ABC-Norm Regularization for Fine-Grained and Long-Tailed Image Classification

Yen-Chi Hsu*, Cheng-Yao Hong*, Ming-Sui Lee, Davi Geiger and Tyng-Luh Liu

***Abstract*—Image classification for real-world applications often involves complicated data distributions such as fine-grained and long-tailed. To address the two challenging issues simultaneously, we propose a new regularization technique that yields an adversarial loss to strengthen the model learning. Specifically, for each training batch, we construct an *adaptive batch prediction* (ABP) matrix and establish its corresponding *adaptive batch confusion norm* (ABC-Norm). The ABP matrix is a composition of two parts, including an adaptive component to class-wise encode the imbalanced data distribution, and the other component to batch-wise assess the softmax predictions. The ABC-Norm leads to a norm-based regularization loss, which can be theoretically shown to be an upper bound for an objective function closely related to rank minimization. By coupling with the conventional cross-entropy loss, the ABC-Norm regularization could introduce adaptive classification confusion and thus trigger adversarial learning to improve the effectiveness of model learning. Different from most of state-of-the-art techniques in solving either fine-grained or long-tailed problems, our method is characterized with its simple and efficient design, and most distinctively, provides a unified solution. In the experiments, we compare ABC-Norm with relevant techniques and demonstrate its efficacy on several benchmark datasets, including (CUB-LT, iNaturalist2018); (CUB, CAR, AIR); and (ImageNet-LT), which respectively correspond to the real-world, fine-grained, and long-tailed scenarios.**

***Index Terms*—Classification, fine-grained, long-tailed, deep neural network, regularization.**

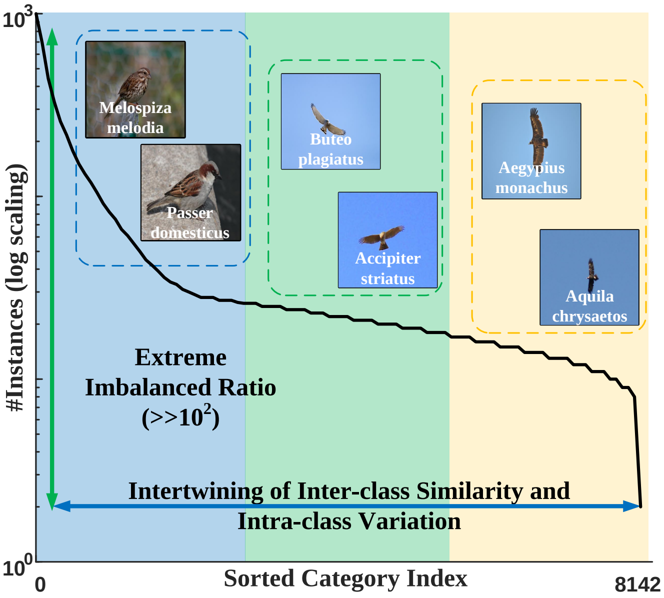


Fig. 1. The distribution of real-world data can include various subtleties such as *fine-grained* and *long-tailed* complications. In terms of algorithm design, these two aspects of challenges can be exemplified by iNaturalist 2018 [1]. As illustrated, the extremely imbalanced numbers of instances among its object classes could derange learning proper features of tail classes for effective classification. Meanwhile, model overfitting could become a major concern in that it is hard to disentangle the fine-grained ambiguities due to the inter-class similarities and intra-class variations in the underlying object categories.

I. Introduction

The performance of an image classification model critically depends on the underlying data distribution, both during the training and the testing stages. For the majority of real-world applications, their underlying data distributions can substantially deviate from those of conventional benchmark collections established solely for research evaluations. Indeed, the distribution of real-world data is often not regular, and for many practical applications, it tends to be more or less fine-grained and even complicated with long-tailed imbalance. To account for such discrepancies in data distribution, recent datasets, *e.g.*, iNaturalist 2018 [1], have been proposed to bridge the gap so that their resulting classification techniques can be widely applied. Figure 1 illustrates two notable and challenging aspects of iNaturalist. First, it exhibits a long-tailed distribution, characterized by extremely imbalanced ratios between head and tail categories. In particular, the almost three orders of magnitude difference in the number of training instances embodied in the long-tailed distribution imposes a difficult task in learning proper representations of tail classes. Second, the object categories in this dataset are also fine-grained, while inter-class similarity and intra-class variations are subtly intertwined. Performing classifications over iNaturalist 2018 is essentially a daunting task, no matter what a specific group (`many`, `medium` or `few`) of fine-grained object classes is under consideration. Motivated by these challenges, we aim to simultaneously address both the fine-grained and long-tailed issues in designing classification techniques for practically dealing with real-world data.

Fine-grained visual classification (FGVC) is an active and challenging problem in computer vision. Such a recognition task differs from the classical problem of large-scale visual classification (LSVC) by focusing on differentiating *similar* sub-categories of the same meta-category. While the inter-class similarity among the object categories is pervasive, the intra-class variations further impose ambiguities in learning a uni-

Y.-C. Hsu, C.-Y. Hong and T.-L. Liu are with the Institute of Information Science, Academia Sinica, Nankang, Taipei 11529, Taiwan e-mail: {yenchi, liutyng}@iis.sinica.edu.tw

M.-S. Lee and Y.-C. Hsu are with the Department of Computer Science and Information Engineering, Taiwan

D. Geiger is with the Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, USA

∗ These two authors contributed equally.

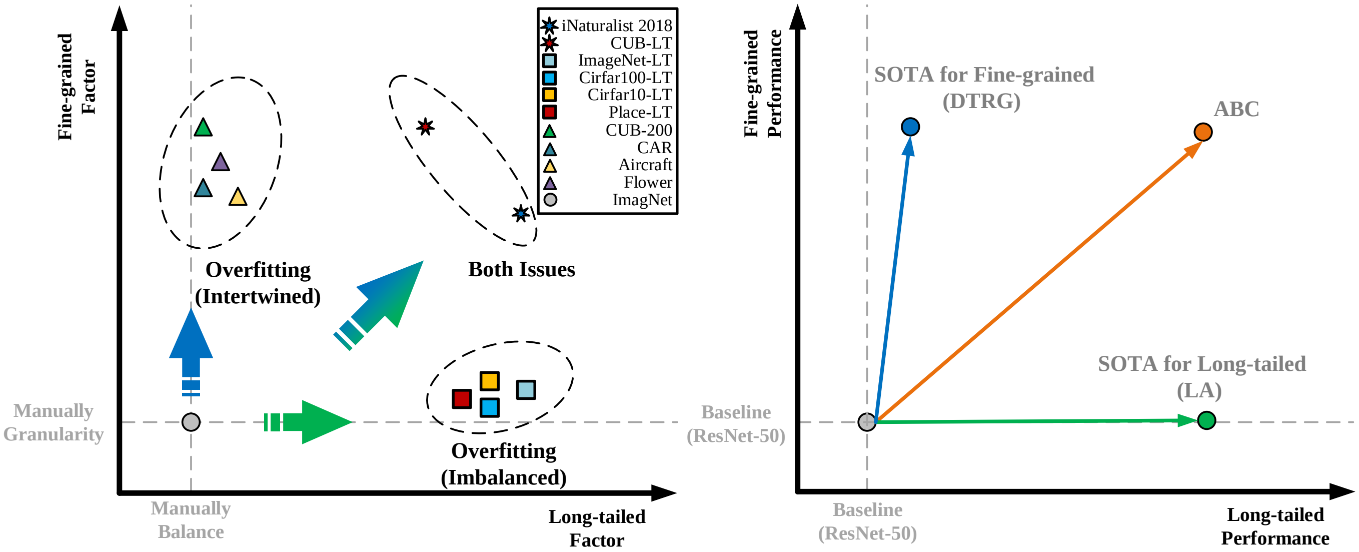


Fig. 2. Left: Different datasets exhibit varying degrees of long-tailed and fine-grained characteristics. Right: Mainstream techniques focus on solving one aspect of the two characteristics, where DTRG [2] (blue dot) and LA [3] (green dot) are respectively current SOTA techniques for tackling the fine-grained and long-tailed classification tasks.

fied and discriminative representation for the FGVC task. On the other hand, considering the issue of long-tailed distribution brings in another aspect of difficulty in developing practical classification techniques. The significantly large numbers of samples from head categories tend to dominate the training procedure. Even with sophisticated learning strategies, the resulting classification model often ends up performing poorly for the tail categories, compared with the expected result on the head counterparts. In fact, the performance curve somewhat resembles the shape of a long-tailed distribution.

We note from existing literature of object classification research that there are only a few attempts to simultaneously solve the two aforementioned challenging issues. Relevant developments mainly focus on tackling either of the two tasks. In FGVC, most of the recent research efforts have converged to learning pivotal local/part details related to distinguishing fine-grained categories *e.g.*, [4]–[6]. Moreover, to improve the classification performance further, a number of these efforts require the fusion of several sophisticated computer vision techniques, such as in [7], [8]. In resolving the long-tailed difficulty, previous approaches have drawn on balanced data sampling to rectify their model training [9]–[11]. For example, the recent technique of [11] first learns the representation and then refines the classifier by balanced sampling. All these different research attempts involve varying degrees of fine-grained and long-tailed factors. As shown in Figure 2 (Left), we take the maximum imbalanced ratio and the normalized feature cosine similarity between object categories as the respective criterion to measure the fine-grained and long-tailed factors and characterize the two aspects of difficulties among popular datasets adopted in object recognition research. Moreover, Figure 2 (Right) indicates that a purely fine-grained state-of-the-art (SOTA) approach does not necessarily perform well for the long-tailed case, and vice versa, while our approach provides a unified solution to tackling the two challenging issues of image classification.

In this work, we focus on establishing a fundamental approach based on exploring the characteristics of the real-world data distributions rather than relying on various data augmentation schemes and sophisticated DNN-based engineering tricks. From the two plots in Figure 3, we observe that when the objective function during training converges very close to zero, the results in testing are often not the best. To avoid being trapped with over-optimizing the underlying model, previous approaches have adopted regularization techniques to resolve this matter. Take, for example, the inclusion of *margin* in the triplet loss [12]. The design principle of triplet loss is to separate positive and negative samples by at least a default margin, say m, which turns out to play a pivotal role in boosting the learning efficacy. Different from typical regularization techniques, it implicitly raises the learning difficulty of the objective function, instead of limiting the model capacity.

The concept of incorporating extra difficulty into training has also been proposed in dealing with the FGVC problem. Pairwise Confusion (PC) [13] and Maximum Entropy (MaxEnt) [14] are two such approaches, closely related to our proposed method. PC argues that slightly confusing the model in training can prevent overfitting problems. MaxEnt observes that the data diversity of FGVC is usually smaller than that of a large-scale classification dataset, *e.g.*, ImageNet. It thus presumes that the entropy of the model's prediction should tend to be higher than that of typical classification scenarios. Both PC and MaxEnt add a confusion-like loss to improve the FGVC performances of their resulting models. Still, there are currently no relevant arguments in addressing fine-grained and long-tailed issues simultaneously.

We are thus motivated to develop a new classification technique, termed *adaptive batch confusion norm* (ABC-Norm), to

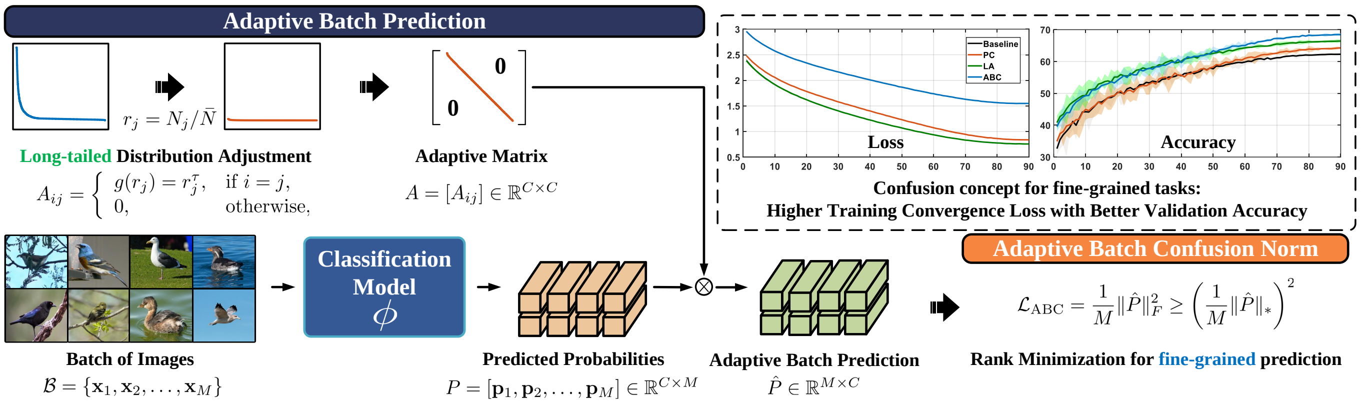

Adaptive Batch Prediction
$r_j = N_j/\bar{N}$
0
0
Long-tailed Distribution Adjustment
$A_{ij} = \begin{cases} g(r_j) = r_j^\tau, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases}$
Adaptive Matrix
$A = [A_{ij}] \in \mathbb{R}^{C\times C}$
Baseline
PC
LA
ABC
3
2.5
2
1.5
1
0.5
0
10
20
30
40
50
60
70
80
90
Loss
70
60
50
40
30
Accuracy
Confusion concept for fine-grained tasks:
Higher Training Convergence Loss with Better Validation Accuracy
Batch of Images
$\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$
Classification Model
ϕ
Predicted Probabilities
$P = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M] \in \mathbb{R}^{C\times M}$
$\otimes$
Adaptive Batch Prediction
$\hat{P} \in \mathbb{R}^{M\times C}$
Adaptive Batch Confusion Norm
$\mathcal{L}_{\text{ABC}} = \frac{1}{M}\|\hat{P}\|_F^2 \geq \left(\frac{1}{M}\|\hat{P}\|_*\right)^2$
Rank Minimization for fine-grained prediction

Fig. 3. Overview of *adaptive batch confusion norm* (ABC-Norm). The adaptive batch prediction $\hat{P}$ can be obtained by class-wise modulating the predicted probabilities P with respect to the adaptive matrix A that encodes the underlying data distribution. Our formulation then adds slight classification confusions to yield an adversarial regularization effect in model training. Despite that ABC-Norm converges to a higher training loss than other techniques, it indeed achieves better validation accuracy.

regularize its corresponding *adaptive batch prediction* (ABP) matrix to better account for real-world data distributions. ABC-Norm can be used to deal with both fine-grained and long-tailed factors and to construct an adversarial loss for enhancing the training efficacy. Optimizing with respect to the ABC-Norm drives the learning process to (class-wise) adaptively add confusions to achieve better classification results. We also provide a mathematical derivation to justify the concept and the ideas it represents. Figure 3 illustrates an overview of ABC-Norm. We characterize the advantages of our method as follows.

- The computation of ABC-Norm regularization is efficient and does not incur significant increase in training time.
- Unlike related techniques, *e.g.*, [11], [15] that decouple representation learning from classification or learn multiple distribution-aware experts, our regularization-based method leads to an end-to-end trainable implementation.
- Without relying on complicated model design or sophisticated data augmentations such as in, *e.g.*, [2], [8], [16], ABC-norm not only provides a unified solution to resolving fine-grained and long-tailed issues but also improves the baselines to achieve competitive classification results.

II. Related Work

In addressing conventional computer vision tasks, the underlying distribution of training data is often relatively balanced. The numbers of samples across various object categories do not differ substantially, and in addition, the diversity among the categories is typically high. However, the data distributions for real-world applications are far more complicated; it could even contain fine-grained and long-tailed complexities at the same time. Recent related work for image classification tends to emphasize either aspect of the two difficulties, but not both. Taking such development into account, we divide the literature survey of relevant techniques into two groups, namely, *fine-grained visual classification* and *long-tailed visual recognition*.

A. Fine-grained visual classification

The Fine-Grained Visual Classification (FGVC) problem is notably characterized by two intriguing properties, significant inter-class similarity and intra-class variations, which cause learning an effective FGVC classifier a challenging task. Driven by impressive research progress, the setting of FGVC has gradually evolved from strong labels to weak labels.

a) Early work: In the initial efforts for tackling FGVC, the developed methods mostly assume that the training datasets are made with comprehensive annotations, such as the part location labels in CUB-200 [17]. Along this line, Berg *et al.* [18] explore the labeled part locations to eliminate highly similar object categories for improving the classifier. Huang *et al.* [19] introduce an approach established based on a two-stream classification network to capture both object-level and part-level information explicitly. However, due to the rapid research advances in visual classification, the most recent FGVC approaches are designed to complete the model learning based on the category labels solely. Hence, without accessing the part location labels, how to learn the discriminative parts automatically becomes the next research direction.

b) Discriminative parts: Existing FGVC approaches usually draw on data augmentations and specific attention mechanisms to effectively learn the discriminative parts. Yang *et al.* [5] propose a self-supervision mechanism to localize informative regions without the need of bounding-box and part annotations. Wang *et al.* [20] present a filter bank within a CNN framework to learn high-quality discriminative patches. Zheng *et al.* [6] introduce a trilinear attention sampling network for fine-grained image recognition, which can learn rich feature representations from hundreds of part proposals. Chen *et al.* [21] propose a *destruction and construction learning* (DCL) framework for fine-grained image recognition. DCL partitions each training image into several local regions and then shuffles them by a *region confusion mechanism* (RCM). It implicitly excludes the global object structure information and forces the model to predict the category label based on local information. Moreover, construction learning can model the semantic correlation among parts of the object. In other words, the ability to identify the object category from local details is expected to be enhanced through shape destruction. Du *et al*. [8] apply a progressive training strategy to address the fine-grained classification task. They formulate a framework named *progressive multi-granularity* (PMG) training with two key components. One is a training strategy that progressively fuses multi-granularity features, and the other is a puzzle generator to form images containing information of different granularity levels. Chang *et al*. [22] propose a *mutual-channel loss* (MCLoss) that drives the model to learn channel diversity and emphasize different discriminative regions. In summary, the above techniques are established based on employing richer augmentations and specialized attention mechanisms. In the case of the top-performing PMG, each iteration requires four different phases of augmentation combined with four classifiers. Although the results are state-of-the-art, PMG requires more training time and extensive model parameters.

c) Auxiliary task variants: Several related approaches include an additional branch to explore auxiliary information. Shu *et al*. [23] propose a self-training framework for FGVC with insufficient data annotation by considering an additional auxiliary task path to generate pseudo labels. They leverage the Grad-CAM technique [24] to generate salient regions for seeking discriminative parts, which can be further extended to yield multiple attention maps for improving the quality of the representation. Chang *et al*. [25] introduce a novel FGVC problem setting by generalizing it from single-label to multiple-label predictions on a predefined label hierarchy. A user study is also provided to show that a multi-granular label hierarchy is more expressive and probably preferred. Their proposed solution shows that the inherent coarse-fine hierarchical relationship can improve FGVC performance.

d) Regularization effects: The regularization-related formulations for dealing with intra-class variations and inter-class similarity in FGVC generally have two main implications. First, it can be applied to alleviate the overfitting problem in learning an FGVC model. Dubey *et al*. [13] propose to divide each batch into two groups and train the model with a loss function including *pairwise confusion* (PC). The design reasons that bringing the class-wise probabilities closer could prevent the learned FGVC model from overfitting. Second, the regularization tactic implicitly maximizes the prediction entropy. MaxEnt [14] assumes that the data diversity of FGVC is intuitively smaller than the large-scale dataset, ImageNet. So the prediction entropy for the FGVC task is reasonable to become more prominent than usual. In other words, regularization approaches escalate the training difficulty on the total loss, which complicates the training convergence and forces the model to search for an ideal local minimum. In [2], Liu *et al*. introduce *dynamic target relation graphs* (DTRG) to address the fine-grained classification problem with a self-supervised regularization. DTRG evaluates every training sample to calculate the class center online. And then, DTRG aims to reduce the intra-class distance between each training feature and its corresponding class center, while keeping the class centers to be away from each other. It can be observed that the regularization principle of DTRG is quite different from the entropy-based confusion view entailed in PC and MaxEnt. In addition, the training process of DTRG is more intricate and also requires substantial augmentation techniques to strengthen the outcome of model learning.

B. Long-tailed visual recognition

a) Distribution re-balancing: Existing techniques for long-tailed visual recognition that consider distribution re-balancing can be divided into two groups: *re-sampling* and *re-weighting*. As described in [26]–[29], re-sampling involves adjusting the sampling frequencies of different categories based on their sample count via under-sampling for head categories and over-sampling for tail categories. The approach of class-balanced sampling [30] weights each image based on the number of samples in its category. In [31], the dynamic-sampling mechanism, termed as *repeat factor sampling* (RFS) by Gupta *et al*. [31], also aims to balance the number of instances across categories. While the goal of re-sampling is to reduce the overfitting of head data, the tactic may not always be a reliable solution. It could cause over-sampling of small amounts of tail data, resulting in insufficient sample diversity and under-sampling of large amounts of head data, leading to insufficient learning.

b) Loss re-weighting: The strategy of re-weighting has been widely utilized in the loss calculation of a classification task. Unlike re-sampling, re-weighting offers greater flexibility and ease of computation, making it a popular choice for resolving the challenge of long-tailed distribution in more complex tasks such as object detection and instance segmentation. When an image contains multiple objects that need to be detected or segmented, it is often more manageable to reweigh the loss at the image level rather than sample by category. Re-weighting implementations range from reverse weighting based on category distribution to more advanced methods such as Hard Example Mining [32], Focal loss [33], and Label-Distribution-Aware Margin (LDAM) loss [34], which adjust the weight according to the classification credibility without the need for category knowledge. Owing to its ease of implementation, re-weighting has been shown to yield competitive results in complex tasks [35]–[37].

c) Model training strategies: Another viewpoint for solving the long-tailed visual recognition problem is that the re-balancing technique should be applied only to the classifier, and the distribution of image features during representation learning should remain unchanged. This two-stage training strategy, in which the classifier is trained with re-balanced data and the representation is learned with the original data, is considered an effective solution for handling the long-tailed distribution. Kang *et al*. [38] divide the training of a long-tailed classification model into two steps, first directly learning a representation model from traditional classification with raw data and then connecting a separate classifier via class-balanced sampling learning. Zhou *et al*. [39] realize the two-step learning with a two-branch model where both branches share parameters and are dynamically weighted, one branch learning from raw data and the other from re-sampled data. Li *et al*. [40] adopt a two-stage learning approach and introduce a balanced group softmax module into the classification head. Meanwhile, Hu *et al.* [41] tackle the long-tailed distribution scenario through incremental learning from the head to the tail. Wang *et al.* [42] add a separate classifier to calibrate prediction logits, while Tang *et al.* [43] compute the moving average vector of a feature in the traditional training framework, excluding it from the gradient calculation. Menon *et al.* [3] revisit the classic idea of logit adjustment based on statistical information, encouraging a large relative margin between the logits of rare and dominant labels. Tian *et al.* [16] address long-tailed object recognition with the VL-LTR model, which jointly trains the image and text encoders by considering co-embedding between class-wise linguistic and visual information. Wang *et al.* [44] establish a quantitative measure, defining an overlap coefficient between von Mises-Fisher distributions, to evaluate representation quality for long-tailed learning.

In summary, the majority of the aforementioned methods for long-tailed learning emphasize exploring the aspect of data distribution. Such approaches, as we have just described, can be broadly categorized into three groups: distribution re-sampling, loss re-weighting, and model training strategies. In this work, we introduce a novel approach to addressing fine-grained and long-tailed issues at the same time. By infusing pivotal statistical characteristics of the data distribution into an adaptive matrix, the proposed regularization learning with an adversarial loss is shown to be a promising solution.

III. Our Method

Consider now learning a classification model Φ, as illustrated in Figure 3, with respect to a dataset $\mathcal{D}$ of C object categories, where each sample $\mathbf{x} \in \mathcal{D}$ is specified with a one-hot class label vector $\mathbf{y}$. For an arbitrary training batch $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$ from $\mathcal{D}$ and $M \leq C$, forward propagation via Φ yields M predicted (softmax) probabilities, denoted as

$$P = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M] \in \mathbb{R}^{C\times M}, \tag{1}$$

where $\Phi(\mathbf{x}_i) = \mathbf{p}_i \in \mathbb{R}^C$ is the predicted probability distribution. Let $\mathbf{p}_{i,j}$ be the probability of $\mathbf{x}_i$ being class j. We have $\sum_{j=1}^{C} \mathbf{p}_{i,j} = 1$. The batch prediction matrix P in (1) is central to our approach—its rank property is closely related to how our approach resolves the fine-grained issue.

The data distribution over the C object classes in $\mathcal{D}$ reflects the long-tailed characteristic. Let N_j be the sample size of class j and $\bar{N}$ be the averaged sample size over the C classes. We express the ratio of N_j to $\bar{N}$ as $r_j = N_j/\bar{N}$ and consider a unit-coefficient power function of r_j, namely $g(r_j) = r_j^\tau$, to model the underlying long-tailed distribution. Note that the real-valued power τ is a hyper-parameter of our method, and its value is to be adjusted according to the extent of long-tailed distribution. Specifically, to encode the class-wise imbalance, we define an adaptive matrix $A = [A_{ij}] \in \mathbb{R}^{C\times C}$ by

$$A_{ij} = \begin{cases} g(r_j) = r_j^\tau, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where (as we will explain later) the value of τ, along with r_j, reflects the degree of long-tailed attribute and can be adaptively set to account for different application scenarios.

A. Adaptive Batch Confusion Norm

We aim to introduce a regularization based framework to simultaneously address the fine-grained and long-tailed issues of object classification. Based on (1) and (2), we construct an *Adaptive Batch Prediction* (ABP) matrix $\hat{P} \in \mathbb{R}^{M\times C}$ by

$$\hat{P} = P^\intercal A, \tag{3}$$

where the adjusted softmax prediction of each sample in $\mathcal{B}$ now forms a row vector of $\hat{P}$. Observe from (2) that how the adaptive matrix A modifies the prediction outputs depends on the exponent τ and the *imbalanced* factor r_j of each class j in the training data $\mathcal{D}$. When $r_j \to 1$ or the exponent $\tau \to 0$, A would approach the identity matrix I. In other words, the ABP matrix in (3) will be reduced to $P^\intercal$ when the distribution of training data is class-wise balanced, or τ is set to 0. Both cases exclude the long-tailed consideration of $\hat{P}$.

The main idea of our approach is to establish a unified regularization mechanism from the ABP matrix $\hat{P}$ so that the model training process can effectively improve its inference performance on our targeted classification scenarios. To this end, we propose the *Adaptive Batch Confusion Norm* (ABC-Norm) to assess the corresponding loss, expressed as $\mathcal{L}_{ABC}$, which realizes the desired regularization effects for addressing the fine-grained and long-tailed issues. Specifically, we define the loss term for the ABC-Norm regularization as

$$\mathcal{L}_{ABC} = \frac{1}{M}\|\hat{P}\|_F^2\,, \tag{4}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and M is the batch size as in (1). Unlike other existing techniques that are often developed by integrating several sophisticated classification modules to tackle the fine-grained or long-tailed difficulties, our formulation learns the proposed model by directly optimizing the following objective function:

$$\mathcal{L}_{\text{total}} = (1-\lambda)\,\mathcal{L}_{\text{CE}} + \lambda\,\mathcal{L}_{\text{ABC}}, \tag{5}$$

where $\lambda \in [0, 1]$ is a weight parameter,

$$\mathcal{L}_{\text{CE}} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{C} \mathbf{y}_{i,j}\log \mathbf{p}_{i,j} \tag{6}$$

is the conventional cross-entropy loss, and

$$\begin{aligned}\mathcal{L}_{\text{ABC}} = \frac{1}{M}||\hat{P}||_F^2 &= \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{C} \hat{\mathbf{p}}_{i,j}^2 \\ &= \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{C} A_{j,j}\mathbf{p}_{i,j}^2. \end{aligned} \tag{7}$$

Empirically, we set $\tau = 0.5$ and consider $\lambda \in \{0.1, 0.3, 0.5\}$ for various datasets to achieve the best performance accounting for different fine-grained and long-tailed characteristics of real-world distributions. Since the regularization term $\mathcal{L}_{\text{ABC}}$ in (5) is the only factor that distinguishes our method from a vanilla classification scheme, the performance gains reported in our experiments are evidently owing to the ABC-Norm efficacy.

B. ABC-Norm: Justifications and Properties

The rank of the ABP matrix $\hat{P}$ in (3) plays a pivotal role in our formulation, and is closely related to the ABC-Norm regularization. Assume for the moment that minimizing $\mathcal{L}_{\text{ABC}}$ can lead to rank minimization of $\hat{P}$. It would then reduce the variability among the M softmax predictions of P from a batch $\mathcal{B}$, and infuse slight classification *confusions* into the training procedure. Whereas correct predictions would always be penalized with the confusion loss as in (5), the training would be driven to further improve the model by reducing the cross-entropy loss as much as possible, and consequently better solve the fine-grained classification problem. Such an adversarial regularization idea is analogous to enhancing the model learning by introducing an extra margin to increase the difficulty of a correct prediction.

It is known that the rank-related minimization problems are often NP-hard. We follow [45] and consider convex relaxation so that the underlying rank minimization of $\hat{P}$ can be reduced to the minimization of its nuclear norm,

$$\|\hat{P}\|_* = \sum_{i=1}^{M} \sigma_i(\hat{P})\,, \tag{8}$$

where $\sigma_i(\cdot)$ yields the ith singular value of the corresponding matrix. However, training a deep neural network with an objective function that involves solving singular values of a non-trivial matrix is not practically feasible. It is also the main reason that we do not establish the ABC-Norm regularization based on the nuclear norm. We instead consider minimizing its upper bound as in (4). In this way, rank minimization of $\hat{P}$ can be efficiently achieved by employing $\mathcal{L}_{ABC}$. To complete the mathematical derivation of our method, we are left to justify the following upper-bound property.

Property 1. *If the batch size M is set as less or equal to the number of classes C, then*

$$\mathcal{L}_{\text{ABC}} = \frac{1}{M}\|\hat{P}\|_F^2 \geq \left(\frac{1}{M}\|\hat{P}\|_*\right)^2 . \tag{9}$$

It follows that minimizing the nuclear norm of $\hat{P}$ can be achieved by including $\mathcal{L}_{\text{ABC}}$ in the total loss. That is, rank minimization is implicitly carried out during the model training of the classifier Φ. To verify the upper-bound property stated in (9), we have, from the matrix norm definitions and Cauchy-Schwarz inequality,

$$\begin{aligned}\mathcal{L}_{\text{ABC}} &= \frac{1}{M}\|\hat{P}\|_F^2 = \frac{1}{M}\sum\nolimits_{i=1}^{M}\sigma_i^2(\hat{P})\\ &= \frac{1}{M^2}\left(\sum\nolimits_{i=1}^{M}\sigma_i^2(\hat{P})\right)\left(\sum\nolimits_{i=1}^{M}1^2\right)\\ &\geq \frac{1}{M^2}\left(\sum\nolimits_{i=1}^{M}\sigma_i(\hat{P})\cdot 1\right)^2\\ &= \left(\frac{1}{M}\sum\nolimits_{i=1}^{M}\sigma_i(\hat{P})\right)^2\\ &= \left(\frac{1}{M}||\hat{P}||_*\right)^2 .\end{aligned}$$

We now turn our attention to explaining how the adaptive matrix $A \in \mathbb{R}^{C\times C}$ in $\hat{P} = P^\intercal A$ is used in dealing with the long-tailed issue. Notice that A is a diagonal matrix whose jth diagonal entry $A_{jj} = r_j^\tau = (N_j/\bar{N})^\tau$ adjusts the predicted probability $\mathbf{p}_{i,j}$ for the jth class. When learning with a long-tailed training dataset $\mathcal{D}$, its head classes are those that include more training samples and thus have $r_j \gg 1$. Hence the adaptive effects on these head classes are to enforce more confusions/difficulties in classify their abundant samples. On the contrary, tail classes are characterized with $r_j \ll 1$ and the adaptive matrix A is used to instead lessen their confusion regularization so that learning with these scarce data can be guided by the cross-entropy loss.

C. ABC-Norm vs. Relevant Regularization

To our knowledge, there are no existing regularization techniques that are developed to simultaneously tackle both the fine-grained and long-tailed classifications. The two most relevant approaches, but focusing on only the fine-grained aspect, are Pairwise Confusion (PC) [13] and Maximum Entropy (MaxEnt) [14]. We describe their design principles and relevance to the ABC-Norm regularization below.

a) PC Regularization: This is the first work [13] that brings in the "confusion" concept to solve the fine-grained classification task. The purpose of confusion energy is twofold. Besides preventing the model training form overfitting, it implicitly increases the learning difficulty to aim for performance gains in testing. PC randomly divides each batch into two equal-size sub-batches. While computing the individual cross-entropy losses for each sample of the whole batch, it evaluates the pairwise confusion loss, denoted as $\mathcal{L}_{\text{PC}}$, by sampling from the two parts. Specifically, we have

$$\mathcal{L}_{\text{PC}} = \frac{2}{M}\sum_{i=1}^{M/2}\mathbb{I}(\mathbf{y}_i = \mathbf{y}_{i+M/2})||\mathbf{p}_i - \mathbf{p}_{i+M/2}||_2\,, \tag{10}$$

where $\mathbb{I}(\cdot)$ is the indicator function to signal whether two paired training samples are of the same category.

b) MaxEnt Regularization: The maximum entropy criterion is proposed in [14] to more effectively address the fine-grained classification problem. As the inter-class variations between fine-grained classes could be subtly minimal, MaxEnt regularization assumes no prior distributions other than the uniform one should be imposed on the softmax predictions. Analogous to PC, the maximum entropy regularization also increases the learning difficulty and therefore drives the optimization process to work harder in tackling the challenging classification scenario. The corresponding loss for MaxEnt regularization is defined as follows:

$$\mathcal{L}_{\text{MaxEnt}} = \frac{-1}{M}\sum_{i=1}^{M}\sum_{j=1}^{C}\mathbf{p}_{i,j}\log\mathbf{p}_{i,j}\,. \tag{11}$$

Comparing the three regularization schemes, ABC-Norm, PC and MaxEnt, their most distinction is that our formulation tackles not only the fine-grained but also the long-tailed difficulty. Furthermore, by setting the adaptive matrix A in (3) to the identity matrix I, we can look further into how their design improves the performance on fine-grained classification. The three techniques resemble each other by imposing adversarial difficulty in the model training to enhance the classification efficacy. For PC versus ABC-Norm, both are established based on the concept of confusion, while ABC-Norm has the advantage of exploring the adversarial measure from an entire batch at the same time, rather than the pairwise mechanism as in PC. For MaxEnt versus ABC-Norm, while the softmax prediction of each sample in the batch $\mathcal{B}$ being a uniform distribution is a minimum for both regularization losses, $\mathcal{L}_{\text{ABC}}$ is more general in accommodating other minima. In our experiments, we replace $\mathcal{L}_{\text{ABC}}$ in the total loss in (5) with $\mathcal{L}_{\text{PC}}$ and $\mathcal{L}_{\text{MaxEnt}}$, respectively to thoroughly compare their performances on various datasets and settings.

IV. Experiments

A. Datasets

We conduct experiments on the six datasets listed in Table I. In particular, our main objective is to evaluate the efficacy of the proposed ABC-Norm approach to the real-world classification challenges over the two datasets, CUB-LT [46] and iNaturalist2018 [1]. To further analyze its performance, we evaluate ABC-Norm on three fine-grained datasets (CUB, CAR, AIR) and a long-tailed dataset (ImageNet-LT), respectively. The results of our experiments demonstrate that ABC-Norm can effectively and efficiently tackle the challenging classification tasks posed by these benchmark datasets.

a) Real-world: We begin by evaluating the ABC-Norm regularization on the two real-world datasets, CUB-LT [46] and iNaturalist2018 [1], each of which includes both fine-grained and long-tailed distribution characteristics. iNaturalist2018 is a large-scale collection. Owing to its challenging nature, as demonstrated in recent literature [11], [47], the performance on this dataset could serve as an objective measure for the effectiveness of each particular method.

b) FGVC: We then compare solely the fine-grained classification results from four different regularization approaches, *adaptive batch confusion Norm* (ABC-Norm), *pairwise confusion* (PC) [13], *maximum entropy* (MaxEnt) [14], and *dynamic target relation graphs* (DTRG) [2] on the three popular fine-grained visual classification datasets, namely, CUB-200-2011 [17], Stanford Cars [48], and FGVC-Aircraft [49]. The size ratio between training and testing sets is about $1:1$ for CUB-200-2011 and Stanford Cars, and about $2:1$ for FGVC-Aircraft. The class distribution of the three datasets is nearly balanced, which can be used to measure the proposed method's performance only in the fine-grained scenario. Notice that the adaptive matrix A will be reduced to an identity matrix I in dealing with the balanced data distribution.

c) Long-tailed: Finally, we carry out experiments on the long-tailed dataset ImageNet-LT [50], which can be considered to have a low fine-grained factor. The study aims to confirm the capability of ABC-Norm to tackle long-tailed learning over purely imbalanced datasets. In line with the definition in [11], we divide the categories into three groups: `Many`, `Medium`, and `Few`, representing the categories with instance numbers in the ranges $(100, +\infty)$, $(20, 100]$, and $(0, 20]$, respectively.

TABLE I
Dataset splits in our experiments.

Dataset	# Train	# Val/Test	# Category
iNaturalist2018	437,513	24,426	8,142
CUB-LT	2,945	2,348	200
CUB	5,994	5,794	200
CAR	8,144	8,041	196
AIR	6,667	3,333	100
ImageNet-LT	115,846	20,000	1,000

B. Implementation Details

We now describe the implementation details of the experiments on the real-world, fine-grained, and long-tailed datasets. All results are obtained from end-to-end training, and the numerical outcomes represent the mean of three runs. We implement our method using the PyTorch framework [51] on a platform with four Nvidia V100 GPUs. The source code will be made available for public use.

a) Real-world: These results pertain to the CUB-LT and iNaturalist2018 datasets. To ensure a fair comparison, our training settings mostly conform to those outlined in [11], [44]. The backbone network is ResNet-50 with an input size of 224×224 and 90 training epochs, optimized using SGD. The batch size is set to 16 for CUB-LT and 128 for iNaturalist2018. The initial learning rate is set to $0.004 \times M$, where M denotes the batch size, and is decreased by a cosine annealing schedule. The regularization weight λ is set to 0.5, and the value of τ for the adaptive matrix A is set to 0.5.

b) FGVC: We evaluate the performance of the ABC-Norm on popular classification architectures, including ResNet series [52] and DenseNet-161 [53], in the fine-grained visual classification task. The training setup for the different regularization terms remains consistent. We adopt the data augmentation strategy from [21], using an input size of 448×448 and randomly applying horizontal flipping. The initial learning rate, the weighting factor λ, and τ are set to 0.008, 0.3, and 0.5, respectively. The training batch size is 16 when the GPU memory allows, and the adopted optimization algorithm is Momentum SGD with cosine annealing [54] for the learning rate decay. Taking account of the smaller scale of FGVC datasets compared to iNaturalist2018, we train the model for 200 epochs to assess the outcomes of different regularization methods.

c) Long-tailed visual recognition: We further evaluate the proposed ABC-Norm on an imbalanced dataset, ImageNet-LT. The implementation details follow the training process described in [11]. We report results for both ResNeXt-50 and ResNeXt-152, and observe consistent behavior between shallow and deep models. Given the substantial imbalance present in the ImageNet-LT with a low fine-grained factor, we set the hyper-parameters λ and τ to 0.1 and 0.5, respectively.

C. Real-world Data

Before we delve into the real-world data, let us quickly look at a small-scale one, CUB-LT, which contains both fine-grained and long-tailed factors. It is an appropriate dataset

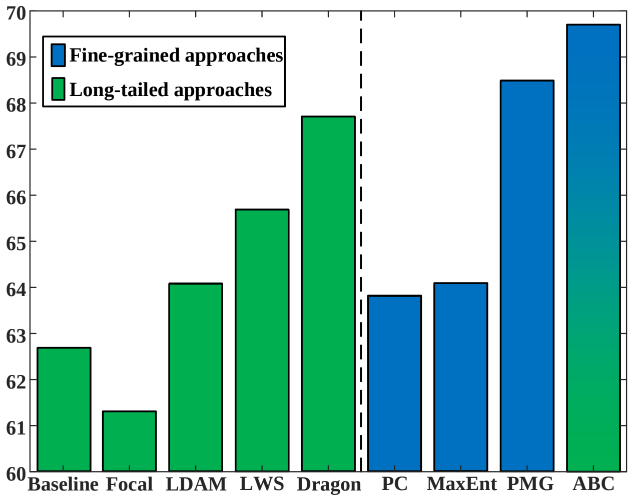


Fig. 4. Compare the proposed ABC-Norm with other long-tailed and fine-grained approaches on CUB-LT.

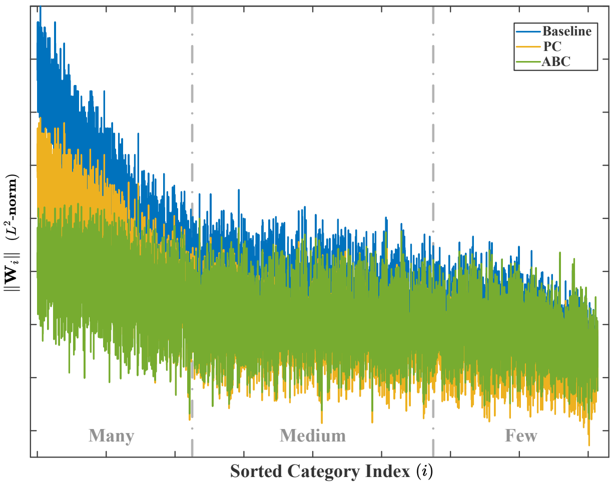


Fig. 5. The distribution of L^2-norm weight magnitude $\|\mathbf{w}_i\|$ for baseline, PC and ABC-Norm, where $\mathbf{w}_i$ is the classifier weight vector of category i.

TABLE II
COMPARE THE ABC-NORM WITH OTHER PRIMARY APPROACHES ON INATURALIST2018. THE BACKBONE MODEL USED IN THIS EXPERIMENT IS VANILLA RESNET-50 BASELINE WITHOUT USING ADDITIONAL PARAMETERS AND ADVANCED AUGMENTATION SCHEMES.

Method	Many	Medium	Few	Total
	90 epochs			
Baseline	**72.2**	63.0	57.2	61.7
Focal [33]	-	-	-	61.1
Re-weighted	-	-	-	64.9
cRT	-	-	-	65.2
PC† [13]	70.9	64.6	59.6	62.1
MaxEnt† [14]	69.8	65.1	59.4	61.9
LDAM [47]	-	-	-	64.6
LDAM w/ DRW [47]	-	-	-	68.0
LWS [11]	65.0	66.3	65.5	65.9
LA [3]	-	-	-	66.4
ABC-Norm	66.6	68.0	68.2	68.4
ABC-Norm‡	66.5	**73.4**	**69.2**	**70.8**
	200 epochs			
Baseline	**75.7**	66.9	61.7	65.8
cRT	73.2	68.8	66.1	68.2
PC† [13]	67.8	64.2	60.2	62.8
MaxEnt† [14]	70.8	65.3	59.1	62.1
DTRG [2]	-	-	-	65.5
DTRG w/ DRW [2]	-	-	-	69.5
LWS [11]	71.0	69.8	68.8	69.5
vMF [44]	72.8	71.7	70.0	71.0
ABC-Norm‡	68.1	**73.2**	**70.4**	**71.4**

† Re-implement with the same setting as ours.
‡ Follow the data augmentation scheme in [44].

for investigating the performances among the respective approaches for fine-grained [8], [13] and for long-tailed [3], [11], [46]. As shown in Figure 4, PC and MaxEnt, which are proposed to account for the fine-grained factor, only show slight improvements for resolving the long-tailed issue. PMG provides a strong performance, but requires more advanced data augmentations and larger model sizes. Meanwhile, LDAM, LWS, vMF, and Dragon demonstrate that addressing the long-tailed issue can also improve performance on real-world data distributions. However, the proposed ABC-Norm significantly outperforms these approaches by explicitly tackling both the fine-grained and long-tailed challenges.

Table II shows the experimental results on the large-scale and real-world distribution dataset, iNaturalist2018. The adaptive matrix A enables the ABC-Norm to emphasize the head categories but scale down the regularization effect on the tail categories. Note that our models are trained not only with the most common way of data sampling, *i.e.*, *instance-balanced sampling*, but also in an end-to-end manner. In contrast, LWS [11] learns the model in two stages, which requires the use of *class-balanced sampling*. Notwithstanding that LA [3] has the same starting point as ours, which also proposes an approach that does not require any extra parameters, strong augmentation schemes, and data sampling strategies, the ABC-Norm regularization does yield a better performance. The main advantage of ABC-Norm over LA on this real-world dataset is that the proposed ABC-Norm provides a unified solution to addressing both long-tailed and fine-grained factors.

Recall that PC [13], MaxEnt [13] and DTRG [2] are introduced to validate that proper regularization is useful for dealing with the fine-grained problem. We, however, observe that the three techniques only yield slight improvements on the real-world dataset. In fact, to properly tackle the long-tailed difficulty, DTRG [2] has adopted the DRW [34] schedule. Compared with DTRG, ABC-Norm achieves better performance without relying on additional schemes such as Mixup and DRW. (We have also reported in Table II the result of ABC-Norm using the data augmentation scheme from [44].) Overall, our ABC-Norm method provides a general and flexible approach to solving real-world classification tasks.

D. Model Analysis

We begin by evaluating the effect of regularization on the magnitude of the classifier weight $\mathbf{w}_i$ for each category i, as depicted in Figure 5. While the L^2-norm magnitude distribution $\|\mathbf{w}_i\|$ of the baseline method exhibits a long-tailed

TABLE III
HEAD-TO-HEAD COMPARISONS AMONG FOUR DIFFERENT REGULARIZATION APPROACHES, ABC-NORM, PC, MAXENT AND DTRG, ON THE STANDARD FGVC DATASETS CUB-200-2011 (CUB), STANFORD CARS (CAR), AND FGVC-AIRCRAFT (AIR).

Model	ResNet-50			ResNeXt-50			ResNeXt-101			DenseNet-161		
	CUB	CAR	AIR	CUB	CAR	AIR	CUB	CAR	AIR	CUB	CAR	AIR
Baseline	85.5	92.7	90.3	86.3	93.1	90.9	87.3	93.5	91.6	87.5	93.4	92.7
PC [13]	87.0	92.4	90.1	87.5	93.2	91.2	88.2	93.7	92.4	88.2	93.6	92.9
MaxEnt [14]	87.2	91.9	90.3	87.6	92.8	91.3	88.2	93.4	92.5	88.3	93.3	93.0
DTRG$^-$ [2]	**88.3**	**94.8**	93.0	-	-	-	-	-	-	89.0	**94.8**	**94.0**
ABC-Norm	87.8	94.3	**93.2**	**88.1**	**94.4**	**93.3**	**88.6**	**94.5**	**93.5**	**89.2**	**94.8**	93.5

The notation $^-$ indicates the results by DTRG without using Mixup, as reported in the original paper [2].

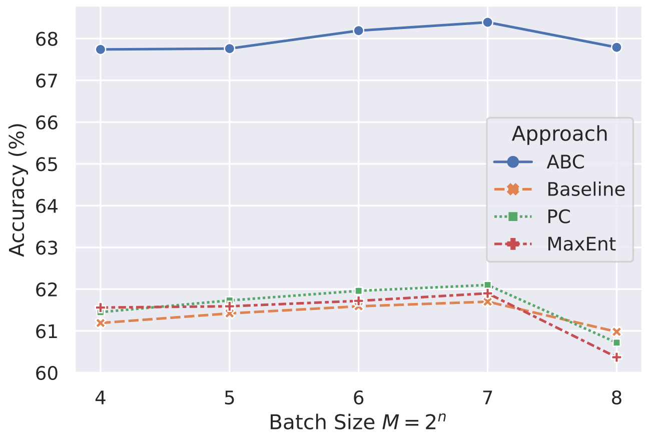


Fig. 6. Accuracy versus batch size on iNaturalist2018 for the three regularization methods: ABC-Norm, PC and MaxEnt.

pattern, the proposed ABC-Norm instead produces a smoother magnitude distribution for the head categories, reducing their dominance. In comparison, PC also lessens the dominance of head categories, but the distribution remains largely unchanged, indicating the persistence of the long-tailed issue.

Next, we conduct an ablation study on the iNaturalist2018 dataset to assess the impact of different batch sizes on the various regularization approaches, including ABC-Norm, PC, and MaxEnt. Figure 6 shows that the performance variations among different batch sizes are similar across all regularization methods as well as the baseline. This suggests that the influence of batch size stems from the use of “batch normalization” and the correlation between the performance of ABC-Norm and batch size is weak. Hence, choosing a specific batch size for ABC-Norm is generally not an issue of concern.

The long-tailed issue often requires the selection of an appropriate value of hyper-parameter τ to incorporate the statistical information embodied in the training data. Figure 7 shows the results of varying the τ value, where the resulting curve exhibits a downward parabolic trend from $\tau = 0.1$ to $\tau = 1.0$, with the best performance achieved at $\tau = 0.5$.

We also investigate the optimal regularization weight λ between cross-entropy loss and ABC-Norm in (5). Figure 8 displays the probing result of such search. The classification performance gradually improves as λ increases, reaching a sweet spot at $\lambda = 0.5$. Beyond this point, further increasing the λ value leads to a decrease in performance, suggesting that the suitable range of λ is $[0.1, 0.5]$.

TABLE IV
ACCURACY VERSUS BATCH SIZE ON THE CUB DATASET.

Batch Size	2	4	8	16	32
Baseline	80.9	84.5	85.4	85.5	85.5
PC [13]	81.1	85.8	86.7	87.0	86.9
MaxEnt [14]	**81.4**	85.9	86.9	87.2	87.0
ABC-Norm	81.2	**86.1**	**87.5**	**87.8**	**87.7**

We conclude our analysis by providing a qualitative comparison of the baseline, PC, and ABC-Norm methods using the class activation mapping (Grad-CAM) [24] on the CUB dataset. As shown in Figure 9, the results reveal that PC and ABC-Norm correctly predict more samples than the baseline. The redder an area is, the more significant the model’s prediction is, while the bluer the area indicates the opposite. For instance, PC and ABC-Norm focus on the appropriate regions to identify the object rather than the background. Moreover, as in the right panel of Figure 9, ABC-Norm can correctly classify even the challenging samples that the PC and baseline fail to recognize. This is because that ABC-Norm further exploits the inter-class similarity information to ensure the resulting classifier to focus on the most discriminative parts.

E. More on Fine-grained

To investigate the compatibility of the proposed ABC-Norm on the FGVC datasets, we conduct experiments with different backbones against PC [13], MaxEnt [14] and DTRG [2], respectively. The backbones are chosen from shallow to deep, including ResNet-50, ResNeXt-50, ResNeXt101 and DenseNet-161. We re-implement PC and MaxEnt with the same training condition. Table III shows the head-to-head comparison; the experimental results imply that ABC-Norm outperforms both PC and MaxEnt. Compared to DTRG, although the performances are about even, the training process of ABC-Norm is simple and essentially the same as the baseline case. Moreover, we also conduct an ablation study to gauge the batch-size influence. Table IV shows that the batch-size influence is still similar to that in Figure 6. It suggests that we only need to pay more attention to the hyper-parameters λ and τ. Among the confusion-based techniques, ABC-Norm is not only more

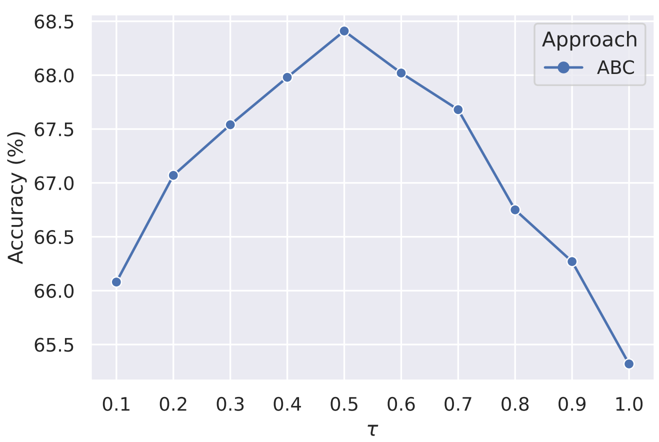


Fig. 7. An ablation study about various τ values for the adaptive matrix A on iNaturalist2018. The sweet point for τ on the real-world dataset, iNaturalist2018, locates at 0.5.

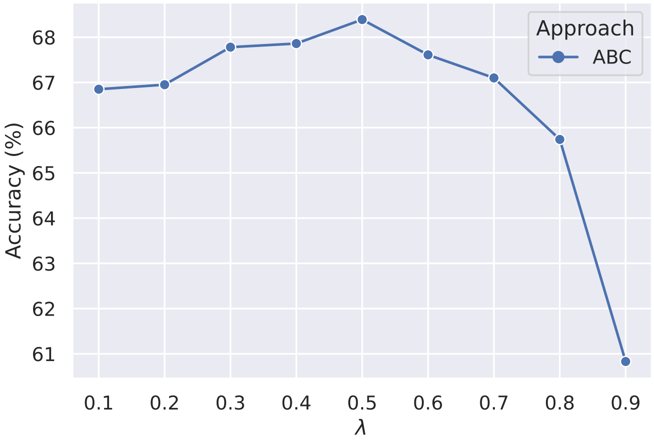


Fig. 8. An ablation study of the regularization weight λ on iNaturalist2018. We observe that the suitable value for λ locates in the range $[0.1, 0.6]$.

TABLE V
FOLLOWING [11], THE EXPERIMENT ON IMAGENET-LT IS CARRIED OUT WITH RESNEXT-50.

Method	Many	Median	Few	Total
Baseline	**65.9**	37.5	7.7	44.4
with LWS	60.2	47.2	30.3	49.9
PC	63.9	35.5	8.8	42.8
PC + LWS	57.3	46.4	29.8	48.4
MaxEnt	63.4	35.9	8.6	42.5
MaxEnt + LWS	59.3	46.1	29.5	48.9
ABC-Norm	65.5	43.1	10.9	47.5
ABC-Norm + LWS	60.7	**49.7**	**33.1**	**51.7**

TABLE VI
FOLLOWING [11], THE EXPERIMENT ON IMAGENET-LT IS CARRIED OUT WITH RESNEXT-152.

Method	Many	Median	Few	Total
Baseline	**69.1**	41.4	10.4	47.8
with LWS	63.5	50.4	34.2	53.3
PC	66.9	37.3	10.2	45.1
PC + LWS	59.5	48.7	32.6	50.6
MaxEnt	66.1	37.8	10.1	44.8
MaxEnt + LWS	61.9	47.8	30.8	50.9
ABC-Norm	68.7	45.5	12.3	49.9
ABC-Norm + LWS	63.6	**51.8**	**35.5**	**54.2**

effective in fine-grained classification than PC and MaxEnt but also valid when dealing with the long-tailed issue.

F. More on Long-tailed

Since ImageNet-LT has a low fine-grained factor but poses strong long-tailed difficulty, we perform experiments on it to confirm the effectiveness of ABC-Norm for purely long-tailed learning. Following the training formulation in [11], we decompose the training process into two stages, representation learning and classifier learning. In stage one for representation learning, the data sampling strategy is instance-balanced which can also be called a baseline. Next, the sampling strategy turns class-balanced to fine-tune the classifier at stage two. Table V first shows the results based on ResNeXt-50. At stage one, with end-to-end training, the baseline, PC, and MaxEnt are prone to overfit the head categories since these methods do not take account of the imbalanced distribution of the training set. On the contrary, the results show that the ABC-Norm significantly improves and alleviates the domination problem of head categories. Furthermore, through stage two, fine-tuning the classifier with LWS can improve the representation model learned from ABC-Norm. In conclusion, tackling the long-tailed distribution with our method can learn a better representation model than PC and MaxEnt.

To further verify the robustness of the proposed ABC-Norm regularization, we re-evaluate the experiment with the same setting but using a deeper backbone, ResNeXt-152. The experimental results are presented in Table VI. We see that no matter how deep or shallow the model is, ABC-Norm still achieves consistent improvements, which again confirms the compatibility of the ABC-Norm approach.

G. Additional Results

Finally, to demonstrate that it is no coincidence that ABC-Norm improves via sufficient training, we also explore the experiment on both a deep network and longer training epochs. Following the training procedure from previous work [3], [11], we apply the ABC-Norm to the ResNet-152 backbone and report the experimental results trained with 90 and 200 epochs on the real-world dataset, iNaturalist2018. This additional experiment is designed to justify the robustness of our method and address the concern that the ABC-Norm is effective only for a specific setting. The experimental results are shown in Table VII. For the same data augmentation

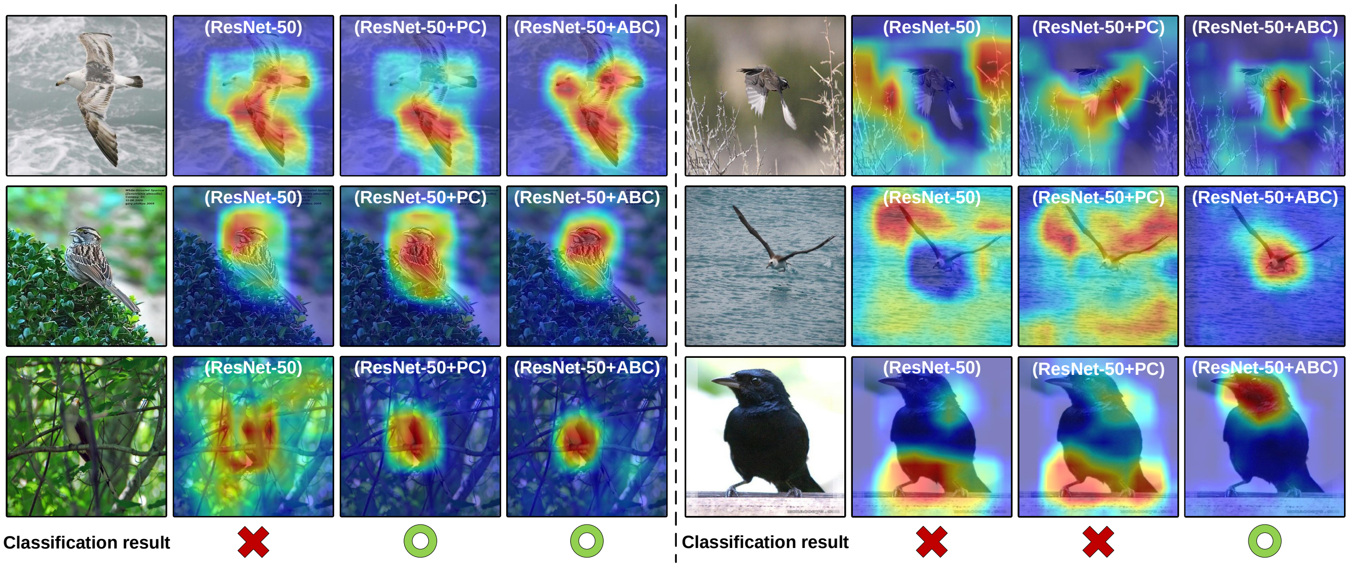


Fig. 9. Grad-CAM heatmap visualization for six testing images. In each example, the resulting heatmap is specified by the corresponding model (ResNet-50, ResNet-50+PC, ResNet-50+ABC). Results of correct classification are marked with " ✖ " and otherwise, " ◎ ".

TABLE VII
iNATURALIST2018 CLASSIFICATION BY DIFFERENT METHODS WITH RESNET-152, BASED ON 90 AND 200 TRAINING EPOCHS.

Setting	90 epochs	200 epochs
Baseline	65.0	69.0
Re-weighted	68.1	69.9
PC† [13]	66.9	69.3
MaxEnt† [14]	66.6	69.2
LWS [11]	69.1	72.1
LA [3]	68.9	69.9
ABC-Norm	71.7	72.6
ABC-Norm‡	**73.8**	**74.0**

† Re-implement with the same setting as ours.
‡ Follow the data augmentation scheme in [44].

and the complete model, the results are consistent with those of Table II. Meanwhile, the other two regularization-based approaches, PC and MaxEnt, still do not perform well in this experiment, which includes both fine-grained and long-tailed difficulties in the underlying real-world dataset. With all our extensive experimental results, we demonstrate that the proposed approach, ABC-Norm, is generic and not specific.

V. CONCLUSIONS

We introduce Adaptive Batch Confusion Norm (ABC-Norm), a general regularization technique to tackle the challenging problem of fine-grained and long-tailed image classification. Our method is simple in design, consisting of only the cross-entropy and the ABC-Norm regularization terms. During training, the ABC-Norm regularization adaptively generates confusion for each object category and activates an adversarial-like learning mechanism, leading to improved learning efficiency and more discriminative features within regions of interest. Through experiments, we show that ABC-Norm outperforms other relevant (adversarial) regularization-based approaches, such as PC and MaxEnt, and effectively reduces overfitting in training. In future work, we plan to generalize this regularization concept to transformer-based networks and enhance its effectiveness with attention mechanisms.

ACKNOWLEDGMENTS

This work was supported in part by the NSTC grants 111-2221-E-001-015-MY3 and 111-2634-F-007-010 of Taiwan. We are grateful to National Center for High-performance Computing for providing computational resources and facilities.

REFERENCES

[1] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 8769–8778.

[2] K. Liu, K. Chen, and K. Jia, "Convolutional fine-grained classification with self-supervised target relation regularization," *IEEE Trans. Image Process. (TIP)*, pp. 5570–5584, 2022.

[3] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," in *Int. Conf. Learn. Represent. (ICLR)*, 2021.

[4] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 4438–4446.

[5] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 420–435.

[6] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 5012–5021.

[7] W. Ge, X. Lin, and Y. Yu, "Weakly supervised complementary parts models for fine-grained image classification from the bottom up," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 3034–3043.

[8] R. Du, D. Chang, A. K. Bhunia, J. Xie, Z. Ma, Y. Song, and J. Guo, "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 153–168.

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

[10] Y.-X. Wang, D. Ramanan, and M. Hebert, "Learning to model the tail," in *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2017, pp. 7029–7039.

[11] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *Int. Conf. Learn. Represent. (ICLR)*, 2020.

[12] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," in *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2004, pp. 41–48.

[13] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik, "Pairwise confusion for fine-grained visual classification," in *Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 70–86.

[14] A. Dubey, O. Gupta, R. Raskar, and N. Naik, "Maximum-entropy fine grained classification," in *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2018, pp. 635–645.

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

[16] C. Tian, W. Wang, X. Zhu, J. Dai, and Y. Qiao, "VL-LTR: learning class-wise visual-linguistic representation for long-tailed visual recognition," in *Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 73–91.

[17] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[18] T. Berg, J. Liu, S. Woo Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2014, pp. 2011–2018.

[19] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked cnn for fine-grained visual categorization," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 1173–1182.

[20] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a cnn for fine-grained recognition," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 4148–4157.

[21] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 5157–5166.

[22] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, M. Wu, J. Guo, and Y. Song, "The devil is in the channels: Mutual-channel loss for fine-grained image classification," *IEEE Trans. Image Process. (TIP)*, pp. 4683–4695, 2020.

[23] Y. Shu, B. Yu, H. Xu, and L. Liu, "Improving fine-grained visual recognition in low data regimes via self-boosting attention mechanism," in *Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 449–465.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

[25] D. Chang, K. Pang, Y. Zheng, Z. Ma, Y. Song, and J. Guo, "Your "flamingo" is my "bird": Fine-grained, or not," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 11 476–11 485.

[26] C. Drummond, R. C. Holte *et al.*, "C4. 5, Class imbalance, and cost sensitivity: why under-sampling beats over-sampling," in *Workshop on learning from imbalanced datasets II*, 2003, pp. 1–8.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, pp. 321–357, 2002.

[28] H. Han, W. Wang, and B. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," in *ICIC*, D. Huang, X. S. Zhang, and G. Huang, Eds., 2005, pp. 878–887.

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

[30] L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 467–482.

[31] A. Gupta, P. Dollar, and R. Girshick, "LVIS: A dataset for large vocabulary instance segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 5356–5364.

[32] A. Shrivastava, A. Gupta, and R. B. Girshick, "Training region-based object detectors with online hard example mining," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 761–769.

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

[34] K. Cao, C. Wei, A. Gaidon, N. Aréchiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2019, pp. 1565–1576.

[35] Y. Cui, M. Jia, T. Lin, Y. Song, and S. J. Belongie, "Class-balanced loss based on effective number of samples," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 9268–9277.

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

[37] J. Tan, C. Wang, B. Li, Q. Li, W. Ouyang, C. Yin, and J. Yan, "Equalization loss for long-tailed object recognition," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 11 659–11 668.

[38] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *Int. Conf. Learn. Represent. (ICLR)*, 2020.

[39] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 9719–9728.

[40] Y. Li, T. Wang, B. Kang, S. Tang, C. Wang, J. Li, and J. Feng, "Overcoming classifier imbalance for long-tail object detection with balanced group softmax," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 10 988–10 997.

[41] X. Hu, Y. Jiang, K. Tang, J. Chen, C. Miao, and H. Zhang, "Learning to segment the tail," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 14 042–14 051.

[42] T. Wang, Y. Li, B. Kang, J. Li, J. H. Liew, S. Tang, S. C. H. Hoi, and J. Feng, "The devil is in classification: A simple framework for long-tail instance segmentation," *CoRR*, 2020.

[43] K. Tang, J. Huang, and H. Zhang, "Long-tailed classification by keeping the good and removing the bad momentum causal effect," in *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2020.

[44] H. Wang, S. Fu, X. He, H. Fang, Z. Liu, and H. Hu, "Towards calibrated hyper-sphere representation via distribution overlap coefficient for long-tailed learning," in *Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 179–196.

[45] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, pp. 471–501, 2010.

[46] D. Samuel, Y. Atzmon, and G. Chechik, "From generalized zero-shot learning to long-tail with class descriptors," in *IEEE Winter Conf. App. Comput. Vis. (WAVC)*, 2021, pp. 286–295.

[47] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2019, pp. 1567–1578.

[48] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D Object representations for fine-grained categorization," in *Int. Conf. Comput. Vis. Worksh. (ICCVW)*, 2013, pp. 554–561.

[49] S. Maji, E. Rahtu, J. Kannala, M. B. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *CoRR*, vol. abs/1306.5151, 2013.

[50] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," *CoRR*, 2019.

[51] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Adv. Neural Inform. Process. Syst. Worksh. (NeurIPSW)*, 2017.

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

[53] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 4700–4708.

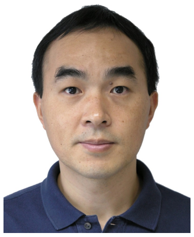
[54] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," in *Int. Conf. Learn. Represent. (ICLR)*, 2017.

Yen-Chi Hsu received the B.S. degree from National Chen Kung University (Tainan, Taiwan) in 2013 and the M.S. degree from National Chen Kung University in 2016, both in mathematics. And now he is a Ph.D. student in computer science from National Taiwan University. His research interests include computer vision, artificial intelligence, and machine learning.

Cheng-Yao Hong received the B.S. degree from National Chiao Tung University, Hsinchu, Taiwan, in 2012, and the M.S. degree from National Taiwan University, Taipei, Taiwan, in 2017, all in electrical engineering. His research interests include computer vision, artificial intelligence, and machine learning.

Ming-Sui Lee is an Associate Professor in the Graduate Institute of Networking and Multimedia and the Department of Computer Science and Information Engineering, National Taiwan University. She received Ph.D. degree from University of Southern California, and M.S. degree from University of California, Los Angeles, both majoring in Electrical Engineering. Her research expertise focuses on image/video processing, computer vision, deep learning, and medical image processing.

Davi Geiger received the BS degree in physics from Pontificia Universidade Católicia do Rio de Janeiro (PUC-Rio), Brazil, in 1980, and the PhD degree in physics from Massachusetts Institute of Technology (MIT), Cambridge, in 1990, having developed his thesis at the Artificial Intelligence Laboratory at MIT. He is now an associate professor of computer science and neural science at New York University. He received the Career Award from the US National Science Foundation in 1998. The central theme of his research is the development of a theory of how the brain works and, more generally, how information is processed. The domain of interests include computer vision, learning theory, memory and its applications.



Tyng-Luh Liu received the Ph.D. degree in computer science from New York University in 1997. He is currently a Research Fellow at the Institute of Information Science, Academia Sinica, Taiwan. From 2019 to 2022, he was a Chief Scientist of the medical imaging team at Taiwan AI Labs. He received the Academia Sinica Early-Career Investigator Research Achievement Award in 2006. His research interests include computer vision, artificial intelligence, machine learning, and medical imaging.