DARL: Mitigating Gradient Conflicts in Long-Tailed Out-of-Distribution Learning

Xuan Zhang*
zhangxua22@tsinghua.org.cn
Tsinghua University
Shenzhen International Graduate School
Shenzhen, China

Sinchee Chin*
imsinchee@gmail.com
Tsinghua University
Shenzhen International Graduate School
Shenzhen, China

Jing-Hao Xue
jinghao.xue@ucl.ac.uk
University College London
Department of Statistical Science
London, United Kingdom

Xiaochen Yang
xiaochen.yang@glasgow.ac.uk
University of Glasgow
School of Mathematics and Statistics
Glasgow, United Kingdom

Wenming Yang†
yang.wenming@sz.tsinghua.edu.cn
Tsinghua University
Shenzhen International Graduate School
Shenzhen, China

Abstract

Long-tailed out-of-distribution learning aims to reduce performance bias in long-tailed in-distribution (ID) data while rejecting out-of-distribution (OOD) samples, which are often mistaken for under-represented tail classes. To achieve OOD detection, existing methods incorporate an outlier exposure (OE) term into the long-tailed recognition (LTR) loss. However, as we prove in this paper, the OE term induces a gradient conflict with the ID objectives, especially for tail classes, thereby contradicting the core motivation of LTR. To avoid the ID-OOD dilemma, we propose Dynamic Ambiguity-aware Recalibration for Logits (*DARL*), an ambiguity-guided long-tailed OOD learning approach, grounded on two theoretical insights. First, we show that the mixed ID data can mitigate the conflict in OE training and exhibits higher intrinsic ambiguity than the original ID data, thus able to serve as a surrogate for real OOD data. Second, we introduce an ambiguity-aware logit adjustment that can dynamically calibrate the class margins using energy-based ambiguity metrics, effectively reducing early-stage bias while avoiding late-stage overfitting. Extensive experiments show that DARL achieves the overall state-of-the-art performance of long-tailed OOD learning. Moreover, compared with the OE methods, DARL trains solely on the ID data, which can reduce the data requirements by 80%. The code is available here.

CCS Concepts

• **Computing methodologies** → **Learning from implicit feedback**.

*Equal contribution.
†Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755127

Keywords

Long-tailed Recognition, Out-of-distribution Detection, ID-OOD Gradient Conflicts, Ambiguity-Aware Logit Adjustment

1 Introduction

Real-world deployment of deep learning models remains challenged by the coexistence of long-tailed data and out-of-distribution (OOD) samples. It can fail catastrophically if these two issues are addressed in isolation. On the one hand, long-tailed distributions bias models toward over-represented head classes, suppressing correct recognition of under-represented in-distribution (ID) tail classes. On the other hand, OOD samples (i.e. samples far from the training distribution) induce overconfident mispredictions that jeopardize reliability in safety-critical applications [49]. Crucially, these two issues are not independent: Tail-class samples, already scarce in training, can be easily misclassified as OOD due to their low-confidence features, while OOD samples can be misclassified into head classes due to their overconfidence obtained in classification [32]. This duality stems from a fundamental dilemma: (a) Long-tailed recognition (LTR) requires increased sensitivity to tail-class features, but (b) OOD detection demands reduced sensitivity to rare features.

Existing methods for joint LTR and OOD detection typically follow the outlier exposure (OE) paradigm, which trains the model with known OOD samples via a regularization term. Early OE methods push OOD logits toward a uniform distribution to model uncertainty [26], while recent approaches use metric learning to better separate tail and OOD features [32, 44], or introduce absence classes to identify OOD [49]. These designs, however, overlook two practical pitfalls: (1) the *absence of real OOD samples* to train the model, due to the difficulty of attaining them [3, 54]. (2) shared features or spurious correlations between ID and OOD samples, causing *gradient conflict*. Specifically, when ID and OOD samples exhibit overlapping features, e.g., cars (ID) and trucks (OOD) sharing common features such as tires and exhibiting similar backgrounds, the OE and ID training objectives simultaneously assert their influence on the shared feature representation. This results in the ID-OOD trade-off, where the optimization for OOD detection interferes with the learning of accurate classification boundaries, especially for

tail classes that are already under-represented. Consequently, this conflict undermines the primary goal of long-tailed learning, which is to balance performance across all classes, thereby limiting the overall efficacy of conventional OE methods in LTR tasks.

To address these limitations, we propose Dynamic Ambiguity-aware Recalibration for Logits (*DARL*), an ambiguity-guided method for long-tailed OOD learning that seamlessly integrates three key components with no need for real OOD samples for training: (1) the use of mixed ID data as effective *pseudo-OOD* samples, (2) Ambiguity-Aware Logits Adjustment (ALA) that dynamically recalibrates classification margins based on model training dynamics, and (3) a gradient-driven OOD detection mechanism that leverages ODIN.

More specifically, we reinterpret mixed data not merely as a data augmentation strategy but as viable pseudo-OOD samples. Our theoretical analysis from an energy perspective shows that mixed samples inherently exhibit higher energy than pure ID data, thus serving as a credible proxy for OOD inputs. Moreover, we show that, compared with conventional OE, mixed data induce substantially less gradient conflict between ID and OOD objectives, effectively mitigating the inherent trade-off in joint optimization.

We also introduce Ambiguity-Aware Logits Adjustment (ALA), which dynamically recalibrates the traditional static class priors by incorporating an energy function (see Eq. 6). ALA harnesses the model's training dynamics (posterior information) to adjust the predefined class priors on the fly, thereby establishing clear and adaptive decision margins for both under-represented classes and pseudo-OOD samples.

Finally, we show that ODIN [28]'s adversarial perturbation mechanism naturally aligns with our DARL training strategy by amplifying gradients in high-energy regions that correspond to ambiguous pseudo-OOD samples, thus creating a self-supervised mechanism for enhanced OOD separation.

In summary, our principal contributions are as follows.

- We provide theoretical insights demonstrating that outlier exposure can induce gradient conflicts with the ID objectives in LTR.
- We prove that mixed ID data alleviates the inherent trade-off in OE training and exhibits higher intrinsic ambiguity than raw ID data, making it a viable substitute for real OOD data.
- We propose ambiguity-aware logit adjustment, introducing model's ambiguity to enhance the logit adjustment and OOD detection.
- Extensive experiments demonstrate that our DARL achieves overall state-of-the-art performance in both LTR and OOD detection, while reducing data requirements by 80%.

2 Related Work

2.1 Long-Tailed Recognition (LTR)

Current approaches to LTR focus on two aspects: data and algorithms. Data-focused strategies aim to rebalance the data distribution through either resampling [4, 16, 19, 29, 35, 47] or data augmentation[9, 10, 15, 62]. Algorithm-focused approaches include category-sensitive learning and transfer learning. Most category-sensitive learning methods adjust the training loss for each category, using techniques such as reweighting [2, 7, 13, 27, 37, 38, 51] or remargining (logit adjustment) [2, 16, 18, 31, 35, 40, 51, 57].

Transfer learning, on the other hand, leverages knowledge from one area to strengthen model training in another, including three main tactics: (a) Two-stage training [24, 63], which starts with training the model on an imbalanced dataset and then retrains the classifier on a balanced dataset. (b) Model ensemble [12, 12, 39, 40, 46, 50, 60, 61, 64], which merges insights from various experts with distinct capabilities to produce a more balanced output. (c) Head-to-tail transfer [6, 55, 62], which aims to leverage the knowledge of the head classes to improve the performance of the tail classes.

2.2 Out-of-Distribution (OOD) Detection

Existing OOD detection methods for classification can be broadly categorized into two types: post-hoc methods and outlier exposure methods. Post-hoc methods detect OOD samples using features or logits without altering the training process [21, 28, 30, 34, 36, 43]. MSP [21] uses the maximum softmax probability to distinguish ID and OOD samples. ODIN [28] enhances separation via temperature scaling and input perturbations. EBO [30] replaces softmax with an energy-based score, reducing overconfidence on OOD inputs. kNN-OOD [36] introduces a non-parametric approach using deep nearest neighbors, while NNGuide [34] adjusts confidence scores based on similarity to training samples. ViM [43] combines feature residuals and logits to compute a virtual OOD logit, addressing limitations of relying solely on logits or features. In contrast, outlier exposure methods incorporate auxiliary OOD samples during training to improve detection [22]. EBO [30] extends its energy-based approach with energy regularization to separate ID and OOD energy scores. ATOM [5] mines harder outliers to better shape the decision boundary. FSOOD [53] addresses both semantic and covariate shifts by constructing a semantic score based on deep and shallow features. DAL [45] models a Wasserstein ball around auxiliary OOD data and optimizes performance against worst-case OOD distributions within this ball. However, these methods may degrade ID classification performance [26], and collecting OOD datasets is often costly or impractical [3].

2.3 Joint LTR and OOD Detection

Current long-tailed OOD learning often combine conventional long-tailed learning methods with outlier exposure. Some of them explicitly separate OOD samples from tail classes. For instance, PASCL [44] disentangles tail classes from OOD data through asymmetric contrastive learning. COCL [32] enhances tail-OOD separation via debiased margin allocation and anomaly-aware logit calibration. [23] models OOD likelihood using ID class priors instead of uniform assumptions. [17] introduces a three-branch framework with a dedicated "missing class" branch and prototype-guided contrastive loss to improve tail-OOD separability. Other methods leverage OOD data to enrich tail-class representations. Open-Sampling [48] dynamically samples OOD data similar to ID tail classes. COLT [1] selects OOD samples via neighborhood sparsity in feature space and employs distribution-aware contrastive learning. EAT [49] introduces multiple "missing classes" and augments tail samples using CutMix [58]. PATT [20] employs von Mises-Fisher semantic augmentation, temperature scaling for confidence enhancement, and

attention-based calibration to differentiate tail features from OOD data.

However, these methods have three main limitations: (1) they require real OOD datasets, which are costly and often impractical to construct; (2) as we demonstrate below, introducing outlier data can degrade ID performance [3, 8, 26, 41, 44], particularly for tail data, which contradicts the core motivation of LTR;

and (3) they depend on static training priors that fail to capture training dynamics. Therefore, a method is needed that (a) detects OOD samples without real OOD datasets and (b) accounts for training dynamics—this is precisely the focus of our work.

3 Issues of OE Methods in LTR

3.1 Notation

The long-tailed ID training set, denoted by $\mathcal{D}_{\text{in}} = \{x_i, y_i\}_{i=1}^{N}$, consists of N training samples with y_i representing the ground-truth label for the image x_i. The total number of training samples is represented by $N = \sum_{c=1}^{C} n_c$, in which C is the total number of classes, and n_c is the number of samples for the c-th class. In accordance with [40, 46, 60], we define the class prior as $\Phi = \{\varphi_c\}_{c=1}^{C}$, where $\varphi_c = n_c/N$. The imbalance ratio is calculated as the maximum n_j divided by the minimum n_c, i.e., $\max\{n_c\}/\min\{n_c\}$. Let θ represent the model parameters. For each input x_i, we denote the final logits produced by the model as $\mathbf{z}_\theta(x_i)$ (or simply $\mathbf{z}_i$), where $z_i^{(c)}$ is the c-th element of the logits vector. Applying the softmax function, the corresponding probability vector is given by $\mathbf{p}_\theta(x_i) = \text{softmax}(\mathbf{z}_\theta(x_i))$, or simply $\mathbf{p}_i$.

3.2 Gradient Conflict Introduced by OE

In practice, ID samples and OOD samples often share some common features (e.g., the wheels present in both cars and trucks) or exhibit similar backgrounds. In these circumstances, the training objectives for LTR and OOD detection incur a *gradient conflict*.

We use the classical OE method in [26] as an example, and denote the overall loss by $\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{OE}}$, where $\mathcal{L}_{\text{cls}}$ is the cross entropy loss. For an ID sample x with ground-truth label y and an OOD sample x', the inner product between the gradient of the classification loss and that of the OE loss, computed with respect to the classification head parameters θ_{cls}, is expressed as

$$\begin{aligned}&\langle \nabla_{\theta_{\text{cls}}} \mathcal{L}_{\text{cls}},\ \nabla_{\theta_{\text{cls}}} \mathcal{L}_{\text{OE}} \rangle \\ &= \sum_{c=1}^{C} \Big(\mathbf{p}_\theta^{(c)}(x) - \mathbf{e}_y^{(c)}\Big)\Big(\frac{1}{C} - \mathbf{p}_\theta^{(c)}(x')\Big) \langle \nabla_{\theta_{\text{cls}}} \mathbf{z}_\theta^{(c)}(x),\ \nabla_{\theta_{\text{cls}}} \mathbf{z}_\theta^{(c)}(x') \rangle,\end{aligned} \tag{1}$$

where $\mathbf{e}_y$ is the the one-hot label of class y. When x' shares features with x, the network tends to classify it towards the correct class of x, i.e., $\mathbf{p}_\theta^{(y)}(x') > \frac{1}{C}$. Consequently, the term $\frac{1}{C} - \mathbf{p}_\theta^{(y)}(x')$ becomes negative, leading to a negative gradient inner product:

$$\langle \nabla_{\theta_{\text{cls}}} \mathcal{L}_{\text{cls}},\ \nabla_{\theta_{\text{cls}}} \mathcal{L}_{\text{OE}} \rangle < 0. \tag{2}$$

This negative inner product between gradients, termed *gradient conflict* in this paper, indicates misalignment between standard classification and outlier exposure objectives during optimization, consequently impeding effective classifier updates. More importantly, in the long-tailed recognition setting, due to the under-representation of tail classes and the logit adjustment that punishes tail classes more, the adverse impact is more pronounced for tail classes, which goes against the core motivation of long-tailed learning.

Notably, this issue can be mitigated substantially by replacing real OOD data used in the OE approach with mixed ID data. Let $\tilde{x}$ denote a mixture of x and another ID sample x_k, associated with a two-hot label $\mathbf{q} = \lambda_m \mathbf{e}_y + (1 - \lambda_m)\mathbf{e}_k$, where $\mathbf{e}_y$ and $\mathbf{e}_k$ are the one-hot labels of class y and class k, and λ_m denotes the mixing proportion; any data mixture strategy may be adopted, with two options presented in Sec. 4.2. The loss function is defined as

$$\mathcal{L}_{\text{mix}}(\tilde{x}, \mathbf{q}) = \lambda_m\, \mathcal{L}_{\text{cls}}(\tilde{x}, \mathbf{e}_y) + (1 - \lambda_m)\, \mathcal{L}_{\text{cls}}(\tilde{x}, \mathbf{e}_k). \tag{3}$$

The corresponding gradient is given by

$$\nabla_{\theta_{\text{cls}}} \mathcal{L}_{\text{mix}}(\tilde{x}, \mathbf{q}) = \lambda_m\, \nabla_{\theta_{\text{cls}}} \mathcal{L}_{\text{cls}}(\tilde{x}, \mathbf{e}_y) + (1 - \lambda_m)\, \nabla_{\theta_{\text{cls}}} \mathcal{L}_{\text{cls}}(\tilde{x}, \mathbf{e}_k). \tag{4}$$

It can be shown that the inner product between the gradient of the standard classification loss for an ID sample and that of the mixed data loss remains positive:

$$\langle \nabla_{\theta_{\text{cls}}} \mathcal{L}_{\text{cls}}(x, \mathbf{e}_y),\ \nabla_{\theta_{\text{cls}}} \mathcal{L}_{\text{mix}}(\tilde{x}, \mathbf{q}) \rangle > 0. \tag{5}$$

Thus, the gradient contribution from mixed data is well aligned with that of the standard classification loss, ensuring stable and effective parameter updates. This alignment contrasts sharply with the conflicting gradients observed in OOD-based OE, and it helps mitigate the harmful impact on tail classes. More detailed derivations and analysis are provided in the Supplementary Material.

4 Methodology

4.1 Overview of DARL

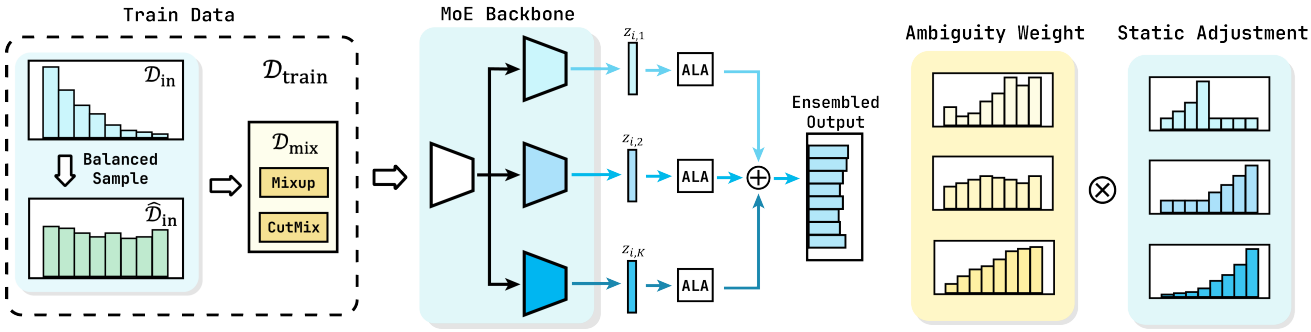
The aim of DARL is to propose an ambiguity-guided joint approach to LTR and OOD detection, leveraging only known long-tailed ID samples. Fig. 1a illustrates the overall structure of DARL, which mainly consists of two parts: (a) generating pseudo-OOD samples by mixing the original ID data $\mathcal{D}_{\text{in}}$ and its balanced version $\hat{\mathcal{D}}_{\text{in}}$; the original, balanced, and mixed datasets together form the training set $\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{in}} \cup \hat{\mathcal{D}}_{\text{in}} \cup \mathcal{D}_{\text{mix}}$; and (b) harnessing a mixture-of-experts (MoE) backbone f_θ to address class imbalance through expert-specific partitions guided by group-wise priors [40].

Fig. 1b illustrates the dynamic recalibration in the Ambiguity-Aware Logit Adjustment (ALA) module. It combines an energy-based ambiguity weight, which reflects sample uncertainty, with a prior-based static logit offset capturing class imbalance. Their product adaptively compensates ambiguous samples, addressing both gradient conflicts and tail under-representation.

4.2 Mixed ID Data Serving as Pseudo-OOD Data

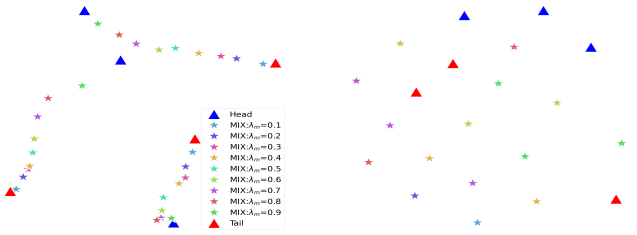
In this section, we provide both qualitative and quantitative evidence to show that mixed data are highly suitable as pseudo-OOD samples. The t-SNE visualization of mixed data and source data is shown in Fig. 2. The features of mixed data reside in proximity to the ID manifold, rendering it a highly suitable option of near-OOD.

Mixup. Mixup generates synthetic samples by $\tilde{x}_{\text{mixup}} = \lambda_m x_i + (1 - \lambda_m) x_j$, where $i \neq j$, $x_i, x_j \sim \mathcal{D}_{\text{in}}$, and $\lambda_m \sim \text{Beta}(\alpha, \alpha)$. These linear interpolations position the synthesized features in *transitional regions* between two distinct ID classes, as illustrated in Fig. 2a.



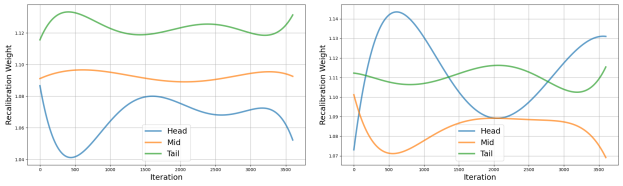
(a) The pipeline of our framework. (b) Ambiguity-Aware Logits Adjustment.

Figure 1: The overview of our framework DARL.



(a) t-SNE of Mixup (b) t-SNE of Cutmix

Figure 2: Effect of varying λ_m on the positions of MixUp [59] and CutMix [58] samples relative to original data.



(a) LGLA [40]. (b) ALA.

Figure 3: Trend of the ambiguity $\hat{E}(z)$ for head, mid, and tail classes on CIFAR10-IR100.

Specifically, the star-shaped markers denote Mixup-generated features corresponding to varying λ_m values. As λ_m transitions from 0 to 1, the generated feature smoothly migrates between the two original classes.

CutMix. In contrast, CutMix synthesizes samples according to: $\tilde{x}_{\text{cutmix}} = M \odot x_i + (1-M) \odot x_j$, where M is a binary mask and $\odot$ denotes element-wise multiplication. Unlike Mixup, CutMix produces *semantically inconsistent* samples, thus violating the smoothness prior inherent in natural images. As depicted in Fig. 2b, CutMix-generated features do not lie along a direct interpolation between two source features. Due to random cropping, their representation in the feature space does not exhibit linear movement from one class to another, but rather random displacement within proximity to the original features.

For quantitative analysis, we further examine the ambiguity of mixed data through the decomposition of the energy function (for detailed deduction, please refer to the supplementary material):

$$-E(z_x) = \underbrace{\tau \mathrm{KL}(\mathbf{q} \| \mathbf{p}_\theta(x))}_{\text{Training Objective}} + \underbrace{\tau H(\mathbf{q})}_{\text{Label Uncertainty}} + \underbrace{\mathbb{E}_{\mathbf{q}}[z_x]}_{\text{Ground Truth Logit Magnitude}}, \quad (6)$$

where τ is the temperature parameter, z_x is the logits of data, KL term is the loss function of ID data, $\mathbf{p}_\theta(x)$ is the output possibility, and the label entropy term $H(\mathbf{q})$ quantitatively reflects the intrinsic uncertainty of mixed data labels. While for original ID samples with one-hot label $H(\mathbf{q}) = 0$, for mixed data with two-hot label $H(\mathbf{q}) > 0$.

Therefore, *the ambiguity of mixed data is inherently higher than that of ID data, validating the role of mixed data as effective pseudo-OOD samples.*

4.3 Ambiguity-Aware Logit Adjustment

Traditional logit adjustment methods combat class imbalance by rebalancing classifier outputs through a static prior φ_c, reflecting training-set frequencies. This manifests as a fixed margin term $T(c) = \log(\varphi_c)$ in the softmax logits:

$$p(x_i) = \frac{\exp\left(z_i^{(y_i)} + T(y_i)\right)}{\sum_{c=1}^{C} \exp\left(z_i^{(c)} + T(c)\right)}. \quad (7)$$

While this static rebalancing offers a straightforward solution to class imbalance, it *fails to adapt to the training dynamics.*

Fig. 3a reveals that during early training the head classes (blue curve) exhibit a rapid decrease in energy, indicating strong learning, while the tail classes (green curve) remain with relatively high energy, reflecting limited learning progress. This suggests that at the beginning, tail classes require additional compensation to catch up. However, the static prior based on constant training-set frequencies fails to provide sufficient adjustment (i.e., it under-compensates the tail classes). In contrast, as training advances and the tail classes begin to improve, the same static prior provides an excessive level of compensation, leading to over-compensation. This dynamics illustrates the inherent limitation of fixed margins, which cannot adapt to the evolving learning state of different classes.

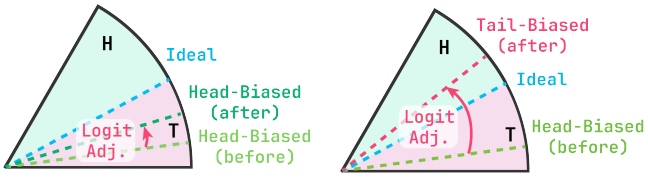


Figure 4: The under-compensation (left) and over-compensation (right) of existing logit adjustment methods. 'H' and 'T' represent the head and tail, respectively.

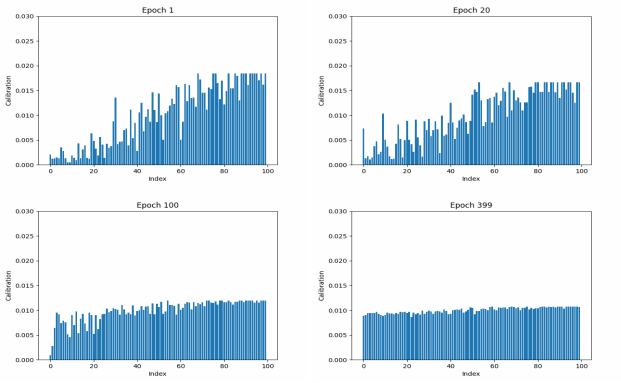


Figure 5: The calibration factor of ALA on CIFAR10-IR100. The horizontal axis is the index of the classes, sorted by the number of samples in each class. The vertical axis is $\hat{E}(z)$, but subtracting 1 for optimal visualization, i.e., $\hat{E}(z) - 1$.

This dynamics is also illustrated in Fig. 4, where the blue line (representing the ideal decision boundary) is flanked by head- and tail-biased separations (green and pink dashed lines, respectively). Consequently, incorporating a dynamic adjustment, which can leverage real-time model feedback, becomes essential to achieve balanced, context-sensitive compensation throughout the training.

To resolve this, we propose *Ambiguity-Aware Logit Adjustment (ALA)* to dynamically recalibrate the static prior via an energy function as follows:

$$E(z_i) = -\tau \log \sum_{j=1}^{C} \exp\left(z_i^{(j)}/\tau\right), \tag{8}$$

where τ is a temperature scaling factor and $E(z_i)$ measures the model's ambiguity through the log-sum-exp of the feature embeddings. Higher energy $E(z_i)$ corresponds to greater uncertainty, as the log-sum-exp term diminishes when logits lack a dominant class. Specifically, we replace the logit adjustment $T(j)$ with $\hat{T}(j, z_i)$ as

$$\hat{T}(c, z_i) = \underbrace{\hat{E}(z_i)}_{\text{Ambiguity Weight}} \cdot \underbrace{\log \varphi_c}_{\text{Static Adjustment}}, \tag{9}$$

in which $\hat{E}(z_i) = 1 + \text{softmax}\left(E(z_i)\right)$, where the softmax operation normalizes energy scores across samples, while the +1 offset ensures head classes retain non-vanishing margins. The $\hat{E}(z_i)$ scales the margin proportionally to the sample's ambiguity.

Fig. 5 demonstrates the dynamic process by which ALA adjusts the decision boundary on CIFAR100-LT. It plots the ambiguity weight, $\hat{E}(z) - 1$, on the vertical axis against class indices sorted by sample count on the horizontal axis, with classes 0–99 spanning from head (frequent) to tail (rare). At initialization, tail-class samples (right side of the horizontal axis) exhibit significantly higher $\hat{E}(z)$, reflecting their under-representation and the model's initial inability to align logits with labels. This aligns with the energy decomposition (Eq. 6):

$$-E(z_x) = \tau \text{KL}(\mathbf{q}||\mathbf{p}_\theta(x)) + \mathbb{E}_\mathbf{q}[z_x] \tag{10}$$

where the KL divergence dominates due to poor logit alignment ($\mathbb{E}_\mathbf{q}[z_x] \approx 0$). High $E(z_x)$ amplifies margins for tail classes, prioritizing their learning. Notably, some head classes also benefit marginally, as their proximity to tail regions induces ambiguity.

As training progresses, logits align more with labels ($\mathbb{E}_\mathbf{q}[z_x]$ increases), reducing KL divergence. The energy distribution converges to uniformity (Fig. 5, bottom right), signifying a balanced alignment margin, where the ambiguity weights equilibrate across all class groups, preventing over-compensation for tail or head samples. This convergence of $\hat{E}(z)$ values demonstrates ALA's ability to mitigate under-representation during early training, leveraging ambiguity weighting to counteract tail-class bias. During late training, the model achieved a stable state, where the ambiguity function approximates a uniform distribution, preventing over-compensation. This process is also shown in Fig. 3b. Besides, the stabled energy function signifies the model is well-trained for ID samples, enhancing the ability of OOD detection of ODIN, as described in Sec. 4.4.

4.4 Leveraging Ambiguity-Aware Energy for OOD Detection

The ambiguity-aware energy function $E(z)$ derived from ALA naturally bridges ID classification and OOD detection. During training, this energy function explicitly encodes uncertainty within the feature space, exhibiting lower values for confident ID samples and higher values for ambiguous or pseudo-OOD samples. Consequently, $E(z)$ effectively distinguishes well-learned ID regions from ambiguous (potentially OOD) regions, inherently making it suitable for gradient-based OOD detection methods like ODIN [28].

ODIN detects OOD samples by employing adversarial perturbations to maximize the softmax confidence of the predicted class:

$$\hat{x} = x - \epsilon \cdot \text{sign}\left(-\nabla_x \log \mathbf{p}(x)\right), \tag{11}$$

where ϵ controls the perturbation magnitude, and $\log p(y|x)$ is the log-softmax probability for the predicted class y. By decomposing $\log p(y|x)$, we obtain

$$\log p^{(y)}(x) = z^{(y)} - \log \sum_{c=1}^{C} \exp(z^{(c)}), \tag{12}$$

which allows substituting the energy function, simplifying the gradient to

$$\nabla_x \log p^{(c)}(x) = \nabla_x z^{(y)} - \nabla_x E(\mathbf{z}). \tag{13}$$

This reveals that ODIN's perturbation direction is governed by two opposing gradients: the class-specific logit gradient $\nabla_x z^{(y)}$, and the ambiguity-driven energy gradient $\nabla_x E(\mathbf{z})$.

Table 1: Comparison with LTR methods on CIFAR10-LT. The best is in bold; the second best is underlined.

Method	LTR: ACC (↑)			OOD Detection: AUROC (↑)			OOD Detection: FPR95 (↓)		
	IR10	IR50	IR100	IR10	IR50	IR100	IR10	IR50	IR100
Focal Loss [37]	90.91	83.10	79.43	71.10	63.15	63.02	74.61	87.36	87.91
LDAM+DRW [2]	89.16	83.10	79.41	54.64	53.23	53.07	100.0	100.00	100.00
RIDE [46]	90.18	84.53	80.91	59.82	56.75	56.34	96.47	97.06	95.32
SADE [60]	93.03	89.84	87.92	75.81	71.68	69.30	66.52	73.46	75.57
LGLA [40]	92.86	90.20	87.80	75.91	71.92	68.95	64.29	72.49	76.87
Ours	**94.10**	**90.97**	**89.06**	**97.28**	**91.08**	**89.76**	**10.78**	**38.60**	**40.35**

Table 2: Comparison with LTR methods on CIFAR100-LT.

Method	LTR: ACC (↑)			OOD Detection: AUROC (↑)			OOD Detection: FPR95 (↓)		
	IR10	IR50	IR100	IR10	IR50	IR100	IR10	IR50	IR100
Focal Loss [37]	62.65	50.38	45.02	65.29	62.86	62.85	80.31	82.05	81.84
LDAM+DRW [2]	60.15	48.93	42.67	59.28	60.00	58.26	95.33	93.84	90.73
RIDE [46]	64.47	52.38	47.01	63.06	62.94	60.84	94.32	90.40	88.42
SADE [60]	69.39	58.93	54.12	68.20	65.15	62.28	74.03	77.01	81.26
LGLA [40]	69.88	60.60	56.50	67.22	66.32	64.77	75.75	75.25	78.65
Ours	**71.25**	**62.19**	**57.58**	**80.85**	**79.60**	**80.62**	**53.62**	**55.06**	**55.32**

With ALA training, the energy function intrinsically encodes feature-space ambiguity, which is low in confident ID regions but high in ambiguous OOD regions, resulting in a tension between these gradients. Specifically, for ID samples, the gradient $\nabla_x z^{(y)}$ dominates, guiding perturbations towards class prototypes. In contrast, for OOD samples, the gradient $\nabla_x E(\mathbf{z})$ dominates, pushing perturbations away from the ID manifold. This gradient-based mechanism naturally enhances the separation between ID and OOD samples, thereby effectively resolving the ID-OOD dilemma without requiring external OOD samples.

5 Results and Analysis

5.1 Experimental Setups

Datasets. We compare our LTR results on the CIFAR10-LT and CIFAR100-LT datasets [2] across different imbalance ratios (IR) of 10, 50, and 100. Note that these datasets also serve as the ID data in OOD detection. For OOD detection, we also use six datasets: SVHN [33], Texture [11], Places365 [65], Tiny ImageNet [25], iNaturalist2018 [42], and LSUN [56] (for LSUN, two distinct variants are used: LSUNCropped and LSUNResized). For results of ImageNet-LT [14] and ImageNet-Extra [44], please refer to the supplementary material. Note that we only use ID data to train, which reduces the data requirement by approximately 80% compared to OE methods.

Baselines. We compare our approach with state-of-the-art methods in both conventional LTR and long-tailed OOD detection. For conventional LTR, we include simple loss reweighting approaches (Focal Loss [37] and LDAM [2]) and more sophisticated multi-expert methods (RIDE [46], SADE [60], and LGLA [40]). For long-tailed OOD, we benchmark against state-of-the-art approaches including PASCL [44], COCL [32], and EAT [49]. We employ ODIN [28], a standard and widely-used detector, to assess OOD detection performance during testing.

Evaluation protocols. Following [52], we use standard metrics for evaluating OOD detection and ID classification: FPR95, AUROC, and ACC, the classification accuracy on ID data. Note that only average OOD detection results are shown in the main text. We categorize the classes into three distinct groups: head, middle, and tail classes. For CIFAR10-LT and CIFAR100-LT, we set the thresholds based on the 1/3 and 2/3 points of the class number list, following [32, 49]. For ImageNet, the thresholds are set at 100 and 10 samples, respectively.

Implementation details. Following LGLA [40], we train our method on CIFAR10-LT and CIFAR100-LT using three ResNet-32 expert networks, and on ImageNet-LT using a ResNet-50 expert network. The initial learning rate is set to 0.1, scaled by 0.1 during a warm-up period of the first 5 epochs. We train the model for 400 epochs with the SGD optimizer (momentum=0.9), using 8 Nvidia RTX 3090 GPUs. All LTR baselines are rerun under the setting same as ours. For long-tailed OOD baselines, we retained their original configurations due to significant structural differences. To ensure fairness, we *substituted the auxiliary dataset with Gaussian noise*, as our approach does not use auxiliary data.

5.2 Comparison with Current Methods

Comparison with Current LTR Methods. Table 1 and Table 2 present comprehensive comparisons with current LTR baselines on CIFAR10-LT and CIFAR100-LT. For CIFAR10-LT, our method surpasses the state-of-the-art LGLA [40] by approximately 1.2, 0.8, and 1.3 for IR10, IR50, and IR100, respectively. Similarly, on CIFAR100-LT our approach achieves improvements of roughly 1.4, 1.6, and 1.1 over LGLA. Notably, our approach delivers dramatic gains in OOD

Table 3: Comparison with LT-OOD methods on CIFAR10-IR100. Methods marked with * indicate that the auxiliary dataset in the original method is replaced with Gaussian noise.

Method	ID ACC (↑)				OOD AUROC (↑)				OOD FPR (↓)
	Overall	Head	Middle	Tail	Overall	Head	Middle	Tail	
COCL* [32]	72.37	93.90	71.22	52.37	81.16	**91.36**	78.76	74.17	53.56
PASCL* [44]	77.51	92.63	73.38	67.90	76.61	84.58	74.17	71.89	63.53
EAT* [49]	75.99	90.87	72.05	66.37	78.07	84.79	74.64	75.93	70.17
LGLA [40]	87.87	90.97	86.30	85.30	68.95	74.16	66.20	67.40	76.87
Ours	**89.06**	**94.93**	**86.52**	**86.57**	**89.76**	84.70	**94.41**	**88.63**	**40.35**

Table 4: Comparison with LT-OOD methods on CIFAR100-IR100.

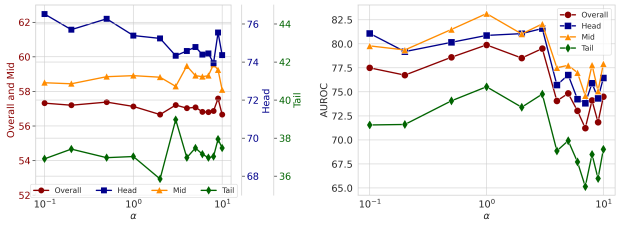
Method	ID ACC (↑)				OOD AUROC (↑)				OOD FPR (↓)
	Overall	Head	Middle	Tail	Overall	Head	Middle	Tail	
COCL* [32]	41.17	69.94	39.62	12.25	64.98	74.06	66.24	54.61	77.33
PASCL* [44]	44.55	65.65	43.29	23.47	59.37	69.30	59.55	49.25	82.46
EAT* [49]	37.43	57.62	37.35	16.06	60.71	63.46	62.54	56.09	82.61
LGLA [40]	56.15	68.58	56.76	**43.09**	64.77	71.81	64.77	57.72	78.65
Ours	**57.58**	**76.24**	**58.65**	37.82	**80.62**	**81.06**	**83.76**	**76.94**	**55.32**

Figure 6: t-SNE on CIFAR10-IR10. (L) LGLA; (R) DARL.

detection on CIFAR10-LT: it attains AUROC (%) of 97.28, 91.08, and 89.76 for IR10, IR50, and IR100—corresponding to improvements of approximately 21, 19, and 21 from LGLA's 75.91, 71.92, and 68.95. Moreover, our method substantially lowers FPR95 (%) from 64.29, 72.49, and 76.87 down to 10.78, 38.60, and 40.35 for IR10, IR50, and IR100, respectively. These results affirm that our method effectively enhances both ID recognition and OOD detection, mitigating the gradient conflict discussed in Sec. 3.2 and yielding more discriminative and compact feature representations. This conclusion is further corroborated by t-SNE in Fig. 6, where our ALA produces markedly more distinct clusters than those generated by LGLA [40].

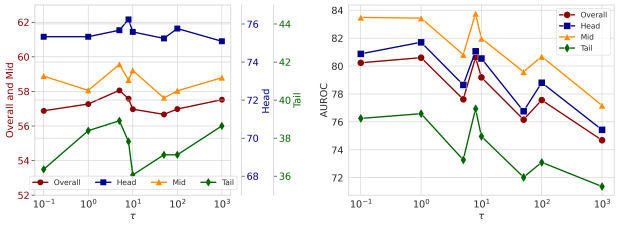
Comparison with Current LT-OOD Methods. Tables 3 and 4 compare our DARL with recent long-tailed OOD methods on CIFAR10-LT and CIFAR100-LT with an imbalance ratio of 100. On CIFAR10-LT, DARL significantly improves the overall OOD AUROC by approximately 8.6% over the previous best method (COCL), clearly demonstrating the effectiveness of leveraging mixed pseudo-OOD samples for enhanced ID-OOD separation. Notably, DARL consistently outperforms other approaches across middle and tail scenarios, underscoring the robustness and efficacy of our ALA in dynamically recalibrating class boundaries based on logits ambiguity. Similar trends emerge on CIFAR100-LT, where DARL achieves a remarkable enhancement in overall OOD AUROC by approximately 15.6%, surpassing prior methods across all class shots. This substantial improvement validates the potency of our mixed data strategy and ALA's role in mitigating the traditional dilemma between ID classification accuracy and OOD detection capability, aligning perfectly with our previous theoretical assertions.

Analysis for Different Class Groups. ALA effectively attains fairness across different class groups by dynamically adapting margins based on logits ambiguity, as evidenced in Tables 3 and 4. **ID Accuracy:** DARL sets a new state-of-the-art ID accuracy across all class groups, surpassing all prior methods. On CIFAR10-LT, DARL achieves an overall ID accuracy of 89.06%, outperforming LGLA with notable gains in head (+3.96%), middle (+0.22%), and tail (+1.27%) classes. This improvement arises from ALA's adaptive calibration strategy, which prevents under-representation of tail classes during early training and avoids overcompensation in later stages. Conversely, OE-based methods such as COCL, PASCL, and EAT significantly lag behind DARL and LGLA in ID accuracy (by over 11%). **OOD Detection:** DARL notably closes the AUROC gap between tail and OOD samples, unlike LGLA, which suffers a significant tail-class disadvantage. While COCL achieves slightly better many-shot OOD detection, DARL substantially surpasses all methods overall (+8.6% AUROC improvement on CIFAR10-LT), showcasing the effectiveness of ambiguity-aware margins in enhancing OOD discrimination across all class groups.



(a) ID Accuracy (b) AUROC

Figure 7: Ablation study of α in ALA on CIFAR100-IR100.



(a) ID Accuracy (b) AUROC

Figure 8: Ablation study of τ in ALA on CIFAR100-IR100.

5.3 Ablation Studies

Effectiveness of ALA and Mixed Data. As shown in Table 5, mixed data and Ambiguity-Aware Logit Adjustment (ALA) provide distinct and synergistic contributions. Without augmentation or ALA, the baseline shows significant limitations, achieving only 64.77% overall OOD AUROC, 56.15% ID accuracy, poor tail-class detection (61.04% AUROC), and high OOD false positives (78.65% FPR). Introducing Mixup alleviates LTR issues by reducing imbalance, as evidenced by improvements in overall, middle, and tail AUROCs, along with increased LTR accuracy. However, this method also introduces over-compensation, resulting in a 1.31% decline in head AUROC. In contrast, incorporating CutMix notably boosts OOD detection but simultaneously the improvement in ID accuracy is not as high as with Mixup. This trade-off underscores the limitation in relying solely on a single augmentation technique. Combining Mixup and CutMix resolves this issue, boosting OOD AUROC to 77.77% through complementary spatial (CutMix) and feature-space (Mixup) ambiguities. However, they still lack the awareness of training dynamics. Finally, introducing ALA helps *both OOD detection and long-tailed recognition*, offering the best performance in all cases in Table 5, which showcases the benefit of ALA from dynamically adapting to the training process.

Table 5: Ablation Study of ALA and Mixed Data.

Mixup	Cutmix	ALA	OOD AUROC (↑)				OOD FPR (↓)	ID ACC (↑)
			Overall	Head	Middle	Tail		
✗	✗	✗	64.77	71.81	64.77	61.04	78.65	56.15
✓	✗	✗	72.16	70.50	76.62	69.24	67.77	57.13
✗	✓	✗	77.22	77.65	79.12	74.81	60.94	57.01
✓	✓	✗	77.77	76.27	81.27	75.64	58.88	57.37
✓	✓	✓	**80.62**	**81.06**	**83.76**	**76.94**	**55.32**	**57.58**

Effect of α on Data Mixing. The mixing ratio $\lambda_m \sim \text{Beta}(\alpha, \alpha)$ controls both the diversity of the mixed data and its deviation from the source classes, significantly influencing model performance. (1) When α is small ($\alpha < 1$), the Beta distribution becomes bimodal, favoring values near 0 or 1. This results in mixed samples close to the original ID data, enhancing the diversity of ID samples without significant deviation from their original distribution. This setting primarily improves ID accuracy, particularly for head classes. (2) When $\alpha = 1$, the Beta distribution becomes uniform, making all mixing ratios equally probable. This maximizes the diversity of mixed samples and leads to the best AUROC across all shots, effectively enhancing OOD detection. (3) For larger α ($\alpha > 1$), the Beta distribution peaks around 0.5, producing highly ambiguous mixed samples. While this ambiguity can reduce overall ID accuracy, it helps improve mid and tail class performance due to increased diversity. However, if α becomes too large (e.g., $\alpha \approx 10$), the samples become overly ambiguous, degrading performance across all metrics as the model struggles to learn from such uncertain data. These patterns can be observed from Fig. 7.

Effect of τ on Energy Function. From Eq.8, we observe that the temperature parameter τ controls the dominant component within the energy function $E(z_i)$. Thus, we analyze how varying τ (ranging from 0.1 to 100) influences model performance, as illustrated in Fig. 8. (1) Low τ: At smaller values, the energy function primarily reflects the magnitude of the ground-truth logits ($\mathbb{E}\mathbf{q}[z_x]$), i.e., $E(z_x) \approx -\mathbb{E}\mathbf{q}[z_x]$, favoring well-represented classes and failing to address over-confidence in under-represented ones. (2) Medium τ: As τ increases, KL and entropy terms gain influence. Since they vary inversely, they may cancel each other at an intermediate τ, making the entropy term $\tau H(\mathbf{q})$ dominant. Given the higher uncertainty in our mixed data with two-hot labels, this can lead the model to misclassify ambiguous pseudo-OOD data as ID, boosting ID accuracy while hurting AUROC. (3) High τ: Further increase allows the KL term to dominate (e.g., at $\tau = 8$), balancing ID accuracy and AUROC. Yet overly large τ leads to overfitting on ID data, again reducing AUROC.

6 Conclusion

In this paper, we tackle two core challenges in LTOOD learning: the scarcity of OOD training data and the gradient conflicts introduced by OE methods. Our analysis shows that OE can cause harmful conflicts, especially when ID and OOD samples share features—a problem that's more pronounced for tail classes. To address this, we propose using mixed ID data as pseudo-OOD, which exhibits higher ambiguity and helps ease the trade-off. Building on this, we introduce DARL, a dynamic logit recalibration framework that leverages energy-based ambiguity to align competing objectives and boost tail performance. We further show that ODIN's adversarial perturbations naturally complement our approach, enhancing OOD detection. Extensive experiments confirm DARL's state-of-the-art results on both ID classification and OOD detection.

7 Acknowledgement

This work is partly supported by National Key R&D Program of China(No.2023YFB4302200) and the Special Foundations for the Development of Strategic Emerging Industries of Shenzhen (No.KJZD20231023094700001).

References

[1] Jianhong Bai, Zuozhu Liu, Hualiang Wang, Jin Hao, Yang Feng, Huanpeng Chu, and Haoji Hu. 2022. On the Effectiveness of Out-of-Distribution Data in Self-Supervised Long-Tail Learning. In *International Conference on Learning Representations*.

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

[3] Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. 2020. Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts. In *Advances in Neural Information Processing Systems* (2020), Vol. 33. 1356–1367.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.

[5] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. 2021. Atom: Robustifying out-of-distribution detection using outlier mining. In *Machine Learning and Knowledge Discovery in Databases*. Springer, 430–445.

[6] Jiahao Chen and Bing Su. 2023. Transfer Knowledge from Head to Tail: Uncertainty Calibration under Long-tailed Distribution. In *IEEE Conference on Computer Vision and Pattern Recognition*. 19978–19987.

[7] Xiaohua Chen, Yucan Zhou, Dayan Wu, Chule Yang, Bo Li, Qinghua Hu, and Weiping Wang. 2023. AREA: Adaptive Reweighting via Effective Area for Long-Tailed Classification. In *International Conference on Computer Vision*. 19277–19287.

[8] Yongqiang Chen, Kaiwen Zhou, Yatao Bian, Binghui Xie, Bingzhe Wu, Yonggang Zhang, MA KAILI, Han Yang, Peilin Zhao, Bo Han, et al. 2023. Pareto Invariant Risk Minimization: Towards Mitigating the Optimization Dilemma in Out-of-Distribution Generalization. In *International Conference on Learning Representations*.

[9] Yuan-Chih Chen and Chun-Shien Lu. 2023. RankMix: Data Augmentation for Weakly Supervised Learning of Classifying Whole Slide Images With Diverse Sizes and Imbalanced Categories. In *IEEE Conference on Computer Vision and Pattern Recognition*. 23936–23945.

[10] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. 2020. Remix: Rebalanced Mixup. In *European Conference on Computer Vision*. 95–110.

[11] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3606–3613.

[12] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. 2023. ResLT: Residual Learning for Long-Tailed Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2023), 3695–3706.

[13] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9268–9277.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 248–255.

[15] Fei Du, Peng Yang, Qi Jia, Fengtao Nan, Xiaoting Chen, and Yun Yang. 2023. Global and Local Mixture Consistency Cumulative Learning for Long-Tailed Visual Recognitions. In *IEEE Conference on Computer Vision and Pattern Recognition*. 15814–15823.

[16] Chengjian Feng, Yujie Zhong, and Weilin Huang. 2021. Exploring Classification Equilibrium in Long-Tailed Object Detection. In *International Conference on Computer Vision*. 3397–3406.

[17] Shuai Feng and Chongjun Wang. 2024. When an Extra Rejection Class Meets Out-of-Distribution Detection in Long-Tailed Image Classification. *Neural Networks* 178 (2024), 106485.

[18] Boyu Han, Qianqian Xu, Zhiyong Yang, Shilong Bao, Peisong Wen, Yangbangyan Jiang, and Qingming Huang. 2024. Aucseg: Auc-oriented pixel-level long-tail semantic segmentation. *Advances in Neural Information Processing Systems* 37 (2024), 126863–126907.

[19] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. 2005. Borderline-SMOTE: A New over-Sampling Method in Imbalanced Data Sets Learning. In *International Conference on Intelligent Computing*. 878–887.

[20] Yina He, Lei Peng, Yongcun Zhang, Juanjuan Weng, Zhiming Luo, and Shaozi Li. 2025. Long-Tailed Out-of-Distribution Detection: Prioritizing Attention to Tail. In *AAAI Conference on Artificial Intelligence*.

[21] Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* (2016).

[22] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. 2019. Deep Anomaly Detection with Outlier Exposure. In *International Conference on Learning Representations*.

[23] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. 2023. Detecting out-of-distribution data through in-distribution class prior. In *International Conference on Machine Learning*. 15067–15088.

[24] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2019. Decoupling Representation and Classifier for Long-Tailed Recognition. In *International Conference on Learning Representations*.

[25] Ya Le and Xuan Yang. 2015. Tiny imagenet visual recognition challenge. *CS 231N* 7, 7 (2015), 3.

[26] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. 2018. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *International Conference on Learning Representations* (2018).

[27] Buyu Li, Yu Liu, and Xiaogang Wang. 2019. Gradient Harmonized Single-Stage Detector. In *AAAI Conference on Artificial Intelligence*, Vol. 33. 8577–8584.

[28] Shiyu Liang, Yixuan Li, and R. Srikant. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*.

[29] X.W. Liang, A.P. Jiang, T. Li, Y.Y. Xue, and G.T. Wang. 2020. LR-SMOTE — An Improved Unbalanced Data Set Oversampling Based on K-means and SVM. *Knowledge-Based Systems* 196 (2020), 105845.

[30] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-Based Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*, Vol. 33. 21464–21475.

[31] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2021. Long-Tail Learning via Logit Adjustment. In *International Conference on Learning Representations*. arXiv:2007.07314 [cs, stat]

[32] Wenjun Miao, Guansong Pang, Xiao Bai, Tianqi Li, and Jin Zheng. 2024. Out-of-Distribution Detection in Long-Tailed Recognition with Calibrated Outlier Class Learning. In *AAAI Conference on Artificial Intelligence*, Vol. 38. 4216–4224.

[33] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.

[34] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. 2023. Nearest neighbor guidance for out-of-distribution detection. In *International Conference on Computer Vision*. 1686–1695.

[35] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. 2020. Balanced Meta-Softmax for Long-Tailed Visual Recognition. In *Advances in Neural Information Processing Systems*, Vol. 33. 4175–4186.

[36] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-Distribution Detection with Deep Nearest Neighbors. In *International Conference on Machine Learning*. 20827–20840.

[37] T. -Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. 0022/2017-10-29. Focal Loss for Dense Object Detection. In *International Conference on Computer Vision*. 2999–3007.

[38] Jingru Tan, Bo Li, Xin Lu, Yongqiang Yao, Fengwei Yu, Tong He, and Wanli Ouyang. 2023. The Equalization Losses: Gradient-Driven Training for Long-tailed Object Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 11 (2023), 13876–13892.

[39] Zichang Tan, Jun Li, Jinhao Du, Jun Wan, Zhen Lei, and Guodong Guo. 2024. NCL++: Nested Collaborative Learning for Long-Tailed Visual Recognition. *Pattern Recognition* 147 (2024), 110064.

[40] Yingfan Tao, Jingna Sun, Hao Yang, Li Chen, Xu Wang, Wenming Yang, Daniel Du, and Min Zheng. 2023. Local and Global Logit Adjustments for Long-Tailed Learning. In *International Conference on Computer Vision*. 11783–11792.

[41] Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. 2023. ID and OOD Performance Are Sometimes Inversely Correlated on Real-world Datasets. In *Advances in Neural Information Processing Systems*, Vol. 36. 71703–71722.

[42] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The INaturalist Species Classification and Detection Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*. 8769–8778.

[43] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. 2022. ViM: Out-of-Distribution With Virtual-Logit Matching. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4921–4930.

[44] Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J. Smola, and Zhangyang Wang. 2022. Partial and Asymmetric Contrastive Learning for Out-of-Distribution Detection in Long-Tailed Recognition. In *International Conference on Machine Learning*. 23446–23458.

[45] Qizhou Wang, Zhen Fang, Yonggang Zhang, Feng Liu, Yixuan Li, and Bo Han. 2023. Learning to Augment Distributions for Out-of-distribution Detection. *Advances in Neural Information Processing Systems* 36 (2023), 73274–73286.

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

[47] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. 2019. Dynamic Curriculum Learning for Imbalanced Data Classification. In *International Conference on Computer Vision*. 5016–5025.

[48] Hongxin Wei, Lue Tao, Renchunzi Xie, Lei Feng, and Bo An. 2022. Open-Sampling: Exploring Out-of-Distribution Data for Re-balancing Long-tailed Datasets. In *International Conference on Machine Learning*. 23615–23630.

[49] Tong Wei, Bo-Lin Wang, and Min-Ling Zhang. 2024. EAT: Towards Long-Tailed Out-of-Distribution Detection. In *AAAI Conference on Artificial Intelligence*, Vol. 38. 15787–15795.

[50] Liuyu Xiang, Guiguang Ding, and Jungong Han. 2020. Learning From Multiple Experts: Self-paced Knowledge Distillation for Long-Tailed Classification. In *European Conference on Computer Vision (Lecture Notes in Computer Science, Vol. 12350)*. 247–263.

[51] Zhengzhuo Xu, Ruikang Liu, Shuo Yang, Zenghao Chai, and Chun Yuan. 2023. Learning Imbalanced Data With Vision Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*. 15793–15803.

[52] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. 2022. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems* 35 (2022), 32598–32611.

[53] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. 2023. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision* 131, 10 (2023), 2607–2622.

[54] Taocun Yang, Yaping Huang, Yanlin Xie, Junbo Liu, and Shengchun Wang. 2023. MixOOD: Improving Out-of-distribution Detection with Enhanced Data Mixup. *ACM Transactions on Multimedia Computing, Communications, and Applications* 19, 5 (2023), 155:1–155:18.

[55] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. 2019. Feature Transfer Learning for Face Recognition With Under-Represented Data. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5697–5706.

[56] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365* (2015).

[57] Zhuoning Yuan, Dixian Zhu, Zi-Hao Qiu, Gang Li, Xuanhui Wang, and Tianbao Yang. 2023. Libauc: A deep learning library for x-risk optimization. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*. 5487–5499.

[58] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *International Conference on Computer Vision*. 6023–6032.

[59] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. Mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

[60] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. 2022. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. *Advances in Neural Information Processing Systems* 35 (2022), 34077–34090.

[61] Qihao Zhao, Chen Jiang, Wei Hu, Fan Zhang, and Jun Liu. 2023. MDCS: More Diverse Experts with Consistency Self-distillation for Long-tailed Recognition. In *International Conference on Computer Vision*. 11597–11608.

[62] Hongwei Zheng, Linyuan Zhou, Han Li, Jinming Su, Xiaoming Wei, and Xiaoming Xu. 2024. BEM: Balanced and Entropy-based Mix for Long-Tailed Semi-Supervised Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[63] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. 2021. Improving Calibration for Long-Tailed Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 16484–16493.

[64] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. 2020. Bbn: Bilateral-branch Network with Cumulative Learning for Long-Tailed Visual Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9719–9728.

[65] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2017), 1452–1464.