, in order to eliminate the preference of data quantity of majority classes without class weight tuning, each sample probability in CE-loss [i.e., p_i in (1)] is replaced by the average sample probability of each class. It is obvious that no matter how many samples the class has, its average sample probability is in the range of (0, 1). Hence, the proposed loss function treats each class equally regardless of its data quantity.

In imbalanced data, minority samples are easily misclassified and the average sample probability of minority class is usually very low, which hinders from generating a low loss value. Hence, the proposed loss function will dynamically pay more attention to minority classes and effectively resolve a class-imbalanced problem. The formula is expressed as

$$\mathrm{PF} = -\frac{1}{M}\sum_{j=1}^{M} \log(q_j) \tag{9}$$

and

$$q_j = \frac{1}{N_j}\sum_{i=1}^{N_j} p_i \tag{10}$$

where N_j is the number of samples in class j in one mini-batch during training. Noteworthy, in highly imbalanced data, N_j for some minority class is easily equal to zero in one mini-batch. In our experiments, $q_j = (1/N_j + \mathrm{eps})(\sum_{i=1}^{N_j} p_i + \mathrm{eps})$ to avoid dividing by zero (eps represents a mathematic small delta). If $N_j = 0$, $q_j = 1$ and the loss for class j is zero. The model will not concentrate on class j (even it is a minority one) but other classes with $N_j \neq 0$, which makes sense.

Example: Under the example in Section II-A, $\mathrm{PF}^k = -(1/2)\log(0.4) - (1/2)\log(0.8+0.7+0.4)/(3) = 0.298$.

When *majority* probability increases, $\mathrm{PF}^{k+1}_{\mathrm{ma}} = -(1/2)\log(0.4) - (1/2)\log(0.8+0.7+0.9)/(3) = 0.2475$.

When *minority* probability increases, $\mathrm{PF}^{k+1}_{\mathrm{mi}} = -(1/2)\log(0.9) - (1/2)\log(0.8+0.7+0.4)/(3) = 0.122$.

The example results show that the proposed PF-loss focuses on minority classes (large decrease of PF-loss occurs when the minority sample is correctly classified) and effectively resolves a class-imbalanced problem without hyperparameters' tuning.

B. Difference Between PF-Loss and WCE-Loss

In WCE-loss, if the class weight α_i is set to $1/N_j$ when the ith sample belongs to class j, the formula of WCE-loss is revised to

$$\begin{aligned}\mathrm{WCE} &= -\frac{1}{M}\sum_{j=1}^{M}\frac{1}{N_j}\sum_{i=1}^{N_j}\log(p_i) \\ &= -\frac{1}{M}\sum_{j=1}^{M}\log\left(\prod_{i=1}^{N_j} p_i\right)^{\frac{1}{N_j}} \\ &= -\log\left(\prod_{j=1}^{M}\left(\prod_{i=1}^{N_j} p_i\right)^{\frac{1}{N_j}}\right)^{\frac{1}{M}}.\end{aligned} \tag{11}$$

In contrast, by substituting (10) into (9), the PF-loss is obtained by

$$\begin{aligned}\mathrm{PF} &= -\frac{1}{M}\sum_{j=1}^{M}\log\left(\frac{1}{N_j}\sum_{i=1}^{N_j} p_i\right) \\ &= -\log\left(\prod_{j=1}^{M}\left(\frac{1}{N_j}\sum_{i=1}^{N_j} p_i\right)\right)^{\frac{1}{M}}.\end{aligned} \tag{12}$$

By comparing (11) with (12), we can clearly observe that the geometric mean of class accuracies [i.e., $\left(\prod_{j=1}^{M}(\bullet)\right)^{(1/M)}$ in (11) and (12)] is used in both methods to focus on minority classes. However, the geometric mean of sample probabilities in each class [i.e., $\left(\prod_{i=1}^{N_j} p_i\right)^{(1/N_j)}$ in (11)] is also calculated in WCE-loss, whereas in PF-loss, the arithmetic mean of samples probabilities in each class [i.e., $(1/N_j)\sum_{i=1}^{N_j} p_i$ in (12)] is computed, which is more reasonable than geometric mean for classification in practice (detailed in Section III-E).

C. Difference Between PF-Loss and Focal Loss

In order to compare PF-loss and focal loss, we have derived their gradients for analysis. For binary classification, PF-loss can be represented by

$$\mathrm{PF}_b = -\log\left(\frac{1}{N_1}\sum_{i}^{N} t_i y_i\right) - \log\left(\frac{1}{N_0}\sum_{i}^{N}(1-t_i)(1-y_i)\right) \tag{13}$$

where N_1 is the number of samples in positive/minority class and N_0 is the number of samples in negative/majority class; $t_i = 1$ if sample i belongs to positive class; otherwise, $t_i = 0$. y_i is the estimated output of sample i being positive for all N samples.

The gradient of PF_b to estimated output y_i for sample i is calculated by

$$\mathrm{PF}_b^{\mathrm{grad}} = -\frac{t_i y_i (1-y_i)}{\sum_i^N t_i y_i} + \frac{(1-t_i) y_i (1-y_i)}{\sum_i^N (1-t_i)(1-y_i)}. \tag{14}$$

Similarly, the gradient of focal loss to estimated output y_i for sample i is calculated by

$$\begin{aligned}\mathrm{FL}_b^{\mathrm{grad}} = &-\left[t_i (1-y_i)^\gamma (\gamma\, y_i \log(y_i) + 1 - y_i)\right] \\ &-\left[(1-t_i) p_i^\gamma (\gamma\, (1-y_i) \log(1-y_i) - y_i)\right].\end{aligned} \tag{15}$$

From (14) and (15), we observe that the gradient update for one individual sample is based on the overall accuracy of one class in PF-loss, whereas in focal loss, the gradient update for one individual sample is only based on its accuracy. If one minority sample is correctly classified while the overall minority class accuracy is low, the sample is still useful for model update in PF-loss but useless in the focal loss. Eventually, the focal loss will update the model only based on the hard samples resulting in hard margin because hard samples usually almost lie on the classification boundary. On the contrary, our PF-loss concentrates on the overall class accuracy and will get a soft margin.

D. Robustness to Outliers

Recently, the robustness to noisily labeled data or outliers has been shown to have great practical importance in deep learning [29], [30]. In this work, with arithmetic mean of sample probabilities, the PF-loss is more robust to the outlier samples than WCE-loss and focal loss. For instance, there are three majority samples with probabilities of 0.001, 0.8, and 0.9 and one minority sample with a probability of 0.2. The geometric mean of sample probabilities in majority is 0.09, whereas the arithmetic mean is 0.57. Hence, with geometric mean of sample probabilities, WCE-loss will take a lot of time and effort on some outlier samples (e.g., the sample with a probability of 0.001), whereas the PF-loss concentrates on maximizing every class accuracy rather than every sample accuracy. As shown in Fig. 1, with the increase of outlier probability, both WCE-loss and focal loss are decreased, while PF-loss is almost unchanged. Only when the minority probability is increased, PF-loss will be decreased, as detailed in Section III-A.

Consequently, compared to both WCE-loss and focal loss, the proposed PF-loss is much more robust to outliers, which reduces the negative influence of outliers and improves the accuracies of minority classes.

E. Why PF-Loss Works and Better

Although the arithmetic mean of samples probabilities may reduce the sensitivity to hard samples in majority classes, it increases the generalization ability of PF-loss. This is because the hard majority samples usually almost lie on the classification boundary. If the model is highly sensitive to hard majority samples, the obtained classification boundary will be very complicated and overfitting easily occurs. To verify this statement, another PF-loss is designed to focus on hard majority samples (i.e., pay less attention on easy majority samples) by setting the sample probability higher than 0.5 to a constant (e.g., 1.0). Under this way, the model will focus on hard majority samples (i.e., the samples with low probability, e.g., smaller than 0.5). This model is named PF-loss for hard majority samples (PFH-loss). The results show that PFH-loss is worse than PF-loss on almost all compared datasets (the detailed results are included in the supplementary materials), which is consistent with the analysis in Section III-C.

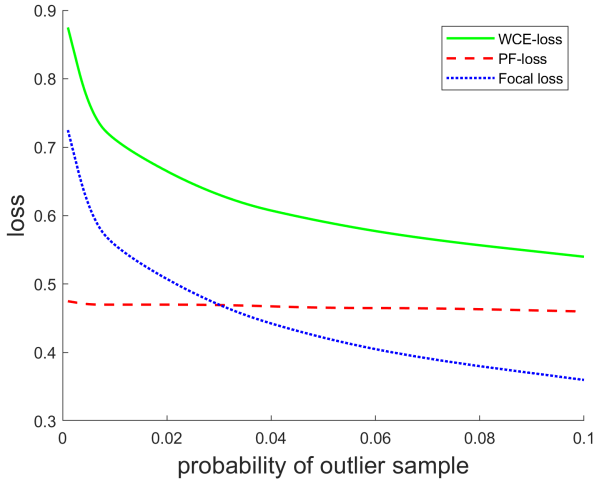

Fig. 1. Comparison on robustness to outlier samples.

TABLE I

WHOLE TRAINING TIME (S) FOLLOWED BY AVERAGE ONE-TIME TRAINING TIME OF MODEL WITH DIFFERENT LOSSES

Method	Binary		Multi-class	
	CIFAR-10	CIFAR-100	Fashion MNIST	CIFAR-10
WCE	2,052 (108)	1,140 (60)	3,401 (179)	1,387 (73)
FL	1,134 (162)	2,828 (404)	2,191 (313)	665 (95)
FL(α)	15,428 (116)	9,044 (68)	59,451 (447)	4,655 (35)
LDAM	3,880 (388)	3,350 (335)	2,630 (263)	2,670 (267)
WLDAM	1,110 (111)	1,220 (122)	2,320 (232)	4,170 (417)
PF-loss	**428** (428)	**594** (594)	**401** (401)	**387** (387)

Moreover, although the arithmetic mean of samples probabilities in one class makes the model less sensitive to hard majority samples, the geometric mean of class accuracies in PF-loss does not allow too much sacrifice of majority accuracy. It is because the geometric mean is low as long as there is one class with low accuracy.

IV. EXPERIMENTS

In these experiments, the proposed PF-loss is compared with the popular WCE-loss, focal loss (with and without class weight α), and LDAM on both training time and testing accuracy. LDAM is selected as a representative of the losses using other hyperparameters (e.g., C). In the literature [18], LDAM and weighted LDAM (WLDAM) are both used in their proposed deferred reweighting training schedule. Hence, in this work, LDAM and WLDAM are both compared for a fair comparison. In terms of training time, the whole training time, including hyperparameters tuning and one-time training time, is compared. For the comparison on testing accuracy, two popular evaluation metrics, G-mean and area under receiver operating characteristic (ROC) curve (AUC), for imbalance learning [6] are used.

The comparisons of PF-loss with WCE-loss and focal loss were conducted in four aspects: 1) training efficiency of model with different losses; 2) overall effectiveness evaluation in G-mean and AUC for binary imbalance classification; 3) overall effectiveness evaluation for multiclass-imbalanced learning under different IRs; and 4) evaluation on average class accuracy.

A. Comparison Datasets

CIFAR-10: It is one of the most widely used image datasets for deep learning, which contains 60 000 color images with size of 32×32 in 10 different classes. There are 50 000 training images and 10 000 test images [31].

TABLE II
COMPARISON ON BINARY IMBALANCED DATASETS WITH IR = 9

	CIFAR-10-1 (%)		CIFAR-10-3 (%)		CIFAR-10-5 (%)		CIFAR-10-7 (%)		CIFAR-10-9 (%)	
Method	G-mean	AUC	G-mean	AUC	G-mean	AUC	G-mean	AUC	G-mean	AUC
WCE-loss	90.51	90.53	77.28	77.29	**81.86**	**81.91**	86.88	86.89	**89.94**	**89.99**
FL	84.54	85.39	61.61	67.12	71.03	74.17	81.72	82.74	86.36	86.93
FL(α)	90.73	90.74	77.92	77.93	81.58	81.59	86.57	86.62	89.93	89.94
LDAM	87.74	88.17	63.80	68.72	72.68	75.07	82.91	83.83	87.24	87.70
WLDAM	90.60	90.38	**78.47**	**78.52**	81.60	81.60	**87.59**	**87.60**	89.61	89.61
PF-loss	**91.08**	**91.08**	78.05	78.06	81.47	81.47	87.06	87.06	89.81	89.83

CIFAR-100: It is similar to CIFAR-10 except that CIFAR-100 consists of 100 classes [31].

Fashion-Modified National Institute of Standards and Technology (MNIST): It is a direct drop-in replacement for the original MNIST dataset, which contains a training set of 60 000 samples and a test set of 10 000 samples. Each sample is a 28×28 grayscale image, associated with a label from ten classes [32].

For evaluating the performance on binary imbalanced datasets, CIFAR-10 and CIFAR-100 are used, in which one class is considered as minority and all samples in the remaining classes are majority. In detail, odd classes in CIFAR-10 are used in turn as minority, and hence, IR $= 9$. For CIFAR-100, classes 19, 39, 59, 79, and 99 are chosen as minority, and hence, IR $= 99$. It is noting that the experiments about other classes (e.g., 0 and 9) chosen as minority were also conducted. Due to space limitation, we select five classes as representative. This operation is performed in both training and test data.

For multiclass-imbalanced learning, Fashion MNIST and CIFAR-10 are used, in which the samples in even classes are randomly sampled to construct multiclass-imbalanced datasets with different IRs. For instance, a dataset with IR $= 10$ is constructed by selecting 10% samples from every even class and keep 100% samples in odd classes. This operation is performed in both training and test data.

B. Experimental Setting

In order to show the performance of loss function that does not depend on the ability of network model, multiple network structures are used in our experiments. In detail, for binary classification of CIFAR-10 and CIFAR-100, we use the simple structure with two convolution layers, followed by three fully connected layers (FCs) to verify the effectiveness of loss functions. For multiclass classification of Fashion MNIST, a little bit complicated structure of Conv-rectified linear unit (ReLU)-Conv-ReLU-MaxPool-FC-ReLU-Dropout-FC is adopted. For multiclass classification of CIFAR-10, ResNet18 is used to get better performance for all losses. To train these models, the number of epochs is 200, the learning rate is 0.001, the batch size is 500, the weighting decay is 0, and the Adam algorithm is used as an optimizer.

In these experiments, the class weight in both WCE-loss and focal loss is set by $\alpha_i = 1/N_j$ if the ith sample belongs to class j because most studies have verified its effectiveness and $\alpha_i = 1/N_j$ is the optimal one in the range of $\left[0.1/N_i, 1/N_j\right]$ with step of $0.1/N_j$ and $\left[1/N_j, 10/N_j\right]$ with step of $1/N_j$ for minority class and α_i for majority samples that are unchanged. Moreover, the hyperparameter γ follows the default setting (i.e., $\gamma = 2$) in the focal loss [13], [14] because it is the optimal one in the set of $\{0, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0\}$. For LDAM and WLDAM, C is tuned in the range of [0.1, 1] with a step of 0.1.

C. Training Efficiency

Based on the setting of hyperparameters tuning in Section IV-B, the model with WCE-loss, FL (i.e., focal loss without class weight), FL(α) (i.e., focal loss with class weight), LDAM, and WLDAM was trained 19, 7, (19×7), 10, and ten times to search for optimal α, γ, (α, γ), C, and C, respectively, whereas the model with our PF-loss only needs one time of training due to no tuning on hyperparameters. The whole training time, including hyperparameters' tuning, followed by the average one-time training time of models under different hyperparameters and variants of each dataset (e.g., Fashion MNIST with different IRs), are shown in Table I [e.g., 2052 (108)].

In terms of average one-time training time in Table I, our PF-loss shows slower than other compared losses. This is because the gradient update for one individual sample is based on the overall accuracy of one class in PF-loss [shown in (14)], which is difficult to converge earlier than that based on single sample accuracy (e.g., all other compared losses). Consequently, PF-loss requires more convergence time (i.e., one-time training time) than other compared losses. Although PF-loss takes more convergence time, it does not need to tune hyperparameters to fit for data every time, this significantly reduces the whole training time (shown in Table I).

The whole training time in Table I verifies the training efficiency of our PF-loss. Compared to FL(α) that has two hyperparameters, PF-loss only takes about 1/148 of training time spent by FL(α) on Fashion MNIST. Although FL takes less training time than FL(α), its accuracy is not satisfactory (detailed in Sections IV-D and IV-E). Actually, the hyperparameters in experiments are not exhaustively tuned but selected from a small but reasonable range. If exhaustive tuning is used on these hyperparameters, the model with WCE-loss, focal loss, and LDAM will take much more time for the whole training than that shown in Table I.

D. Accuracy on Binary Imbalance Learning

In Table II for CIFAR-10 with IR $= 9$, all losses except FL and LDAM get similar performances. Hence, for CIFAR-10, the proposed PF-loss achieves comparable performance with the popular WCE-loss, FL(α), and WLDAM (i.e., WLDAM), even though the hyperparameters in WCE-loss, FL(α), and WLDAM are all optimally tuned.

On the other hand, for CIFAR-100 with IR $= 99$ in Table III, the proposed PF-loss totally outperforms WCE-loss and FL(α), and the improvement is up to 2.73% (CIFAR-100-99) in terms of G-mean. Compared to WLDAM, PF-loss achieves better performance in 3/5 datasets. This gives an interesting conclusion that our PF-loss is more effective than other losses on the imbalanced datasets with higher IR. Hence, the experiments of multiclass datasets with different IRs are conducted, as detailed in Section IV-E.

In addition, the results of FL in Tables II and III can verify that the focal loss cannot effectively resolve the class-imbalanced problem without using class weight. Similarly, WLDAM achieves better performance than the original LDAM. This is mainly because after the weighting scheme, the classification boundary has a high probability of lying in between positive and negative classes. Otherwise, the

TABLE III
COMPARISON ON BINARY IMBALANCED DATASETS WITH IR = 99

	CIFAR-100-19 (%)		CIFAR-100-39 (%)		CIFAR-100-59 (%)		CIFAR-100-79 (%)		CIFAR-100-99 (%)	
Method	G-mean	AUC	G-mean	AUC	G-mean	AUC	G-mean	AUC	G-mean	AUC
WCE-loss	75.14	75.41	82.73	82.77	86.62	86.69	82.63	82.63	80.46	80.47
FL	41.13	58.25	64.77	70.94	57.22	66.10	50.65	62.33	55.54	65.25
FL(α)	75.96	76.14	82.70	82.79	86.94	86.96	82.29	82.41	80.87	80.92
LDAM	46.82	60.83	69.88	74.33	56.42	65.73	59.04	67.29	66.83	72.13
WLDAM	**77.58**	**77.60**	**83.44**	**83.56**	87.07	87.08	82.70	82.91	81.96	82.51
PF-loss	76.82	76.98	83.00	83.10	**87.97**	**88.00**	**83.84**	**83.85**	**83.19**	**83.19**

TABLE IV
COMPARISON ON MULTICLASS-IMBALANCED DATASET FASHION MNIST WITH DIFFERENT IRS

Fashion MNIST	Method	G-mean(%)	AUC(%)
IR=10	WCE-loss	88.43	94.32
	FL	88.30	94.35
	FL(α)	88.13	94.24
	LDAM	88.15	94.32
	WLDAM	88.34	94.28
	PF-loss	**88.74**	**94.44**
IR=100	WCE-loss	82.35	92.03
	FL	77.13	89.87
	FL(α)	80.19	91.37
	LDAM	80.51	91.50
	WLDAM	82.58	92.02
	PF-loss	**84.85**	**92.93**
IR=200	WCE-loss	77.98	90.35
	FL	75.93	89.48
	FL(α)	78.53	90.07
	LDAM	78.88	91.41
	WLDAM	78.98	90.78
	PF-loss	**82.52**	**91.56**

TABLE V
COMPARISON ON MULTICLASS-IMBALANCED DATASET CIFAR-10 WITH DIFFERENT IRS

CIFAR-10	Method	G-mean(%)	AUC(%)
IR=10	WCE-loss	71.10	85.05
	FL	70.22	84.76
	FL(α)	**72.20**	**85.14**
	LDAM	69.74	84.52
	WLDAM	68.11	83.05
	PF-loss	71.18	84.58
IR=50	WCE-loss	52.10	78.38
	FL	51.32	78.16
	FL(α)	57.52	78.90
	LDAM	60.22	79.90
	WLDAM	62.84	80.22
	PF-loss	**62.99**	**80.28**
IR=100	WCE-loss	55.24	76.82
	FL	45.42	77.06
	FL(α)	55.94	77.33
	LDAM	55.79	77.96
	WLDAM	56.92	78.27
	PF-loss	**61.28**	**79.59**

optimal classification boundary may lie out of both classes (detailed explanation is included in the supplementary materials).

E. Accuracy on Multiclass-Imbalanced Learning

In Table IV for Fashion MNIST, when IR = 10, all losses achieve similar performances. However, when the data are more skewed (i.e., IR = 100 and 200), the proposed PF-loss totally outperforms other five compared class-imbalanced losses. In detail, compared to WCE-loss, FL, FL(α), LDAM, and WLDAM, the improvement of PF-loss is up to 4.54% (IR = 200), 7.72% (IR = 100), 4.66% (IR = 100), 4.34% (IR = 100), and 3.54% (IR = 200) in G-mean.

Similarly, in Table V for CIFAR-10 with IR = 50 and 100, compared to WCE-loss, FL, FL(α), LDAM, and WLDAM, the improvement of PF-loss is up to 10.89% (IR = 50), 15.86% (IR = 100), 5.47% (IR = 50), 5.49% (IR = 100), and 4.36% (IR = 100) in G-mean, respectively. For AUC, our PF-loss also achieves the highest results when the data are highly skewed.

The reason is that the data distribution of each mini-batch is dynamically changing during training. Hence, the IR changes seriously during training in multiclass-imbalanced datasets with higher IR. Nevertheless, the proposed PF-loss is designed to dynamically pay more attention to the minority classes without hyperparameters, and hence, it is adaptive to current mini-batch of training data, whereas in WCE-loss, FL(α), and WLDAM, the class weight is fixed during training and is not adaptive to the changes in each mini-batch. For more experiments and explanations, kindly refer to the supplementary materials.

Since FL and LDAM are designed without class weight, hence, FL and LDAM should be adaptive to the changes in each mini-batch. However, FL focuses on the hard samples regardless of their classes, and hence, it may indirectly focus on majority classes, which is explained in Section IV-F. LDAM may easily get the wrong optimal classification boundary (as detailed in Section IV-D).

TABLE VI
AVERAGE MINORITY AND MAJORITY CLASS ACCURACIES FOR FASHION MNIST WITH DIFFERENT IRS

Fashion MNIST	Method	Minority(%)	Majority(%)
IR=10	WCE-loss	**81.00**	97.22
	FL	80.20	**98.00**
	FL(α)	80.40	97.48
	LDAM	80.20	97.92
	WLDAM	81.40	96.74
	PF-loss	81.40	97.30
IR=100	WCE-loss	72.00	96.90
	FL	62.00	**98.04**
	FL(α)	68.00	98.02
	LDAM	70.00	96.84
	WLDAM	72.00	96.88
	PF-loss	**76.00**	96.58
IR=200	WCE-loss	64.00	97.96
	FL	60.00	**98.38**
	FL(α)	64.00	97.04
	LDAM	68.00	98.14
	WLDAM	68.00	96.10
	PF-loss	**72.00**	95.50

F. Average Class Accuracy

In Table VI of average minority and majority class accuracies for Fashion MNIST as an example, the focal loss without class weight

(i.e., FL) gets the highest average majority accuracy, but the lowest average minority accuracy is under all IR. This observation verifies the conclusion in Section II-C that without class weight, the focal loss indirectly focuses on majority class accuracy. In other words, FL cannot effectively resolve a class-imbalanced problem because minority class accuracy is the main concern in class-imbalanced deep learning.

On the other hand, since the proposed PF-loss is more robust to outliers and focuses its attention on minority classes rather than outliers (detailed in Section III-D), the proposed PF-loss achieves higher accuracies on minority and comparable accuracies on majority than other compared losses, as shown in Table VI. Similar results are also observed in CIFAR-10 with different IRs. Note that, due to the space limitation, the experiments to verify the robustness to outliers are included in the supplementary materials.

V. Conclusion

In this brief, we proposed a PF-loss function for the class-imbalanced problem in deep learning. PF-loss provides the following advantages: 1) there are NO hyperparameters (e.g., class weight), and therefore, NO hyperparameter tuning is necessary, which significantly reduces the whole training time; 2) it can dynamically pay more attention to minority classes during training and adaptive to the change of class-imbalanced ratio within the mini-batch of training data; and 3) it is robust to the outliers and, hence, attention is focused on minority samples rather than outliers, which more effectively addresses the class-imbalanced problem. Experiments were conducted on image datasets CIFAR-10, CIFAR-100, and Fashion-MNIST with different IRs for both binary- and multiclass-imbalanced classification and the IR can be up to 200. Experimental results show that our PF-loss provides a significant reduction (down to 1/148) of the training time of compared state-of-the-art losses. Simultaneously, even without hyperparameter tuning, our PF-loss outperforms WCE-loss up to 10.89% and 1.9% in terms of G-mean and AUC, respectively. Compared to the focal loss with class weight, the improvement of PF-loss is up to 5.47% and 1.89% in G-mean and AUC, respectively. Compared to WLDAM, the improvement of PF-loss is up to 4.36% and 1.53% in G-mean and AUC, respectively.

References

[1] N. Wang, X. Zhao, Y. Jiang, and Y. Gao, "Iterative metric learning for imbalance data classification," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 2805–2811.

[2] W. Zhang, Y. Chen, W. Yang, G. Wang, J.-H. Xue, and Q. Liao, "Class-variant margin normalized softmax loss for deep face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 28, 2020, doi: 10.1109/TNNLS.2020.3017528.

[3] S. Wang, L. L. Minku, and X. Yao, "A systematic study of online class imbalance learning with concept drift," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4802–4821, Oct. 2018.

[4] M. Bader-El-Den, E. Teitei, and T. Perry, "Biased random forest for dealing with the class imbalance problem," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 2163–2172, Jul. 2019.

[5] S. H. Dumpala, R. Chakraborty, and S. K. Kopparapu, "A novel data representation for effective learning in class imbalanced scenarios," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2018, pp. 2100–2106.

[6] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[7] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, p. 27, 2019.

[8] X. Yuan, L. Xie, and M. Abouelenien, "A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data," *Pattern Recognit.*, vol. 77, pp. 160–172, May 2018.

[9] J. Wang, W. Fu, H. Lu, and S. Ma, "Bilayer sparse topic model for scene analysis in imbalanced surveillance videos," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5198–5208, Dec. 2014.

[10] J. Kim, J. Jeong, and J. Shin, "M2m: Imbalanced classification via major-to-minor translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13896–13905.

[11] Y. Cui, M. Jia, T.-Y. 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.

[12] 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.

[13] T.-Y. 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. 318–327.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[15] Q. Qian, L. Chen, H. Li, and R. Jin, "DR loss: Improving object detection by distributional ranking," 2019, *arXiv:1907.10156*. [Online]. Available: http://arxiv.org/abs/1907.10156

[16] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8577–8584.

[17] K. Chen *et al.*, "Towards accurate one-stage object detection with ap-loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, Apr. 2019, pp. 5119–5127.

[18] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," 2019, *arXiv:1906.07413*. [Online]. Available: http://arxiv.org/abs/1906.07413

[19] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *J. Electron. Sci. Technol.*, vol. 17, no. 1, pp. 26–40, 2019.

[20] Y. Lu, Y.-M. Cheung, and Y. Y. Tang, "Bayes imbalance impact index: A measure of class imbalanced data set for classification problem," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3525–3539, Sep. 2020.

[21] B. Richhariya and M. Tanveer, "A reduced universum twin support vector machine for class imbalance learning," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107150.

[22] Y. Xu, "Maximum margin of twin spheres support vector machine for imbalanced data classification," *IEEE Trans. Cybern.*, vol. 47, no. 6, pp. 1540–1550, Jun. 2017.

[23] Y. Lu, Y.-M. Cheung, and Y. Y. Tang, "Dynamic weighted majority for incremental learning of imbalanced data streams with concept drift," in *Proc. 26th Int. Joint Conf. Artif. Intell. (IJCAI)*, Melbourne, VIC, Australia, Aug. 2017, pp. 2393–2399.

[24] K. Malialis, C. G. Panayiotou, and M. M. Polycarpou, "Online learning with adaptive rebalancing in nonstationary environments," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 22, 2020, doi: 10.1109/TNNLS.2020.3017863.

[25] 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.

[26] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[27] C. Huang, Y. Li, C. L. Chen, 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, doi: 10.1109/TPAMI.2019.2914680.

[28] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3573–3587, Aug. 2018.

[29] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 322–330.

[30] A. Akbari, M. Awais, Z. Feng, A. Farooq, and J. Kittler, "Distribution cognisant loss for cross-database facial age estimation with sensitivity analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 7, 2020, doi: 10.1109/TPAMI.2020.3029486.

[31] A. Krizhevsky *et al.*, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2009.

[32] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*. [Online]. Available: http://arxiv.org/abs/1708.07747