Confusion-Aware Spectral Regularizer for Long-Tailed Recognition

Ziquan Zhu[1,*] Gaojie Jin[2,*] Hanruo Zhu[1,*] Si-Yuan Lu[3,*] Yunxiao Zhang[2]
Zeyu Fu[2] Ronghui Mu[2] Guoqiang Zhang[2] Zhao Sun[4] Yuhang Xia[5]
Jiaxing Shang[2,6] Xiang Li[7] Lu Liu[2] Tianjin Huang[2,8,†]

[1]University of Leicester [2]University of Exeter [3]Nanjing University of Posts and Telecommunications
[4]Zhengzhou University [5]Chengdu University of Technology [6]Chongqing University
[7]University of Bristol [8]Eindhoven University of Technology
[*]Equal contribution [†]Corresponding author: T.Huang2@exeter.ac.uk.

Abstract

*Long-tailed image classification remains a long-standing challenge, as real-world data typically follow highly imbalanced distributions where a few head classes dominate and many tail classes contain only limited samples. This imbalance biases feature learning toward head categories and leads to significant degradation on rare classes. Although recent studies have proposed re-sampling, reweighting, and decoupled learning strategies, the improvement on the most underrepresented classes still remains marginal compared with overall accuracy. In this work, we present a confusion-centric perspective for long-tailed recognition that explicitly focuses on worst-class generalization. We first establish a new theoretical framework of class-specific error analysis, which shows that the worst-class error can be tightly upper-bounded by the spectral norm of the frequency-weighted confusion matrix and a model-dependent complexity term. Guided by this insight, we propose the **C**onfusion-**A**ware Spectral **R**egularizer (`CAR`) that minimizes the spectral norm of the confusion matrix during training to reduce inter-class confusion and enhance tail-class generalization. To enable stable and efficient optimization, `CAR` integrates a Differentiable Confusion Matrix Surrogate and an EMA-based Confusion Estimator to maintain smooth and low-variance estimates across mini-batches. Extensive experiments across multiple long-tailed benchmarks demonstrates that `CAR` substantially improves both worst-class accuracy and overall performance. When combined with ConCutMix augmentation, `CAR` consistently surpasses exisiting state-of-the-art long-tailed learning methods under both the training-from-scratch setting (by $2.37\% \sim 4.83\%$) and the fine-tuning-from-pretrained setting (by $2.42\% \sim 4.17\%$) across ImageNet-LT, CIFAR100-LT, and iNaturalist datasets. Code is available at* `https://github.com/misswayguy/CAR`.

1. Introduction

Image classification has achieved remarkable progress in recent years, mainly due to the success of deep learning and large-scale balanced datasets [44]. However, data collected from real-world scenarios rarely follow such balanced distributions. Instead, they usually exhibit a long-tailed distribution, where a few head classes contain abundant samples while most tail classes have only limited instances [60, 62]. This severe imbalance causes biased feature learning—models tend to focus on head classes while failing to capture meaningful representations for tail ones [14, 21]. This skewed distribution induces optimization bias toward head classes and degrades performance on tail categories, making long-tailed recognition a central challenge in practice. To mitigate the negative effects of long-tailed distributions, numerous approaches have been explored in recent years. Data-level strategies such as resampling [4, 25, 53] aim to balance the number of samples per class by over- or under-sampling the training data. Loss-level adjustments, including re-weighting [8, 33, 45] and logit adjustment [39, 51, 57], attempt to compensate for class imbalance by assigning adaptive weights or biases to the loss function. At the model level, representation enhancement and decoupled learning methods [9, 48] focus on separating feature learning from classifier optimization to reduce the dominance of head classes.

Despite steady progress in long-tailed learning, performance on the worst-performing classes remains substantially inferior to overall accuracy. As shown in Figure 1, we observe two critical gaps: ❶ worst-class test accuracy lags significantly behind overall test accuracy, and ❷ worst-class test accuracy falls substantially short of worst-class training accuracy. These disparities reveal a fundamental limitation that existing approaches fail to effectively generalize to the most challenging tail categories, even when they fit the training data well.

To bridge this gap, we propose a new confusion-centric perspective that explicitly regularizes spectral norm of the weighted confusion matrix and focuses on improving worst-class performance. Specifically, we introduce a weighted worst-class error metric that integrates frequency priors to amplify the influence of minority classes. We further develop a generalization upper bound for the class-specific error based on the PAC-Bayesian framework [38, 40, 41]. This generalization bound reveals that the worst-class error can be bounded by two key components: (i) the spectral norm of the weighted empirical confusion matrix, and (ii) a model- and data-dependent complexity term. Building on this theoretical insight, we propose a confusion-aware spectral regularizer (`CAR`) that directly minimizes the spectral norm of the frequency-weighted confusion matrix during training. To enable efficient optimization, we introduce a differentiable confusion matrix surrogate combined with an exponential moving average (EMA) mechanism that maintains stable and efficient estimates across mini-batches.

Our efforts are unfolded with the following four thrusts:

- ⋆ **(Theoretical Analysis)** We establish a novel confusion-centric perspective for long-tailed recognition. Building on the PAC-Bayesian theory, we derive a new upper bound showing that the *worst-class error* can be tightly controlled by the spectral norm of the frequency-weighted confusion matrix, offering a principled route toward improving worst-class generalization.
- ⋆ **(Algorithm)** Guided by the theoretical insights, we propose the practical Confusion-Aware Spectral Regularizer (`CAR`). `CAR` introduces two key components: **(i)** a *Differentiable Confusion Matrix Surrogate* that replaces non-differentiable indicators with smooth approximations, and **(ii)** an *EMA-based Confusion Estimator* that stabilizes optimization by maintaining low-variance estimates of confusion matrix.
- ⋆ **(Experiments)** We conduct extensive experiments across diverse long-tailed benchmarks, architectures, and imbalance factors. Results consistently demonstrate that `CAR` achieves superior head–tail balance, and improves both *worst-class* accuracy and *overall* performance. For example, when combined with ConCutMix augmentation, `CAR` surpasses the previous state-of-the-art LOS [49] by $2.37\% \sim 4.83\%$ under the training-from-scratch setting across CIFAR100-LT, ImageNet-LT, and iNaturalist.
- ⋆ **(Extra Findings)** We further observe that `CAR` can be seamlessly integrated with existing data-augmentation-based long-tailed learning methods, leading to additional performance gains on both head and tail classes.

2. Preliminaries

Notation. Let $\mathcal{X} \subset \mathbb{R}^d$ denote the input space and $\mathcal{Y} = \{1, \dots, K\}$ the label set. A training set $\mathcal{S} = \{(x_q, y_q)\}_{q=1}^m$ is drawn from the imbalanced distribution $\mathcal{D}$. m is the number of training samples and K is the number of classes. A classifier $f : \mathcal{X} \to \mathbb{R}^K$ outputs predicted probability vector $f(x)$, and the prediction is $\hat{y}(x) = \arg\max_i f(x)[i]$. For a matrix $\mathbf{A} = (a_{ij})$, we use $\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$ and $\|\mathbf{A}\|_2$ for the spectral norm.

Confusion Matrice. The population off-diagonal confusion matrix $\mathbf{C}_{\mathcal{D}}^f \in \mathbb{R}^{K\times K}$ is defined as

$$c_{ij} = \mathbb{P}_{(x,y)\sim\mathcal{D}}(\hat{y}(x) = i \mid y = j)\,, \qquad c_{jj} = 0. \tag{1}$$

The column sum $\sum_i c_{ij}$ corresponds to the class-conditional error for class j. Its empirical analogue $\mathbf{C}_S^f$ replaces probabilities with sample averages:

$$\hat{c}_{ij} = \frac{1}{m_j} \sum_{q:y_q=j} \mathbf{1}(\hat{y}(x_q) = i)\,, \qquad \hat{c}_{jj} = 0, \tag{2}$$

where m_j is the number of samples with label j in the training set $\mathcal{S}$. In previous works such as [13, 41], the PAC-Bayesian generalization analysis for a DNN is conducted on the margin loss. Following the margin setting, we consider any positive margin γ in this work and define the empirical margin confusion matrix $\mathbf{C}_{S,\gamma}^f = (\hat{c}_{ij}^\gamma)$ as:

$$\hat{c}_{ij}^\gamma = \begin{cases} 0, & i = j, \\ \displaystyle\sum_{q:y_q=j} \frac{1}{m_j} \mathbf{1}\big(f_w(\mathbf{x}_q)[y_q] \le \gamma + f_w(\mathbf{x}_q)[i]\big) & \\ \qquad\qquad \times\, \mathbf{1}\big(\arg\max_{i'\neq y_q} f_w(\mathbf{x}_q)[i'] = i\big), & \text{else.} \end{cases}$$

where $\mathbf{1}[a \le b] = 1$ if $a < b$, else $\mathbf{1}[a \le b] = 0$.

3. Weighted Worst-Class Error

Rationale. Long-tail datasets exhibit severe imbalance. Frequent classes dominate risk minimization while rare classes are poorly controlled. To counteract this, we introduce a frequency-dependent weighting that amplifies the influence of minority classes in the confusion-based analysis. We define the class-wise weight as $\lambda_j = (m_j + r_0)^{-1/2}$, where $r_0 > 0$ is a smoothing factor and m_j denotes the relative frequency of class j, i.e., the ratio between the number of samples in class j and the total number of samples in the dataset. The diagonal weighting matrix is then given by $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_K)$.

Definition 3.1 (weighted worst-class error)**.**

$$\texttt{WCE}(f) = \left\|\mathbf{C}_{\mathcal{D}}^f \mathbf{\Lambda}\right\|_1 = \max_j \; \lambda_j \sum_i c_{ij}. \tag{3}$$

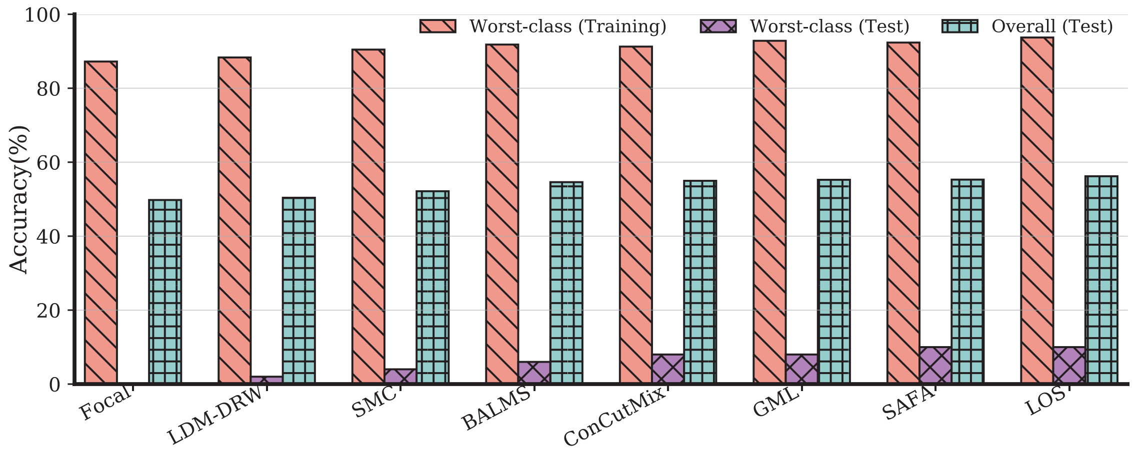

Figure 1. Poor generalization of worst-class performance in existing long-tailed learning methods. Experiments are conducted on ImageNet-LT using ViT-Small as the backbone. The three bars for each method correspond to the worst-class accuracy on the training set (left), the worst-class accuracy on the test set (middle), and the overall test accuracy (right).

This criterion emphasizes classes with fewer samples by scaling their conditional errors with λ_j.

Proposition 3.2 (**Upper Bound**). *Consider a training set $\mathcal{S}$ with m samples drawn from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Let B represent the largest ℓ_2 norm of input samples. For any $B, n, h > 0$, let the base classifier $f_w : \mathcal{X} \to \mathcal{Y}$ be an n-layer feedforward network with h units per layer and ReLU activation. Then, for any $\delta \in (0,1)$ and margin $\gamma > 0$, with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$, we have upper bound for class-specific error e_j:*

$$\begin{aligned} e_j &\leq \frac{1}{\lambda_j} \left\| \mathbf{C}^f_{\mathcal{D}} \boldsymbol{\Lambda} \right\|_1 \\ &\leq \frac{\nu}{\lambda_j} \underbrace{\left\| \mathbf{C}^f_{\mathcal{S},\gamma} \boldsymbol{\Lambda} \right\|_2}_{\textit{Empirical spectral norm}} + \underbrace{\mathcal{E}(f, \mathcal{S}, \gamma, \delta)}_{\substack{\textit{Model and training} \\ \textit{set dependence}}}, \forall j, \end{aligned} \tag{4}$$

where ν is a positve constant which depend on K. $\mathcal{E}(f, \mathcal{S}, \gamma, \delta)$ is

$$\mathcal{O}\left(\sqrt{\frac{K}{(m_{min} - 8K)\gamma^2} \left[\Psi(f_w) + \ln\left(\frac{n m_{min}}{\delta} \right) \right]} \right), \tag{5}$$

where K is the number of classes, m_{min} represents the minimal number of examples from $\mathcal{S}$ which belong to the same class, $\Psi(f_w) = B^2 n^2 h \ln(nh) \prod_{l=1}^n \|W_l\|_2^2 \sum_{l=1}^n \frac{\|W_l\|_F^2}{\|W_l\|_2^2}$, W_l represents the l-th weight matrix.

Proof. See Appendix 10. □

Remark 1. *The above result demonstrates that **(i)** the class-specific error is upper-bounded by the weighted worst-class error, **(ii)** this upper bound is tighter for classes with higher errors and looser for those with lower errors, and is exactly tight for the worst-class error; and **(iii)** the weighted worst-class error itself can be further bounded by two terms: ❶ the spectral norm of the weighted empirical confusion matrix, and ❷ a model- and data-dependent complexity term $\Psi(f_w)$. While the second term has been extensively studied through techniques such as spectral normalization of weight matrices [13, 58], our contribution is to highlight the role of the spectral norm of the confusion-matrix. Controlling this spectral norm offers a complementary mechanism for improving worst-class generalization, particularly in imbalanced settings.*

4. CAR:Confusion-Aware Spectral Regularizer

The above analysis indicates that reducing the upper bound of the weighted worst-class error effectively decreases the error across all classes, particularly for the worst-performing class. Motivated by this, we introduce the empirical spectral norm term as a regularizer, while omitting the leading constant factor $\frac{\nu}{\lambda_j}$ for simplicity. Specifically, it is formulated as follows:

$$\mathcal{R}(f) = \|\mathbf{C}^f_{\mathcal{S},\gamma} \boldsymbol{\Lambda}\|_2 \tag{6}$$

However, optimizing the model with respect to this regularizer presents two key challenges: (1) the empirical margin confusion matrix is non-differentiable; and (2) computing it over the entire training set $\mathcal{S}$ in real time is computationally infeasible. To overcome these challenges, we introduce a `Differentiable Confusion Matrix Surrogate` and an `EMA-based Confusion Estimator`.

Differentiable Confusion Matrix Surrogate. The empirical confusion matrix $\mathbf{C}^f_{\mathcal{S},\gamma}$ involves non-differentiable indicator functions, making it unsuitable for gradient-based optimization. To address this, we introduce a differentiable surrogate formulation $\tilde{\mathbf{C}}^f_{\mathcal{S},\gamma} = (\tilde{c}_{ij})$:

$$\begin{aligned} \tilde{c}_{ij} = \frac{1}{m_j} \sum_{q:y_q=j} & \underbrace{\sigma(\gamma + f_w(x_q)[i] - f_w(x_q)[j])}_{\textbf{soft margin gate}} \\ & \times \underbrace{\mathbf{S}(f_w(x_q) - f_w(x_q)[j])[i]}_{\textbf{soft argmax over non-j}} \end{aligned} \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\mathbf{S}(\cdot)$ the softmax function. The **soft margin gate** term softly evaluates whether class i surpasses the ground-truth class j by a margin γ, while the **soft argmax over non-*j*** term provides a differentiable approximation of the most competitive class other than the true one.

EMA-based Confusion Estimator. Considering that training models is conducted with batchsize, a directly way to avoid the infeasible trainset level confusion matrix calcuation is to calcuate batchsize level confusion matrix. However, a single mini-batch may yields a high-variance estimate of the confusion matrix. To stablilize training without recomputing overal full trainset, we maintain an *exponential moving average* of the *batch-level differentiable estimate*. Specficially, let $\mathcal{B}_t$ denote the mini-batch at iteration t and $\beta \in [0, 1)$ be a momentum parameter , the EMA of the batch-level confusion matrix is defined as:

$$\hat{\mathbf{C}}_{\mathbf{t}} \;=\; \beta\,(\hat{\mathbf{C}}_{t-1}) \;+\; (1-\beta)\,\tilde{\mathbf{C}}^{f}_{\mathcal{B}_t,\gamma}, \qquad \hat{\mathbf{C}}_{\mathbf{0}} = 0. \quad (8)$$

Only the current term $\tilde{\mathbf{C}}^{f}_{\mathcal{B}_t,\gamma}$ carries gradients w.r.t. w; the history $\hat{\mathbf{C}}_{t-1}$ is treated as a constant, preserving differentiability while reducing variance (See Appendix 11 for a detailed stability analysis.)

The overall training objective $\mathcal{L}(f)$ combines the standard cross-entropy loss with the proposed confusion-aware regularization term, where $\alpha > 0$ controls the strength of regularization.

$$\mathcal{L}(f) = \frac{1}{m}\sum_{q=1}^{m} \mathtt{CE}\big(f(x_q), y_q\big) + \alpha\, \|\big(\hat{\mathbf{C}}_{\mathbf{t}}\boldsymbol{\Lambda}\big)\|_2, \quad (9)$$

where $\mathtt{CE}(\cdot)$ denote the cross-entropy loss.

5. Experiments

5.1. Experimental Setup

Datasets. We conduct extensive experiments on four widely used long-tailed benchmarks: CIFAR100-LT [28], ImageNet-LT [34], Tiny-ImageNet-LT [29], and iNaturalist2018 [52]. Following [32], we adopt the same data construction and evaluation protocols. For CIFAR-100-LT, ImageNet-LT, and Tiny-ImageNet-LT, the imbalanced training sets are generated by sampling the original balanced datasets according to an exponential distribution, with imbalance factors of 200, 100, and 50, respectively. The test sets remain class-balanced. The iNaturalist2018 dataset contains 437.5K natural images from 8,142 categories, exhibiting a naturally long-tailed distribution without synthetic resampling. Consistent with [63], we report results on three subsets according to the number of training samples per class: Head (more than 100 images), Medium (20∼100 images), and Tail (less than 20 images). More details abour datasets would be discussed in the Appendix 12.1.1.

Implementation Details. We adopt ViT-Small [10] as the main backbone, while additional results on different model sizes (Tiny, Base, and Large) and architectures (ResNet [18] and Swin Transformer [35]) are also included to demonstrate model generalization.

For training from scratch, we follow the standard protocol used in long-tailed recognition [19]: models on CIFAR100-LT and ImageNet-LT are trained for 200 epochs, while iNaturalist2018 uses 300 epochs due to its greater intra-class variability and inherent natural imbalance. For fine-tuning from pre-trained models, we adopt a shorter schedule of 100 epochs, which is consistent with common practice for adapting large-scale pretrained backbones to long-tailed settings. All experiments use the AdamW optimizer. The batch size is fixed to 128 across all datasets. Additional training settings are provided in the Appendix 12.1.2.

Baselines. We compare our method with a comprehensive set of baselines, covering standard classification, long-tailed data augmentation, and representative long-tailed learning approaches. (1) *Standard Methods.* To provide a fair comparison, we first include widely used classification baselines: Cross-Entropy (CE). (2) *LT Data Augmentation Methods.* We further compare with a series of augmentation-based long-tailed strategies, including ReMix [6], MetaSAug [31], CMO [43], ConCutMix [42], and SAFA [20]. (3) *Other Long-Tailed Recognition Methods.* We evaluate against representative long-tailed learning methods, including Class-Balanced Loss (CB) [8], Focal Loss [33], BALMS [46], GML [11], Weight Balancing (WB) [1], BBN [65], LDAM-DRW [5], and LOS [49].

5.2. Superior Performance on Long-tailed Datasets

Training from Scratch. We evaluate the proposed `CAR` under a strict training-from-scratch setting. Experiments are conducted on three representative long-tailed benchmarks: CIFAR100-LT [28], ImageNet-LT [34], and iNaturalist2018 [52], using ViT-Small as the backbone. The results are summarized in Table 1. *We observe* that `CAR` consistently achieves the best overall performance across all three long-tailed benchmarks, outperforming all competing methods. *In addition*, `CAR` improves the best tail accuracy by 1.59% ∼ 4.61%, demonstrating its strong effectiveness in enhancing tail-class performence. *Furthermore*, when combined with the ConCutMix augmentation, `CAR` surpasses the current state-of-the-art LOS [49] by 2.37% ∼ 4.83% and 3.28% ∼ 7.98% in overall and tail accuracy, respectively, across the three datasets. For example, on ImageNet-LT, our method improves the previous best overall accuracy from 56.20% to 60.07%, and boosts the best tail accuracy from 32.73% to 38.07%.

Table 1. Top-1 accuracy (%) comparison on ImageNet-LT, CIFAR100-LT (IF=100), and iNaturalist using ViT-Small. Results are reported for Head, Medium, Tail, and Overall. The best results are in **bold**, and the second-best are underlined.

Methods	Venue	ImageNet-LT				CIFAR100-LT				iNaturalist			
		Head	Medium	Tail	Overall	Head	Medium	Tail	Overall	Head	Medium	Tail	Overall
CE	–	69.71	43.91	16.30	46.51	68.20	45.37	15.17	41.40	68.66	63.54	58.82	59.56
Focal [33]	ICCV'2017	67.57	49.80	22.83	49.78	66.43	47.49	21.87	43.13	67.74	68.55	64.96	66.86
CB [8]	CVPR'2019	69.69	47.69	19.90	48.05	67.63	45.77	16.37	42.10	67.69	66.66	62.39	64.71
LDAM-DRW [5]	NeurIPS'2019	68.06	47.14	25.90	50.39	68.97	46.40	25.07	45.40	67.15	68.30	63.64	65.57
BALMS [46]	NeurIPS'2020	69.80	53.14	30.23	54.61	68.80	58.57	28.03	50.74	70.48	70.19	69.53	70.55
ReMix [6]	ECCV'2020	69.49	46.94	25.67	50.05	69.06	46.97	20.10	43.34	68.14	66.49	62.55	64.27
BBN [65]	CVPR'2020	69.11	50.26	29.37	52.79	**69.21**	57.71	29.42	50.85	68.58	69.29	65.72	67.44
MetaSAug [31]	CVPR'2021	66.63	47.86	30.57	51.94	67.31	47.91	27.57	45.55	69.15	68.14	67.82	68.00
CMO [43]	CVPR'2022	67.40	50.20	28.97	52.15	68.34	49.37	28.53	46.01	68.57	70.31	69.60	69.41
SAFA [20]	ECCV'2022	67.20	52.03	32.03	55.29	67.29	54.94	29.47	48.02	68.86	72.03	70.42	70.86
WB [1]	CVPR'2022	70.46	49.80	31.90	53.71	68.86	49.94	29.60	47.81	70.10	69.74	67.95	69.35
GML [11]	CVPR'2023	69.03	53.66	32.17	55.24	68.63	55.20	30.97	50.23	**70.71**	70.62	69.38	70.85
ConCutMix [42]	TIP'2024	69.23	48.94	31.00	54.97	67.83	53.54	30.60	47.86	68.64	71.48	70.32	70.24
LOS [49]	ICLR'2025	**70.79**	55.31	32.73	56.20	**69.21**	57.71	29.42	50.85	68.50	71.43	72.14	71.01
CAR (Ours)	–	67.00	54.57	35.77	57.48	65.37	58.60	34.03	51.85	68.52	71.32	73.73	71.56
CAR (Ours) + ConCutMix	–	69.56	**55.94**	**38.07**	**60.07**	68.89	**60.14**	**37.40**	**55.68**	69.10	**72.05**	**75.42**	**73.38**

Table 2. Top-1 accuracy (%) comparison on Tiny-ImageNet-LT, CIFAR100-LT (IF=100), and iNaturalist using pre-trained ViT-Small. Results are reported for Head, Medium, Tail, and Overall. The best results are in **bold**, and the second-best are underlined.

Methods	Venue	Tiny-ImageNet-LT				CIFAR100-LT				iNaturalist			
		Head	Medium	Tail	Overall	Head	Medium	Tail	Overall	Head	Medium	Tail	Overall
CE	–	87.31	71.66	39.90	63.91	90.46	73.37	50.50	70.49	75.47	73.95	71.82	73.45
Focal [33]	ICCV'2017	88.57	74.11	44.23	67.31	91.69	77.66	54.63	73.66	75.15	78.21	78.93	79.59
CB [8]	CVPR'2019	85.60	75.31	43.58	66.99	91.83	76.17	51.57	72.27	75.36	75.84	75.93	77.08
LDAM-DRW [5]	NeurIPS'2019	85.91	75.29	43.90	65.28	91.31	75.66	52.87	72.40	75.87	77.84	77.86	79.05
BALMS [46]	NeurIPS'2020	89.54	76.69	46.77	70.11	**93.80**	78.37	56.93	77.44	**77.60**	80.24	82.56	82.34
ReMix [6]	ECCV'2020	85.49	75.51	43.73	65.19	92.09	73.63	53.37	72.56	74.57	77.59	76.51	78.22
BBN [65]	CVPR'2020	86.27	75.78	44.94	66.47	92.23	76.00	53.37	74.89	75.07	79.57	79.83	80.88
MetaSAug [31]	CVPR'2021	86.26	74.40	47.40	67.40	92.77	75.46	55.20	74.34	75.62	78.86	79.84	80.06
CMO [43]	CVPR'2022	85.94	75.97	45.70	67.38	91.57	76.14	57.07	75.37	75.45	79.12	81.38	81.52
SAFA [20]	ECCV'2022	86.09	73.20	49.97	68.99	92.57	77.23	59.50	77.88	76.39	80.09	83.07	82.77
WB [1]	CVPR'2022	88.97	74.87	45.82	67.51	93.54	77.63	53.23	75.78	76.04	80.09	79.93	81.33
GML [11]	CVPR'2023	**89.87**	75.93	48.87	70.43	92.09	79.11	57.10	77.40	76.95	80.69	83.10	82.79
ConCutMix [42]	TIP'2024	88.23	76.54	46.27	69.15	92.40	76.83	58.77	76.46	75.86	79.55	83.32	82.06
LOS [49]	ICLR'2025	88.36	76.83	49.86	71.67	93.02	80.06	58.14	78.50	76.43	80.58	83.18	83.02
CAR (Ours)	–	87.60	75.91	52.47	72.97	92.89	80.46	61.17	79.37	75.38	81.71	84.26	83.74
CAR (Ours) + ConCutMix	–	88.71	**76.84**	**54.23**	**75.84**	93.26	**81.24**	**63.33**	**82.12**	76.32	**82.08**	**85.40**	**85.44**

Fine-tuning from Pre-trained Models. Our success extends beyond the training-from-scratch setting. We further evaluate CAR and all baselines under the fine-tuning setting, using a pre-trained ViT-Small as the backbone. Results are reported in Table 2 *Consistent observation can be made across all the three datasets*: ❶ CAR delivers substantial improvements in both tail accuracy and overall accuracy, highlighting its effectiveness in adapting pre-trained models to long-tailed distributions; ❷ when combined with ConCutMix, CAR achieves the strongest performance, boosting tail accuracy by $2.22\% \sim 5.19\%$ and overall accuracy by $2.42\% \sim 4.17\%$, significantly outperforming existing state-of-the-art methods. For example, on iNaturalist, the best overall accuracy improves from 83.02% to 85.44% and the best tail accuracy increases from 83.18% to 85.40%.

Worst-Class Performance. To evaluate the effectiveness of CAR in enhancing the generalization of the most underrepresented category, we report the worst-class accuracy on both the training and test sets, along with the Worst-class Ratio (WR = Test/Training), which quantifies the generalization ability of the worst-performing class. Experiments are conducted on ImageNet-LT and CIFAR100-LT using ViT-Small as the backbone. The results in Table 3 demonstrate that the proposed CAR substantially enhances worst-class generalization. ❶ Existing methods exhibit extremely poor worst-class accuracy, typically below 10%. ❷ In contrast, CAR improves worst-class accuracy by 8% on ImageNet-LT and 6% on CIFAR100-LT, while CAR +

ConCutMix further boosts the gains to over 10% on both datasets. ❸ Moreover, CAR yields a notable increase in the Worst-class Ratio (WR), indicating stronger generalization from training to test time.

Table 3. Worst-class accuracy on training/test sets and the Worst-class Ratio (WR = Test/Training) based on ViT-Small. All results are presented as percentages. The best results are highlighted in **bold**, and the second-best are <u>underlined</u>.

Methods	ImageNet-LT			CIFAR100-LT		
	Training(%)	Test(%)	WR	Training(%)	Test(%)	WR
Focal [33]	87.22	0	0.00	85.45	2	0.02
CB [8]	84.59	0	0.00	80.00	0	0.00
BALMS [46]	91.81	6	0.07	90.91	5	0.05
CMO [43]	90.45	4	0.04	86.00	6	0.07
SAFA [20]	92.35	10	0.11	91.09	8	0.09
GML [11]	92.83	8	0.09	90.91	8	0.09
ConCutMix [42]	91.26	8	0.09	87.82	8	0.09
LOS [49]	93.72	10	0.11	91.23	8	0.09
CAR (Ours)	<u>94.24</u>	<u>18</u>	<u>0.19</u>	<u>92.73</u>	<u>14</u>	<u>0.15</u>
CAR (Ours) + ConCutMix	**94.85**	**22**	**0.23**	**93.17**	**18**	**0.19**

5.3. Generalization Across Backbones

To verify the effectiveness of our method across different architectures, we further evaluate CAR on five representative backbones, including ViT-Tiny, ViT-Base, ViT-Large, ResNet [18] and Swin Transformer [35], as summarized in Table 4. CAR consistently achieves the highest top-1 accuracy across all evaluated backbones, demonstrating that its effectiveness is not tied to any specific architectural design. Notably, on larger transformer backbones, CAR yields consistent gains, including +1.0% on ViT-Large and +0.8% on Swin, suggesting that the method integrates effectively with self-attention architectures. The improvements observed on ResNet further confirm that the proposed confusion-aware spectral regularization also benefits traditional CNN architectures. Overall, the consistent performance gains across diverse model families verify that CAR is a backbone-agnostic regularization method that can be seamlessly incorporated into a wide range of architectures to deliver stable and transferable improvements. Additional results are reported in Appendix 12.3.

5.4. Generalization Across Imbalance Factors

We further examine the adaptability of CAR under different imbalance factors (IF = 50 and 200) on ImageNet-LT and CIFAR100-LT using ViT-Small as the backbone. As shown in Table 5, our method achieves the highest overall accuracy across all settings, demonstrating its strong ability to handle varying levels of class skewness. While the performance of existing methods degrades noticeably as the imbalance increases, CAR maintains a more stable trend, outperforming the second-best approach by clear margins under both moderate (IF=50) and extreme (IF=200) conditions. This consistent behavior indicates that the proposed method effectively mitigates head-class dominance and improves representation alignment for tail categories. Furthermore, the smaller performance gap between IF=50 and IF=200 shows that CAR preserves balanced learning dynamics even when minority classes are extremely rare. More extensive results and analyses across additional imbalance settings can be found in the Appendix 12.4.

Table 4. Top-1 accuracy (%) on ImageNet-LT across different backbones. Results include ViT variants (Tiny/Base/Large), ResNet, and Swin. The best results are highlighted in **bold**.

Methods	ViT-Tiny	ViT-Base	ViT-Large	ResNet	Swin
Focal [33]	37.98	54.44	60.66	42.31	50.92
CB [8]	35.09	52.65	59.14	40.06	48.79
BALMS [46]	45.71	62.55	67.92	48.73	55.17
ReMix [6]	37.13	54.10	62.10	41.83	49.08
MetaSAug [31]	40.80	56.70	64.22	44.47	52.52
CMO [43]	40.07	57.36	64.71	45.67	52.60
SAFA [20]	43.00	60.34	67.46	46.33	54.43
GML [11]	45.24	61.86	68.63	48.77	55.19
ConCutMix [42]	43.26	58.48	66.17	45.73	54.48
LOS [49]	45.66	62.54	68.26	49.54	55.62
CAR (Ours)	**46.35**	**63.79**	**69.26**	**50.27**	**56.38**

5.5. Visualization

To further illustrate the effect of our method on inter-class discrimination, we visualize the class-wise confusion matrices of different approaches on CIFAR100-LT using ViT-Small, as shown in Figure 2. Specifically, we randomly select *ten* categories spanning head, medium, and tail regions to provide a balanced view of class-wise interactions under long-tailed distributions. The left group corresponds to models trained from scratch, while the right group represents fine-tuning from a pre-trained model. Compared with WB and BALMS, Our CAR yields substantially fewer high-intensity off-diagonal responses on both training-from-scratch and fine-tuning-from-pre-trained settings, suggesting that it effectively suppresses inter-class confusion and enhances the separability between categories.

5.6. Complementary to Data Augmentation

We further examine whether CAR can complement existing long-tailed augmentation strategies as a general regularization module. To this end, we combine CAR with five representative approaches under different imbalance factors on ImageNet-LT and CIFAR100-LT. All models are trained under identical configurations to ensure a fair comparison across varying augmentation paradigms. As presented in Table 6, CAR consistently boosts the performance of all augmentation methods across datasets and imbalance levels. The gains are most pronounced when integrated with ConCutMix and SAFA, achieving the highest overall accuracies under every setting. These consistent improvements verify that the proposed spectral regularization yields com-

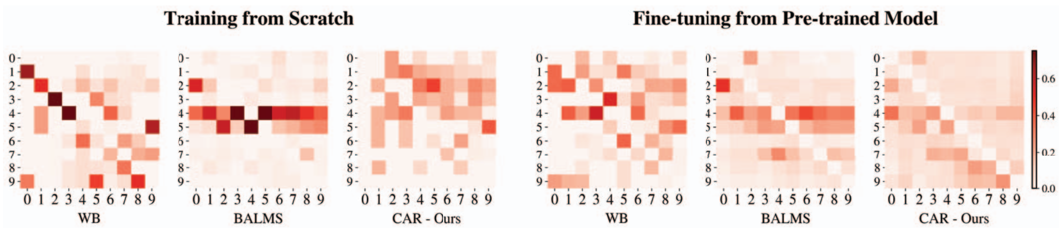

Figure 2. Class-wise confusion matrices on CIFAR100-LT using ViT-Small. Left: training from scratch. Right: fine-tuning from pre-trained model.

Table 5. Top-1 accuracy (%) across different imbalance factors (IF) on ImageNet-LT and CIFAR100-LT based on the ViT-Small. The best results are highlighted in **bold**.

Methods	ImageNet-LT		CIFAR100-LT	
	IF=50	IF=200	IF=50	IF=200
CE	49.33	35.99	44.54	37.19
Focal [33]	52.14	37.27	47.60	40.08
CB [8]	50.65	36.13	45.38	39.03
LDAM-DRW [5]	54.96	40.78	49.57	40.96
BALMS [46]	59.57	42.18	53.34	43.85
ReMix [6]	53.32	37.62	48.34	39.79
BBN [65]	55.26	41.89	51.79	41.43
MetaSAug [31]	55.97	38.23	50.94	42.37
CMO [43]	57.27	42.66	51.17	43.44
SAFA [20]	58.77	43.30	53.16	43.72
WB [1]	57.93	41.76	51.67	42.68
GML [11]	59.52	43.65	54.40	45.23
ConCutMix [42]	58.14	43.20	50.40	44.80
LOS [49]	60.11	44.14	54.82	45.62
CAR (Ours)	**61.10**	**46.88**	**55.93**	**46.30**

Table 6. Top-1 accuracy (%) of ViT-Small on ImageNet-LT and CIFAR100-LT under different imbalance factors (IF). "+" indicates the combination of our method with other long-tailed data augmentation methods.

Methods	ImageNet-LT			CIFAR100-LT		
	IF=50	IF=100	IF=200	IF=50	IF=100	IF=200
CAR (Ours)	61.10	57.48	46.88	55.93	51.85	46.30
+ ReMix [6]	62.05	58.74	48.13	56.77	53.64	48.57
+ MetaSAug [31]	64.25	59.28	49.98	57.64	53.75	47.70
+ CMO [43]	64.80	59.72	50.74	58.20	54.77	49.45
+ SAFA [20]	65.98	60.43	51.20	59.80	55.71	49.89
+ ConCutMix [42]	66.73	60.07	50.03	58.77	55.68	50.19

plementary benefits, enhancing model performance without overlapping effects with existing augmentation techniques. These results suggest that CAR promotes better class-wise feature separation and facilitates more effective exploitation of augmented samples. The observed synergy between model-level regularization and data-level transformations confirms that CAR provides a generalizable mechanism that can be seamlessly integrated with various augmentation frameworks to further enhance long-tailed recognition. More detailed results and visual comparisons are presented in the Appendix 12.5.

6. Ablation Studies

Ablation for Class-wise Weight. We examine the influence of the class-wise weight Λ on spectral regularization, as shown in Table 7. All models are trained under identical settings using ViT-Small on ImageNet-LT and CIFAR100-LT. Incorporating Λ consistently improves overall accuracy on both datasets, indicating that frequency-aware weighting mitigates the dominance of head categories and stabilizes optimization under imbalance. The absence of Λ leads to skewed gradients toward frequent classes, which degrades tail performance and hinders convergence. These observations confirm that Λ effectively rebalances gradient contributions across categories, encouraging uniform learning dynamics and improving recognition of minority classes in long-tailed visual recognition.

Ablation for EMA. We evaluate the effect of the EMA mechanism on stabilizing spectral estimation, as reported in Table 7. Introducing EMA leads to consistent improvements on both datasets, confirming that exponential averaging effectively smooths confusion updates and enhances training stability under imbalance. Without EMA, the estimation becomes highly variant across iterations, resulting in unstable convergence and reduced overall performance. These results demonstrate that EMA plays a vital role in maintaining robust spectral regularization. By maintaining smoother updates of the confusion matrix, it enables more robust convergence and contributes to overall performance gains in long-tailed recognition.

Hyperparameter Analysis. We analyze the sensitivity of CAR to four key hyperparameters on CIFAR100-LT using ViT-Small, as shown in Figure 3. From left to right, the plots correspond to the EMA factor β, smoothing radius r_0, regularization weight α, and margin gate γ. Overall, CAR exhibits stable behavior across a wide range of configurations, suggesting that the model is not overly sensi-

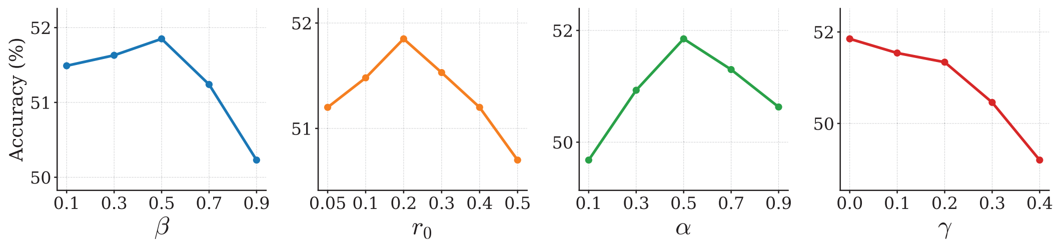

Figure 3. Ablation on four hyperparameters on CIFAR100-LT with ViT-Small (Top-1 accuracy). From left to right: EMA factor β, smoothing radius r_0, regularization weight α, and margin gate γ.

tive to precise parameter choices. A moderate EMA factor ($\beta = 0.5$) provides the most stable moving average for spectral estimation, while $r_0 = 0.2$ achieves a good balance between suppressing noise and preserving discrimination. The regularization weight α reaches an optimum near 0.5, and smaller γ values yield better results by preventing boundary distortion under long-tailed conditions. In summary, the analysis highlights how different hyperparameters cooperatively shape the trade-off between stability and discriminative ability in `CAR`.

Table 7. Ablations for Λ and EMA. Experiments are conducted on ImageNet-LT and CIFAR100-LT based on the ViT-Small.

Factor	Setting	ImageNet-LT	CIFAR100-LT
Λ	w/o Λ	54.39	49.62
	w/ Λ	57.48	51.85
EMA	w/o EMA	55.77	50.20
	w/ EMA	57.48	51.85

7. Related Work

Long-Tailed Learning. Long-tailed learning aims to address the severe class imbalance that commonly occurs in real-world datasets, where a few head classes dominate the training distribution while numerous tail classes have scarce samples [54]. Such imbalance leads to biased decision boundaries and degraded generalization on rare categories. To mitigate this issue, a wide range of strategies have been developed [19]. Re-sampling and re-weighting approaches (e.g., CB [8], RS [47], RW [27], LDAM-DRW [5]) rebalance the data distribution or modify loss weights to emphasize tail instances. Decoupled and representation-oriented learning methods (e.g., MiSLAS [64], DisAlign [61], ResLT [7], BALMS [46]) refine feature spaces through balanced fine-tuning or class-specific calibration. Augmentation and meta-learning schemes such as Remix [6], MetaSAug [31], and ConCutMix [42] further enhance minority representation by generating more diverse visual patterns. Despite these advances, existing approaches mainly focus on adjusting sample frequency or loss weighting, while the confusion spectrum, a critical indicator of model bias, remains largely underexplored and motivates our confusional spectral regularization framework.

Confusion Matrix-based Learning. The confusion matrix has long been a core diagnostic tool for analyzing model predictions, capturing both class-wise accuracy and inter-class misclassification patterns [26, 37, 40]. Beyond evaluation, it has been employed to characterize classifier bias [30, 56], improve fairness [23, 50, 55], and support robust optimization [59]. Calibration-based methods adjust posterior probabilities to reduce bias [36], while relation-based approaches build class-correlation graphs to regularize representations and encourage balanced decision boundaries [2, 16]. More recently, spectral analysis of the confusion matrix has revealed that its singular value spectrum reflects the dominance of major confusion directions [12]. Building on these insights [17], confusional spectral regularization constrains the spectral norm of the confusion matrix to suppress biased eigen-directions and enhance classifier fairness.

8. Conclusion

In this work, we introduced a confusion-centric perspective for long-tailed recognition and established a new generalization upper bound that tightly connects the class-specific error to the spectral norm of a *frequency-weighted confusion matrix*. This analysis reveals that controlling the spectral structure of inter-class confusions provides a principled route for improving generalization on the underrepresented categories. Guided by these insights, we proposed `CAR`, a *Confusion-Aware Spectral Regularizer* that integrates a differentiable confusion matrix surrogate with an EMA-based estimator to enable stable and scalable optimization within standard training pipelines. Extensive experiments across CIFAR100-LT, ImageNet-LT, and iNaturalist demonstrate that `CAR` consistently improves both *worst-class* and *overall* accuracy, achieving state-of-the-art performance under both training-from-scratch and fine-tuning regimes. Moreover, `CAR` complements existing data-augmentation strategies such as ConCutMix, yielding further gains on both head and tail classes.

9. Acknowledgment

The authors acknowledge the use of resources provided by the UKRI SLAIDER project, the MRC SLAIDER-QA project, the Isambard-AI National AI Research Resource (AIRR), and the Dutch national e-infrastructure, supported by the SURF Cooperative (Project EINF-17091). Isambard-AI is operated by the University of Bristol and funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023]. Finally, we thank the anonymous reviewers for their insightful comments, which significantly improved the quality of this paper.

References

[1] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6897–6907, 2022. 4, 5, 7, 3

[2] Anna Arias-Duart, Ettore Mariotti, Dario Garcia-Gasulla, and Jose Maria Alonso-Moral. A confusion matrix for evaluating feature attribution methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3709–3714, 2023. 8

[3] Afonso S Bandeira and March T Boedihardjo. The spectral norm of gaussian matrices with correlated entries. *arXiv preprint arXiv:2104.02662*, 2021. 2

[4] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018. 1

[5] Kaidi Cao, Chunjing Wei, Aurelien Gaidon, Nicolas Arechiga, and Trevor Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019. 4, 5, 7, 8

[6] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *European conference on computer vision*, pages 95–110. Springer, 2020. 4, 5, 6, 7, 8

[7] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3695–3706, 2022. 8

[8] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 1, 4, 5, 6, 7, 8

[9] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Lpt: Long-tailed prompt tuning for image classification. *arXiv preprint arXiv:2210.01033*, 2022. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 4

[11] Yingxiao Du and Jianxin Wu. No one left behind: Improving the worst categories in long-tailed learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15804–15813, 2023. 4, 5, 6, 7

[12] Johan Erbani, Pierre-Édouard Portier, Elöd Egyed-Zsigmond, and Diana Nurbakova. Confusion matrices: A unified theory. *IEEE Access*, 2024. 8

[13] Farzan Farnia, Jesse M Zhang, and David Tse. Generalizable adversarial training via spectral normalization. *arXiv preprint arXiv:1811.07457*, 2018. 2, 3

[14] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd annual ACM SIGACT symposium on theory of computing*, pages 954–959, 2020. 1

[15] Georg Frobenius, Ferdinand Georg Frobenius, Ferdinand Georg Frobenius, Ferdinand Georg Frobenius, and Germany Mathematician. Über matrizen aus nicht negativen elementen. 1912. 1

[16] Jochen Görtler, Fred Hohman, Dominik Moritz, Kanit Wongsuphasawat, Donghao Ren, Rahul Nair, Marc Kirchner, and Kayur Patel. Neo: Generalizing confusion matrix visualization to hierarchical and multi-output labels. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2022. 8

[17] Muhammad Hasnain, Muhammad Fermi Pasha, Imran Ghani, Muhammad Imran, Mohammed Y Alzahrani, and Rahmat Budiarto. Evaluating trust prediction and confusion matrix measures for web services ranking. *Ieee Access*, 8: 90847–90861, 2020. 8

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4, 6

[19] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6626–6636, 2021. 4, 8

[20] Yan Hong, Jianfu Zhang, Zhongyi Sun, and Ke Yan. Safa: Sample-adaptive feature augmentation for long-tailed image classification. In *European Conference on Computer Vision (ECCV)*, 2022. 4, 5, 6, 7

[21] Chengkai Hou, Jieyu Zhang, Haonan Wang, and Tianyi Zhou. Subclass-balancing contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5395–5407, 2023. 1

[22] Gaojie Jin, Xinping Yi, Liang Zhang, Lijun Zhang, Sven Schewe, and Xiaowei Huang. How does weight correlation affect generalisation ability of deep neural networks? *Advances in Neural Information Processing Systems*, 33: 21346–21356, 2020. 1

[23] Gaojie Jin, Sihao Wu, Jiaxu Liu, Tianjin Huang, and Ronghui Mu. Enhancing robust fairness via confusional spectral regularization. *arXiv preprint arXiv:2501.13273*, 2025. 8, 1

[24] Gaojie Jin, Xinping Yi, Wei Huang, Sven Schewe, and Xiaowei Huang. S^{2}o: Enhancing adversarial training with second-order statistics of weights. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8630–8641, 2025. 1

[25] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 1

[26] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems*, 34:4421–4434, 2021. 8

[27] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017. 8

[28] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[29] Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 4

[30] Boqi Li and Weiwei Liu. Wat: improve the worst-class robustness in adversarial training. In *Proceedings of the AAAI conference on artificial intelligence*, pages 14982–14990, 2023. 8

[31] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5212–5221, 2021. 4, 5, 6, 7, 8

[32] Sicong Li, Qianqian Xu, Zhiyong Yang, Zitai Wang, Linchao Zhang, Xiaochun Cao, and Qingming Huang. Focal-sam: Focal sharpness-aware minimization for long-tailed classification. *arXiv preprint arXiv:2505.01660*, 2025. 4

[33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 4, 5, 6, 7

[34] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019. 4

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 4, 6

[36] David Lovell, Dimity Miller, Jaiden Capra, and Andrew Bradley. Never mind the metrics–what about the uncertainty? visualising confusion matrix metric distributions. *arXiv preprint arXiv:2206.02157*, 2022. 8

[37] Pierre Machart and Liva Ralaivola. Confusion matrix stability bounds for multiclass classification. *arXiv preprint arXiv:1202.6221*, 2012. 8

[38] David A McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170, 1999. 2

[39] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 1

[40] Emilie Morvant, Sokol Koço, and Liva Ralaivola. Pac-bayesian generalization bound on confusion matrix for multi-class classification. *arXiv preprint arXiv:1202.6228*, 2012. 2, 8, 1

[41] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017. 2, 1

[42] Haolin Pan, Yong Guo, Mianjie Yu, and Jian Chen. Enhanced long-tailed recognition with contrastive cutmix augmentation. *IEEE Transactions on Image Processing*, 2024. 4, 5, 6, 7, 8

[43] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6887–6896, 2022. 4, 5, 6, 7

[44] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017. 1

[45] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020. 1

[46] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Neural Information Processing Systems (NeurIPS)*, 2020. 4, 5, 6, 7, 8

[47] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pages 467–482. Springer, 2016. 8

[48] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Jie-Jing Shao, Xin-Yan Han, and Yu-Feng Li. Long-tail learning with foundation model: Heavy fine-tuning hurts. *arXiv preprint arXiv:2309.10019*, 2023. 1

[49] Siyu Sun, Han Lu, Jiangtong Li, Yichen Xie, Tianjiao Li, Xiaokang Yang, Liqing Zhang, and Junchi Yan. Rethinking classifier re-training in long-tailed recognition: Label over-smooth can balance. In *The Thirteenth International Conference on Learning Representations*. 2, 4, 5, 6, 7

[50] Tianyu Sun, Hengrui Zhang, Weitao Liu, Xinyu Ren, and Zhen Zhang. Enhancing robust fairness via confusional spectral regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16732–16741, 2024. 8

[51] Junjiao Tian, Yen-Cheng Liu, Nathaniel Glaser, Yen-Chang Hsu, and Zsolt Kira. Posterior re-calibration for imbalanced datasets. *Advances in neural information processing systems*, 33:8101–8113, 2020. 1

[52] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 4

[53] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *European conference on computer vision*, pages 728–744. Springer, 2020. 1

[54] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5017–5026, 2019. 8

[55] Zeming Wei, Yifei Wang, Yiwen Guo, and Yisen Wang. Cfa: Class-wise calibrated fair adversarial training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8193–8201, 2023. 8

[56] Ming-Te Wu. Confusion matrix and minimum cross-entropy metrics based motion recognition system in the classroom. *Scientific Reports*, 12(1):3095, 2022. 8

[57] Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. Adversarial robustness under long-tailed distribution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8659–8668, 2021. 1

[58] Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017. 3

[59] Xinli Yue, Mou Ningping, Qian Wang, and Lingchen Zhao. Revisiting adversarial robustness distillation from the perspective of robust fairness. *Advances in Neural Information Processing Systems*, 36:30390–30401, 2023. 8

[60] Chongsheng Zhang, George Almpanidis, Gaojuan Fan, Binquan Deng, Yanbo Zhang, Ji Liu, Aouaidjia Kamel, Paolo Soda, and João Gama. A systematic review on long-tailed learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 1

[61] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021. 8

[62] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10795–10816, 2023. 1

[63] Qihao Zhao, Yalun Dai, Shen Lin, Wei Hu, Fan Zhang, and Jun Liu. Ltrl: Boosting long-tail recognition via reflective learning. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 4

[64] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021. 8

[65] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020. 4, 5, 7