GEOBIAS: GEOMETRIC BIAS CORRECTION VIA DECISION BOUNDARY TRANSLATION FOR LONG-TAILED RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world visual recognition tasks often exhibit long-tailed label distributions, where a few head classes dominate training while most tail classes are underrepresented. Standard empirical risk minimization biases classifiers toward head classes, severely degrading tail-class generalization. Existing decoupling methods mainly adjust classifier weights, overlooking the bias term as an independent lever for decision-boundary correction. We propose GeoBias, a geometry-driven bias correction method that fixes the learned weight directions and analytically computes bias shifts from class centroids and pairwise boundary geometry. Multi-class conflicts are resolved by weighted aggregation, and a time-decaying anchoring regularizer prevents SGD from eroding the corrected biases. GeoBias is single-stage and plug-and-play, requiring no data re-sampling, architectural modification, or complex training schedule. Experiments on CIFAR-10-LT, CIFAR-100-LT, and ImageNet-LT show that GeoBias consistently improves tail-class accuracy and achieves competitive or superior performance relative to single-stage and two-stage methods. Code will be released upon acceptance

1 INTRODUCTION

Real-world visual recognition tasks typically exhibit a long-tailed label distribution(Liu et al., 2019), where a few head classes occupy the vast majority of training samples, while most tail classes have only a limited number of instances(Van Horn & Perona, 2017). Such imbalanced distributions pose severe challenges for training deep neural networks(Buda et al., 2018). When trained with standard empirical risk minimization (ERM)(Vapnik & Chervonenkis, 1991), the model inherently favors head classes, since their samples dominate the total loss, whereas the contribution from tail classes is diluted(He & Garcia, 2009; Cui et al., 2019), leading to significantly degraded generalization performance on tail classes(Zhang et al., 2023).

To alleviate this issue, a substantial body of research has been proposed, which can be broadly categorized into three groups: data re-sampling(Chawla et al., 2002), loss function modification(Lin et al., 2017), and classifier decoupling(Kang et al., 2019). Data re-sampling balances the training distribution by over-sampling tail classes or under-sampling head classes, yet each approach has inherent drawbacks(Shi et al., 2023; Buda et al., 2018). Loss modification methods increase the influence of tail classes via re-weighting(Cui et al., 2019) or logit adjustment(Menon et al., 2020). However, the adjustment magnitude of these methods relies primarily on prior assumptions based on sample counts(Cao et al., 2019), representing a global and static correction that fails to account for the actual geometric distribution of each class in the feature space(Yang et al., 2022). Consequently, they are inherently insufficient to rectify the systematic geometric distortion suffered by the classifier bias term.

Under the decoupling paradigm(Kang et al., 2019), the classifier itself, rather than the feature extractor, has been identified as the primary source of bias in long-tailed classification. Representative works such as cRT and LWS(Kang et al., 2019) follow this rationale, yet they almost exclusively intervene on the weight matrix $\mathbf{W}$ of the classifier(Zhang et al., 2021). These approaches adjust the decision boundary normals by distorting the weight space, which may disrupt the already learned

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

discriminative directions(Liu et al., 2016). Critically, none of them explicitly treat the bias term b as an independent object for correction.

In this paper, we argue that this "weight-only" perspective is incomplete. The position of the decision boundary is jointly determined by the weight-feature coupling term $\Delta\mathbf{W} \cdot \boldsymbol{\mu}_i$ and the bias difference Δb(Dai & Jin'no, 2026). Modifying the former risks introducing discrepant logit scales that degrade generalization (Mai et al., 2024), while adjusting the latter only shifts the boundary position without altering the learned feature geometry (Dai & Jin'no, 2026). Thus, the bias term provides a simpler and safer lever for boundary correction, and directly computing the required adjustment via geometric analysis offers a more effective rectification strategy that is agnostic to data distribution and preserves learned representations (Menon et al., 2021; Anonymous, 2026).

Beyond weight fine-tuning methods, researchers have also explored architectural designs and training strategies, such as GCL(Li et al., 2022), MiSLAS(Zhong et al., 2021), BBN(Zhou et al., 2020), and DisAlign(Zhang et al., 2021). While these methods achieve notable results, they typically require modifying network architectures, relying on complex multi-stage training procedures, or both(Fu et al., 2022). Notably, although LDAM(Cao et al., 2019) and ALA Loss(Zhao et al., 2022) are formulated as loss functions, their effective implementation relies on replacing the fully-connected layer with a normalized linear layer (NormedLinear)(Cao et al., 2019), and LDAM is often paired with its two-stage DRW training strategy(Cao et al., 2019) to achieve optimal performance.

Motivated by the above analyses, we propose GeoBias, a geometry-driven classifier bias correction method that explicitly decouples weight directions from bias positions, with a time-decaying bias anchoring regularizer to prevent SGD from washing out the correction. GeoBias adopts a single-stage, plug-and-play design requiring no data re-sampling, architectural modification, or complex training strategies.

- We reveal a blind spot in existing decoupling methods: they mainly adjust the classifier weight matrix $\mathbf{W}$, while overlooking the bias term $\mathbf{b}$ as an independent lever that translates the decision boundary without altering the learned weight directions.
- We propose GeoBias, a heuristic geometry-driven method that fixes the learned weight directions and analytically computes bias shifts from class centroids and pairwise boundary geometry, with weighted averaging to resolve multi-class conflicts.
- GeoBias is single-stage and plug-and-play, requiring no data re-sampling, architectural modification, or complex training schedule. Experiments show competitive or superior performance to two-stage methods with high training efficiency.

2 PRELIMINARY

2.1 BASIC SETUP

Let the training set be $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathcal{Y} = \{1, \dots, C\}$. Let $n_c = \sum_{i=1}^N \mathbf{1}[y_i = c]$ satisfy $n_1 \geq n_2 \geq \cdots \geq n_C$ and $n_1 \gg n_C$. The model consists of a feature extractor $\Phi : \mathbb{R}^d \to \mathbb{R}^D$ and a linear classifier

$$f(\mathbf{x}) = \mathbf{W}^\top \Phi(\mathbf{x}) + \mathbf{b}, \tag{1}$$

where $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_C] \in \mathbb{R}^{D \times C}$ and $\mathbf{b} = [b_1, \dots, b_C]^\top \in \mathbb{R}^C$. The logit of class c is $z_c = \mathbf{w}_c^\top \Phi(\mathbf{x}) + b_c$. Standard ERM minimizes

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(z_{y_i})}{\sum_{c=1}^C \exp(z_c)}. \tag{2}$$

Under long-tailed distributions, minimizing $\mathcal{L}_{\mathrm{CE}}$ biases the classifier toward head classes(Cao et al., 2019).

2.2 GEOMETRIC CHARACTERIZATION OF THE CLASSIFIER

For a class pair (i, j), define

$$\Delta\mathbf{W}_{ij} = \mathbf{w}_i - \mathbf{w}_j, \qquad \Delta b_{ij} = b_i - b_j. \tag{3}$$

The logit difference is

$$a_{ij}(\mathbf{x}) = \Delta\mathbf{W}_{ij}^{\top}\Phi(\mathbf{x}) + \Delta b_{ij}. \tag{4}$$

The decision boundary is $a_{ij}(\mathbf{x}) = 0$, where $\Delta\mathbf{W}_{ij}$ controls the boundary orientation and Δb_{ij} controls its translation(Bishop, 2006). The geometric signed distance is

$$d_{ij}(\mathbf{x}) = \frac{a_{ij}(\mathbf{x})}{\|\Delta\mathbf{W}_{ij}\|} = \frac{\Delta\mathbf{W}_{ij}^{\top}\Phi(\mathbf{x}) + \Delta b_{ij}}{\|\Delta\mathbf{W}_{ij}\|}. \tag{5}$$

The sign of $d_{ij}(\mathbf{x})$ indicates the side of the boundary.

Define the class centroid(Snell et al., 2017)

$$\boldsymbol{\mu}_c = \frac{1}{n_c}\sum_{k:y_k=c}\Phi(\mathbf{x}_k) \in \mathbb{R}^D. \tag{6}$$

For pair (i, j), let

$$\mathbf{v}_{ij} = \boldsymbol{\mu}_j - \boldsymbol{\mu}_i, \qquad \mathbf{h}(\lambda) = \boldsymbol{\mu}_i + \lambda\mathbf{v}_{ij}, \quad \lambda \in \mathbb{R}. \tag{7}$$

Here $\lambda = 0$ corresponds to $\boldsymbol{\mu}_i$, $\lambda = 1$ to $\boldsymbol{\mu}_j$, and $\lambda = 0.5$ to the geometric midpoint. Substituting $\mathbf{h}(\lambda)$ into equation 4 yields

$$a_{ij}(\lambda) = a_{ij}^{(i)} + \lambda q_{ij}, \qquad a_{ij}^{(i)} = \Delta\mathbf{W}_{ij}^{\top}\boldsymbol{\mu}_i + \Delta b_{ij}, \qquad q_{ij} = \Delta\mathbf{W}_{ij}^{\top}\mathbf{v}_{ij}. \tag{8}$$

The boundary intersection with the center line satisfies $a_{ij}(\lambda) = 0$, giving

$$\lambda_{\text{curr}}^{ij} = -\frac{\Delta\mathbf{W}_{ij}^{\top}\boldsymbol{\mu}_i + \Delta b_{ij}}{q_{ij}}, \qquad a_{ij}(\lambda) = q_{ij}(\lambda - \lambda_{\text{curr}}^{ij}). \tag{9}$$

Along the center line, if $q_{ij} > 0$, class i occupies $\lambda > \lambda_{\text{curr}}^{ij}$ and class j occupies $\lambda < \lambda_{\text{curr}}^{ij}$; if $q_{ij} < 0$, the occupancies are reversed. The value $\lambda_{\text{curr}}^{ij} = 0.5$ is the geometric midpoint of the center line. We use it only as a reference to measure center-line occupancy imbalance and to identify the head/tail direction. It is not assumed to be the optimal boundary position. For convenience, we denote by $\lambda_{\text{target}}^{ij}$ the desired center-line boundary position for the ordered pair (i, j), i.e., the value of λ at which the decision boundary $a_{ij}(\lambda) = 0$ should intersect the center line after bias correction. Its concrete construction is deferred to Section 3.

Let Δb_{ij} be increased by δ_{ij} while keeping $\Delta\mathbf{W}_{ij}$ fixed. From equation 9,

$$\lambda_{\text{new}}^{ij} = \lambda_{\text{curr}}^{ij} - \frac{\delta_{ij}}{q_{ij}}, \qquad \delta_{ij} = q_{ij}\left(\lambda_{\text{curr}}^{ij} - \lambda_{\text{target}}^{ij}\right), \tag{10}$$

where $\lambda_{\text{target}}^{ij}$ is the desired boundary position introduced above. Thus bias adjustment translates the boundary linearly without perturbing $\mathbf{W}$. Equivalently, for any fixed $\mathbf{x}$, the signed distance shifts uniformly by $\delta_{ij}/\|\Delta\mathbf{W}_{ij}\|$. This fact will be used in Section 3 to instantiate $\lambda_{\text{target}}^{ij}$ from boundary-channel samples. Since δ_{ij} measures the increment of $\Delta b_{ij} = b_i - b_j$, it changes sign under order reversal:

$$\delta_{ji} = -\delta_{ij}. \tag{11}$$

For a head-tail pair (h, t), the update is

$$b_h \leftarrow b_h + \frac{\delta_{ht}}{2}, \qquad b_t \leftarrow b_t - \frac{\delta_{ht}}{2}, \tag{12}$$

where δ_{ht} is the increment of $\Delta b_{ht} = b_h - b_t$.

3 METHOD

GeoBias is a heuristic geometry-driven bias correction method. It runs in parallel with standard single-stage training and consists of periodic geometric calibration and a bias anchoring regularizer. It does not aim to find the Bayes-optimal decision boundary. No data re-sampling, network modification, or multi-stage training is required.

3.1 GEOMETRIC QUANTITIES COLLECTION

Each calibration performs one full forward pass over the training set to collect features $\Phi(\mathbf{x})$, labels, and the Symmetric Probability Difference (SPD). For a sample $\mathbf{x}$, let $\mathbf{t}(\mathbf{x}) = [t_1(\mathbf{x}), \ldots, t_C(\mathbf{x})]^\top \in [0,1]^C$ be its label distribution and $\mathbf{p}(\mathbf{x}) = \mathrm{softmax}(\mathbf{z}(\mathbf{x}))$. Define

$$m(\mathbf{x}) = \sum_{c=1}^{C} t_c(\mathbf{x}) p_c(\mathbf{x}) - \max_{c:t_c(\mathbf{x})=0} p_c(\mathbf{x}), \qquad m_{\mathrm{opt}}(\mathbf{x}) = \sum_{c=1}^{C} t_c(\mathbf{x})^2, \tag{13}$$

and

$$\mathrm{SPD}(\mathbf{x}) = \begin{cases} m(\mathbf{x})/m_{\mathrm{opt}}(\mathbf{x}), & m(\mathbf{x}) \le m_{\mathrm{opt}}(\mathbf{x}), \\ m_{\mathrm{opt}}(\mathbf{x})/m(\mathbf{x}), & m(\mathbf{x}) > m_{\mathrm{opt}}(\mathbf{x}). \end{cases} \tag{14}$$

If $\{c : t_c(\mathbf{x}) = 0\}$ is empty, we set $\max_{c:t_c(\mathbf{x})=0} p_c(\mathbf{x}) = 0$. For one-hot labels, $\mathrm{SPD}(\mathbf{x}) > 0$ indicates a correct prediction; for soft labels, it measures the matching between the prediction and the label distribution. $|\mathrm{SPD}(\mathbf{x})|$ reflects the soft margin. We also compute

$$\kappa_c = \frac{1}{n_c} \sum_{k:y_k=c} \|\Phi(\mathbf{x}_k) - \boldsymbol{\mu}_c\|. \tag{15}$$

3.2 PAIRWISE GEOMETRIC ANALYSIS AND BIAS ADJUSTMENT

For each pair (i, j), we compute $\lambda_{\mathrm{curr}}^{ij}$ and identify the head and tail classes according to their center-line occupancy. Let (h, t) denote the resulting ordered head-tail pair, and re-index all boundary quantities by (h, t). That is, $\Delta\mathbf{W}_{ht} = \mathbf{w}_h - \mathbf{w}_t$, $\Delta b_{ht} = b_h - b_t$, and $\lambda_{\mathrm{curr}}^{ht}$ denotes the corresponding center-line intersection. For brevity, we omit the elementary index conversion; it follows directly from $\Delta b_{ht} = -\Delta b_{th}$.

We define the boundary channel as the region within distance

$$\Delta_{\mathrm{channel}} = r_{ht} \|\boldsymbol{\mu}_t - \boldsymbol{\mu}_h\| \tag{16}$$

from the decision boundary, where

$$r_{ht} = \frac{\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_h\| - \kappa_h - \kappa_t}{\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_h\|}, \tag{17}$$

clipped to a lower bound $r_{\min} = 0.1$. Within the boundary channel, we collect head-class samples that are correctly classified and have $\mathrm{SPD} > 0$:

$$S_{ht} = \left\{ \mathbf{x} : y = h,\ \arg\max_c p_c(\mathbf{x}) = h,\ \mathrm{SPD}(\mathbf{x}) > 0,\ |d_{ht}(\mathbf{x})| \le \Delta_{\mathrm{channel}} \right\}. \tag{18}$$

For soft labels, the correctness condition can be replaced by $\mathrm{SPD}(\mathbf{x}) > 0$. If $S_{ht} = \varnothing$, the pair is skipped. Otherwise, let $\bar{d}_{ht}$ be the average signed geometric distance of samples in S_{ht} to the boundary $a_{ht}(\mathbf{x}) = 0$. Since $\Delta\mathbf{W}_{ht}$ is fixed during bias correction, changing Δb_{ht} by δ_{ht} shifts every signed distance uniformly by $\delta_{ht}/\|\Delta\mathbf{W}_{ht}\|$. Therefore, making the average signed distance of S_{ht} equal to zero requires

$$\delta_{ht} = -\|\Delta\mathbf{W}_{ht}\| \bar{d}_{ht}. \tag{19}$$

Using the relation between Δb_{ht} and $\lambda_{\mathrm{curr}}^{ht}$, this corresponds to the center-line target

$$\lambda_{\mathrm{target}}^{ht} = \lambda_{\mathrm{curr}}^{ht} + \frac{\|\Delta\mathbf{W}_{ht}\|}{q_{ht}} \bar{d}_{ht}, \tag{20}$$

and the increment of $\Delta b_{ht} = b_h - b_t$ is

$$\delta_{ht} = q_{ht} \left(\lambda_{\mathrm{curr}}^{ht} - \lambda_{\mathrm{target}}^{ht} \right). \tag{21}$$

This derivation does not require the selected samples to lie on the center line. It is a heuristic target that pushes the boundary toward the head-class side, i.e., toward the region occupied by the selected head-class samples, thereby compressing the head-class center-line occupancy, until these samples are no longer deeply inside the head side. It is not assumed to be optimal. The bias is then updated as

$$b_h \leftarrow b_h + \frac{\delta_{ht}}{2}, \qquad b_t \leftarrow b_t - \frac{\delta_{ht}}{2}. \tag{22}$$

3.3 Global Aggregation via Weighted Averaging

Each bias b_c participates in $C-1$ pairwise boundaries, so the pairwise demands may conflict. For pair (i,j), define the center-line occupancies

$$\mathrm{occ}_i = \begin{cases} 1-\lambda^{ij}_{\mathrm{curr}}, & q_{ij}>0, \\ \lambda^{ij}_{\mathrm{curr}}, & q_{ij}<0, \end{cases} \qquad \mathrm{occ}_j = 1-\mathrm{occ}_i, \qquad \rho_{ij} = \max(\mathrm{occ}_i, \mathrm{occ}_j). \tag{23}$$

For normal pairs with $0 \le \lambda^{ij}_{\mathrm{curr}} \le 1$, the weight is

$$w_{ij} = \frac{\rho_{ij}}{1-\rho_{ij}}, \tag{24}$$

where ρ_{ij} is clipped to $1-\epsilon$ with a small $\epsilon>0$ to avoid singularity. Thus $w_{ij}=1$ at $\lambda^{ij}_{\mathrm{curr}}=0.5$ and increases as the imbalance becomes more severe. Let $w_{\max}$ be the maximum weight among normal pairs, computed before assigning abnormal pairs. The aggregated adjustment for class c is

$$\delta b_c = \frac{\sum_{(i,j):\, i<j,\, c\in\{i,j\}} w_{ij}\cdot \mathrm{contrib}_{ij\to c}}{\sum_{(i,j):\, i<j,\, c\in\{i,j\}} w_{ij}}, \qquad \mathrm{contrib}_{ij\to c} = \begin{cases} \delta_{ht}/2, & c=h, \\ -\delta_{ht}/2, & c=t, \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

Here (h,t) is the head-tail pair corresponding to (i,j), and δ_{ht} is computed by equation 21. Severely imbalanced pairs receive larger weights and dominate the aggregation. This aggregation is a heuristic conflict-resolution step, not an exact multi-class geometric optimization.

3.4 Protective Measures

Abnormal cases. During early training, r_{ht} and $\lambda^{ij}_{\mathrm{curr}}$ may be unstable. We clip r_{ht} to $r_{\min}=0.1$. Pairs with $\lambda^{ij}_{\mathrm{curr}}<0.1$ or >0.9 are treated as abnormal. To prevent them from affecting the weight scale of reliable pairs, abnormal pairs are deferred: $w_{\max}$ is computed only over normal pairs, and deferred pairs are aggregated afterwards with $w=w_{\max}$. If a deferred pair has no valid δ_{ht}, it is skipped.

Head/tail verification. For each pair (i,j), we verify $n_h > n_t$; otherwise the pair is skipped. Under mild imbalance, the model may naturally reach a balanced solution, and manual adjustment could be harmful.

Bias anchoring regularization. Between calibrations, SGD may erode the corrected biases. We anchor the bias to the last calibrated reference $\mathbf{b}^{\mathrm{ref}}$:

$$\mathcal{L}_{\mathrm{anchor}} = \gamma(\tau)\frac{1}{C}\sum_{c=1}^{C}\alpha_c(b_c-b_c^{\mathrm{ref}})^2, \qquad \gamma(\tau)=\max\left(0,1-\frac{\tau}{T}\right), \tag{26}$$

where τ is the number of epochs since the last calibration, T is the calibration frequency, and

$$\alpha_c = \frac{\|\mathbf{w}_c\|}{\frac{1}{C}\sum_{k=1}^{C}\|\mathbf{w}_k\|}\|\boldsymbol{\mu}_c\|^2. \tag{27}$$

The design of α_c follows a gradient-magnitude rationale: near the class centroid, a weight update perturbs the logit by an amount of order $\|\boldsymbol{\mu}_c\|^2$, whereas a bias update perturbs it only by a constant; meanwhile, a larger $\|\mathbf{w}_c\|$ indicates stronger weight drift for head classes. Thus α_c adaptively strengthens bias anchoring for classes with larger centroid norm and weight magnitude, preventing SGD from washing out the calibrated bias while keeping the correction mild for tail classes. This is a magnitude-based heuristic rather than a strict equality. The anchoring is strongest immediately after calibration and fades gradually.

4 Experiments

4.1 Datasets

We evaluate our method on three widely used long-tailed recognition benchmarks: CIFAR-10-LT, CIFAR-100-LT, and ImageNet-LT.

Algorithm 1 GeoBias Training Procedure

Require: Training set $\mathcal{D}$, model $f(\mathbf{x}) = \mathbf{W}^\top \Phi(\mathbf{x}) + \mathbf{b}$, calibration frequency T, warmup period E_w
Ensure: Trained model with rectified biases

```
Initialize model parameters; b_ref ← b; τ ← 0
for epoch = 1 to max_epoch do
    for each mini-batch (x, y) ∈ D do
        Compute L_CE and L_anchor
        L ← L_CE + L_anchor
        Update parameters via back-propagation
    end for
    if epoch ≥ E_w and (epoch − E_w) mod T = 0 then
        Geometric Calibration:
        Collect features Φ(x), labels, and SPD for all training samples
        Compute class centroids μ_c and compactness κ_c
        for each valid pair (i, j) with i < j do
            Compute λ^ij_curr, identify (h, t), and determine r_ht, λ^ht_target
            Compute δ_ht and weight w_ij; skip if n_h ≤ n_t
        end for
        Aggregate: δb_c = (Σ w_ij · contrib_{ij→c}) / (Σ w_ij)
        Update b ← b + δb; b_ref ← b; τ ← 0
    else
        τ ← τ + 1
    end if
end for
return Trained model
```

CIFAR-10-LT and CIFAR-100-LT. The original CIFAR-10 and CIFAR-100 datasets(Krizhevsky et al., 2009) consist of 50,000 training images and 10,000 test images with 10 and 100 classes, respectively. Following the standard protocol established in prior work(Cao et al., 2019), we create long-tailed versions by subsampling the original training set according to an exponential decay function over class indices. Specifically, for class $c \in \{0, 1, \ldots, C-1\}$, the number of retained training samples is:

$$n_c = n_{\max} \cdot \mu^{\frac{c}{C-1}}, \tag{28}$$

where $n_{\max}$ is the original sample count per class (5,000 for CIFAR-10 and 500 for CIFAR-100), and $\mu \in (0, 1)$ controls the imbalance degree. The imbalance ratio is defined as $\gamma = n_{\max}/n_{\min}$, where $n_{\min}$ is the sample count of the least frequent class. We adopt four imbalance ratios: $\gamma \in \{10, 50, 100, 200\}$, corresponding to $\mu = 0.1, 0.02, 0.01, 0.005$, respectively(Cui et al., 2019). The original test sets are kept unchanged for evaluation.

ImageNet-LT. ImageNet-LT is a long-tailed version of the large-scale ImageNet-2012 dataset(Russakovsky et al., 2015), constructed by sampling a subset following a Pareto distribution with power value $\alpha = 6$(Liu et al., 2019). The dataset contains 115.8K training images from 1,000 categories, with the number of samples per class ranging from 5 to 1,280. The imbalance ratio is approximately 1,280 / 5 = 256. As with CIFAR-LT, the original balanced validation set is used for testing.

4.2 Implementation Details

For CIFAR-10-LT and CIFAR-100-LT, we use ResNet-32 as the backbone(He et al., 2016). Models are trained for 200 epochs with SGD momentum 0.9 and weight decay 5×10^{-4}. The initial learning rate is 0.2 for CIFAR-10-LT and 0.1 for CIFAR-100-LT, decayed by 0.01 at epochs 120 and 160. Batch sizes are 256 and 128, respectively. A 5-epoch linear warmup is applied.

For ImageNet-LT, we use ResNet-50 trained for 90 epochs with batch size 256, SGD momentum 0.9, and weight decay 1×10^{-4}. The learning rate starts at 0.1 and decays by 0.1 at epochs 30 and 60, with a 5-epoch warmup.

Table 1: Top-1 accuracy (%) on CIFAR-10-LT under different imbalance ratios. Each cell reports "Overall / Head / Medium / Tail". Best and second best overall results are shown in **bold** and underline, respectively. Methods above the line are single-stage; below the line are two-stage.

(a) $\gamma = 10$

Method	Top1	Head	Medium	Tail
CE	87.82	**97.20**	85.45	85.55
Focal	86.03	96.40	83.50	83.25
LDAM	86.81	94.10	84.08	87.70
ALA	87.95	96.75	85.60	86.20
GCL	87.85	93.45	85.13	90.40
GeoBias(ours)	87.61	92.05	85.00	**91.00**
LDAM-DRW	87.96	93.85	85.33	89.95
cRT	88.06	95.00	84.98	90.35
LWS	**88.66**	96.15	**86.08**	88.90

(b) $\gamma = 50$

Method	Top1	Head	Medium	Tail
CE	78.61	**97.70**	77.75	62.10
Focal	77.99	96.80	76.48	63.70
LDAM	77.78	90.30	74.25	75.85
ALA	79.80	97.05	78.38	66.80
GCL	81.73	95.90	79.35	74.70
GeoBias(ours)	**83.97**	90.90	**80.95**	**86.10**
LDAM-DRW	81.99	96.10	79.47	75.45
cRT	81.25	95.05	78.50	75.70
LWS	81.48	96.40	79.28	73.15

(c) $\gamma = 100$

Method	Top1	Head	Medium	Tail
CE	72.49	**97.85**	72.07	48.40
Focal	70.16	97.75	69.65	44.10
LDAM	71.29	89.45	68.72	60.85
ALA	73.82	97.55	73.87	49.95
GCL	78.23	96.50	**76.78**	64.30
GeoBias(ours)	**79.78**	88.10	75.42	**84.55**
LDAM-DRW	77.93	96.50	75.97	65.25
cRT	77.15	94.90	74.53	67.25
LWS	76.92	96.60	74.45	64.65

(d) $\gamma = 200$

Method	Top1	Head	Medium	Tail
CE	66.44	97.55	67.55	32.00
Focal	64.26	**97.75**	64.70	29.45
LDAM	63.41	83.65	59.75	54.15
ALA	68.33	**97.75**	68.97	37.00
GCL	72.90	96.95	71.93	51.75
GeoBias(ours)	**76.03**	90.00	**72.75**	**71.90**
LDAM-DRW	70.92	96.20	69.63	49.50
cRT	72.11	95.30	71.47	50.85
LWS	71.71	96.15	71.10	49.10

For GeoBiasLoss, we use consistent hyperparameters across all datasets: warmup = 40 and bias_adjust_freq = 20.

4.3 Main Results

We evaluate GeoBias against single-stage loss-based methods (CE, Focal Loss(Lin et al., 2017), LDAM(Cao et al., 2019), ALA(Zhao et al., 2022), GCL(Li et al., 2022)) and two-stage reference methods (LDAM-DRW(Cao et al., 2019), cRT(Kang et al., 2019), LWS(Kang et al., 2019)). All methods share the same backbone and training schedule; for LDAM, ALA and GCL, we follow their original implementations by replacing the final layer with a normalized linear layer, while GeoBias operates on the standard linear classifier without architectural modification.

CIFAR-10-LT. As shown in Table 1, GeoBias consistently outperforms all single-stage competitors, with the gap widening as imbalance intensifies. At $\gamma = 200$, it surpasses the best single-stage baseline GCL by 3.13% in Top-1 and by a substantial 20.15% on tail classes. Against two-stage methods, GeoBias remains highly competitive, outperforming LDAM-DRW and cRT across most settings, while slightly trailing LWS only at the mildest imbalance ratio ($\gamma = 10$).

CIFAR-100-LT. Table 2 shows that GeoBias achieves the best overall accuracy at $\gamma = 100$ and $\gamma = 200$, surpassing all single-stage methods. The advantage on tail classes is even more pronounced: at $\gamma = 200$, GeoBias reaches 8.85%, more than doubling the tail accuracy of GCL (5.75%) and far exceeding ALA (3.40%). Against two-stage methods, GeoBias outperforms LDAM-DRW and cRT across all settings, and the gap with the strongest two-stage baseline LWS is marginal (40.14% vs. 41.06% at $\gamma = 200$).

Table 2: Top-1 accuracy (%) on CIFAR-100-LT under different imbalance ratios. Each cell reports "Overall / Head / Medium / Tail". Best and second best overall results are shown in **bold** and underline, respectively. Methods above the line are single-stage; below the line are two-stage.

(a) $\gamma = 10$

Method	Top1	Head	Medium	Tail
CE	58.93	**74.50**	59.70	41.05
Focal	57.75	72.55	58.78	39.85
LDAM	57.12	61.55	58.95	47.20
ALA	58.87	<u>72.80</u>	59.95	41.70
GCL	**60.85**	64.30	**62.08**	**53.70**
GeoBias(ours)	<u>60.12</u>	64.50	<u>61.12</u>	52.75
LDAM-DRW	59.93	70.90	59.67	49.75
cRT	58.75	64.80	58.82	52.50
LWS	**60.85**	68.55	60.93	<u>52.90</u>

(b) $\gamma = 50$

Method	Top1	Head	Medium	Tail
CE	45.82	74.55	46.58	14.80
Focal	45.55	<u>75.15</u>	46.23	13.90
LDAM	40.67	55.35	43.63	17.10
ALA	47.54	**75.30**	48.05	18.25
GCL	48.49	69.80	49.22	25.00
GeoBias(ours)	<u>49.02</u>	66.55	**49.67**	<u>29.55</u>
LDAM-DRW	47.52	71.50	48.33	21.10
cRT	47.37	63.75	48.25	28.35
LWS	**49.15**	68.10	<u>49.30</u>	**29.75**

(c) $\gamma = 100$

Method	Top1	Head	Medium	Tail
CE	41.85	**75.25**	42.40	6.80
Focal	41.57	73.75	42.25	7.35
LDAM	32.46	49.70	34.73	8.40
ALA	43.04	<u>74.50</u>	43.37	10.60
GCL	44.67	70.30	46.55	13.40
GeoBias(ours)	<u>45.50</u>	66.95	<u>46.58</u>	**20.80**
LDAM-DRW	42.98	71.00	43.72	12.75
cRT	43.58	62.85	45.22	<u>19.40</u>
LWS	**45.64**	69.30	**46.60**	19.10

(d) $\gamma = 200$

Method	Top1	Head	Medium	Tail
CE	36.87	**74.10**	36.38	1.10
Focal	36.92	<u>73.35</u>	36.48	1.80
LDAM	24.75	34.70	28.40	3.85
ALA	37.23	72.85	36.63	3.40
GCL	39.78	71.00	40.72	5.75
GeoBias(ours)	<u>40.14</u>	68.10	<u>41.25</u>	8.85
LDAM-DRW	37.12	70.25	36.82	4.90
cRT	38.76	61.60	40.07	**12.00**
LWS	**41.06**	68.00	**41.85**	<u>11.75</u>

Table 3: Top-1 accuracy (%) on ImageNet-LT. Best and second best results are shown in **bold** and underline, respectively. Methods above the line are single-stage; below the line are two-stage.

Method	Top1	Head	Medium	Tail
CE	38.39	64.04	37.88	14.26
Focal	37.20	61.83	36.81	13.75
LDAM	34.01	51.07	34.12	16.62
ALA	<u>40.27</u>	**65.55**	<u>39.84</u>	16.29
GCL	39.90	64.38	39.47	16.72
GeoBias(ours)	39.27	62.24	38.11	19.79
LDAM-DRW	**40.96**	<u>64.44</u>	**40.03**	20.26
cRT	36.83	56.45	35.30	<u>21.81</u>
LWS	39.43	62.11	37.59	**22.27**

ImageNet-LT. Table 3 reports results on the large-scale ImageNet-LT dataset. GeoBias achieves 39.27% overall accuracy, outperforming CE, Focal, LDAM, and GCL, while being slightly behind ALA and LDAM-DRW. Notably, on tail classes, GeoBias attains 19.79%, surpassing all single-stage methods and performing competitively with two-stage approaches. Considering its single-stage nature and lack of architectural modifications, this result is highly encouraging.

Summary. Across all datasets, GeoBias consistently outperforms single-stage loss-based methods, especially on tail classes. It achieves results comparable to or better than two-stage methods, yet requires no re-sampling, complex training schedules, or architectural modifications — offering a simple, plug-and-play solution for long-tailed recognition.

Table 4: Ablation study on CIFAR-100-LT. Results report Top-1 accuracy (%). Default configuration: reg=True, warmup=40, freq=20. With careful tuning of these hyperparameters, GeoBias can surpass the best two-stage method.

(a) $\gamma = 100$

Configuration	Top1	Head	Medium	Tail
Default	45.50	66.95	46.58	20.80
w/o reg ($\lambda = 0$)	45.87	65.90	46.93	22.65
freq=10	45.27	64.35	46.77	21.70
warmup=35, freq=15	45.67	69.80	46.83	18.05
warmup=30, freq=10	45.58	62.70	47.53	22.60

(b) $\gamma = 200$

Configuration	Top1	Head	Medium	Tail
Default	40.14	68.10	41.25	8.85
w/o reg ($\lambda = 0$)	40.80	66.85	41.75	11.90
freq=10	41.07	63.60	43.43	11.45
warmup=35, freq=15	40.88	64.90	42.65	11.55
warmup=30, freq=10	41.56	64.15	43.52	13.10

4.4 Ablation Study

We conduct ablation experiments on CIFAR-100-LT to analyze the impact of key components: bias anchoring regularization (reg), calibration frequency (freq), and warmup period. Table 4 summarizes the results under $\gamma = 100$ and $\gamma = 200$.

Effect of Bias Anchoring Regularization. Removing the regularization (reg=False) leads to slightly higher overall accuracy in some settings, but this is not consistently observed. In practice, enabling regularization plays a crucial role in stabilizing and improving tail-class accuracy, which is the primary focus of long-tailed learning.

Effect of Calibration Frequency and Warmup Period. Varying the calibration frequency and warmup period does not exhibit a deterministic trend. Nonetheless, fine-tuning these hyperparameters can further improve performance, surpassing two-stage methods and achieving the best overall results among all compared methods. With the default configuration (reg=True, warmup=40, freq=20), GeoBias already achieves highly competitive performance, and the gap between the default and tuned configurations remains marginal.

5 Conclusion

In this paper, we identify a fundamental yet overlooked issue in long-tailed recognition: existing methods predominantly focus on adjusting the classifier weight matrix $\mathbf{W}$, while neglecting the bias term $\mathbf{b}$ as a critical geometric lever for decision boundary correction. To address this, we propose GeoBias, a geometry-driven bias rectification method that explicitly decouples weight directions from bias positions. By analytically computing bias adjustments using class centers and inter-class normals, and resolving multi-class coupling via weighted global least squares, GeoBias directly shifts decision boundaries toward geometric balance without perturbing the learned discriminative structures. Extensive experiments on CIFAR-10-LT, CIFAR-100-LT, and ImageNet-LT demonstrate that GeoBias consistently outperforms single-stage loss-based methods and achieves competitive performance against two-stage approaches, while requiring no data re-sampling, architectural modifications, or complex training schedules. Its plug-and-play nature makes it readily applicable to existing long-tailed learning pipelines.

AI USE STATEMENT

We have not used generative AI tools for any tasks with required disclosure, including but not limited to generating synthetic datasets, developing theoretical models, formulating mathematical propositions, assisting in proof writing, proposing or refining hypotheses, designing experimental methods, implementing methods, data cleaning and formatting, or interpreting results.

We used generative AI tools for the following tasks with recommended disclosure: (1) searching for and reading related literature; (2) polishing English sentences to ensure fluency and compliance with scientific writing style; and (3) generating LaTeX code for equations, the algorithm (Algorithm 1), and the data tables in the experimental section. We have reviewed all AI-assisted work and take full responsibility for the final content of this work.

REFERENCES

Anonymous. Coin: Dual-filter consensus for open-set recognition, class imbalance, and noisy labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. To appear.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

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

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.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9260–9269. IEEE, 2019.

Mizuki Dai and Kenya Jin'no. Geometric analysis of the influence of bias terms and weight norms on decision borders in low-dimensional softmax classifiers. In *General Conference of the Institute of Electronics, Information and Communication Engineers*, 2026. Presentation N-1-23.

Yu Fu, Liuyu Xiang, Yumna Zahid, Guiguang Ding, Tao Mei, Qiang Shen, and Jungong Han. Long-tailed visual recognition with deep models: A methodological survey and evaluation. *Neurocomputing*, 509:290–309, 2022.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

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

Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6919–6928. IEEE, 2022.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*, 2016.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2532–2541. IEEE, 2019.

Zheda Mai, Arpita Chowdhury, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Vardaan Pahuja, Tanya Berger-Wolf, Song Gao, Charles Stewart, Yu Su, and Wei-Lun Chao. Fine-tuning is fine, if calibrated. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-4323.

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

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 (ICLR)*, 2021.

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.

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.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

G Van Horn and P Perona. The devil is in the tails: Fine-grained classification in the wild, arxiv. *arXiv preprint arXiv:1709.01450*, 2017.

Vladimir N Vapnik and A Chervonenkis. The necessary and sufficient conditions for consistency of the method of empirical risk minimization. *Pattern Recognition and Image Analysis*, 1(3): 284–305, 1991.

Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *International Journal of Computer Vision*, 130(7):1837–1872, 2022.

Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2361–2370. IEEE, 2021.

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.

Yan Zhao, Weicong Chen, Xu Tan, Kai Huang, and Jihong Zhu. Adaptive logit adjustment loss for long-tailed visual recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 3472–3480, 2022.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16484–16493. IEEE, 2021.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9716–9725. IEEE, 2020.