Reframing Long-Tailed Learning via Loss Landscape Geometry

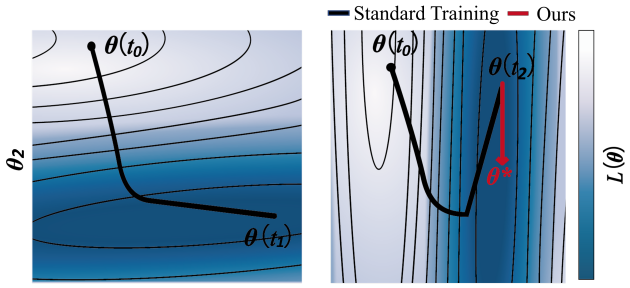
Shenghan Chen[1*] Yiming Liu[1*] Yanzhen Wang[1] Yujia Wang[2] Xiankai Lu[1†]
[1]Shandong University [2]Zhejiang Sci-Tech University

Abstract

Balancing performance trade-off on long-tail (LT) data distributions remains a long-standing challenge. In this paper, we posit that this dilemma stems from a phenomenon called "tail performance degradation" (the model tends to severely overfit on head classes while quickly forgetting tail classes) and pose a solution from a loss landscape perspective. We observe that different classes possess divergent convergence points in the loss landscape. Besides, this divergence is aggravated when the model settles into sharp and non-robust minima, rather than a shared and flat solution that is beneficial for all classes. In light of this, we propose a continual learning inspired framework to prevent "tail performance degradation". To avoid inefficient per-class parameter preservation, a Grouped Knowledge Preservation module is proposed to memorize group-specific convergence parameters, promoting convergence towards a shared solution. Concurrently, our framework integrates a Grouped Sharpness Aware module to seek flatter minima by explicitly addressing the geometry of the loss landscape. Notably, our framework requires neither external training samples nor pre-trained models, facilitating the broad applicability. Extensive experiments on four benchmarks demonstrate significant performance gains over state-of-the-art methods. The code is available at:https://gkp-gsa.github.io/.

1. Introduction

Deep learning has achieved remarkable success across numerous computer vision tasks [2, 9, 20]. This success is often predicated on the availability of large-scale, well-curated, and balanced datasets, such as MS-COCO [31]. However, data encountered in real-world scenarios frequently exhibits a highly imbalanced or long-tailed distribution [17, 62]. Models trained on such datasets tend to develop a strong bias towards the data-abundant head classes, resulting in significantly degraded performance on the tail classes.



(a)Tail-Class Loss Geometry (b)Full Dataset Loss Geometry

Figure 1. "Tail performance degradation" from the loss landscape view. Starting from a randomly initialized point $\theta(t_0)$ of LT model, (a) Training only on tail classes converges to $\theta(t_1)$ in a flat region, while (b) standard training on the long-tailed dataset converges to $\theta(t_2)$ in a sharp region. The optimization trajectory settles in $\theta(t_2)$, which causes tail performance degradation by diverging from the tail convergence point $\theta(t_1)$. In contrast, our optimization (red line) steers the model towards a solution θ^* that remains closer to the tail-class minimum $\theta(t_1)$ and resides in a flatter region (θ_1 and θ_2 on the axes denote projection directions for 2D visualization [23]).

A variety of approaches have been proposed to mitigate this challenge, broadly categorized into three groups [67]: class re-balancing [6, 15, 34], information augmentation [25, 57] and module improvement [5, 60, 71] approaches. Despite their demonstrated effectiveness, these methods often encounter a fundamental trade-off dilemma [28, 29, 36]: enhancing the performance on tail classes leads to performance degradation on head classes, and vice versa (*i.e.*, seesaw dilemma). While a recent trend involving more samples through external data [64, 68] has proven to effectively address the dilemma, this approach is often infeasible in real-world scenarios (*e.g.*, medical) where data is private. This limitation motivates a central research question: **What are the key factors in resolving the head-tail trade-off?** To answer this question, we delve LT learning into the loss landscape view and analyze the optimization trajectories of model parameters during training [3, 23].

In this work, we first visualize the optimization trajectories of a representative long-tailed classifier, BCL [75], by

*Equal contribution.
†Corresponding author: Xiankai Lu (luxiankai@sdu.edu.cn)

projecting model parameters onto the loss landscape. Fig. 1 reveals two critical observations from the loss landscape view. First, the model suffers from "tail performance degradation": the standard optimization converges to $\theta(t_2)$ in Fig. 1 (b) that diverges significantly from the optimal convergence point for tail classes $\theta(t_1)$ in Fig. 1(a) [10]. This plot means the learned model focuses on head classes while rapidly forgetting the tail classes. Accordingly, the learned model with parameters $\theta(t_2)$ yields inferior performance for the tail classes than $\theta(t_1)$. Second, the model with standard training converges to a sharp minimum: compared to (a), the model settles at $\theta(t_2)$ (b) within a sharper region in the loss landscape. Accordingly, the model with $\theta(t_2)$ is inherently sensitive to the underlying perturbation and not robust for generalization across disparate classes.

These two reasons make the current LT training paradigm fail to locate an optimal solution (*i.e.*, θ^*) yielding balanced recognition performance for both head and tail classes [27].

Although standard LT involves joint training without explicit task boundaries, dominant head-class gradients eventually pull the model from tail-friendly flat minima. Recognizing this implicit "forgetting", we formulate long-tailed learning as a continual learning (CL) task from the head classes to the tail classes. Thus, we transfer the head-tail balance issue into the *knowledge preservation and acquisition* balance in continual learning. Building on the insights gained from our investigation, we propose a framework of preserving Knowledge and flattening landscapes that is composed of two branches: the *Grouped Knowledge Preservation* (GKP) branch and the *Grouped Sharpness Aware* (GSA) branch. The GKP branch mitigates tail performance degradation while GSA branch directs the optimization towards a flat convergence region, promoting convergence towards a unified solution beneficial for all classes. Subsequently, an adaptive parameter, scheduled according to the training epoch, is used to aggregate the losses from these branches.

The main contributions of this paper are as follows:

- We investigate the head-tail seesaw dilemma from the loss landscape view and posit the underlying factors, *i.e.*, "tail performance degradation" and sharpness region.
- We transfer long-tailed recognition into a continual learning task to diagnose these issues.
- We propose a new long tail learning framework, including a Grouped Knowledge Preservation (GKP) branch to preserve existing knowledge and a Grouped Sharpness Aware (GSA) branch that flattens the loss landscape.
- Extensive experiments show our method achieves state-of-the-art performance on four long-tailed visual recognition benchmarks.

2. Related Work

Long-tailed Learning: Long-tailed recognition presents a significant challenge in computer vision, arising from the inherent imbalanced distribution of real-world data [6, 15, 30, 39, 76]. Prevailing strategies to address this issue can be broadly categorized into three families: 1) Class Re-balancing [15, 34, 44, 69], which aims to counteract the optimization bias caused by class imbalance, typically through re-sampling or re-weighting; 2) Information Augmentation [25, 33, 46, 59], which enriches data-scarce classes by synthesizing or augmenting information; and 3) Module Improvement [8, 16, 45, 60], which involves designing specialized network architectures or components to inherently better handle the class imbalance.

Recently, quite a few works have focused on leveraging external sources to incorporate additional training samples [7, 41, 48, 64] or large pre-trained models [50, 70]. However, a significant limitation of this approach lies in the heavy reliance on external data or models. This requirement is often infeasible in practical scenarios where data privacy is paramount (*e.g.*, medical applications). Yet, these methods pay little attention to the underlying reason for the trade-off dilemma. This paper reframes LT from the loss landscape view and proposes a new solution to alleviate the "tail performance degradation".

Sharpness of Loss Landscape: Research on model generalization increasingly focuses on loss landscape geometry. Extensive work has established that models converging to flatter minima exhibit superior generalization [18, 21, 35]. This principle underpins optimization methods like Sharpness-Aware Minimization (SAM) [10] and Friendly Sharpness-Aware Minimization [27].

In long-tailed learning, SAM has been adapted to improve tail-class generalization. Early methods typically combined SAM with standard re-balancing techniques [42]. Recognizing different class requirements, subsequent works proposed more granular strategies, such as selectively applying SAM only to tail classes [73] or implementing fine-grained, per-class controls [26, 74]. These works have proven effective for imbalanced data. This work also extends the SAM framework for long-tailed classification by removing the head-dominated global perturbation direction to improve the performance of tail classes.

3. The Feature Quality and Landscape: A Motivation Study

In this section, we conduct experiments from two dimensions: feature level and loss level to assess the influence of "tail performance degradation". All studies utilize BCL [75] and our framework with CIFAR100-LT dataset [4].

In the feature level, we define the *Feature Quality* for head classes and tail classes **during** LT learning. Feature Quality Q is based on two components [38]: inter-class separation, which measures the separation between different classes, and intra-class variance, which quantifies the dispersion of

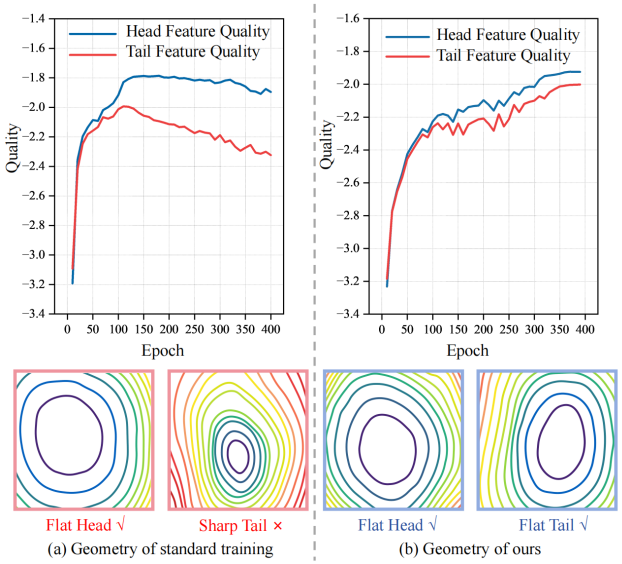


Figure 2. (a) Standard training results in a sharp loss landscape for tail classes, where the corresponding feature quality peaks and then declines. (b) In contrast, our method flattens the landscape and preserves high feature quality for both head and tail classes.

features within each class. In the loss level, we plot the loss landscape geometry for head classes and tail classes **after** model learning. Results are summarized in Fig. 2.

- **Observation 1**: For BCL [75], the feature quality of the tail classes (red line) rises sharply in the initial phase, peaking at epoch 120, after which performance declines. While the feature quality of the head classes improves and maintains greater stability. For our framework, both head and tail classes exhibit a similar upward trend and stable enhancement in their feature quality.
  Analysis: These results support our claim that "the LT model tends to severely overfit on head classes while quickly forgetting tail classes with model training". The superiority of our method confirms that knowledge preservation is a promising alternative.
- **Observation** 2: For BCL, the tail classes have sharp loss landscape geometry, while our method has a flat region for both tail classes and head classes.
  Analysis: This provides direct evidence for our claim that current optimization "is sensitive to the underlying perturbation and not robust for generalization across disparate classes". The Grouped Sharpness Aware strategy, by design, controls the flatness of different classes to balance the performance between head and tail classes.

4. Methodology

4.1. Overall Framework

This section presents our whole framework, as illustrated in Fig. 3. Given a long-tailed training dataset $\mathcal{D} = \{(x_i, y_i)_{i=1}^{N}\}$, where x_i denotes a sample and $y_i \in \mathcal{C}$ represents its corresponding label with a total of C classes. Our aim is to learn a function f_θ with parameter θ mapping from an input space to the label space. The function f_θ is implemented as the composition of an encoder $a_i = f(\theta_{enc}, x_i)$ with parameter θ_{enc} and a fully-connected layer as classifier $W : x_i \rightarrow \hat{y}_i$.

Different from most previous works [25, 36, 75] that focus primarily on an effective encoder to enhance the long tail learning, this work rethinks the LT learning from a landscape view and devises a *grouped knowledge preservation* module and a *grouped sharpness-aware* module to jointly ameliorate the encoder and the linear classifier.

4.2. Grouped Knowledge Preservation Module

The design of the GKP module is inspired by the aim of continual learning (CL) [52, 54], which addresses performance degradation: the tendency for knowledge of previously learned tasks (*e.g.*, Task A) to be lost as information relevant to the current task (*e.g.*, Task B) is incorporated.

Unlike imbalanced Class-Incremental Learning (CIL), which tackles explicit sequential tasks with dual intra/inter-phase imbalances [12], long-tailed learning operates under a single joint objective. Thus, the key challenge in applying our CL-inspired scheme is how to define tasks. A naive per-class preservation strategy (treating each class as a task) is computationally prohibitive for datasets with many classes and severely hinders knowledge acquisition, as the optimization is constrained by the excessive number of preservation targets. Conversely, a simple head-tail task split is too coarse, ignoring the diverse convergence needs within the tail or head classes and thus failing to effectively preserve knowledge. Therefore, our GKP employs a memory-based grouping strategy to balance preservation and acquisition by clustering multiple classes based on their shared optimal parameters and subsequently treating each group as a task.

4.2.1. Memory-based Grouping Strategy

Memory Construction. Memory-based Grouping Strategy first constructs a memory bank [49] $\mathcal{M}$ to dynamically store the encoder parameters θ_{enc}^{c} that achieved the highest feature quality for each class c during training.

During model training, this memory bank is updated dynamically: at each epoch t, the current encoder parameters $\theta_{enc}^{(t)}$ replace the stored θ_{enc}^{c} if it yields a higher feature quality Q (identified in Sec. 3) for that class c:

$$\theta_{enc}^{c} \leftarrow \begin{cases} \theta_{enc}^{(t)} & \text{if } Q(\theta_{enc}^{(t)}, c) > Q(\theta_{enc}^{c}, c) \\ \theta_{enc}^{c} & \text{otherwise} \end{cases}. \quad (1)$$

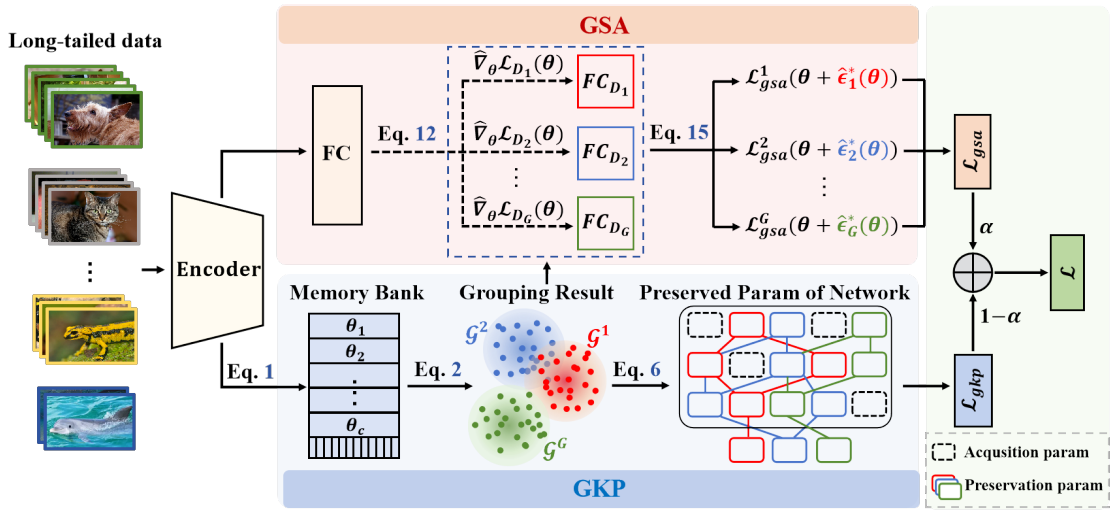


Figure 3. Our framework consists of two key components: (a) The Grouped Sharpness Aware (GSA) module, which minimizes group-specific sharpness to find flat minima. (b) The Grouped Knowledge Preservation (GKP) module, which prevents tail performance degradation of other groups' optimal parameters.

Once the memory bank $\mathcal{M}$ is populated with the optimal encoder parameter sets $\mathcal{M} = \{\theta^1_{enc}, \cdots, \theta^C_{enc}\}$, we proceed to partition the class set $\mathcal{C} = \{1, \cdots, C\}$ into G groups which would be detailed in the following section.

Grouping Operation. Grouping Operation aims to group classes that exhibit similar encoder parameters. Based on $\mathcal{M}$, we leverage spectral clustering [40] method, *i.e.*, Normalized Cuts (NCut) algorithm [47] to implement the operation:

$$\{\mathcal{G}^1, ...\mathcal{G}^g, ..., \mathcal{G}^G\} = \text{NCut}(\mathcal{G}, G), \tag{2}$$

where G means the group number, $\mathcal{G}$ is a weighted, undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ which is built upon $\mathcal{M}$. $\mathcal{G}^g$ denotes the g-th sub-graph. During ablation studies (§5.3), we perform experiments to assess the effect of group number.

Once the group partitioning is obtained, we compute a shared encoder parameter θ^*_g for each group $\mathcal{G}^g$:

$$\theta^*_g = \frac{1}{|\mathcal{G}^g|} \sum_{c \in \mathcal{G}^g} \theta^c_{enc}. \tag{3}$$

In this way, we can obtain the group-wise parameters, as illustrated in Fig. 3.

4.2.2. Parameter Preservation

Regularization-based CL approaches focus primarily on mitigating tail performance degradation by penalizing changes to parameters vital for previously learned tasks. A typical solution is Elastic Weight Consolidation (EWC) [22], which estimates the importance of each parameter using the Fisher Information Matrix as corresponding quadratic penalty:

$$\begin{aligned} \mathcal{L}_{EWC} &= \mathcal{L}_D(\theta) + \mathcal{L}_{penalty}(\theta) \\ &= \mathcal{L}_D(\theta) + \frac{\lambda}{2} \sum_i F_i(\theta_i - \theta^*_{t-1,i})^2, \end{aligned} \tag{4}$$

where $\mathcal{L}_D$ works for knowledge acquisition while $\mathcal{L}_{penalty}(\theta)$ works for knowledge preservation. λ controls the strength of the regularization, F_i is the diagonal of the Fisher Information Matrix, θ^*_{t-1} are the optimal parameters for previous tasks and i denotes each parameter component.

Based on the group-wise parameters in Eq. 3, the penalty term $\mathcal{L}_{penalty}(\theta)$ in Eq. 4 is reformulated as a dynamic, group-wise constraint. Specifically, when the model is training on the current group g (*i.e.*, knowledge acquisition), we revise $\mathcal{L}_{penalty}(\theta)$ to simultaneously preserve the knowledge of all other groups ($j \neq g$) (*i.e.*, knowledge preservation):

$$\mathcal{L}_{penalty}(\theta) = \frac{\lambda}{2} \sum_i \sum_{j \neq g} F_{j,i}(\theta_i - \theta^*_{j,i})^2, \tag{5}$$

where $F_{j,i}$ denotes the i-th diagonal element of the (approximated) Fisher Information Matrix for group j [22].

Considering the sample sizes for each group are various, we balance the importance of each group in Eq. 5 by normalizing group size $|\mathcal{G}^j|$ and obtain the final GKP loss:

$$\mathcal{L}^g_{gkp} = \frac{\lambda}{2} \sum_i \sum_{j \neq g} \left(\frac{1}{|\mathcal{G}^j|} F_{j,i}(\theta_i - \theta^*_{j,i})^2 \right). \tag{6}$$

We use the proposed GKP loss as the penalty term in the EWC loss function (Eq. 4) to implement knowledge preservation. Meanwhile, for $\mathcal{L}_D$ in Eq. 4, we introduce a new optimization solution called grouped sharpness aware module to facilitate knowledge acquisition.

4.3. Grouped Sharpness Aware Module

The Grouped Sharpness Aware (GSA) module is the second component of the framework, responsible for tailoring the

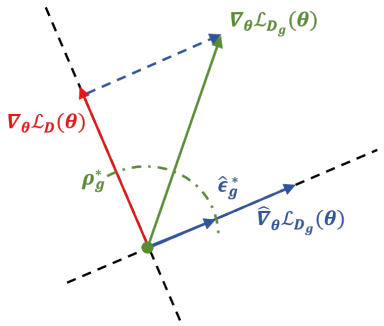


Figure 4. Investigation on the perturbation direction. We decompose the original gradient $\nabla_\theta \mathcal{L}_{\mathcal{D}_g}(\theta)$ (green) into two components: the head-dominated global gradient $\nabla_\theta \mathcal{L}_D(\theta)$ (red) and the beneficial, group-specific gradient $\hat{\nabla}_\theta \mathcal{L}_{\mathcal{D}_g}(\theta)$ (blue).

knowledge acquisition objective $\mathcal{L}_D$ (Eq. 4) to obtain flat sharpness. Before elaborating our GSA, we first retrospect existing Sharpness-aware Minimization (SAM).

Sharpness-aware Minimization. SAM is theoretically motivated by the PAC-Bayesian generalization bound [11, 43], which establishes a link between a model's generalization error $\mathcal{L}_\mathcal{T}(\theta)$ and the sharpness of the loss landscape:

$$\mathcal{L}_\mathcal{T}(\theta) \leq \max_{||\epsilon||_2 \leq \sqrt{d}\rho} \mathcal{L}_D(\theta+\epsilon) + \sqrt{\frac{\frac{||\theta||_2^2}{4\rho^2} + \log(\frac{N}{\delta}) + \mathcal{O}(1)}{N-1}}, \tag{7}$$

where $\mathcal{L}_D(\cdot)$ comes from Eq. 4, $\mathcal{T}$ means the distribution of D. This bound holds for any perturbation radius $\rho > 0$ and $\delta \in (0, 1)$, with a probability of $1-\delta$. Moreover, N represents the number of training samples, and $d := \dim(\theta)$ denotes the dimensionality of the parameter space.

The objective of SAM is to find an optimal radius ρ^* that minimizes this bound (Eq. 7):

$$\rho^* \approx \left(\frac{\|\theta\|_2}{2\|\nabla_\theta \mathcal{L}_D(\theta)\|_2}\right)^{\frac{1}{2}} d^{-\frac{1}{4}}(N-1)^{-\frac{1}{4}}, \tag{8}$$

With the approximated optimal generalization bound, Eq. 7 can be rewritten as:

$$\hat{\mathcal{L}}_\mathcal{T}(\theta) \approx \max_{\|\epsilon\|_2 \leq \sqrt{d}\rho^*} \mathcal{L}_\mathcal{D}(\theta+\epsilon) + \frac{1}{2\sqrt{N-1}} \frac{\|\theta\|_2}{\rho^*}. \tag{9}$$

The first term of this bound quantifies the sharpness of $\mathcal{L}_\mathcal{D}$, while the second term serves as a regularizer on the model parameters θ. The optimal perturbation vector $\hat{\epsilon}^*(\theta)$ can be computing as:

$$\begin{aligned} \hat{\epsilon}^*(\theta) &\approx \arg\max_{\|\epsilon\|_2 \leq \sqrt{d}\rho^*} \left[\mathcal{L}_\mathcal{D}(\theta) + \epsilon^T \nabla_\theta \mathcal{L}_\mathcal{D}(\theta)\right] \\ &= \sqrt{d}\rho^* \frac{\nabla_\theta \mathcal{L}_\mathcal{D}(\theta)}{\|\nabla_\theta \mathcal{L}_\mathcal{D}(\theta)\|_2}. \end{aligned} \tag{10}$$

Grouped Sharpness-aware Minimization. However, the standard perturbation in Eq. 10 is dependent upon the global gradient $\nabla_\theta \mathcal{L}_\mathcal{D}(\theta)$, which is dominated by head classes. This renders standard SAM insensitive to the elevated sharpness of tail classes. To address this limitation, we propose the Grouped Sharpness-Awareness (GSA) module.

Firstly, upon the grouping result in Eq. 2, we shift from using the single global gradient (Eq. 10) to calculating group-wise gradients $\nabla_\theta \mathcal{L}_{D_g}(\theta)$ for each group $\mathcal{G}^g$:

$$\nabla_\theta \mathcal{L}_{D_g}(\theta) \leftarrow [\nabla_\theta \mathcal{L}_\mathcal{D}(\theta), \mathcal{G}_g]. \tag{11}$$

After that, considering our goal is to define a perturbation direction that is not biased by the head classes. Therefore, as shown in Fig. 4, via the gradient decomposition analysis [27], we decompose $\nabla_\theta \mathcal{L}_{D_g}(\theta)$ (Eq. 11) into two components:

$$\hat{\nabla}_\theta \mathcal{L}_{\mathcal{D}_g}(\theta) = \nabla_\theta \mathcal{L}_{\mathcal{D}_g}(\boldsymbol{\theta}) - \mathrm{Proj}_{\nabla_\theta \mathcal{L}_D(\boldsymbol{\theta})} \nabla_\theta \mathcal{L}_{\mathcal{D}_g}(\boldsymbol{\theta}), \tag{12}$$

where $\nabla_\theta \mathcal{L}_D(\boldsymbol{\theta})$ is the global gradient over $\mathcal{D}$ and $\mathrm{Proj}_{\nabla_\theta \mathcal{L}_D(\boldsymbol{\theta})}(\cdot)$ denotes the projection operator on $\nabla_\theta \mathcal{L}_{D_g}(\theta)$ along the direction of the global gradient. In this way, we can obtain the group-specific gradient $\hat{\nabla}_\theta \mathcal{L}_{\mathcal{D}_g}(\theta)$.

Moreover, to balance the perturbation radius across groups, we refine Eq. 8 by leveraging the group size $|\mathcal{G}^g|$:

$$\rho_g^* \approx \left(\frac{\|\theta\|_2}{2\|\hat{\nabla}_\theta \mathcal{L}_{\mathcal{D}_g}(\theta)\|_2}\right)^{\frac{1}{2}} d^{-\frac{1}{4}}(|\mathcal{G}^g|-1)^{-\frac{1}{4}}. \tag{13}$$

By substituting this group-specific radius ρ_g^* (Eq. 13) and the corresponding group gradient $\hat{\nabla}_\theta \mathcal{L}_{\mathcal{D}_g}(\theta)$ (Eq. 12) into the optimal perturbation formula (Eq. 10), we derive the final GSA perturbation vector $\hat{\epsilon}_g^*(\theta)$:

$$\hat{\epsilon}_g^*(\theta) = \sqrt{d}\rho_g^* \frac{\hat{\nabla}_\theta \mathcal{L}_{\mathcal{D}_g}(\theta)}{\|\hat{\nabla}_\theta \mathcal{L}_{\mathcal{D}_g}(\theta)\|_2}. \tag{14}$$

By using a group gradient (Eq. 12) that removes the head-dominated global gradient, GSA improves tail classes flatness. Consequently, the training objective of GSA module that works for knowledge acquisition is formulated as:

$$\mathcal{L}_{gsa}^g(\theta) = \mathcal{L}_{\mathcal{D}_g}(\theta + \hat{\epsilon}_g^*(\theta)) + \frac{1}{2\sqrt{|\mathcal{G}^g|-1}} \frac{\|\theta\|_2}{\rho_g^*}. \tag{15}$$

4.4. Training Objects

Our framework is jointly optimized by grouped sharpness-aware objective $\mathcal{L}_{gsa}^g$ (Eq. 15) and grouped knowledge preservation objective $\mathcal{L}_{gkp}^g$ (Eq. 6):

$$\mathcal{L} = \sum_{g=1}^{G} \left[\alpha \mathcal{L}_{gsa}^g + (1-\alpha)\mathcal{L}_{gkp}^g\right], \tag{16}$$

where α is an adaptive parameter scheduled according to the training epoch [71].

Table 1. Results on CIFAR100-LT [4] and CIFAR10-LT datasets [4]. The imbalance ratio r is set to 100, 50 and 10. Additionally, we present the results for different groups ("Many", "Med." and "Few") in CIFAR100-LT with r = 100. † denotes methods that utilize Large Language Models.

Method	CIFAR100-LT			CIFAR10-LT			Statistic(r=100)		
	r=100↑	r=50↑	r=10↑	r=100↑	r=50↑	r=10↑	Many↑	Med.↑	Few↑
CE (Baseline)	38.3	43.9	55.7	70.4	74.8	86.4	65.2	37.1	9.1
Focal Loss [32] (ICCV'17)	38.4	44.3	55.8	70.4	76.7	86.7	65.3	38.4	8.1
LDAM-DRW [4] (NeurIPS'19)	42.0	46.6	58.7	77.0	81.0	88.2	61.5	41.7	20.2
cRT [19] (ICLR'20)	42.3	46.8	58.1	75.7	80.4	88.3	64.0	44.8	18.1
BBN [71] (CVPR'20)	42.6	47.0	59.1	79.8	82.2	88.3	-	-	-
RIDE (3 experts) [56] (ICLR'21)	48.0	-	-	-	-	-	68.1	49.2	23.9
CAM-BS [66] (AAAI'21)	41.7	46.0	-	75.4	81.4	-	-	-	-
DiVE [14] (ICCV'21)	45.4	51.1	62.0	-	-	-	-	-	-
SAM [42] (NeurIPS'22)	45.4	-	-	81.9	-	-	64.4	46.2	20.8
BCL [75] (CVPR'22)	51.9	56.6	64.9	84.3	87.2	91.1	67.2	53.1	32.9
CUDA [1] (ICLR'23)	47.6	51.1	58.4	-	-	-	67.3	50.4	21.4
ADRW [58] (NeurIPS'23)	46.4	-	61.9	83.6	-	90.3	-	-	-
H2T [24] (AAAI'24)	48.9	53.8	-	-	-	-	-	-	-
GBG [28] (AAAI'24)	52.3	57.2	-	85.1	87.7	-	-	-	-
DiffuLT [46] (NeurIPS'24)	51.5	56.3	63.8	84.7	86.9	90.7	69.0	51.6	29.7
DiffuLT + BBN [46] (NeurIPS'24)	51.9	56.7	64.0	85.0	87.2	90.9	**69.5**	51.9	30.2
SEL [17] (ICCV'25)	52.3	57.3	68.4	84.4	86.3	90.2	-	-	-
Heuristic-CALA[72] (AAAI'25)	50.5	-	64.3	83.9	-	91.7	-	-	-
Meta-CALA[72] (AAAI'25)	52.3	-	65.5	84.7	-	92.4	-	-	-
FeatRecon [63] (ICLR'25)	52.5	57.0	65.3	85.2	87.8	91.6	-	-	-
LLM-AutoDA† [55](NeurIPS'24)	51.0	54.8	-	-	-	-	66.6	50.6	33.1
Ours	**53.2**	**57.6**	**68.7**	**86.3**	**88.2**	**92.5**	67.3	**54.9**	**34.9**

5. Experiment

5.1. Experiment Setup

Datasets. Our proposed framework is evaluated on four long-tailed benchmarks: CIFAR10-LT [4], CIFAR100-LT [4], ImageNet-LT [37] and iNaturalist 2018 [53].

Metrics. Following standard evaluation protocols, we report Top-1 accuracy for comparison with state-of-the-art methods. To evaluate our method's effectiveness across varying imbalance levels, we report Top-1 accuracy under three imbalance ratios $r \in \{100, 50, 10\}$ for CIFAR10-LT and CIFAR100-LT. Additionally, we report results for "Many" (classes with over 100 samples), "Med." (classes with 20 to 100 samples), and "Few" (classes with fewer than 20 samples) categories separately to enable in-depth analysis, following the methodology described in [4]. For ImageNet-LT, we present results with different feature backbones for a comprehensive evaluation of our method.

Implementation Details. For both CIFAR10-LT and CIFAR100-LT, we use the ResNet-32 as the backbone following [4]. For ImageNet-LT datasets, we use ResNet-50 [13] and ResNeXt50 [61] as backbone. For iNaturalist 2018, we use ResNet-50 [13] as backbone. For all datasets, we set the batch size to 256 and train all models on NVIDIA GeForce RTX 3090 GPU. For further implementation details, please refer to the Appendix.

5.2. Main Results

CIFAR10-LT and CIFAR100-LT [4]. The comparison results between the proposed method and other existing methods on long-tailed CIFAR are shown in Table 1. Our proposed method surpasses recent state-of-the-art approaches across the datasets under various imbalance ratios. On CIFAR100-LT, our method surpasses competing models, achieving accuracy improvements of 14.9%, 13.7%, and 13.0% compared with the baseline for r = 100, 50, and 10, respectively. On CIFAR10-LT, our model also demonstrates strong competitiveness, enhancing accuracy by 15.9%, 13.4%, and 6.1% for r = 100, 50, and 10, respectively, further validating the effectiveness of our method.

As the main counterpart, our method yields better performance than BCL [75] (CIFAR100-LT: 53.2 vs. 51.9, 57.6 vs. 56.6, 68.7 vs. 64.9; CIFAR10-LT: 86.3 vs. 84.3, 88.2 vs. 87.2, 92.5 vs. 91.1). Considering both methods use the same backbone and loss functions, we attribute the performance improvement solely to the proposed grouped

Table 2. Top-1 accuracy of ResNet-50 on ImageNet-LT [37] and iNaturalist 2018 [53]. † denotes methods that utilize Large Language Models.

Method	ImageNet-LT	iNature2018
CE(Baseline)	41.6	61.0
Focal Loss [32](ICCV'17)	-	61.1
cRT [19](ICLR'20)	47.3	65.2
RIDE (3 experts) [56](ICLR'21)	54.9	72.7
BCL [75] (CVPR'22)	56.0	71.8
SAM [42](NeurIPS'22)	53.1	70.1
CUDA [1](ICLR'23)	51.4	72.2
ADRW [58](NeurIPS'23)	54.1	70.7
GBG [28] (AAAI'24)	57.6	71.9
DiffuLT [46](NeurIPS'24)	56.4	-
DiffuLT + RIDE (3 experts) [46] (NeurIPS'24)	56.9	-
SEL [17] (ICCV'25)	56.3	71.3
Heuristic-CALA [72] (AAAI'25)	54.1	73.2
Meta-CALA [72] (AAAI'25)	55.1	74.0
FeatRecon [63] (ICLR'25)	56.8	72.9
LLM-AutoDA† [55] (NeurIPS'24)	57.5	74.2
Ours	**57.9**	**74.4**

knowledge preservation module and the grouped sharpness aware module.

For CIFAR100-LT with an imbalanced ratio of 100, we also assess performance across three categories: ("Many": 67.3%, "Med.": 54.9%, "Few": 34.9%), effectively addressing the performance trade-off between head and tail classes in long-tailed learning. These results collectively demonstrate our method's effectiveness in handling the fundamental challenges of long-tailed distributions, especially the problem of extreme class imbalance.

Notably, our proposed method also surpasses the LLM-based LLM-AutoDA [55] by 2.2% in performance, despite using only the standard training set without any external data or pre-trained models.

ImageNet-LT [37]. Table 2 and Table 3 report the results on ImageNet-LT with different backbones for comprehensive results comparison. We report the overall Top-1 accuracy as well as the Top-1 accuracy on "Many", "Medium", and "Few" groups. Utilizing ResNet-50 backbone (Table 2), our method registers 57.9% accuracy, outperforming the state-of-the-art LLM-AutoDA [55] by 0.4%.

With the stronger ResNeXt-50 backbone (Table 3), the accuracy of our method further escalates to 58.9%. Specifically, our method achieves 68.7%, 56.8%, and 38.6% on the "Many", "Medium", and "Few" splits, outperforms the top-leading FeatRecon [63] by 1.8%, 1.3%, and 0.8%, respectively. These significant gains, particularly on tail classes, highlight the effectiveness of our proposed grouped knowledge preservation module and the grouped sharpness aware module in addressing extreme data distribution imbalances.

Table 3. Top-1 accuracy of ResNeXt-50 on ImageNet-LT [37].

Method	Many↑	Med.↑	Few↑	All↑
CE (Baseline)	-	-	-	44.4
FocalLoss [32] (ICCV'17)	64.3	37.1	8.2	43.7
τ-norm [19] (ICLR'20)	59.1	46.9	30.7	49.4
BalancedSoftmax [44] (NeurIPS'20)	62.2	48.8	29.8	51.4
Casualmodel [51] (NeurIPS'20)	62.7	48.8	31.6	51.8
LWS [19] (ICLR'20)	60.2	47.2	30.3	49.9
LADE [15] (CVPR'21)	62.3	49.3	31.2	51.9
DisAlign [65] (CVPR'21)	62.7	52.1	31.4	53.4
RIDE (2 experts) [56] (ICLR'21)	-	-	-	55.9
BCL [75] (CVPR'22)	-	-	-	56.7
GBG [28] (AAAI'24)	69.6	55.8	38.1	58.7
FeatRecon [63] (ICLR'25)	67.9	54.7	37.8	57.5
Ours	**69.7**	**56.0**	**38.6**	**58.9**

Again, these gains come from using only standard training set without any external data.

iNaturalist 2018 [53]. Table 2 (right column) shows the experimental results on the real-world large-scale imbalanced iNaturalist 2018. Employing a ResNet-50 backbone, our method achieves an accuracy of 74.4%, demonstrating a significant performance promotion of 13.4% over the baseline.

All of the above improvement on imbalanced recognition confirms the effectiveness of our proposed grouped knowledge preservation and grouped sharpness aware module which learns in a compositional manner, and is informed by parameter preservation and flatten landscape to effectively address "tail performance degradation".

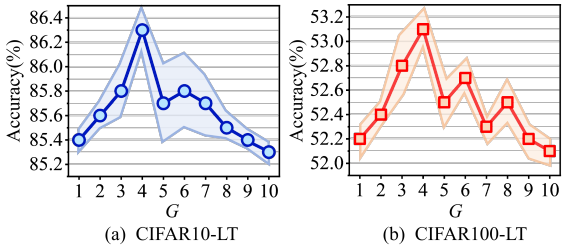
5.3. Ablation Study

In this section, we perform several ablation studies to characterize the proposed the method.

Component Analysis. We first verify the effectiveness of main components of our proposed framework. We specifically choose CIFAR100-LT (r=100) for experiment evaluation and use ResNet-32 as the backbone.

As shown in Table 4, using GKP modules brings +3.7% accuracy on *CE* baseline. Furthermore, combining GKP and GSA modules then achieves a total 4.4% improvement. On the stronger BCL [75] baseline, our GSA and GKP modules yield 0.8% and 0.5% gain, respectively, while the full model achieves a 1.3% total improvement, finally reaching 53.2%. These results confirm the crucial and highly additive contribution of each component.

Group Numbers. In Fig. 5, we investigate the impact of different group numbers within our grouping strategy (*i.e.*, G in Eq. 2). This study was conducted on CIFAR10-LT (blue, left) and CIFAR100-LT (red, right), both with an imbalance factor $r = 100$. The results reveal that $G = 4$ groups



(a) CIFAR10-LT (b) CIFAR100-LT

Figure 5. Ablation Study on group numbers of our method. The shaded area indicates the fluctuation of accuracy.

yield optimal performance as employing an excess number of groups does not necessarily enhance performance in the GKP modules. We report the detailed ablation studies for G on other datasets in the Appendix.

Table 4. Ablation study on main components of our method.

Method	Many↑	Med.↑	Few↑	All↑
CE	65.2	37.1	9.1	38.4
+GKP	71.1	41.2	9.3	42.1
+GKP + GSA	71.8	42.1	9.7	42.8
BCL [75]	67.2	53.1	32.9	51.9
+GKP	67.4	53.8	33.2	52.4
+GSA	67.3	54.0	34.1	52.7
Ours	**67.3** ↑0.1	**54.9** ↑1.8	**34.9** ↑2.0	**53.2** ↑1.3

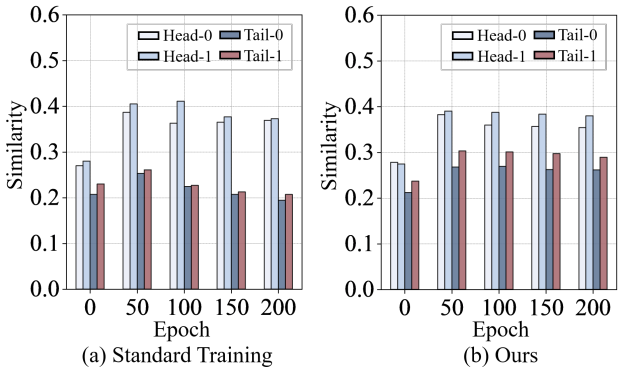
Importance of Gradient Decomposition in GSA. To validate the gradient decomposition (Eq. 12), we compare the performance of different perturbation directions in Table 5. Our GSA, using the group-specific gradient surpasses all other methods, achieves an accuracy 53.2%. This compares favorably to a standard SAM baseline, which does not process the gradient direction and achieves 52.1%. While using only the projected component $\text{Proj}_{\nabla_\theta \mathcal{L}_D(\boldsymbol{\theta})} \nabla_\theta \mathcal{L}_{\mathcal{D}_g}(\boldsymbol{\theta})$ termed GSA-proj yields obvious performance degradation of 46.4%. This performance gap demonstrates that the head-dominated global gradient is a harmful perturbation direction in long-tailed learning.

Table 5. Importance of Gradient Decomposition in GSA.

Method	Many↑	Med.↑	Few↑	All↑
SAM	66.3	53.0	34.5	52.1
GSA-proj	64.7	43.8	28.1	46.4
GSA (Ours)	**67.3** ↑1.0	**54.9** ↑1.9	**34.9** ↑0.4	**53.2** ↑1.1

5.4. Further Analysis from the Gradient Similarity

To further validate the effectiveness of our method, we analyze the training dynamics from the gradient similarity [28] perspective. We measure the mean gradient similarity between class and whole gradient in each batch in different



(a) Standard Training (b) Ours

Figure 6. Gradient imbalance in long-tail learning. The bars denote the mean similarity between class-level and batch-level gradients in each batch. This experiment compares our framework against the Cross-Entropy (CE) baseline on the CIFAR10-LT dataset [4]. We show the result of the top two head classes and the last two tail classes.

epochs, comparing our method against the Cross-Entropy (CE) baseline. The similarity measures the contribution of gradients from different classes to the gradient descent process, and a larger similarity means a larger contribution.

As illustrated in Fig. 6, while the head classes gradients exhibit high similarity in both methods, their behavior on tail classes diverges significantly. For the baseline method (a), the gradient similarity of the tail classes declines after 50 epochs. In contrast, our method (b) consistently maintains a larger gradient similarity for the tail classes throughout the training, demonstrating the effectiveness in preserving the tail classes knowledge.

6. Conclusion

Long-tail (LT) recognition remained a fundamental challenge in deep learning due to severe class imbalance. To address the "seesaw dilemma" in LT, we presented a novel framework from loss landscape view and traced the "tail performance degradation". Inspired by continual learning, our method introduced two key innovations: a Grouped Knowledge Preservation module to mitigate "tail performance degradation", and a Grouped Sharpness Aware module to seek shared, flat minima. Both modules were enabled by our Memory-based Grouping strategy, which dynamically clusters classes based on their convergence characteristics. Extensive experiments on four benchmark datasets demonstrated our method's superior performance, achieving significant performance gains over state-of-the-art methods without requiring external data. Our proposed method was a general framework for handling "tail performance degradation" in data learning, and can be readily extended to other long tail tasks, such as trajectory prediction and object detection.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. T2541059 and 62302046), the Shandong Excellent Young Scientists Fund (ZR2024YQ006), the Taishan Scholars Program (Grant No. tsqn202507014) and the National Key R&D Program of China (Program Nos. 2024YFC2707500).

References

[1] Sumyeong Ahn, Jongwoo Ko, and Se-Young Yun. CUDA: Curriculum of data augmentation for long-tailed recognition. In *ICLR*, 2023. 6, 7

[2] Khaled Bayoudh. A survey of multimodal hybrid deep learning for computer vision: Architectures, applications, trends, and challenges. *Information Fusion*, 105:102217, 2024. 1

[3] John V. Breakwell. The optimization of trajectories. *Journal of the Society for Industrial and Applied Mathematics*, 7(2): 215–247, 1959. 1

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019. 2, 6, 8

[5] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *ICCV*, 2021. 1

[6] 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, 2

[7] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Lpt: Long-tailed prompt tuning for image classification. In *ICLR*, 2022. 2

[8] Qi Dong, Shaogang Gong, and Xiatian Zhu. Class rectification hard mining for imbalanced deep learning. In *ICCV*, 2017. 2

[9] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):5, 2021. 1

[10] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2020. 2

[11] Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. Pac-bayesian theory meets bayesian inference. In *NeurIPS*, 2016. 5

[12] Jiangpeng He. Gradient reweighting: Towards imbalanced class-incremental learning. In *CVPR*, 2024. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[14] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling virtual examples for long-tailed recognition. In *ICCV*, 2021. 6

[15] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, 2021. 1, 2, 7

[16] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, 2016. 2

[17] Zhongquan Jian, Yanhao Chen, Yancheng Wang, Junfeng Yao, Meihong Wang, and Qingqiang Wu. Supervised exploratory learning for long-tailed visual recognition. In *ICCV*, 2025. 1, 6, 7

[18] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *ICLR*, 2019. 2

[19] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020. 6, 7

[20] Sushant Kaushal, Dushyanth Kumar Tammineni, Priya Rana, Minaxi Sharma, Kandi Sridhar, and Ho-Hsien Chen. Computer vision and deep learning-based approaches for detection of food nutrients/nutrition: New insights and advances. *Trends in Food Science & Technology*, 146:104408, 2024. 1

[21] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *ICLR*, 2017. 2

[22] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526, 2017. 4

[23] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *NeurIPS*, 2018. 1

[24] Mengke Li, HU Zhikai, Yang Lu, Weichao Lan, Yiu-ming Cheung, and Hui Huang. Feature fusion from head to tail for long-tailed visual recognition. In *AAAI*, 2024. 6

[25] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *CVPR*, 2021. 1, 2, 3

[26] 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. In *NeurIPS*, 2025. 2

[27] Tao Li, Pan Zhou, Zhengbao He, Xinwen Cheng, and Xiaolin Huang. Friendly sharpness-aware minimization. In *CVPR*, 2024. 2, 5

[28] Weiqi Li, Fan Lyu, Fanhua Shang, Liang Wan, and Wei Feng. Long-tailed learning as multi-objective optimization. In *AAAI*, 2024. 1, 6, 7, 8

[29] Yuhang Li, Zhuying Li, and Yuheng Jia. Boosting class representation via semantically related instances for robust long-tailed learning with noisy labels. In *ICCV*, 2025. 1

[30] Zhixin Li and Yuheng Jia. Conmix: Contrastive mixup at representation level for long-tailed deep clustering. In *ICLR*, 2025. 2

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 6, 7

[33] Bo Liu, Haoxiang Li, Hao Kang, Gang Hua, and Nuno Vasconcelos. Gistnet: a geometric structure transfer network for long-tailed recognition. In *ICCV*, 2021. 2

[34] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008. 1, 2

[35] Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *CVPR*, 2022. 2

[36] Yuting Liu, Liu Yang, and Yu Wang. Long-tailed classification with multi-granularity semantics. In *ICCV*, 2025. 1, 3

[37] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 6, 7

[38] Anay Majee, Suraj Nandkishor Kothawade, Krishnateja Killamsetty, and Rishabh K Iyer. Score: Submodular combinatorial representation learning. In *ICML*, 2023. 2

[39] Kartik Narayan, Vibashan Vs, and Vishal M Patel. Segface: Face segmentation of long-tail classes. In *AAAI*, 2025. 2

[40] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NeurIPS*, 2001. 4

[41] Vignesh Ramanathan, Rui Wang, and Dhruv Mahajan. Dlwl: Improving detection for lowshot classes with weakly labelled data. In *CVPR*, 2020. 2

[42] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, et al. Escaping saddle points for effective generalization on class-imbalanced data. In *NeurIPS*, 2022. 2, 6, 7

[43] David Reeb, Andreas Doerr, Sebastian Gerwinn, and Barbara Rakitsch. Learning gaussian processes by minimizing pac-bayesian generalization bounds. In *NeurIPS*, 2018. 5

[44] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. In *NeurIPS*, 2020. 2, 7

[45] Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *ICCV*, 2021. 2

[46] Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. Diffult: Diffusion for long-tail recognition without external knowledge. In *NeurIPS*, 2024. 2, 6, 7

[47] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8):888–905, 2000. 4

[48] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Xin-Yan Han, Jie-Jing Shao, and Yufeng Li. Parameter-efficient long-tailed recognition. *CoRR*, 2023. 2

[49] James Seale Smith, Lazar Valkov, Shaunak Halbe, Vyshnavi Gutta, Rogerio Feris, Zsolt Kira, and Leonid Karlinsky. Adaptive memory replay for continual learning. In *CVPR*, 2024. 3

[50] Mingyang Song, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. From head to tail: Towards balanced representation in large vision-language models through adaptive data calibration. In *CVPR*, 2025. 2

[51] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020. 7

[52] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019. 3

[53] 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 *CVPR*, 2018. 6, 7

[54] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE TPAMI*, 46(8):5362–5383, 2024. 3

[55] Pengkun Wang, Zhe Zhao, Haibin Wen, Fanfu Wang, Binwu Wang, Qingfu Zhang, and Yang Wang. Llm-autoda: Large language model-driven automatic data augmentation for long-tailed problems. In *NeurIPS*, 2024. 6, 7

[56] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2021. 6, 7

[57] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *NeurIPS*, 2017. 1

[58] Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning. In *NeurIPS*, 2023. 6, 7

[59] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *CVPR*, 2021. 2

[60] Tz-Ying Wu, Pedro Morgado, Pei Wang, Chih-Hui Ho, and Nuno Vasconcelos. Solving long-tailed recognition with deep realistic taxonomic classifier. In *ECCV*, 2020. 1, 2

[61] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 6

[62] Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *IJCV*, 130(7):1837–1872, 2022. 1

[63] Lingjie Yi, Jiachen Yao, Weimin Lyu, Haibin Ling, Raphael Douady, and Chao Chen. Geometry of long-tailed representation learning: Rebalancing features for skewed distributions. In *ICLR*, 2025. 6, 7

[64] Cheng Zhang, Tai-Yu Pan, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. Mosaicos: A simple and effective use of object-centric images for long-tailed object detection. In *ICCV*, 2021. 1, 2

[65] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, 2021. 7

[66] Yongshun Zhang, Xiu-Shen Wei, Boyan Zhou, and Jianxin Wu. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *AAAI*, 2021. 6

[67] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE TPAMI*, 45(9):10795–10816, 2023. 1

[68] Yifan Zhang, Daquan Zhou, Bryan Hooi, Kai Wang, and Jiashi Feng. Expanding small-scale datasets with guided imagination. In *NeurIPS*, 2023. 1

[69] Zizhao Zhang and Tomas Pfister. Learning fast sample reweighting without reward data. In *ICCV*, 2021. 2

[70] Qihao Zhao, Yalun Dai, Hao Li, Wei Hu, Fan Zhang, and Jun Liu. Ltgc: Long-tail recognition via leveraging llms-driven generated content. In *CVPR*, 2024. 2

[71] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, 2020. 1, 5, 6

[72] Xiaoling Zhou, Ou Wu, and Nan Yang. Class and attribute-aware logit adjustment for generalized long-tail learning. In *AAAI*, 2025. 6, 7

[73] Yixuan Zhou, Yi Qu, Xing Xu, and Hengtao Shen. Imbsam: A closer look at sharpness-aware minimization in class-imbalanced recognition. In *ICCV*, 2023. 2

[74] Zhipeng Zhou, Lanqing Li, Peilin Zhao, Pheng-Ann Heng, and Wei Gong. Class-conditional sharpness-aware minimization for deep long-tailed recognition. In *CVPR*, 2023. 2

[75] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *CVPR*, 2022. 1, 2, 3, 6, 7, 8

[76] Ke Zhu, Minghao Fu, Jie Shao, Tianyu Liu, and Jianxin Wu. Rectify the regression bias in long-tailed object detection. In *ECCV*, 2024. 2