OFBD: Object-Focused Background Debiasing for Long-Tailed Learning

Shenghan Chen[1,2] * **Yiming Liu**[2] **Zhipeng Deng**[1] **Haolin Wang**[3]
Jiale Zhou[1] **Zhijian Wu**[1] **Xiankai Lu**[2,6] **Yafei Ou**[5] † **Yefeng Zheng**[1] †

[1]Westlake University, Hangzhou, China
[2]Shandong University, Jinan, China
[3]Hokkaido University, Sapporo, Japan
[5]RIKEN, Japan
[6]Inspur Cloud Information Technology Co., Ltd., Jinan, 250000, Shandong, China

Abstract

Balancing performance trade-offs on long-tailed data distributions remains a long-standing challenge in visual recognition. Existing methods mainly improve tail classes through re-balancing, representation learning, or data augmentation, but the underlying cause of tail class degradation is still insufficiently explored. In this paper, we find that standard long-tailed training induces background-biased representation and optimization: tail classes suffer larger background distribution shifts and become increasingly driven by background gradients. This reveals that tail degradation is not merely caused by insufficient samples, but also by the learning of irrelevant background features. To tackle this issue, we propose Object-Focused Background Debiasing (OFBD), a framework that mitigates background bias from both distribution and optimization perspectives. Specifically, Foreground-guided CutMix preserves target-related foregrounds while diversifying complementary backgrounds, and Background-guided Feature Rectification suppresses background-biased features without learnable parameters or additional training. Extensive experiments show that our method improves overall accuracy, achieves significant tail-class gains, and can serve as a plug-in for mainstream long-tailed methods without external data or pretrained recognition models. The code is available at: https://ofbd-neurips2026-longtail-learning.github.io/

1 Introduction

The prevailing paradigms in visual recognition [37, 80, 16] owe much of their success to large-scale, carefully curated datasets [66, 47] with balanced class distributions. However, real-world data naturally exhibits highly imbalanced or long-tailed distributions [97]. When optimized on such skewed distributions, standard training protocols inevitably become biased toward data-abundant head classes, resulting in severe performance degradation on data-scarce tail classes.

Existing methods addressing the class imbalance issue mainly fall into three paradigms. First, class re-balancing strategies such as resampling [50, 60] or reweighting [21, 11], amplify tail signals but may overfit tail classes [74]. Second, architectural improvements correct feature and classifier bias via contrastive representation learning [110, 83, 32]; and via logit calibration [53] or expert routing [87, 9, 100]. Third, information augmentation techniques synthesize or transfer semantic variations for tail classes [84, 70], beyond conventional augmentations [1] (*e.g.*, cropping, flipping).

*This work was done while Shenghan Chen was a visiting student at Westlake University.
†Yafei Ou and Yefeng Zheng are co-corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

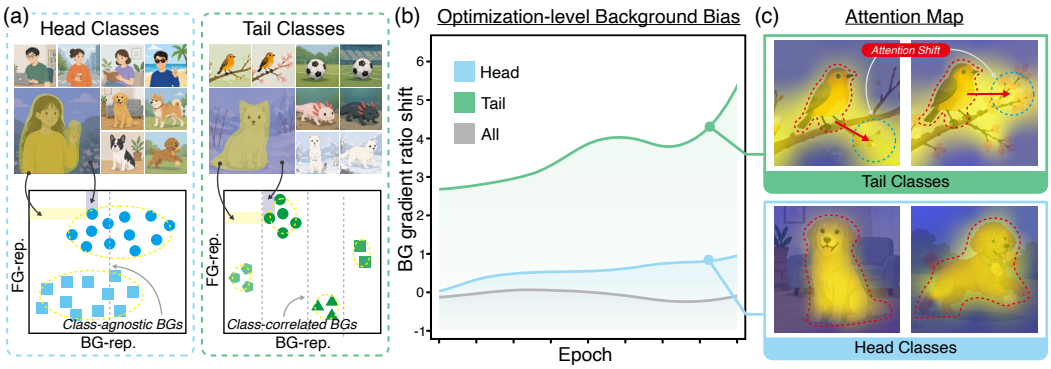


Figure 1: Motivation of OFBD. Tail-class degradation stems from both sample scarcity and background bias in distribution and optimization. (a) Head classes possess diverse, class-agnostic backgrounds (BGs), making background representations (BG-Rep.) uninformative for classification. Conversely, tail classes often co-occur with class-correlated BGs (*e.g.*, snow fox in snowy scenes), which makes BG-Rep. a predictive shortcut and encourages reliance on BG cues over sparse foregrounds. Gray dashed lines denote BG-based decision boundaries. (b) Optimization-level background bias. "Head" and "Tail" represent the evaluation results under the long-tailed setting, while "All" indicates the overall result on the full balanced dataset. (c) Attention maps reveal that tail predictions co-activate foregrounds and backgrounds, whereas head predictions remain object-focused.

Despite these improvements, existing methods mainly compensate for data quantity, leaving the degradation of tail representations during standard optimization [40] largely unexplored.

Interestingly, recent advances in self-supervised [28, 49] and few-shot learning [88] demonstrate that models can learn highly transferable representations from extremely limited data [15]. Such findings indicate that sample scarcity alone is insufficient to explain the profound representational collapse observed in tail classes. This paradox raises a critical question: **If limited data is not an inherent barrier, why do long-tailed models fail to capture the discriminative features of tail classes?**

To investigate the underlying mechanism, we visualize the spatial representations learned by representative long-tailed networks [53, 110]. As illustrated in Fig. 1(c), we observe a striking phenomenon: standard imbalanced training severely compromises the spatial inductive bias of the network, causing it to co-activate target foregrounds with irrelevant background regions. This empirical evidence reveals that tail-class degradation extends beyond simple data scarcity; crucially, it induces a profound background bias in both representation and optimization. Consequently, discriminative foreground features are marginalized by spurious contextual cues, leading to degradation on tail classes [77, 55]. Therefore, resolving the long-tailed dilemma necessitates a dual approach: enhancing target-related foreground while explicitly suppressing the exploitation of irrelevant background evidence.

Suppressing irrelevant background evidence requires understanding its entanglement with targets. Building upon recent insights [55, 73], we attribute this background bias to two structural forms of input-level co-occurrence: object-scene (*e.g.*, fish with water) and object-object (*e.g.*, racket with person) entanglements (Fig. 1(a)). These co-occurrences bias optimization toward spurious background shortcuts, resulting in background-biased representations and skewed BG feature distributions (Fig. 1(b)). We formalize this contextual and structural analysis in Sec. 3. Although spatial mixup strategies like CutMix [98] can disrupt these co-occurrences, prior works have shown that naive random replacements may introduce irrelevant noise [58, 29]. To tackle this issue, we propose Object-Focused Background Debiasing (OFBD), a framework that mitigates background bias from both distribution and optimization perspectives. Specifically, *Foreground-guided CutMix* (FG-CutMix) uses a reinforcement-learning selector to handle discrete and non-differentiable region selection, preserving target-related foregrounds while replacing backgrounds to break foreground-background co-occurrence. Meanwhile, *Background-guided Feature Rectification* (BFR) estimates background contribution from channel-wise feature statistics and down-weights background-biased spatial features in a parameter-free manner, avoiding head-class-dominated learnable rectification.

Overall, the main contributions of this paper are summarized as follows:

- **Novel Perspective:** We systematically investigate the inherent background bias in long-tailed recognition. Through structural analysis and empirical validation, we reveal how this bias shifts feature distributions and misguides optimization dynamics.

- **Dual Debiasing Framework:** We propose FG-CutMix and BFR to address the two biases. FG-CutMix integrates reinforcement learning [71] to preserve target foregrounds, while BFR rectifies features to reduce background reliance in a parameter-free manner.
- **Superior Performance & Plug-and-Play Versatility:** Extensive experiments demonstrate that without relying on external data or pretrained vision models, OFBD significantly boosts tail-class performance and overall accuracy. Furthermore, it inherently serves as a highly efficient plug-in module for existing mainstream long-tailed models.

2 Related Work

Long-tailed Learning: Real-world data are often long-tailed [11, 111, 57, 45, 21]. Existing methods include class re-balancing [21, 102, 50, 65, 75], information augmentation [70, 41, 48, 91, 104], and architectural improvements [24, 13, 68, 93, 76]. Yet they mainly compensate class quantity: re-balancing changes sample/loss weights, augmentation adds diversity without separating foreground features from background features, and architectures rarely rectify feature-level background bias. We instead mitigate distribution- and optimization-level background bias by decoupling target-related foreground features from spurious background features.

Background Bias in Visual Recognition: Deep neural networks are notoriously prone to background bias, often relying on spurious contextual correlations (shortcuts) rather than intrinsic object semantics [18, 17, 95, 56]. While various debiasing strategies have been proposed [54, 82, 34, 38, 5], they predominantly assume balanced data distributions, leaving their behavior under severe class imbalance unexplored. Our work bridges this critical gap by revealing that standard long-tailed optimization inherently exacerbates these background biases. In response, we introduce a dual debiasing framework, offering a principled approach to robust and unbiased long-tailed recognition.

RL-driven Data Augmentation: Reinforcement Learning (RL) has been used to search augmentation policies [10, 46, 84], select transformation operations or magnitudes [19, 12], and identify informative regions [89, 90, 44]. These methods optimize validation performance, augmentation strength, or region informativeness for general robustness, but are not tailored to long-tailed recognition, where tail classes are more vulnerable to spurious background reliance. In contrast, our RL selector is designed for FG-CutMix under long-tailed training. By selecting foreground regions that preserve target semantics after background replacement, the selector disrupts foreground-background co-occurrence and reduces tail-class background reliance.

3 Analysis and Motivation

In this section, we analyze background bias from two complementary perspectives and provide empirical evidence. At the distribution level, we study how the learned background feature distribution deviates from that of the full balanced dataset. At the optimization level, we examine how training becomes increasingly driven by background-related gradients.

3.1 Foreground-Background Decomposition

Following prior works [72, 61, 86], we decompose an input image x into the target-related foreground and the target-irrelevant background, including co-occurring scenes and objects. Let $F = \Phi(x) \in \mathbb{R}^{D\times H\times W}$ be the feature map extracted by the encoder. Since foreground and background are spatial concepts, we first separate foreground and background regions on F according to the decomposition of x, and then aggregate them into f_{fg} and f_{bg}, respectively. The image-level features can be written as $f = f_{fg} + f_{bg}$, where f_{fg} and f_{bg} denote foreground and background features.

3.2 Distribution-level Background Bias

For each class y, we use p_y and q_y to denote the background distributions under long-tailed and full-dataset training, respectively. The corresponding background feature means are $\mu_{bg}(y) = \mathbb{E}_{b\sim p_y}[f_{bg}]$ and $\bar{\mu}_{bg}(y) = \mathbb{E}_{b\sim q_y}[f_{bg}]$. We then define the background bias of class y as $\Delta D_{bg}(y) := \mu_{bg}(y) - \bar{\mu}_{bg}(y)$, which measures the shift of the learned background feature mean from the full-dataset background feature mean. Since tail classes have far fewer samples than head classes, their

empirical background distributions are more likely to deviate from q_y. Accordingly, we assume $\|p_y^{\text{tail}} - q_y\|_1 > \|p_y^{\text{head}} - q_y\|_1$, which yields a larger background distribution shift for tail classes:

$$\|\Delta D_{bg}^{\text{tail}}\|_1 > \|\Delta D_{bg}^{\text{head}}\|_1. \tag{1}$$

3.3 Optimization-level Background Bias

We next consider background bias from the optimization perspective. Let g_t denote the gradient signal at epoch t. Consistent with the feature decomposition above, we first separate foreground and background regions on F, and then aggregate their gradient magnitudes into $g_{fg,t}$ and $g_{bg,t}$, respectively. We define the background gradient ratio and its shift for class y as:

$$R_t(y) = \frac{G_{bg,t}(y)}{G_{fg,t}(y) + G_{bg,t}(y)}, \quad R_0(y) = \frac{1}{T_0}\sum_{t=1}^{T_0} R_t(y), \quad \Delta R_t(y) = R_t(y) - R_0(y), \tag{2}$$

where $G_{fg,t}(y) = \|g_{fg,t}(y)\|_1$ and $G_{bg,t}(y) = \|g_{bg,t}(y)\|_1$. Since the absolute value of $R_t(y)$ can be affected by the initial foreground-background composition of each class, $\Delta R_t(y)$ better captures whether optimization progressively leans toward background features. Under long-tailed training, tail classes receive weaker foreground supervision, while background cues are easier to exploit and continue to contribute to the loss. We therefore assume:

$$\Delta R_t^{\text{tail}} > \Delta R_t^{\text{head}}. \tag{3}$$

Detailed derivations and assumptions motivating these hypotheses (Eq. 1 and Eq. 3) are provided in the Appendix B. We empirically validate these claims below.

3.4 Empirical Evidence

To validate the above analysis, we empirically examine background bias from both distribution and optimization perspectives. Visualized results and experimental details are provided in Appendix D.3.

Observation 1: Tail classes suffer larger distribution-level shift. Fig. 5a shows larger tail-class background distribution shift than head classes due to limited samples and less diverse backgrounds, motivating FG-CutMix to diversify backgrounds while preserving target foregrounds.

Observation 2: Long-tailed training amplifies optimization-level bias. Fig. 5b shows a larger background-gradient-ratio increase under long-tailed training, especially for tail classes. This indicates that insufficient and imbalanced foreground supervision shifts optimization toward easily exploitable background features, motivating BFR to suppress background-biased gradients.

Observation 3: Tail-class background influence is more unstable. Fig. 5b further shows stronger fluctuations in tail-class background-gradient ratios, motivating parameter-free BFR and the reference selector in FG-CutMix for stable debiasing and foreground selection.

4 Method

Guided by Sec. 3, our framework addresses background bias from two aspects. FG-CutMix targets the distribution-level shift ΔD_{bg} by preserving foregrounds while replacing backgrounds. BFR targets the optimization-level shift ΔR_t by down-weighting background-biased features. Together, they mitigate biased background distributions and background-driven optimization in long-tailed training. We provide a theoretical interpretation of the effectiveness of both modules in Appendix B.4.

4.1 Foreground-guided CutMix

Unlike naive CutMix, which may exacerbate bias by sampling background-only regions [55], our Foreground-guided CutMix (FG-CutMix) explicitly preserves target foregrounds. Furthermore, its RL-based selector introduces minimal computational overhead, ensuring practicality for large-scale training. Detailed experiments evaluating the computational cost are provided in Appendix D.8.

4.1.1 RL-based Foreground Region Selection

The key challenge in FG-CutMix is selecting foreground regions effective for background replacement. CAM-based [106, 31] and SAM-based [35, 64] strategies can localize complete foregrounds, but accurate localization may not yield the best mixing region. Since CutMix benefits from partial-view recognition and regional perturbation [98], effective regions should balance target preservation and background replacement, as overly complete regions may weaken background variation and regularization. Since foreground-region selection is discrete and non-differentiable, we use an RL-based selector to learn an adaptive policy, capture class-specific foreground importance.

RL-based foreground candidate generation. Given the feature map $F = \Phi(x) \in \mathbb{R}^{D\times H\times W}$, we first randomly generate a region proposal set $\mathcal{P}(x) = \{b_1, \dots, b_N\}$ with different locations and scales. Each proposal b_i corresponds to a candidate crop region on the input image. Instead of ranking these proposals by local energy [85, 26], we introduce an RL-based region selector π_θ to select K proposals from $\mathcal{P}(x)$ as foreground candidates. Specifically, π_θ predicts a selection probability over $\mathcal{P}(x)$, and K proposals are sampled without replacement to form the foreground candidate pool (FG candidate pool) $\mathcal{R}_\theta(x) = \{r_1, \dots, r_k, \dots, r_K\}$. This allows the selector to learn which randomly generated regions are more suitable for preserving target-related foregrounds in FG-CutMix.

Reward design. Given the selected candidate pool, we construct mixed samples $\{\tilde{x}_1, \dots, \tilde{x}_K\}$ and obtain class prediction probabilities $p_k \in [0, 1]^C$ from the classifier. Let y and y^{bg} denote the target label and the background-source label, respectively. The reward for candidate r_k is defined as:

$$\rho_k = \mathbf{1}\big(p_k[y] > p_k[y^{\text{bg}}]\big)\,. \tag{4}$$

Here, p_k provides the reward signal and $o_k = \pi_\theta(r_k \mid F)$ denotes the selector confidence, encouraging regions that preserve target semantics over background-induced semantics.

Policy learning. The region selector π_θ defines a selection distribution over the proposal set $\mathcal{P}(x)$ and samples K proposals $\mathcal{R}_\theta(x)$ as foreground candidates. Since proposal selection is discrete and non-differentiable, we optimize π_θ via policy gradient [36, 69, 94]. Since selected candidates within each image are not directly comparable due to varying classifier confidence [101] and potential background bias, we follow GRPO [71] and normalize rewards within each candidate pool. Since tail-class background gradients fluctuate more (Sec. 3.4), we use the previous-epoch selector as a frozen reference π_{ref} to stabilize foreground selection. The policy learning objective is:

$$\begin{aligned}
\mathcal{L}_{\text{RL}} &= -\frac{1}{K}\sum_{k=1}^{K}\left[\min(\hat{s}_{1,k}A_k, \hat{s}_{2,k}A_k) - \beta\mathcal{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}}) - \alpha\text{CE}(o_k, \rho_k)\right], \\
\mathcal{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}}) &= \frac{\pi_{\text{ref}}(r_k \mid F)}{\pi_\theta(r_k \mid F)} - \log\frac{\pi_{\text{ref}}(r_k \mid F)}{\pi_\theta(r_k \mid F)} - 1, \\
\hat{s}_{1,k} = \frac{\pi_\theta(r_k \mid F)}{\pi_{\text{ref}}(r_k \mid F)}, &\qquad \hat{s}_{2,k} = \text{clip}\left(\frac{\pi_\theta(r_k \mid F)}{\pi_{\text{ref}}(r_k \mid F)}, 1-\epsilon_c, 1+\epsilon_c\right), \\
A_k = \frac{\rho_k - \text{mean}(\rho_1, \dots, \rho_K)}{\text{std}(\rho_1, \dots, \rho_K) + \epsilon}, &\qquad o_k = \pi_\theta(r_k \mid F),
\end{aligned} \tag{5}$$

where θ denotes the parameters of the region selector π_θ, ϵ_c clamps the policy ratio, and β controls the KL penalty. Following prior works [69, 71], the frozen reference selector π_{ref} and KL term stabilize policy learning by limiting abrupt policy shifts, while the clipped objective prioritizes high-advantage foreground candidates without overly large updates. The auxiliary term, weighted by α, encourages region selector confidence to align with the reward signal.

4.1.2 FG-CutMix Sample Construction

Given a selected foreground candidate $r_k \in \mathcal{R}(x)$, we preserve the target-related region in x and replace its complement with content from another image x'. Let M_k be the binary mask of r_k, where $M_k = 1$ denotes the preserved region, and $M_k = 0$ denotes the complement. The mixed sample is $\tilde{x}_k := M_k \odot x + (1 - M_k) \odot x'$, where $\odot$ denotes element-wise multiplication. This preserves target foregrounds, varies surrounding backgrounds, and reduces foreground-background co-occurrence. Following prior works [58, 51], we correct the mixed label as $\tilde{y}_k$ for reliable supervision.

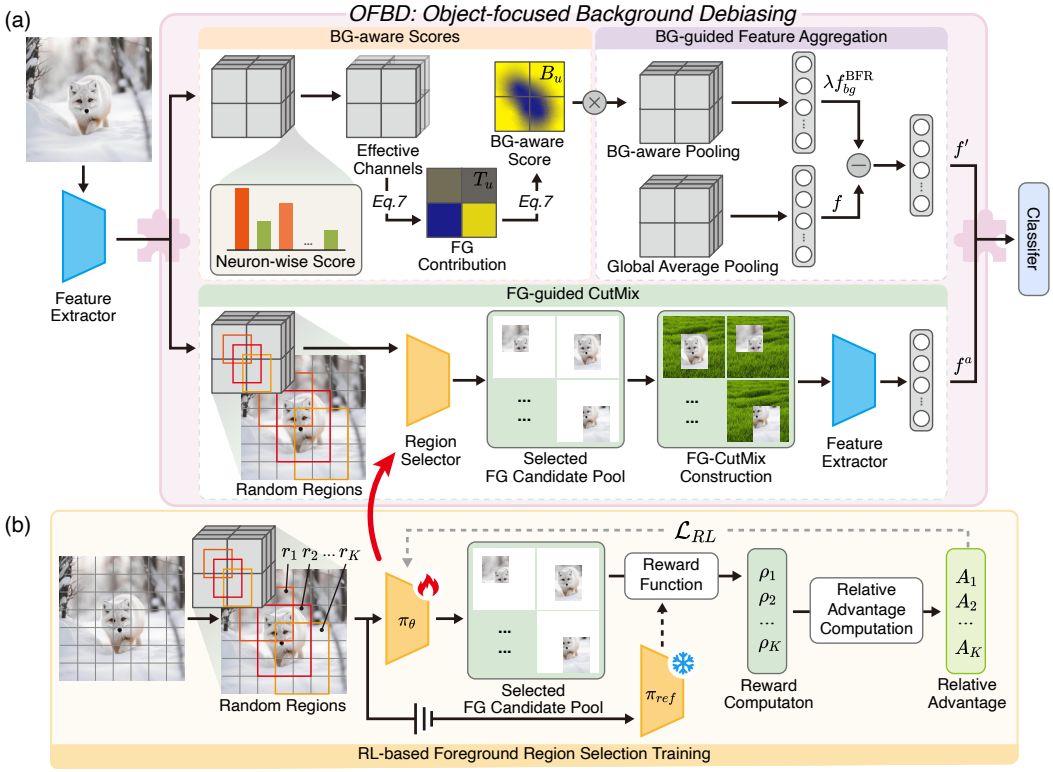


Figure 2: Overview of OFBD. The pink box denotes OFBD as a plug-in module for mainstream long-tailed models. (a) FG-CutMix selects target-related foregrounds from random regions and replaces complementary backgrounds, while BFR estimates background-aware scores to rectify features before classification. (b) The region selector π_θ is optimized via reinforcement learning, and a frozen reference selector π_{ref} stabilizes the training of the selector.

4.2 Background-guided Feature Rectification

Although FG-CutMix alleviates distribution-level background bias by restoring background diversity, optimization may still favor irrelevant background cues that are easier to exploit than sparse tail-class foreground evidence. We therefore introduce Background-guided Feature Rectification (BFR), which uses background-aware scores to down-weight biased feature locations in the final representation. To avoid head-class-dominated gradients from learnable parameters, BFR is strictly parameter-free. Inspired by SimAM [96] and LaSt-ViT [73], BFR estimates foreground contribution from channel-wise feature statistics: larger normalized deviations from stable channel means indicate higher foreground contribution. Effective channel aggregation further prevents sparse tail-class foreground channels from being diluted and derives background-aware scores for rectification.

4.2.1 Background-aware Scores

Given the feature map $F = \Phi(x) \in \mathbb{R}^{D\times H\times W}$, let u denote a spatial location and $F_{d,u}$ denote the activation at channel d and location u, we compute a neuron-wise score $S_{d,u}$:

$$S_{d,u} = \text{sigmoid}\left(\frac{(F_{d,u}-\mu_d)^2}{4(\sigma_d^2+\epsilon)}\right), \qquad \bar{S}_u = \frac{1}{D}\sum_{d=1}^{D} S_{d,u}, \tag{6}$$

where μ_d and σ_d^2 are the mean and variance of channel d over spatial locations, respectively. The detailed derivation is provided in Appendix B.3. We then retain effective channels $\mathcal{C}_u = \{d \mid S_{d,u} \geq \bar{S}_u\}$ and compute the foreground contribution T_u and background-aware score B_u as:

$$T_u = \frac{1}{|\mathcal{C}_u|}\sum_{d\in\mathcal{C}_u} S_{d,u}, \qquad B_u = \text{sigmoid}\left(\frac{\bar{T}-T_u}{\sigma_T+\epsilon}\right), \tag{7}$$

where $\bar{T}$ and σ_T are the mean and standard deviation of T_u over all feature locations, respectively. A larger B_u indicates stronger background bias at location u.

Table 1: Comparisons on CIFAR-LT [4] datasets with varying imbalance ratios (r). Results are reported across all classes as well as different groups based on training sample sizes ("Many", "Med", and "Few"). For each method, the first row reports the original baseline accuracy, and the second row shows the performance after applying OFBD.

	CIFAR-100-LT			CIFAR-10-LT			Statistic($r = 100$)		
Method	r=100↑	r=50↑	r=10↑	r=100↑	r=50↑	r=10↑	Many↑	Med.↑	Few↑
CE	38.32	43.94	55.71	70.45	74.82	86.44	65.19	37.06	9.14
+OFBD	$44.83_{\uparrow 6.51}$	$47.79_{\uparrow 3.85}$	$58.22_{\uparrow 2.51}$	$74.33_{\uparrow 3.88}$	$78.68_{\uparrow 3.86}$	$87.30_{\uparrow 0.86}$	$65.48_{\uparrow 0.29}$	$42.09_{\uparrow 5.03}$	$23.93_{\uparrow 14.79}$
LDAM-DRW [4] (NeurIPS'19)	42.09	46.57	58.64	77.06	81.09	88.02	61.51	41.67	20.22
+OFBD	$45.81_{\uparrow 3.72}$	$49.55_{\uparrow 2.98}$	$59.09_{\uparrow 0.45}$	$79.61_{\uparrow 2.55}$	$82.43_{\uparrow 1.34}$	$88.76_{\uparrow 0.74}$	$64.93_{\uparrow 3.42}$	$45.81_{\uparrow 4.14}$	$26.61_{\uparrow 6.39}$
BBN [107] (CVPR'20)	42.45	47.13	59.23	79.78	81.24	88.32	42.63	50.70	32.60
+OFBD	$44.81_{\uparrow 2.36}$	$48.32_{\uparrow 1.19}$	$60.57_{\uparrow 1.34}$	$80.69_{\uparrow 0.91}$	$84.41_{\uparrow 3.17}$	$89.28_{\uparrow 0.96}$	$43.94_{\uparrow 1.31}$	$51.83_{\uparrow 1.13}$	$37.64_{\uparrow 5.04}$
BCL [110] (CVPR'22)	51.84	56.57	64.03	84.31	87.21	90.92	66.85	52.54	33.07
+OFBD	$54.19_{\uparrow 2.35}$	$57.94_{\uparrow 1.37}$	$64.91_{\uparrow 0.88}$	$86.34_{\uparrow 2.03}$	$88.12_{\uparrow 0.91}$	$91.63_{\uparrow 0.71}$	$66.79_{\downarrow 0.06}$	$53.46_{\uparrow 0.92}$	$40.34_{\uparrow 7.27}$
SBCL [22] (CVPR'23)	44.93	48.65	57.83	74.84	80.05	84.48	64.35	45.27	22.14
+OFBD	$46.92_{\uparrow 1.99}$	$49.08_{\uparrow 0.43}$	$58.44_{\uparrow 0.61}$	$76.52_{\uparrow 1.68}$	$81.43_{\uparrow 1.38}$	$84.89_{\uparrow 0.41}$	$65.26_{\uparrow 0.91}$	$45.35_{\uparrow 0.08}$	$27.36_{\uparrow 5.22}$
GBG [43] (AAAI'24)	51.92	56.92	64.32	85.12	87.64	91.23	66.94	53.15	32.96
+OFBD	$54.23_{\uparrow 2.31}$	$58.37_{\uparrow 1.45}$	$64.98_{\uparrow 0.66}$	$86.57_{\uparrow 1.45}$	$88.25_{\uparrow 0.61}$	$91.61_{\uparrow 0.38}$	$66.81_{\downarrow 0.13}$	$53.57_{\uparrow 0.42}$	$40.32_{\uparrow 7.36}$
MKP [8] (CVPR'26)	53.21	57.63	68.74	86.31	88.26	92.53	67.28	54.86	34.91
+OFBD	$54.42_{\uparrow 1.21}$	$58.41_{\uparrow 0.78}$	$68.12_{\downarrow 0.62}$	$86.78_{\uparrow 0.47}$	$88.42_{\uparrow 0.16}$	$92.02_{\downarrow 0.51}$	$67.57_{\uparrow 0.29}$	$55.95_{\uparrow 1.09}$	$37.29_{\uparrow 2.38}$

4.2.2 Background-guided Feature Aggregation

Based on the background-aware score B_u, we estimate and remove the background-related component from the image-level feature. Let f denote the standard image-level feature:

$$f = \frac{1}{HW}\sum_u F_u, \qquad f_{bg}^{\mathrm{BFR}} = \frac{1}{HW}\sum_u B_u F_u. \tag{8}$$

Here, f_{bg}^{BFR} denotes the estimated background-biased component. The rectified feature is then obtained by removing this estimated background-biased feature:

$$f' = f - \lambda f_{bg}^{\mathrm{BFR}} = \frac{1}{HW}\sum_u (1 - \lambda B_u) F_u, \tag{9}$$

where λ controls the strength of background rectification. In this way, BFR softly down-weights background-biased feature locations instead of relying on a hard foreground-background mask.

Notably, BFR introduces no learnable parameters and **does not require additional training**, making it less affected by label imbalance in long-tailed recognition. By relying on feature statistics rather than extra supervision, BFR provides a lightweight way to rectify background-biased features.

4.3 Training Objective

For a given training sample (x, y), FG-CutMix generates K augmented samples and aggregates them into a single feature f^a with target label $\tilde{y}$. Meanwhile, BFR produces the rectified feature f' for the original sample x. The overall objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \gamma \mathcal{L}_{\mathrm{RL}}, \quad \mathcal{L}_{\mathrm{cls}} = \frac{1}{2}\left[\ell_{\mathrm{LT}}\big(h(f^a), \tilde{y}\big) + \ell_{\mathrm{LT}}\big(h(f'), y\big)\right], \tag{10}$$

where $h(\cdot)$ is the classifier head, ℓ_{LT} denotes the underlying long-tailed objective, $\mathcal{L}_{\mathrm{RL}}$ is the reinforcement learning loss (Eq. 5), and γ controls the strength of policy learning.

5 Experiments

5.1 Experiment Setup

Datasets and Metrics. Our proposed framework is evaluated on four long-tailed benchmarks: CIFAR-10-LT [4], CIFAR-100-LT [4], ImageNet-LT [52], and iNaturalist 2018 [79]. Following standard evaluation protocols [84, 105], we report Top-1 accuracy. For CIFAR datasets, we evaluate effectiveness of OFBD across three imbalance ratios ($r \in \{100, 50, 10\}$) and decompose results into "Many" (> 100 samples), "Med." ($20 \sim 100$), and "Few" (< 20) categories for further analysis. For ImageNet-LT and iNaturalist 2018, we assess various backbones to verify architecture generality.

Compared Methods. To demonstrate the effectiveness and plug-and-play versatility of OFBD, we integrate it into representative, open-source long-tailed methods to ensure reproducibility. These span

Table 2: Top-1 accuracy on ImageNet-LT [52] and iNaturalist 2018 [79].

Method	ImageNet-LT	iNature2018
ResNet-50 [20] backbone		
CE	44.80	65.95
+OFBD	47.39$_{\uparrow 2.59}$	69.76$_{\uparrow 3.81}$
BCL [110] (CVPR'22)	55.67	71.81
+OFBD	56.81$_{\uparrow 1.14}$	72.34$_{\uparrow 0.53}$
GCL [39] (CVPR'22)	55.46	70.76
+OFBD	56.34$_{\uparrow 0.88}$	71.92$_{\uparrow 1.16}$
GBG [43] (AAAI'24)	57.13	71.92
+OFBD	57.64$_{\uparrow 0.51}$	72.36$_{\uparrow 0.44}$
MKP [8] (CVPR'26)	57.92	74.41
+OFBD	58.51$_{\uparrow 0.59}$	74.63$_{\uparrow 0.22}$
ViT-B [14] backbone		
ViT [14] (ICLR'21)	37.51	54.23
+OFBD	39.17$_{\uparrow 1.66}$	56.62$_{\uparrow 2.39}$
DeiT-LT [63] (CVPR'24)	55.64	72.92
+OFBD	56.21$_{\uparrow 0.57}$	74.56$_{\uparrow 1.64}$

Table 3: Ablation study on main components.

Method	Many↑	Med.↑	Few↑	All↑
CE	65.12	36.60	9.06	38.32
+FG-CutMix+BFR	65.48	42.09	23.93	44.83$_{\uparrow 6.51}$
BCL	66.85	52.92	33.07	51.84
+FG-CutMix	67.11	52.11	38.84	53.38$_{\uparrow 1.54}$
+BFR	66.96	53.65	35.86	52.97$_{\uparrow 1.13}$
+FG-CutMix+BFR	66.79	53.46	40.34	54.19$_{\uparrow 2.35}$

Table 4: Further analysis of CutMix region selection. All variants use the same CutMix pipeline but differ in how the preserved region is selected.

CutMix Variant	Acc. (%)
Standard CutMix (random selection)	52.17
CutMix w/ SAM-based foreground selection	52.43
CutMix w/ CAM-based foreground selection	52.63
FG-CutMix w/ RL-based foreground selection	**54.19**

standard training (CE), re-balancing techniques (LDAM-DRW [4], BBN [107]), contrastive learning (SBCL [22], BCL [110]), and advanced representation optimization (GBG [43], MKP [8]). By consistently using standard ResNet backbones, this integration fairly and directly validates OFBD's ability to boost performance by mitigating background bias across diverse learning paradigms.

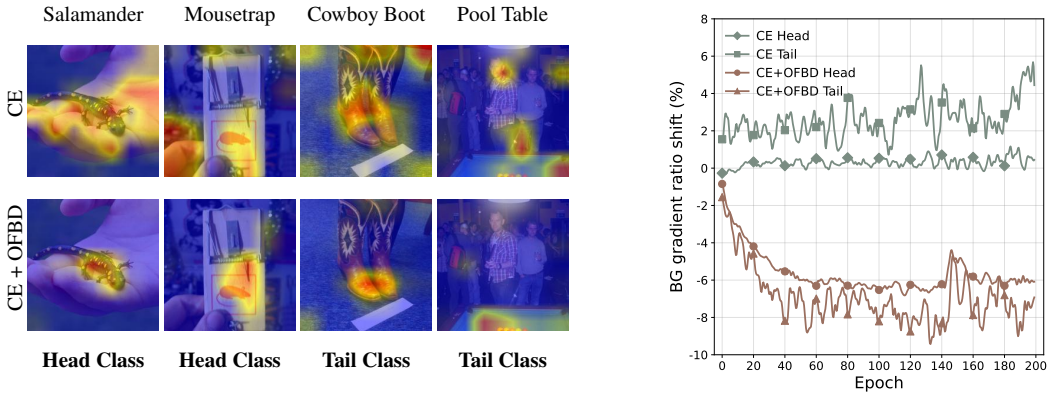
Implementation. For fairness, we reproduce baselines with official codes and fixed seeds. We use ResNet-32 for CIFAR10/100-LT [4], ResNet-50 [20] for ImageNet-LT and iNaturalist 2018, and ViT-B/16 [14] for cross-architecture evaluation. All models are trained on an RTX 4090 GPU with a batch size of 256. For OFBD, we set $\lambda = 0.5$, $\gamma = 0.8$, $\alpha = 1.0$, and $\beta = 0.02$ across datasets. Extensive ablation studies, further analyses, and implementation details are deferred to Appendix D.

5.2 Effectiveness as a Plug-in Module

CIFAR100-LT and CIFAR10-LT [4]. Table 1 reports results on CIFAR-10/100-LT after integrating OFBD into existing methods. Across 42 experimental settings, OFBD improves 40 of them, with the largest gains of $+6.51\%$ on CIFAR-100-LT and $+3.88\%$ on CIFAR-10-LT. For the standard CE baseline, OFBD improves CIFAR-100-LT by $+6.51\%$, $+3.85\%$, and $+2.51\%$ under $r = 100$, 50, and 10, respectively, and improves CIFAR-10-LT by $+3.88\%$, $+3.86\%$, and $+0.86\%$. For re-balancing methods, OFBD improves LDAM-DRW by up to $+3.72\%$ and BBN by up to $+2.36\%$ on CIFAR-100-LT. For contrastive-learning methods, OFBD improves BCL and SBCL by up to $+2.35\%$ and $+1.99\%$, respectively. For representation-optimization methods, OFBD improves GBG by up to $+2.31\%$ and MKP by $+1.21\%$ under high imbalance. These consistent gains verify the plug-in effectiveness of OFBD across diverse long-tailed learning paradigms.

The Many/Med./Few statistics on CIFAR-100-LT with $r = 100$ further show that OFBD mainly benefits data-scarce classes while preserving head-class performance. OFBD improves Few-class accuracy by $+14.79\%$ on CE, $+6.39\%$ on LDAM-DRW, $+5.04\%$ on BBN, $+7.27\%$ on BCL, $+5.22\%$ on SBCL, $+7.36\%$ on GBG, and $+2.38\%$ on MKP. In contrast, Many-class changes remain small for strong baselines (*e.g.*, -0.06% on BCL, -0.13% on GBG, and $+0.29\%$ on MKP). These results indicate that OFBD alleviates tail-class degradation without sacrificing head-class recognition.

ImageNet-LT and iNaturalist 2018 [52, 79]. Table 2 reports results on large-scale long-tailed datasets with both ResNet-50 and ViT-B backbones. With ResNet-50, OFBD consistently improves all baselines on both datasets. The gains are particularly large for the vanilla CE baseline, improving on ImageNet-LT from 44.80% to 47.39% ($+2.59\%$) and on iNaturalist 2018 from 65.95% to 69.76% ($+3.81\%$), showing that OFBD effectively complements standard training. For stronger baselines, OFBD still brings stable gains: $+1.14\%$, $+0.88\%$, $+0.51\%$, and $+0.59\%$ on ImageNet-LT for BCL, GCL, GBG, and MKP, and $+0.53\%$, $+1.16\%$, $+0.44\%$, and $+0.22\%$ on iNaturalist 2018. With ViT-B, OFBD also improves ViT and DeiT-LT by $+1.66\%$/$+0.57\%$ on ImageNet-LT and



(a) Attention Map

(b) Background Gradient Radio Shift

Figure 3: Visualization of background debiasing on CIFAR100-LT ($r = 100$). (a) Pretrained-CAM attention maps of CE and CE+OFBD on three representative head, medium, and tail classes, showing that OFBD reduces background co-activation. More visualization results are provided in Appendix C.1. (b) Background gradient ratio (Eq. 2) shift during training. OFBD reduces optimization-level background bias (Sec. 3.3), especially for tail classes.

$+2.39\%$/$+1.64\%$ on iNaturalist 2018, respectively. These consistent improvements across datasets, baselines, and CNN/Transformer backbones verify the generality of OFBD as a plug-in framework.

5.3 Ablation Study and Analysis

In this section, we present key ablation and analysis experiments to characterize the effectiveness, design choices, and debiasing behavior of OFBD. More experiments are provided in Appendix D.

Component Analysis. We conduct component ablation on CIFAR100-LT ($r = 100$) with ResNet-32. We choose CE and BCL [110] as representative baselines: CE reflects standard long-tailed training, while BCL represents a strong contrastive-learning baseline. As shown in Table 3, adding the full FG-CutMix+BFR framework to CE improves the overall accuracy from 38.32% to 44.83% ($+6.51\%$), mainly due to the large Few-class gain from 9.06% to 23.93%. On BCL, FG-CutMix and BFR improve the overall accuracy by $+1.54\%$ and $+1.13\%$, respectively. Combining them further improves BCL from 51.84% to 54.19% ($+2.35\%$), with Few-class accuracy increasing from 33.07% to 40.34%. These results show that both components are effective, especially for tail classes.

Analysis of RL-based foreground region selection. Built upon the BCL [110], we further analyze the region selection strategy in FG-CutMix on CIFAR100-LT ($r = 100$). As shown in Table 4, standard CutMix with random regions achieves only 52.17%, indicating that random replacement may preserve irrelevant backgrounds or discard target-related evidence. Using pretrained SAM- and CAM-based foreground regions selection improves the accuracy to 52.43% and 52.63%, respectively, but the gains remain limited. In contrast, our RL-based foreground regions selection achieves the best accuracy of 54.19%, outperforming random, SAM-based, and CAM-based selection by $+2.02\%$, $+1.76\%$, and $+1.56\%$, respectively. These results show that accurate foreground localization is not necessarily the most effective mixing strategy, while RL-based selection better balances target preservation and background replacement. Further ablation analysis is provided in Appendix D.7.

Visualization of Background Debiasing. We further visualize whether OFBD reduces background-biased representation and optimization. As shown in Fig. 3a, CE predictions co-activate target foregrounds and surrounding backgrounds. With OFBD, attention becomes more object-focused, indicating reduced reliance on irrelevant background cues. Fig. 3b shows that CE produces a positive and increasing background-gradient-ratio shift, particularly for tail classes, while CE+OFBD suppresses this shift for both head and tail classes, especially for tail classes. Overall, these results confirm that OFBD successfully mitigates background bias, encouraging the model to focus more on target semantics rather than spurious background cues.

Table 5: Validation of B_u on ImageNet-LT.

Group	B_u-FG↓	B_u-BG↑	BG AUROC↑
Overall	0.3657	0.6321	0.7182
Many	0.3639	0.6340	0.7247
Medium	0.3651	0.6315	0.7161
Few	0.3726	0.6292	0.7049

Table 6: Background intervention analysis on ImageNet-LT.

Method	FG-only↑	BG-only↓	BG-swap↑
BCL [110]	40.18	20.15	31.21
+OFBD	41.99$_{\uparrow 1.81}$	18.91$_{\downarrow 1.24}$	38.23$_{\uparrow 7.02}$

Table 7: Foreground selection comparison.

Method	Overall↑	Many↑	Med.↑	Few↑
SaliencyMix [78]	52.85	67.62	53.47	34.90
SnapMix [25]	52.01	67.45	53.94	31.75
PuzzleMix [33]	52.52	67.06	53.80	34.06
ResizeMix [62]	52.03	67.24	53.91	32.96
Attentive CutMix [81]	53.19	**68.92**	53.99	33.91
RL selector	**54.21**	66.51	**54.00**	**40.10**

Table 8: Results with different imbalance ratios.

r	Method	Many↑	Few↑	Overall↑
200	BCL [110]	66.60	26.33	46.30
	+OFBD	65.37$_{\downarrow 1.23}$	30.08$_{\uparrow 3.75}$	47.20$_{\uparrow 0.90}$
300	BCL [110]	68.00	26.09	44.62
	+OFBD	65.10$_{\downarrow 2.90}$	29.59$_{\uparrow 3.50}$	45.78$_{\uparrow 1.16}$
400	BCL [110]	68.33	24.04	42.99
	+OFBD	64.60$_{\downarrow 3.73}$	27.70$_{\uparrow 3.66}$	44.16$_{\uparrow 1.17}$

5.4 Further Analysis of Background Debiasing

Background Score Validation. We evaluate whether B_u captures background-related information using CAM-guided SAM masks. Specifically, CAM responses from a pretrained classifier are used as point prompts for SAM to obtain foreground/background masks. As shown in Table 5, B_u assigns higher average scores to background regions than foreground regions and achieves BG AUROC above 0.70 across all groups, demonstrating its ability to identify background-biased features.

Background Intervention. To verify whether OFBD reduces background shortcut reliance, we conduct intervention-based evaluations on ImageNet-LT [52] using foreground-only, background-only, and background-swapped inputs with fixed post-hoc foreground masks. As shown in Table 6, OFBD improves FG-only accuracy by +1.81 points while reducing BG-only accuracy by 1.24 points, indicating a more object-focused representation. Moreover, it achieves a +7.02-point gain under BG-swap, demonstrating improved robustness against background bias under distribution shifts.

Foreground Selection Analysis. We compare the proposed RL-based foreground selector with representative mixing strategies, including SaliencyMix [78], SnapMix [25], PuzzleMix [33], ResizeMix [62], and Attentive CutMix [81]. As shown in Table 7, the RL selector achieves the best overall accuracy and Few-shot performance, improving Few-shot accuracy by +5.20 points over the strongest heuristic baseline, demonstrating the effectiveness of adaptive foreground preservation for long-tailed recognition. This improvement highlights the advantage of learning region selection over manually designed heuristic strategies.

Performance under Different Imbalance Ratios. We further evaluate OFBD under more severe imbalance settings by increasing the imbalance ratio on CIFAR-100-LT. As shown in Table 8, OFBD consistently improves overall and Few-shot accuracy across different imbalance ratios. Specifically, OFBD improves overall accuracy by +0.90/+1.16/+1.17 points under $r = 200/300/400$, while achieving +3.75/+3.50/+3.66 points gains on Few-shot classes. These results demonstrate that OFBD remains effective under increasingly challenging long-tailed distributions.

6 Conclusion

Revisiting long-tailed recognition via background bias, we reveal that tail degradation arises from sample scarcity combined with background-biased representation and optimization (i.e., larger distribution shifts and gradient reliance). To mitigate this, we propose Object-Focused Background Debiasing (OFBD), a plug-in framework. Specifically, FG-CutMix preserves target foregrounds while diversifying backgrounds, and parameter-free BFR suppresses background-biased features. Experiments across four datasets demonstrate that OFBD improves mainstream methods, boosting tail accuracy without sacrificing head performance on both CNN and Transformer backbones. Its plug-in design inspires extensions to other long-tailed tasks like object detection and trajectory prediction.

Limitation: the performance of OFBD on domains with highly complex scenes (*e.g.*, medical imaging datasets) remains to be explored.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. T2541059 and 62676222), the Shandong Excellent Young Scientists Fund (Grant No. ZR2024YQ006), the Taishan Scholars Program (Grant No. tsqn202507014), the Zhejiang Leading Innovative and Entrepreneur Team Introduction Program (Grant No. 2024R01007), and the "Pioneer" and "Leading Goose" Research and Development Program of Zhejiang (Grant No. 2025C02077).

References

[1] Khaled Alomar, Halil Ibrahim Aysel, and Xiaohao Cai. Data Augmentation in Classification and Segmentation: A Survey and New Strategies. *Journal of Imaging*, 9(2):46, 2023.

[2] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit Regularization in Deep Matrix Factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] Rwiddhi Chakraborty, Yinong Wang, Jialu Gao, Runkai Zheng, Cheng Zhang, and Fernando De la Torre. Visual Data Diagnosis and Debiasing with Concept Graphs. *Advances in Neural Information Processing Systems*, 37:106383–106410, 2024.

[6] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. An Attentive Survey of Attention Models. *ACM Transactions on Intelligent Systems and Technology*, 12(5):1–32, 2021.

[7] Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. A Closer Look at Multimodal Representation Collapse. In *International Conference on Machine Learning*, 2025.

[8] Shenghan Chen, Yiming Liu, Yanzhen Wang, Yujia Wang, and Xiankai Lu. Reframing Long-Tailed Learning via Loss Landscape Geometry. *arXiv preprint arXiv:2603.21217*, 2026.

[9] Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. Adamv-Moe: Adaptive Multi-task Vision Mixture-of-Experts. In *IEEE/CVF International Conference on Computer Vision*, pages 17346–17357, 2023.

[10] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. AutoAugment: Learning Augmentation Strategies from Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[11] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-Balanced Loss Based on Effective Number of Samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[12] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Fang Zeng, Wei Liu, et al. AugGPT: Leveraging Chatgpt for Text Data Augmentation. *IEEE Transactions on Big Data*, 11(3):907–918, 2025.

[13] Qi Dong, Shaogang Gong, and Xiatian Zhu. Class Rectification Hard Mining for Imbalanced Deep Learning. In *IEEE/CVF International Conference on Computer Vision*, 2017.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2020.

[15] Simon Shaolei Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-Shot Learning via Learning the Representation, Provably. In *International Conference on Learning Representations*, 2021.

[16] 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.

[17] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. ImageNet-trained CNNs are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness. In *International Conference on Learning Representations*, 2018.

[18] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[19] Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. Faster AutoAugment: Learning Augmentation Strategies Using Backpropagation. In *European Conference on Computer Vision*, pages 1–16. Springer, 2020.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[21] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling Label Distribution for Long-Tailed Visual Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[22] Chengkai Hou, Jieyu Zhang, Haonan Wang, and Tianyi Zhou. Subclass-balancing Contrastive Learning for Long-tailed Recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 5395–5407, 2023.

[23] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate Attention for Efficient Mobile Network Design. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13708–13717. IEEE, 2021.

[24] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning Deep Representation for Imbalanced Classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[25] Shaoli Huang, Xinchao Wang, and Dacheng Tao. Snapmix: Semantically Proportional Mixing for Augmenting Fine-grained Data. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 1628–1636, 2021.

[26] Ta Duc Huy, Duy Anh Huynh, Yutong Xie, Yuankai Qi, Qi Chen, Phi Le Nguyen, Sen Kim Tran, Son Lam Phung, Anton van den Hengel, Zhibin Liao, Minh-Son To, Johan W. Verjans, and Vu Minh Hieu Phan. Seeing the Trees for the Forest: Rethinking Weakly-Supervised Medical Visual Grounding. In *IEEE/CVF International Conference on Computer Vision*, pages 24445–24455, 2025.

[27] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[28] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A Survey on Contrastive Self-Supervised Learning. *Technologies*, 9(1):2, 2020.

[29] Zhongquan Jian, Yanhao Chen, Yancheng Wang, Junfeng Yao, Meihong Wang, and Qingqiang Wu. Supervised Exploratory Learning for Long-tailed Visual Recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 1870–1880, 2025.

[30] Zheheng Jiang, Hossein Rahmani, Sue Black, and Bryan M Williams. A Probabilistic Attention Model with Occlusion-aware Texture Regression for 3D Hand Reconstruction from a Single RGB Image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 758–767, 2023.

[31] Hyungsik Jung and Youngrock Oh. Towards Better Explanations of Class Activation Mapping. In *IEEE/CVF International Conference on Computer Vision*, pages 1336–1344, 2021.

[32] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring Balanced Feature Spaces for Representation Learning. In *International Conference on Learning Representations*, 2020.

[33] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting Saliency and Local Statistics for Optimal Mixup. In *International Conference on Learning Representations*, pages 5275–5285. PMLR, 2020.

[34] Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and Mitigating Visual Biases through Keyword Explanation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11082–11092, 2024.

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[36] Vijay Konda and John Tsitsiklis. Actor-Critic Algorithms. *Advances in Neural Information Processing Systems*, 12, 1999.

[37] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep Learning. *Nature*, 521(7553): 436–444, 2015.

[38] Haoxin Li, Yuan Liu, Hanwang Zhang, and Boyang Li. Mitigating and Evaluating Static Bias of Action Representations in the Background and the Foreground. In *IEEE/CVF International Conference on Computer Vision*, pages 19911–19923, 2023.

[39] Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed Visual Recognition via Gaussian Clouded Logit Adjustment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6929–6938, 2022.

[40] 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 Conference on Artificial Intelligence*, 2024.

[41] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. MetaSAug: Meta Semantic Augmentation for Long-Tailed Visual Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[42] 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 *International Conference on Machine Learning*, 2025.

[43] Weiqi Li, Fan Lyu, Fanhua Shang, Liang Wan, and Wei Feng. Long-Tailed Learning as Multi-Objective Optimization. In *AAAI Conference on Artificial Intelligence*, 2024.

[44] Yichuan Li, Kaize Ding, Jianling Wang, and Kyumin Lee. Empowering Large Language Models for Textual Data Augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12734–12751, 2024.

[45] Zhixin Li and Yuheng Jia. ConMix: Contrastive Mixup at Representation Level for Long-tailed Deep Clustering. In *International Conference on Learning Representations*, 2025.

[46] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast AutoAugment. *Advances in Neural Information Processing Systems*, 32, 2019.

[47] 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 *European Conference on Computer Vision*, pages 740–755, 2014.

[48] Bo Liu, Haoxiang Li, Hao Kang, Gang Hua, and Nuno Vasconcelos. GistNet: a Geometric Structure Transfer Network for Long-tailed Recognition. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[49] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-Supervised Learning: Generative or Contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876, 2021.

[50] 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.

[51] Yuting Liu, Liu Yang, and Yu Wang. Long-Tailed Classification with Multi-Granularity Semantics. In *IEEE/CVF International Conference on Computer Vision*, pages 4285–4294, 2025.

[52] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-Scale Long-Tailed Recognition in an Open World. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.

[53] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-Tail Learning via Logit Adjustment. In *International Conference on Learning Representations*, 2020.

[54] Matthias Minderer, Olivier Bachem, Neil Houlsby, and Michael Tschannen. Automatic Shortcut Removal for Self-supervised Representation Learning. In *International Conference on Machine Learning*, pages 6927–6937. PMLR, 2020.

[55] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-Aware Contrastive Learning for Debiased Scene Representation. *Advances in Neural Information Processing Systems*, 34:12251–12264, 2021.

[56] Seung Jun Moon, Sangwoo Mo, Kimin Lee, Jaeho Lee, and Jinwoo Shin. MASKER: Masked Keyword Regularization for Reliable Text Classification. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 13578–13586, 2021.

[57] Kartik Narayan, Vibashan Vs, and Vishal M Patel. SegFace: Face Segmentation of Long-tail Classes. In *AAAI Conference on Artificial Intelligence*, 2025.

[58] Haolin Pan, Yong Guo, Mianjie Yu, and Jian Chen. Enhanced Long-Tailed Recognition with Contrastive Cutmix Augmentation. *IEEE Transactions on Image Processing*, 33:4215–4230, 2024.

[59] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck Attention Module. *arXiv preprint arXiv:1807.06514*, 2018.

[60] Minlong Peng, Qi Zhang, Xiaoyu Xing, Tao Gui, Xuanjing Huang, Yu-Gang Jiang, Keyu Ding, and Zhigang Chen. Trainable Undersampling for Class-Imbalance Learning. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 4707–4714, 2019.

[61] Xiaotian Qiao, Quanlong Zheng, Ying Cao, and Rynson WH Lau. Tell Me Where I Am: Object-level Scene Context Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2633–2641, 2019.

[62] Jie Qin, Jiemin Fang, Qian Zhang, Wenyu Liu, Xingang Wang, and Xinggang Wang. Resizemix: Mixing Data with Preserved Object Information and True Labels. *arXiv preprint arXiv:2012.11101*, 2020.

[63] Harsh Rangwani, Pradipto Mondal, Mayank Mishra, Ashish Ramayee Asokan, and R Venkatesh Babu. DeiT-LT: Distillation Strikes Back for Vision Transformer Training on Long-tailed Datasets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23396–23406, 2024.

[64] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment Anything in Images and Videos. In *International Conference on Learning Representations*, 2025.

[65] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced Meta-Softmax for Long-tailed Visual Recognition. In *Advances in Neural Information Processing Systems*, 2020.

[66] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[67] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*, 2019.

[68] Dvir Samuel and Gal Chechik. Distributional Robustness Loss for Long-tail Learning. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[69] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[70] Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. DiffuLT: Diffusion for Long-tail Recognition Without External Knowledge. In *Advances in Neural Information Processing Systems*, 2024.

[71] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

[72] Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not Using the Car to See the Sidewalk–Quantifying and Controlling the Effects of Context in Classification and Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8218–8226, 2019.

[73] Cheng Shi, Yizhou Yu, and Sibei Yang. Vision Transformers Need More Than Registers. *arXiv preprint arXiv:2602.22394*, 2026.

[74] Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. How Re-sampling Helps for Long-tail Learning? *Advances in Neural Information Processing Systems*, 36:75669–75687, 2023.

[75] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Xin-Yan Han, Jie-Jing Shao, and Yufeng Li. Parameter-Efficient Long-Tailed Recognition. *CoRR*, 2023.

[76] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[77] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect. *Advances in Neural Information Processing Systems*, 33:1513–1524, 2020.

[78] AFM Uddin, Mst Monira, Wheemyung Shin, TaeChoong Chung, Sung-Ho Bae, et al. Saliencymix: A Saliency Guided Data Augmentation Strategy for Better Regularization. *arXiv preprint arXiv:2006.01791*, 2020.

[79] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[80] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*, 2018(1):7068349, 2018.

[81] Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. Attentive cutmix: An Enhanced Data Augmentation Approach for Deep Learning based Image Classification. *arXiv preprint arXiv:2003.13048*, 2020.

[82] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning Robust Global Representations by Penalizing Local Predictive Power. *Advances in Neural Information Processing Systems*, 32, 2019.

[83] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive Learning Based Hybrid Networks for Long-Tailed Image Classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 943–952, 2021.

[84] 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 *Advances in Neural Information Processing Systems*, 2024.

[85] Weitian Wang, Rai Shubham, Cecilia De La Parra, and Akash Kumar. MixA-Q: Revisiting Activation Sparsity for Vision Transformers from a Mixed-precision Quantization Perspective. In *IEEE/CVF International Conference on Computer Vision*, pages 22143–22152, 2025.

[86] Xuan Wang and Zhigang Zhu. Context Understanding in Computer Vision: A Survey. *Computer Vision and Image Understanding*, 229:103646, 2023.

[87] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-Tailed Recognition by Routing Diverse Distribution-Aware Experts. In *International Conference on Learning Representations*, 2020.

[88] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys (csur)*, 53(3):1–34, 2020.

[89] Yulin Wang, Kangchen Lv, Rui Huang, Shiji Song, Le Yang, and Gao Huang. Glance and Focus: a Dynamic Approach to Reducing Spatial Redundancy in Image Classification. *Advances in Neural Information Processing Systems*, 33:2432–2444, 2020.

[90] Zaitian Wang, Jinghan Zhang, Xinhao Zhang, Kunpeng Liu, Pengfei Wang, and Yuanchun Zhou. Diversity-oriented Data Augmentation with Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22265–22283, 2025.

[91] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. CReST: A Class-rebalancing Self-training Framework for Imbalanced Semi-supervised Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[92] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. In *European Conference on Computer Vision*, pages 3–19, 2018.

[93] Tz-Ying Wu, Pedro Morgado, Pei Wang, Chih-Hui Ho, and Nuno Vasconcelos. Solving Long-tailed Recognition with Deep Realistic Taxonomic Classifier. In *European Conference on Computer Vision*, 2020.

[94] Zifan Wu, Chao Yu, Deheng Ye, Junge Zhang, Hankz Hankui Zhuo, et al. Coordinated Proximal Policy Optimization. *Advances in Neural Information Processing Systems*, 34: 26437–26448, 2021.

[95] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or Signal: The Role of Image Backgrounds in Object Recognition. In *International Conference on Learning Representations*, 2020.

[96] Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. SimAm: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In *International Conference on Machine Learning*, pages 11863–11874. PMLR, 2021.

[97] Songxiao Yang, Haolin Wang, Yao Fu, Ye Tian, Tamostu Kamishima, Masayuki Ikebe, Yafei Ou, and Masatoshi Okutomi. RAM-W600: A Multi-Task Wrist Dataset and Benchmark for Rheumatoid Arthritis. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

[98] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *IEEE/CVF International Conference on Computer Vision*, October 2019.

[99] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are Adaptive Methods Good for Attention Models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

[100] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised Aggregation of Diverse Experts for Test-agnostic Long-tailed Recognition. *Advances in Neural Information Processing Systems*, 35:34077–34090, 2022.

[101] 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.

[102] Zizhao Zhang and Tomas Pfister. Learning Fast Sample Re-weighting without Reward Data. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[103] Qihao Zhao, Chen Jiang, Wei Hu, Fan Zhang, and Jun Liu. MDCS: More Diverse Experts with Consistency Self-Distillation for Long-tailed Recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 11563–11574. IEEE, 2023.

[104] Qihao Zhao, Yalun Dai, Hao Li, Wei Hu, Fan Zhang, and Jun Liu. LTGC: Long-tail Recognition via Leveraging LLMs-driven Generated Content. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[105] Shizhen Zhao, Xin Wen, Jiahui Liu, Chuofan Ma, Chunfeng Yuan, and Xiaojuan Qi. Learning from Neighbors: Category Extrapolation for Long-tail Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 30483–30492, 2025.

[106] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[107] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch Network with Cumulative Learning for Long-tailed Visual Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[108] Haoyi Zhou, Jianxin Li, Jieqi Peng, Shuai Zhang, and Shanghang Zhang. Triplet Attention: Rethinking the Similarity in Transformers. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2378–2388, 2021.

[109] Zhipeng Zhou, Lanqing Li, Peilin Zhao, Pheng-Ann Heng, and Wei Gong. Class-conditional Sharpness-aware Minimization for Deep Long-tailed Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3499–3509, 2023.

[110] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced Contrastive Learning for Long-Tailed Visual Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[111] Ke Zhu, Minghao Fu, Jie Shao, Tianyu Liu, and Jianxin Wu. Rectify the Regression Bias in Long-Tailed Object Detection. In *European Conference on Computer Vision*, 2024.

OFBD: Object-Focused Background Debiasing for Long-tailed Learning

Overview: This supplementary material provides additional mathematical derivations (Appendix B), extensive visualization results (Appendix C), additional experimental results and ablations (Appendix D), and comprehensive implementation details (Appendix A) to support the main paper.

A More Experiment Protocols

A.1 Implementation Details

OFBD is implemented as a plug-and-play module and can be integrated into various long-tailed recognition frameworks without altering their original backbones or classifier designs.

For CIFAR-10-LT and CIFAR-100-LT, we use ResNet-32 as the backbone. Models are trained for 200 epochs with a batch size of 256 using SGD with momentum 0.9 and weight decay 5×10^{-4}. The learning rate is linearly warmed up to 0.15 during the first 20 epochs and decayed by 0.1 at epochs 160 and 180. When OFBD is plugged into contrastive-learning baselines, *e.g.*, BCL [110] and SBCL [22], we keep their original contrastive objectives unchanged. The projection dimension is 128, the contrastive temperature is 0.1, and the classification and contrastive losses are weighted by 2 and 0.6, following the original baseline settings [22, 110]. For non-contrastive baselines, the projection head and contrastive loss are removed.

For FG-CutMix, we apply the augmentation with probability 0.5. Given an input image, we randomly generate a region proposal set $\mathcal{P}(x) = \{b_1, \ldots, b_N\}$ with different locations and scales. Instead of ranking proposals by local energy, the RL-based region selector π_θ predicts a selection probability over $\mathcal{P}(x)$, and K proposals are sampled without replacement to form the foreground candidate pool $\mathcal{R}_\theta(x) = \{r_1, \ldots, r_K\}$. For each selected candidate, we construct a mixed sample by preserving the selected target-related region and replacing its complementary region with content from another image. The mixed label is corrected in line with prior CutMix-based augmentation methods.

For selector optimization, the classifier prediction on each mixed sample provides the reward signal. Specifically, the reward is set to 1 if the mixed sample has a higher predicted probability for the target foreground-source class than for the background-source class, and 0 otherwise. Rewards are normalized within each candidate pool to compute relative advantages. The selector is optimized by policy gradient with a clipped objective and KL regularization to a frozen reference selector π_{ref}, which is initialized from the previous epoch selector to stabilize policy updates. Unless otherwise specified, we set the RL loss weight $\gamma = 0.8$, the auxiliary selector supervision weight $\alpha = 1.0$, the KL coefficient $\beta = 0.02$, and the clipping range $\epsilon_c = 0.2$.

For BFR, we use the final convolutional feature map as input. Neuron-wise importance scores are computed from channel-wise feature statistics. Effective channels are selected by comparing each neuron-wise score with the spatial mean score, and the resulting foreground contribution is used to derive background-aware scores. The background-biased component is then estimated by background-aware weighted aggregation, and the rectified representation is computed as $f' = f - \lambda f_{\mathrm{bg}}^{\mathrm{BFR}}$, where $\lambda = 0.5$ in all experiments. The rectified representation is fed into the classifier and, when applicable, the projection head. BFR introduces no learnable parameters and requires no external annotations, pretrained localization models, or additional supervision.

For ImageNet-LT and iNaturalist 2018, we use ResNet-50 [20] and ViT-B/16 [14] as backbones. Models are trained with SGD, momentum 0.9, and batch size 256. For ImageNet-LT, we train for 90 epochs with a learning rate of 0.1 and weight decay 5×10^{-4}. For iNaturalist 2018, we train for 100 epochs with a learning rate of 0.2 and a weight decay of 1×10^{-4}. Unless otherwise specified, the same OFBD plug-in hyperparameters are used across datasets.

A.2 Dataset and Evaluation Protocol

We evaluate OFBD on CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018. For CIFAR-10-LT and CIFAR-100-LT, we use imbalance ratios $r \in \{100, 50, 10\}$. The original balanced test sets are used for evaluation. For ImageNet-LT and iNaturalist 2018, we follow the official

train/validation splits [101]. We report Top-1 accuracy for all datasets. For class-frequency analysis, we report Many-, Medium-, and Few-shot accuracy, where classes with more than 100 samples are Many-shot, classes with 20–100 samples are Medium-shot, and classes with fewer than 20 samples are Few-shot. For plug-in comparisons, OFBD keeps the official baseline's backbone, objective, classifier, sampler, and training schedule.

B Theoretical Derivations and Analysis

B.1 Interpretation of Distribution-level Background Bias (Eq. 1)

In Sec. 3, we hypothesized that tail classes suffer from a larger background distribution shift in Eq. 1 and further corroborate this hypothesis through empirical observations. Here, we provide an interpretation to mathematically substantiate this claim based on sample complexity, deeply aligned with findings on spurious causal correlations in imbalanced data [77]. Let N_y denote the number of training samples for class y. In a long-tailed recognition, we have $N_{head} \gg N_{tail}$.

1. Deviation of Empirical Distributions.
Let q_y denote the ideal background distribution for class y constructed from a fully balanced, infinite-sample dataset. Under the long-tailed setting, the model observes an empirical background distribution, denoted as p_y, which is estimated from only N_y samples.

According to statistical learning theory [4, 109, 42], the deviation between an empirical distribution and the true underlying distribution is bounded by the sample size. Specifically, the expected L_1 distance between p_y and q_y converges at a rate of $\mathcal{O}\left(1/\sqrt{N_y}\right)$:

$$\mathbb{E}\left[\|p_y - q_y\|_1\right] \leq \mathcal{O}\left(\frac{1}{\sqrt{N_y}}\right). \tag{11}$$

Since $N_{head} \gg N_{tail}$, the expected distribution deviation for tail classes is significantly larger than that for head classes [4]. This provides the theoretical basis for the inequality:

$$\|p_y^{\text{tail}} - q_y\|_1 \gg \|p_y^{\text{head}} - q_y\|_1. \tag{12}$$

2. Bound on Background Feature Bias.
Next, we relate this distribution-level deviation to the feature-level background bias $\Delta D_{bg}(y)$. By definition, the expected background feature means are

$$\mu_{bg}(y) = \mathbb{E}_{b\sim p_y}[f_{bg}], \quad \bar{\mu}_{bg}(y) = \mathbb{E}_{b\sim q_y}[f_{bg}]. \tag{13}$$

The background bias $\Delta D_{bg}(y)$ can be expanded as

$$\begin{aligned} \|\Delta D_{bg}(y)\|_1 &= \|\mu_{bg}(y) - \bar{\mu}_{bg}(y)\|_1 \\ &= \left\| \int f_{bg}\, p_y(b)\, db - \int f_{bg}\, q_y(b)\, db \right\|_1 \\ &= \left\| \int f_{bg} \left(p_y(b) - q_y(b)\right) db \right\|_1 \end{aligned} \tag{14}$$

Assuming the feature representations f_{bg} extracted by a standard deep neural network are inherently bounded (e.g., strictly constrained by non-linear activation functions or normalization layers [27, 3]), there exists a constant $M > 0$ such that $\sup_b \|f_{bg}\|_1 \leq M$. Using Hölder's inequality, we obtain an upper bound on the feature bias:

$$\|\Delta D_{bg}(y)\|_1 \leq \int \|f_{bg}\|_1 \cdot |p_y(b) - q_y(b)| db \leq M \int |p_y(b) - q_y(b)| db = M\|p_y - q_y\|_1. \tag{15}$$

This bound indicates that larger background distribution deviation can increase the possible magnitude of background feature shift, which supports the hypothesis in Eq. 1. In particular, since tail classes have fewer samples, their empirical background distributions can deviate more from the balanced reference, potentially leading to larger background-biased feature shifts compared with head classes. Hence, this motivates a targeted debiasing mechanism for tail classes, as standard empirical risk minimization (ERM) may overfit spurious background correlations for minority categories [67].

B.2 Formal Analysis of Optimization-level Background Bias (Eq. 3)

We provide a concise justification for Eq. 3 from spectral bias and implicit low-rank regularization [2, 7]. Following Sec. 3.1, we decompose the feature as $f = f_{\text{fg}} + f_{\text{bg}}$, and define:

$$R_t(y) = \frac{1}{1+\gamma_t(y)}, \quad \gamma_t(y) = \frac{G_{\text{fg},t}(y)}{G_{\text{bg},t}(y)}. \tag{16}$$

Spectral bias and low-rank learning dynamics. Under gradient descent, deep networks exhibit spectral bias, favoring directions with larger singular values [2]. Let $a_k(t)$ denote the alignment of the model with the k-th spectral component. Its evolution approximately follows:

$$\frac{d}{dt}a_k(t) \propto \sigma_k^2 a_k(t), \tag{17}$$

which leads to exponentially faster growth along dominant directions. Empirically, deep representations exhibit low-rank structures during training [7], meaning that most energy concentrates in a few principal components. Background features, due to spatial redundancy and shared patterns across samples, naturally form a low-rank structure with dominant singular values. In contrast, foreground features are more diverse and distributed across many weaker components. As a result, gradient descent preferentially fits background-related components early in training, establishing an inherent optimization bias toward low-rank structures.

Effect of sample size on effective rank. For class y, the effective rank of foreground features depends on the sample size N_y. When N_y is large (head classes), diverse foreground patterns are observed, allowing multiple spectral components of f_{fg} to be learned. In contrast, for tail classes with limited samples, the observed foreground variations are restricted, reducing the usable rank of f_{fg}. This further amplifies the low-rank bias: the background components remain stable due to cross-class sharing, while the high-rank foreground structure becomes increasingly difficult to capture. Consequently, optimization for tail classes gradually shifts toward low-rank (background) directions.

Evolution of gradient ratio. Combining spectral bias and effective-rank reduction, the relative strength between foreground and background gradients evolves multiplicatively over time. Following standard analyses of gradient flow in linearized regimes [2], the ratio $\gamma_t(y)$ can be approximated as:

$$\gamma_t(y) \approx \gamma_0(y)\exp\big((\sigma_{\text{fg}}^2 - \sigma_{\text{bg}}^2)t\big), \tag{18}$$

where the exponent reflects the difference in effective learning rates. For tail classes, the reduced effective rank suppresses foreground components, leading to $\sigma_{\text{fg}}^2 < \sigma_{\text{bg}}^2$, and thus $\gamma_t(y) \to 0$. Substituting into $R_t(y)$ gives $R_t(y) \to 1$, implying $\Delta R_t(y) > 0$. For head classes, sufficient samples preserve the foreground rank and prevent $\gamma_t(y)$ from collapsing. Meanwhile, background components are fitted early and their gradients diminish over time, leading to a non-increasing $R_t(y)$ (*i.e.*, $\Delta R_t(y) \simeq 0$). Therefore:

$$\Delta R_t^{\text{tail}} > \Delta R_t^{\text{head}}. \tag{19}$$

B.3 Detailed Derivation of $S_{d,u}$ in Eq. 6

In long-tailed visual recognition, models tend to overfit to background feature, particularly for tail classes. Addressing this optimization-level bias typically requires extra modules. However, introducing learnable parameters often exacerbates class imbalance due to biased gradient updates. To rectify this without additional parameters, we propose estimating the inherent foreground distinctiveness of each neuron based on its feature statistics. Inspired by the energy-based local spatial formulation in SimAM [96], we hypothesize that target-related foreground features exhibit distinctive activation patterns compared to the widely distributed, background feature. To quantify this foreground distinctiveness, we measure the linear separability between the target neuron $F_{d,u}$ and the remaining background neurons $F_{d,v}$ ($v \neq u$). Assigning a target label of 1 to $F_{d,u}$ and -1 to all other neurons, the energy function to optimize the linear transform weights (w, b) is formulated as:

$$e(w,b) = \frac{1}{M-1}\sum_{v\neq u}(-1-(wF_{d,v}+b))^2 + (1-(wF_{d,u}+b))^2 + \eta w^2, \tag{20}$$

where ηw^2 serves as an L_2 regularizer. By taking the partial derivatives of the objective with respect to w and b and setting them to zero, we obtain a fast closed-form solution:

$$w^* = -\frac{2(F_{d,u} - \mu_t)}{(F_{d,u} - \mu_t)^2 + 2\sigma_t^2 + 2\eta}, \quad b^* = -\frac{1}{2}(F_{d,u} + \mu_t)w^*, \tag{21}$$

where μ_t and σ_t^2 are the mean and variance computed over all neurons in the channel except the target $F_{d,u}$. To circumvent the computational overhead of iterative spatial calculations, we approximate them using the global channel mean μ_d and variance σ_d^2 calculated over all M neurons. Substituting w^* and b^* back yields the minimal energy e^*:

$$e^* = \frac{4(\sigma_d^2 + \eta)}{(F_{d,u} - \mu_d)^2 + 2\sigma_d^2 + 2\eta}. \tag{22}$$

A lower minimal energy e^* mathematically indicates that the target neuron $F_{d,u}$ strongly deviates from the background-dominated channel context. While prior work treats this merely as general visual saliency [96, 92, 6], we repurpose it as a highly reliable, unsupervised proxy for foreground information in imbalanced learning. Thus, the inherent foreground contribution of the neuron can be represented by the inverse of the minimal energy ($1/e^*$):

$$\frac{1}{e^*} = \frac{(F_{d,u} - \mu_d)^2 + 2\sigma_d^2 + 2\eta}{4(\sigma_d^2 + \eta)} = \frac{(F_{d,u} - \mu_d)^2}{4(\sigma_d^2 + \eta)} + 0.5. \tag{23}$$

By omitting the constant scaling offset (0.5) and replacing the regularizer η with ϵ for numerical stability, we map this dynamic distinctiveness through a sigmoid function. This directly yields the bounded score $S_{d,u}$ utilized in our main framework (Eq. 6):

$$S_{d,u} = \text{sigmoid}\left(\frac{(F_{d,u} - \mu_d)^2}{4(\sigma_d^2 + \epsilon)}\right). \tag{24}$$

Additionally, we emphasize a fundamental distinction from conventional attention modules [99, 30] regarding **gradient bias and parameterization**. Standard attention modules rely on *learnable* parameters to generate dynamic scores. Under severe long-tailed imbalance, the gradient updates for these parameters are overwhelmingly dominated by head classes. Consequently, to minimize the overall empirical risk, these modules inevitably learn to attend to and universally amplify spurious head-class backgrounds, inadvertently acting as "bias amplifiers" In contrast, our approach derives $S_{d,u}$ purely from the *intrinsic spatial statistics* of the feature map, rendering it strictly parameter-free and immune to class-imbalanced gradient domination. By aggregating these statistical scores into a deterministic background penalty map B_u (Eq. 7), our BFR module shifts from generic feature reweighting to explicit spatial decoupling. It mathematically penalizes spurious background-foreground co-occurrences (Eq. 9), preventing tail-class representations from collapsing into dominant background features.

B.4 Why OFBD Mitigates Background Bias

We further justify that the two components of OFBD directly reduce the two bias terms derived in Sec. B.1 and Sec. B.2. Specifically, FG-CutMix mitigates the distribution-level bias in Eq. 1, while BFR suppresses the optimization-level bias in Eq. 2.

FG-CutMix reduces distribution-level background Bias. From Sec. B.1, the background feature bias is upper-bounded by the background distribution deviation. Thus, it is sufficient to show that FG-CutMix reduces $\|p_y - q_y\|_1$. Let $\tilde{p}_y$ denote the effective background distribution induced by FG-CutMix. Since FG-CutMix preserves the target foreground and replaces the complementary background, $\tilde{p}_y$ can be written as:

$$\tilde{p}_y = (1 - \eta_y)p_y + \eta_y p_{\text{mix}}, \quad \eta_y \in [0, 1], \tag{25}$$

where p_{mix} denotes the background distribution introduced by mixed samples. Since p_{mix} aggregates backgrounds from more samples, it provides a closer estimate of the ideal background distribution:

$$\|p_{\text{mix}} - q_y\|_1 \leq \|p_y - q_y\|_1. \tag{26}$$

By the convexity of the L_1 norm:

$$\|\tilde{p}_y - q_y\|_1 \leq (1 - \eta_y)\|p_y - q_y\|_1 + \eta_y\|p_{\text{mix}} - q_y\|_1 \leq \|p_y - q_y\|_1. \tag{27}$$

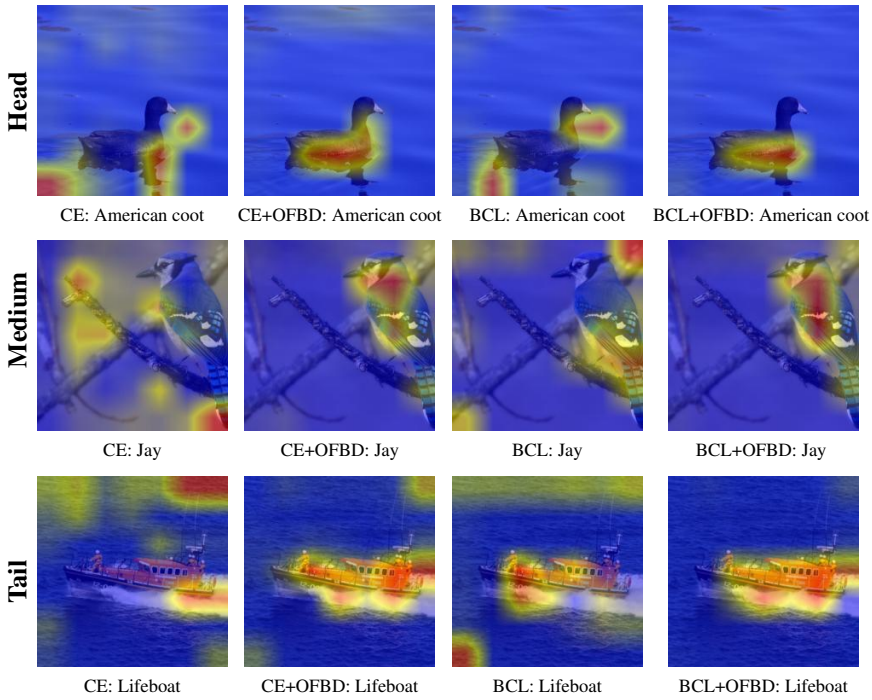


Figure 4: Extended pretrained-CAM attention maps of representative **Head, Medium, and Tail** classes under CE, CE+OFBD, BCL, and BCL+OFBD. OFBD reduces background co-activation and makes attention more object-focused across different baselines and class-frequency groups.

Therefore, using the bound established in Sec. B.1, we obtain:

$$\|\Delta\tilde{D}_{bg}(y)\|_1 \leq \|\Delta D_{bg}(y)\|_1. \tag{28}$$

This shows that FG-CutMix reduces the upper bound of distribution-level background bias. Moreover, the RL reward in Eq. 4 favors regions whose mixed samples remain more predictive of the target label than the background-source label, preventing the mixed distribution from being dominated by spurious background semantics.

BFR reduces optimization-level background bias. From Sec. B.2, the optimization-level bias is characterized by the increase of the background gradient ratio. According to Eq. 9, BFR rectifies the feature by assigning each spatial location a background-aware weight $(1 - \lambda B_u)$. Since a larger B_u indicates stronger background bias, background-dominant locations receive smaller weights than foreground-dominant locations:

$$B_u^{bg} > B_v^{fg} \Rightarrow 1 - \lambda B_u^{bg} < 1 - \lambda B_v^{fg}. \tag{29}$$

Therefore, the background-related feature component, and hence its gradient contribution, is more strongly suppressed than the foreground component:

$$\|\tilde{g}_{bg,t}(y)\|_1 \leq \|g_{bg,t}(y)\|_1, \quad \frac{\|\tilde{g}_{bg,t}(y)\|_1}{\|g_{bg,t}(y)\|_1} < \frac{\|\tilde{g}_{fg,t}(y)\|_1}{\|g_{fg,t}(y)\|_1}. \tag{30}$$

Substituting the rectified gradients into the background gradient ratio defined in Eq. 2 yields

$$\tilde{R}_t(y) = \frac{\|\tilde{g}_{bg,t}(y)\|_1}{\|\tilde{g}_{fg,t}(y)\|_1 + \|\tilde{g}_{bg,t}(y)\|_1} \leq \frac{\|g_{bg,t}(y)\|_1}{\|g_{fg,t}(y)\|_1 + \|g_{bg,t}(y)\|_1} = R_t(y). \tag{31}$$

Thus,

$$\tilde{R}_t(y) \leq R_t(y), \quad \Delta\tilde{R}_t(y) \leq \Delta R_t(y). \tag{32}$$

This shows that BFR suppresses the optimization-level background shift by reducing the relative contribution of background gradients.

Table 9: Ablation on hyper-parameters β and α.

β	**Acc. (%)**	α	**Acc. (%)**
0.005	53.7	0.1	53.6
0.01	53.9	0.5	53.8
0.015	54.0	0.75	54.1
0.02	**54.2**	**1.0**	**54.2**
0.025	54.1	1.25	53.6
0.03	53.8	1.5	53.7

Table 10: Ablation on hyper-parameters λ and γ.

λ	**Acc. (%)**	γ	**Acc. (%)**
0.2	53.6	0.2	53.3
0.3	53.9	0.4	53.7
0.4	53.7	0.6	54.1
0.5	**54.2**	**0.8**	**54.2**
0.6	54.1	1.0	53.8
0.7	53.9	1.2	53.4

Table 11: Background distribution shift on CIFAR-100-LT ($r = 100$). Lower is better.

Method	**Many↓**	**Med.↓**	**Few↓**	**All↓**
CE	0.79	0.84	0.86	0.83
+ FG-CutMix	0.78	0.83	0.80	0.82
+ BFR	**0.79**	0.84	0.81	0.80
+ OFBD	0.80	**0.79**	**0.78**	**0.79**

Table 12: Background gradient ratio shift on CIFAR-100-LT ($r = 100$). Lower is better.

Method	**Many↓**	**Med.↓**	**Few↓**	**All↓**
CE	0.4	0.8	3.6	1.5
+ FG-CutMix	-3.4	-4.7	-6.2	-4.7
+ BFR	-4.6	-5.2	-6.5	-5.0
+ OFBD	**-5.6**	**-6.1**	**-7.0**	**-6.2**

Joint effect. Combining the above results, OFBD reduces both bias terms:

$$\|\Delta\tilde{D}_{bg}(y)\|_1 \leq \|\Delta D_{bg}(y)\|_1, \quad \Delta\tilde{R}_t(y) \leq \Delta R_t(y). \tag{33}$$

Hence, FG-CutMix mitigates the enlarged background distribution deviation of tail classes, while BFR suppresses their amplified background gradient. This provides a formal explanation for why OFBD alleviates both distribution-level and optimization-level background bias in long-tailed recognition.

C Extensive Visualizations

C.1 Extended Background-bias Visualization

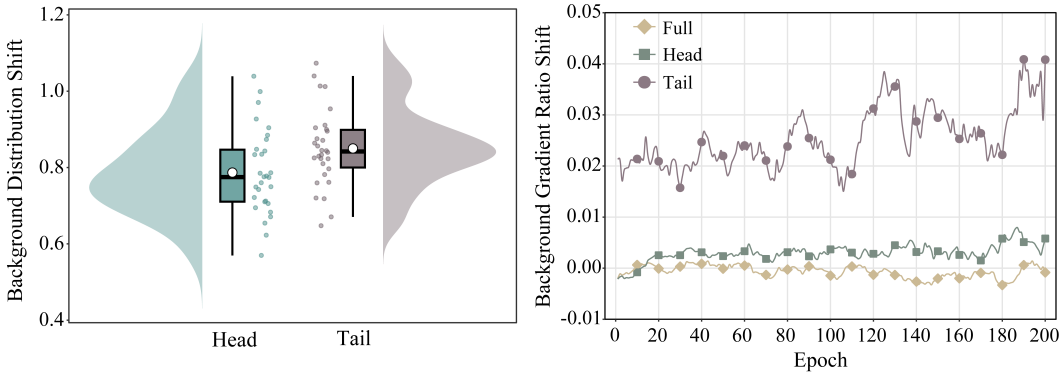
To further support the analysis in Sec. 5.3, we provide extended attention-map visualizations for representative head, medium, and tail classes. As shown in Fig. 4, we compare CE, CE+OFBD, BCL, and BCL+OFBD using pretrained-CAM attention maps. The selected samples cover different class-frequency groups, including head, medium, and tail classes. For both CE and BCL baselines, the attention maps often co-activate target objects and surrounding backgrounds, indicating background-biased representation. After applying OFBD, the activated regions become more concentrated on target-related foregrounds, while irrelevant background responses are reduced. This trend is observed across different baselines and class-frequency groups, suggesting that OFBD consistently improves object-focused representation rather than only benefiting a specific model or class group. These visualizations further validate that OFBD alleviates background reliance.

D Additional Experimental Results

D.1 Hyper-parameters Ablation

We conduct hyper-parameter ablations on CIFAR-100-LT with imbalance ratio 100 to analyze the sensitivity of OFBD. As shown in Tables 9 and 10, OFBD remains stable across a wide range of hyper-parameter values. For the RL-based foreground selector, β controls the KL regularization strength and α controls the auxiliary selector supervision. The best performance is achieved at $\beta = 0.02$ and $\alpha = 1.0$, both reaching 54.2% accuracy. Smaller values provide insufficient regularization or supervision, while overly large values slightly degrade performance.

For the overall OFBD framework, λ controls the strength of background feature rectification and γ controls the weight of policy learning. The best accuracy is obtained at $\lambda = 0.5$ and $\gamma = 0.8$, both reaching 54.4%. When λ is too small, background-biased features are insufficiently suppressed; when



(a) Distribution-level background shift.

(b) Optimization-level background-gradient-ratio shift.

Figure 5: Empirical verification of background bias on CIFAR-100-LT with imbalance ratio 100. Tail classes exhibit both larger background distribution shift and stronger background-gradient-ratio shift than head classes, validating the distribution-level and optimization-level bias analyzed in Sec. 3.

too large, useful foreground information may also be weakened. Similarly, a moderate γ provides the best trade-off between selector learning and classification optimization. These results show that OFBD is not sensitive to exact hyper-parameter choices and maintains competitive performance under nearby settings.

D.2 Ablation on Candidate Pool Size

We further study the effect of the candidate pool size K in FG-CutMix, where K denotes the number of selected foreground candidates used for reward comparison and relative-advantage estimation. Small K values limit candidate diversity, reducing the selector's ability to distinguish regions with different target-preservation quality. Increasing K improves the chance of selecting foregrounds that better preserve target semantics after background replacement. However, excessively large K increases computation cost and may introduce redundant or noisy candidates.

Fig. 6a visualizes the accuracy and per-epoch computation cost for different K on CIFAR-100-LT with imbalance ratio 100. Accuracy improves as K increases from 2 to 6, then slightly decreases for larger K, while per-epoch cost rises monotonically. This indicates that a moderate candidate pool provides sufficient diversity for effective foreground selection. We set $K = 6$ as the default, achieving a favorable trade-off between accuracy and efficiency.

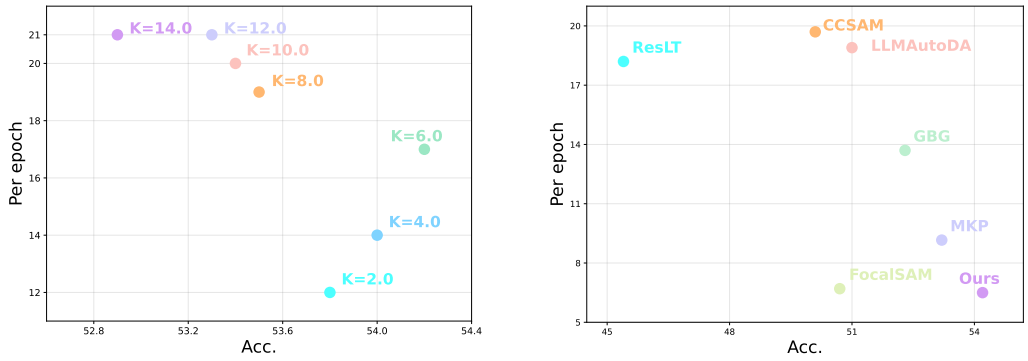
D.3 Empirical Verification of Background Bias

Experiment Setting. We train ResNet-32 [20] with cross-entropy on CIFAR-100-LT [4] and CIFAR-100 under identical settings. The CIFAR-100 model serves as the balanced reference for measuring distribution-level background shift. Background regions are obtained with a pretrained CAM [106, 31].

Results. As shown in Fig. 5, tail classes suffer from stronger background bias than head classes. Specifically, Fig. 5a shows that tail classes have a larger background distribution shift, indicating that their learned background statistics deviate more from the balanced reference distribution. Meanwhile, Fig. 5b shows that the background-gradient-ratio shift of tail classes increases faster and fluctuates more strongly during training, suggesting that tail-class optimization is more easily driven by background-related gradients. These results support our analysis (Sec. 3.4) that tail degradation is not only caused by sample scarcity, but also by background-biased distribution and optimization.

D.4 Effect on Background Distribution Shift

Based on the observed distribution-level bias, we further evaluate whether FG-CutMix can reduce the background distribution shift. We measure the deviation between the background feature distribution learned under long-tailed training and that learned from balanced training. Specifically, we train a



(a) Sensitivity to candidate pool size K.

(b) Efficiency-accuracy comparison.

Figure 6: Efficiency analysis of OFBD on CIFAR-100-LT with imbalance ratio 100. (a) Accuracy and per-epoch (s) computation cost as a function of the candidate pool size K. Accuracy first improves and then slightly decreases for large K, while per-epoch cost rises monotonically. $K = 6$ is selected as the default for a good efficiency-accuracy trade-off. (b) Efficiency-accuracy comparison with representative long-tailed methods, where per-epoch computation cost is measured in seconds. "Ours" denotes BCL+OFBD.

reference model on CIFAR-100 and use it to estimate the class-wise reference background feature mean $\bar{\mu}_{bg}(y)$. For each long-tailed model, we compute the background feature mean $\mu_{bg}(y)$ on the same balanced test set, where background regions are obtained by pretrained CAM masks. The background distribution shift is measured as:

$$\Delta D_{bg}(y) = \|\mu_{bg}(y) - \bar{\mu}_{bg}(y)\|_1. \tag{34}$$

We report the averaged $\Delta D_{bg}(y)$ over Many-, Medium-, and Few-shot classes. Lower values indicate a smaller distribution-level background shift.

As shown in Table 11, standard long-tailed training exhibits a clear background distribution shift, especially on Few-shot classes. Standard CutMix only provides limited reduction, since random region replacement may still preserve or introduce spurious background correlations. In contrast, FG-CutMix consistently reduces $\Delta D_{bg}(y)$, indicating that preserving target-related foregrounds while diversifying backgrounds makes the learned background distribution closer to the balanced reference distribution.

D.5 Effect on Background Gradient Ratio

Based on the observed optimization-level bias, we further examine whether BFR can suppress the increase of the background gradient ratio. We measure the increase of the background gradient ratio defined in Eq. 2. Foreground and background gradient components are computed using the same foreground-background masks as in Sec. D.3. We report the averaged $\Delta R_t(y)$ over Many-, Medium-, and Few-shot classes. Lower values indicate that the optimization process is less dominated by background-related gradients.

As shown in Table 12, both FG-CutMix and BFR substantially reduce the background gradient ratio compared with standard CE, with the effect most pronounced for Few-shot classes. FG-CutMix decreases the gradient shift at the input level by diversifying backgrounds, while BFR further down-weights background-biased spatial features during feature-level rectification. The full OFBD, combining FG-CutMix and BFR, achieves the largest reduction across all class groups, demonstrating the complementarity between input-level background diversification and feature-level rectification in mitigating background-driven optimization.

D.6 Performance Analysis under Long-tailed and Balanced Settings

We further compare the head-, medium-, and tail-class performance to examine where OFBD brings the largest improvement. As shown in Fig. 8, adding OFBD to BCL substantially improves Few-shot accuracy from 32.9% to 40.2%, yielding a $+7.3\%$ gain. Meanwhile, it slightly improves

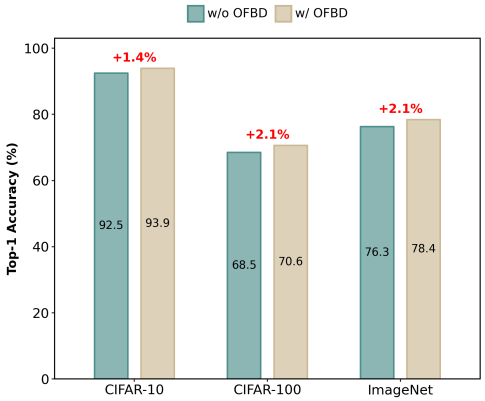


Figure 7: Performance comparison with and without OFBD on balanced datasets ($r = 1$). OFBD consistently improves Top-1 accuracy on CIFAR-10, CIFAR-100, and ImageNet.

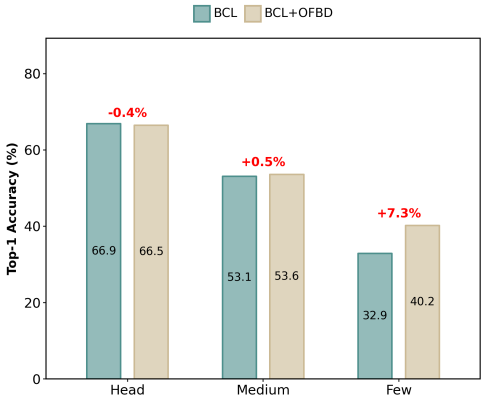


Figure 8: Head-, medium-, and tail-class performance comparison on CIFAR-100-LT with imbalance ratio 100. OFBD mainly improves tail-class accuracy while largely preserving head-class performance.

Medium-shot accuracy from 53.1% to 53.6%, with only a marginal change on Head classes from 66.9% to 66.5%. These results indicate that OFBD mainly benefits data-scarce tail classes while largely preserving head-class performance, demonstrating its effectiveness in alleviating tail-specific background bias.

We also examine whether OFBD is effective under the balanced setting, where the imbalance ratio is 1. As shown in Fig. 7, OFBD consistently improves the baseline without OFBD on balanced CIFAR-10, CIFAR-100, and ImageNet, with gains of $+1.4\%$, $+2.1\%$, and $+2.1\%$, respectively. This demonstrates that OFBD does not rely on long-tailed class distributions and can still provide object-focused regularization benefits when the class distribution is uniform.

D.7 Further Ablation of FG-CutMix

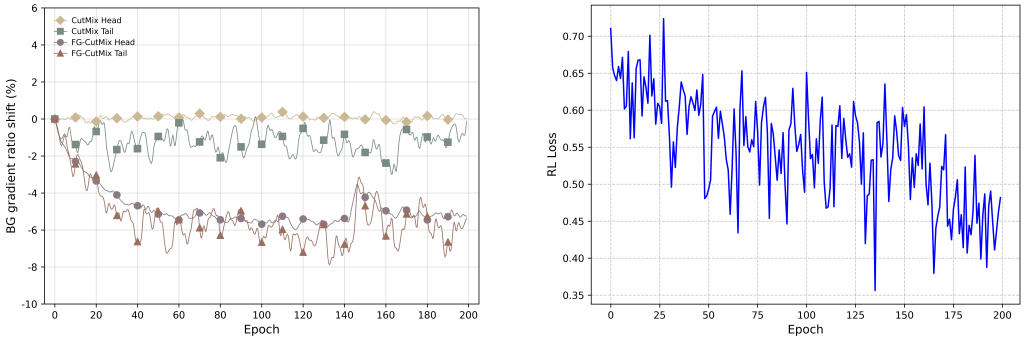
Table 4 shows that FG-CutMix with RL-based foreground selection outperforms standard CutMix with random selection. To further analyze this improvement, we compare their optimization-level background bias by measuring the background-gradient-ratio shift on CIFAR-100-LT with imbalance ratio 100. As shown in Fig. 9a, both CutMix and FG-CutMix keep the head-class shift relatively small, suggesting that head-class optimization is less sensitive to region selection.

For tail classes, however, standard CutMix exhibits a stronger background-gradient-ratio shift, indicating that random replacement may preserve background cues or discard target-related foreground evidence. In contrast, FG-CutMix consistently produces a lower tail-class shift during training. This confirms that RL-based foreground region selection not only improves accuracy, as shown in Table 4, but also reduces background-driven optimization for tail classes.

We further examine the training stability of the RL-based foreground selector. As shown in Fig. 9b, the selector loss exhibits stochastic fluctuations due to random proposal sampling and policy-gradient optimization, but remains bounded and gradually stabilizes throughout training. This suggests that the selector does not suffer from unstable policy updates. Together with the accuracy improvement in Table 4 and the reduced tail-class background-gradient shift in Fig. 9a, these results indicate that the RL-based selector can be optimized stably while learning foreground regions that better mitigate background-driven optimization.

D.8 Efficiency Analysis

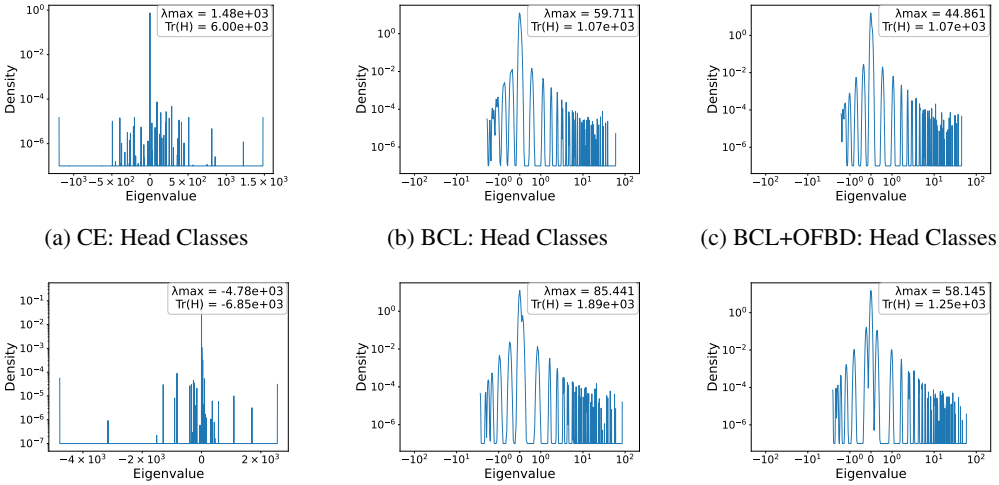
We evaluate the efficiency-accuracy trade-off of BCL+OFBD on CIFAR-100-LT with imbalance ratio 100. As shown in Fig. 6b, BCL+OFBD achieves the highest accuracy while maintaining a low per-epoch training cost. In particular, compared with MKP, BCL+OFBD reduces the computation cost by 2.66 seconds per epoch and improves Top-1 accuracy by 1.0%. These results show that OFBD



(a) BG gradient ratio shift.

(b) RL loss curve of the selector.

Figure 9: Further analysis of FG-CutMix on CIFAR-100-LT with imbalance ratio 100. (a) FG-CutMix reduces tail-class background-gradient-ratio shift compared with standard CutMix, indicating weaker background-driven optimization. (b) The RL loss of the foreground selector remains bounded and gradually stabilizes during training, indicating stable optimization of the RL-based selector.



(a) CE: Head Classes

(b) BCL: Head Classes

(c) BCL+OFBD: Head Classes

(d) CE: Tail Classes

(e) BCL: Tail Classes

(f) BCL+OFBD: Tail Classes

Figure 10: Eigen Spectral Density of Hessian for head and tail classes of ResNet models trained with CE, BCL, and BCL+OFBD on CIFAR-100 LT respectively. A smaller $\lambda_{\max}$ and $Tr(H)$ generally indicate a flatter loss landscape.

provides an efficient plug-in solution for long-tailed recognition, benefiting from its lightweight background-aware feature rectification and training-only foreground-guided augmentation.

D.9 Additional Results for Eigen Spectral Density of Hessian

This section presents additional results on the eigen spectral density of the Hessian for ResNet models trained with CE, BCL, and BCL+OFBD on CIFAR-100-LT. As shown in Fig. 10, we separately analyze the Hessian spectra of head and tail classes. A smaller largest eigenvalue $\lambda_{\max}$ and trace $\mathrm{Tr}(H)$ generally indicate a flatter loss landscape and better optimization stability.

Compared with the CE baseline, BCL tends to reduce the sharpness of the loss landscape to some extent, especially for tail classes, suggesting that contrastive learning helps improve the optimization behavior under long-tailed distributions. However, its Hessian spectrum still shows relatively large eigenvalues, indicating that the optimization landscape remains sharp for certain class groups. In contrast, BCL+OFBD consistently produces lower $\lambda_{\max}$ and $\mathrm{Tr}(H)$ for both head and tail classes. This suggests that BCL+OFBD effectively smooths the loss landscape, rather than only benefiting tail

Table 13: Comparison with attention modules on CIFAR-100-LT ($r = 100$).

Method	Overall↑	Many↑	Med.↑	Few↑
CBAM [92]	50.43	67.80	48.90	31.90
BAM [59]	50.88	68.70	49.30	32.00
Coordinate Attention [23]	50.83	**69.40**	49.10	31.30
Triplet Attention [108]	50.67	67.90	50.00	31.40
BFR (Ours)	**52.97**	66.96	**53.65**	**35.86**

Table 14: Robustness of B_u under input perturbations.

Input	BG AUROC↑	Δ vs. Clean
Clean	0.7182	0.0000
Brightness, factor 1.2	0.7054	-0.0128
Contrast, factor 1.2	0.7051	-0.0131
Gaussian noise, std 0.05	0.7038	-0.0144
Gaussian blur, $k = 5, \sigma = 1.0$	0.7016	-0.0166

Table 15: Integration with additional long-tailed methods.

Method	Overall↑	Many↑	Med.↑	Few↑
MDCS [103]	53.15	67.09	**56.09**	33.47
MDCS + OFBD	**54.39**	**68.31**	55.26	**37.14**
SADE [100]	49.20	**61.31**	51.14	32.80
SADE + OFBD	**51.01**	57.89	**53.49**	**40.10**

Table 16: Feature-layer analysis for the Region Selector.

Layer	Res.	S-FG↑	S-BG↓	AUROC↑	Overall↑	Few↑
layer1	32×32	0.1618	0.1441	0.7285	53.98	41.00
layer2	16×16	0.1694	0.1280	0.7375	54.04	40.97
layer3	8×8	**0.1709**	**0.1026**	**0.8644**	**54.12**	**41.15**

Table 17: Sensitivity to the number of selected proposals K.

K	Overall Acc. (%)↑
2	53.87 ± 0.34
3	53.93 ± 0.22
4	53.94 ± 0.20
5	53.89 ± 0.19
6	**53.96 ± 0.19**

Table 18: Masking analysis based on background-aware score B_u.

Masking	Overall Acc. (%)↑	Δ Acc.
None	58.475	0.000
Random 20%	53.850	-4.625
High-B_u 20%	56.655	-1.820
Low-B_u 20%	45.140	**-13.335**

classes. The flatter Hessian spectrum further supports that BCL+OFBD mitigates background-biased optimization and leads to more stable learning under long-tailed training.

D.10 Visualization of LossLandscape

Fig. 11 visualizes the loss landscapes of head and tail classes for ResNet models trained with CE, BCL, and BCL+OFBD on CIFAR-100-LT. The first row corresponds to head classes, and the second row corresponds to tail classes. Compared with CE and BCL, BCL+OFBD tends to produce a flatter and smoother loss landscape for both class groups. Such a tendency is more pronounced for tail classes, where the loss landscape under long-tailed training is usually sharper and more irregular. The results indicate that OFBD helps stabilize the optimization process and mitigates the unfavorable optimization behavior caused by data imbalance.

D.11 Comparison with Attention Modules

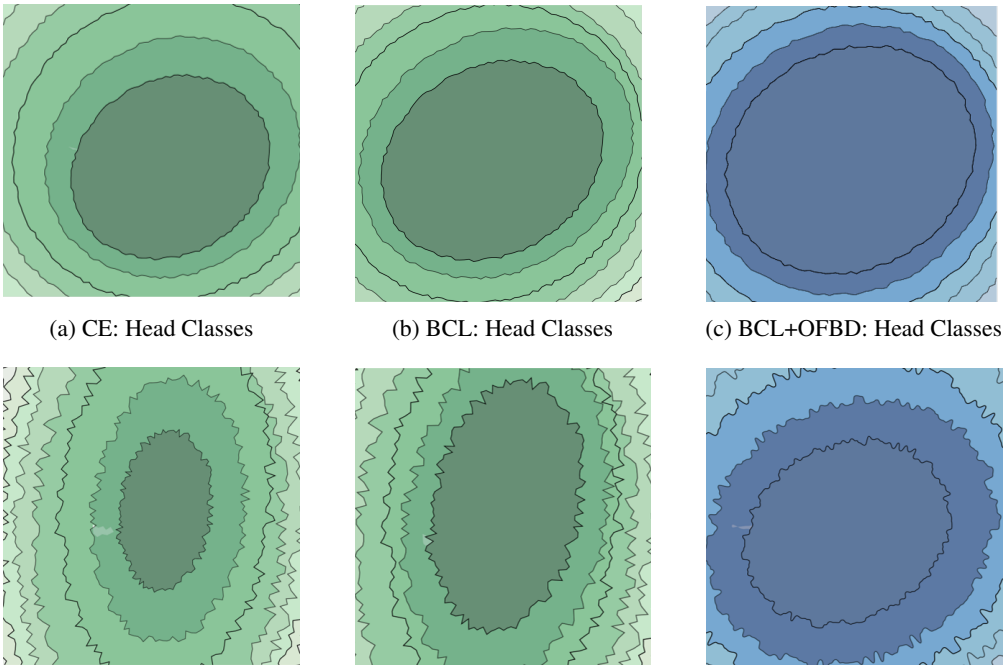
We compare BFR with representative attention modules under the same BCL-based setting on CIFAR-100-LT ($r = 100$). As shown in Table 13, BFR achieves better overall and Few-shot performance, demonstrating its effectiveness in background bias.

D.12 Robustness of Background Score under Input Perturbations

We evaluate the robustness of B_u under different image perturbations on ImageNet-LT. As shown in Table 14, BG AUROC remains above 0.70 across all perturbations, indicating stable background-aware estimation under appearance variations.

D.13 Integration with Multi-Expert Long-tailed Recognition Frameworks

We integrate OFBD into representative multi-expert long-tailed recognition frameworks while keeping their original architectures and training objectives unchanged. As shown in Table 15, OFBD consistently improves different multi-expert baselines, demonstrating its compatibility and effectiveness across existing long-tailed recognition frameworks.



(a) CE: Head Classes (b) BCL: Head Classes (c) BCL+OFBD: Head Classes

(d) CE: Tail Classes (e) BCL: Tail Classes (f) BCL+OFBD: Tail Classes

Figure 11: Visualization of loss landscape for head and tail classes of ResNet models trained with CE, BCL, and BCL+OFBD on CIFAR-100 LT respectively.

D.14 Region Selector Feature Layer Analysis

We analyze the effect of different feature layers used by the Region Selector. As shown in Table 16, deeper layers provide stronger foreground-background separation, and layer3 achieves the best discrimination capability.

D.15 Sensitivity Analysis of Region Proposal Number

We evaluate the sensitivity of K in the Region Selector. As shown in Table 17, OFBD maintains stable performance across different K values, demonstrating robustness to proposal selection.

D.16 Masking Analysis of Background-aware Score

We perform masking experiments based on B_u to analyze the semantic meaning of high-score regions. As shown in Table 18, masking high-B_u regions causes limited degradation, while masking low-B_u regions leads to significant performance drops, validating the effectiveness of BFR.

E Limitation

A key consideration of OFBD lies in its reliance on accurate foreground localization and feature separation. While our current FG-CutMix and BFR designs work effectively on standard visual benchmarks, scenarios with severe occlusion, extremely small objects, or highly entangled foreground-background cues may benefit from enhanced region selection strategies or adaptive rectification. Addressing these cases represents a natural direction for extending OFBD, highlighting opportunities for future work without undermining its demonstrated effectiveness on long-tailed datasets.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the main claims and contributions are explicitly stated in the Abstract and Introduction, which align perfectly with our theoretical and empirical results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Yes, we comprehensively discuss the limitations and underlying assumptions of our proposed method in Appendix D.11

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Yes, all theoretical assumptions are explicitly stated in Section 3, and the complete, step-by-step proofs are provided in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, all experimental details, including network architectures, training procedures, and dataset preprocessing, are fully described in Section 5 and Appendix A to ensure full reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Yes, the source code, pre-trained models, and instructions for reproducing our main experimental results are included in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer:[Yes]

   Justification: Yes, the exact experimental setups, dataset splits, hyperparameter configurations, and the search strategies are thoroughly detailed in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, we report the mean and standard deviation across 20 independent runs with different random seeds in all our main quantitative tables.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, the details regarding the computing infrastructure (e.g., GPU specifications and estimated runtime per experiment) are provided in Appendix A and Section 5.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the research conducted in this paper conforms entirely with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: N/A. This work focuses on fundamental algorithmic improvements and does not have immediate foreseeable negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: N/A. Our paper does not introduce or release models or datasets with a high risk for direct, malicious misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Yes, all datasets and codes used in this work are appropriately cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: N/A. This paper does not introduce or release any distinct new datasets or assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: N/A. This paper does not involve crowdsourcing or any research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: N/A. This paper does not conduct research involving human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: N/A. Large Language Models were not used as a core, original component of the methodology in this research; they were solely used for writing assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.