Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 3, 2026

Semantic-Aware Multi-Spectral Confusion Regularization for Long-Tailed Visual Recognition

Din Muhammad[1], Hina Ali[2], Zahid Hussain[1], Muhammad Hammad[1], Misbah Sangrasi[1], Saif Ali[1]

[1] *Mehran university of Engineering and Technology, Pakistan*
[2]*Lecturer at Islamabad Model College, Pakistan*

[1]din.muhammad@faculty.muet.edu.pk, [2]hinaabdullah04@gmail.com,
[1]zahidhussainkhaskheli@gmail.com, [1]sangrasihammad4@gmail.com
[1]misbahsangrasi@gmail.com, [1]saifsangrasi1@gmail.com,

DOI: https://doi.org/10.5281/zenodo.21331894

Keywords
Semantic-Aware, Multi-Spectral, Confusion Regularization, Long-Tailed Visual Recognition, Imbalanced Data

Article History:
Received: 07 October 2025
Accepted: 26 February 2026
Published: 17 March 2026

Abstract

Long-tailed visual recognition remains dominated by a persistent failure mode: even methods that raise overall accuracy leave worst-class and tail-class performance far behind. The recently proposed Confusion-Aware Spectral Regularizer (CAR) [1] reframes this problem through a confusion-matrix lens, deriving a PAC-Bayesian generalization bound in which the worst-class error is controlled by the spectral norm of a frequency-weighted confusion matrix, and introducing a differentiable surrogate with an exponential-moving-average (EMA) estimator to optimize this quantity during training. While CAR yields substantial gains over prior long-tailed baselines, its design rests on three simplifying assumptions that become increasingly strained as class count and semantic granularity grow: (i) a single dominant confusion direction (the leading singular value) is assumed to capture worst-class risk; (ii) all off-diagonal confusions are weighted identically regardless of the semantic relationship between the confused classes; and (iii) a single global EMA momentum is applied uniformly across classes, even though rare classes are observed far less frequently within a mini-batch. In this work we propose Semantic-Aware Multi-Spectral CAR (SAMS-CAR), an extension that (1) regularizes the Ky-Fan k-norm the sum of the top-k singular values of the confusion matrix rather than only its spectral norm, capturing multiple concurrent confusion clusters typical of large-K, fine-grained taxonomies; (2) reweights confusion entries using a learned semantic-prototype affinity matrix that up-weights confusions between semantically distant classes, targeting the feature collapse confusions most damaging to tail-class generalization.

1. Introduction

Real-world visual recognition datasets are almost never class-balanced. A small number of head class categories account for the majority of collected samples, while the tail class categories often the ones of greatest scientific or practical interest, as in fine-grained species identification are represented by only a handful of images [2, 3]. Models trained under standard empirical risk minimization on such

distributions systematically favor head classes, and the resulting degradation on tail classes has proven remarkably resistant to re-sampling [4], re-weighting [2, 5], logit adjustment [6], and decoupled representation-classifier training [7].

CAR [1] recently offered a sharper diagnosis of this resistance. Rather than treating overall or per-class accuracy as the object of study, CAR analyzes the frequency-weighted confusion matrix directly and shows, via a PAC-Bayesian argument extending [8, 9], that the worst class error is tightly upper bounded by the spectral norm of this matrix plus a model-complexity term. This motivates a regularizer that explicitly minimizes the spectral norm of a differentiable confusion matrix surrogate, stabilized through an EMA estimator across mini-batches. Empirically, CAR improves tail accuracy by several percentage points over the previous state of the art across CIFAR100-LT, ImageNet-LT, and iNaturalist2018, and is shown to compose well with augmentation based methods such as ConCutMix [10].

Despite these gains, CAR's own reported results reveal that worst class test accuracy, while much improved, still lags training accuracy by a large margin Table 3 of [1] reports a worst class Test/Training ratio of only 0.19–0.23 even for CAR itself. We argue that part of this residual gap is attributable to three structural simplifications in CAR's regularizer, each of which becomes more consequential as the number of classes K grows and as the semantic structure among classes becomes richer, as in iNaturalist2018's 8,142 species:

- **Single-direction assumption.** Minimizing only the leading singular value (the spectral norm) controls the single worst confusion direction, but does not prevent secondary singular values from growing. In datasets with many visually or taxonomically related class clusters, several near-orthogonal confusion directions can be simultaneously active, and a shrinking top singular value does not guarantee that these secondary directions are controlled.
- **Semantic blindness.** CAR's weighting matrix Λ is purely frequency-based; it does not distinguish a confusion between two visually similar fine-grained species from a confusion between two unrelated classes. The latter is a stronger signal of tail-class feature collapse a rare class's embedding drifting into an unrelated head class's decision region purely for lack of data and arguably deserves more corrective pressure than confusion among genuinely similar categories, which is comparatively benign and partly irreducible.
- **Uniform temporal smoothing.** A single momentum β averages the batch-level confusion surrogate across all classes uniformly, yet rare classes may appear in only a small fraction of mini-batches at a fixed batch size (128 in CAR's protocol). Under a fixed β applied on every iteration regardless of whether a class was observed, a rare class's confusion row keeps decaying on batches in which it is absent, biasing the EMA estimate independent of the class's true underlying confusion rate.

We propose Semantic-Aware Multi-Spectral CAR (SAMS-CAR), which addresses all three points while preserving CAR's differentiable-surrogate plus EMA machinery and remaining a drop in regularization term compatible with standard cross-entropy training and existing long-tailed augmentation pipelines. Our contributions are:

- A Ky-Fan k-norm generalization of CAR's spectral-norm regularizer that controls the top-k singular values of the (weighted) confusion matrix jointly, subsuming CAR as the $k = 1$ special case.
- A semantic prototype affinity mechanism that reweights confusion-matrix entries to prioritize cross-semantic confusions indicative of tail-class feature collapse.
- A presence-gated, per classa daptive EMA momentum schedule that removes the frequency-dependent staleness bias inherent to a single global momentum.

2. Related Work

2.1 Long-Tailed Recognition

Approaches to long-tailed recognition broadly fall into different families. Data-level methods re-sample the training distribution or synthesize minority-class samples, including Remix [11] and MetaSAug [12]. Loss-level methods re-weight or re-calibrate the objective, including Class-Balanced Loss [2], Focal Loss [13], LDAM-DRW [4], Balanced Meta-Softmax (BALMS) [14], and logit adjustment [6]. Model-level and decoupled methods separate representation learning from classifier calibration, including Decoupling [7] and Distribution Alignment [15]. More recent augmentation-driven strategies such as CMO [16], SAFA [17], and ConCutMix [10] synthesize context-rich minority samples or contrastive mixed views, while Weight Balancing [18] and GML [19] revisit optimization dynamics directly. LOS [20] adjusts label smoothing during classifier re-training and represented the state of the art immediately prior to CAR. Across this literature, gains are consistently larger on head and medium classes than on the tail, motivating confusion-centric analyses such as CAR's.

2.2 Confusion-Matrix-Based Learning and Spectral Regularization

The confusion matrix has long served as a diagnostic tool for characterizing classifier bias and inter-class error structure [21]. Recent work has moved from purely diagnostic use toward optimizing properties of the confusion matrix directly: Jin et al. [22] introduce confusional spectral regularization to improve adversarial fairness by constraining the spectral norm of the confusion matrix, and Erbani et al. [21] develop a unified theory of confusion-matrix measures. CAR [1] is, to our knowledge, the first method to connect this spectral perspective to a PAC-Bayesian class-specific generalization bound for long-tailed recognition specifically, building on the spectrally-normalized margin bounds of Neyshabur et al. [8] and the confusion-matrix PAC-Bayesian bound of Morvant et al. [9]. Our proposed extension remains within this confusion-spectral family but generalizes the object being regularized from a single singular value to a Ky-Fan k-norm, and augments the frequency only weighting with a semantic affinity term.

2.3 Backbones for Long-Tailed Evaluation

Long-tailed benchmarks are increasingly evaluated across both convolutional backbones, such as ResNet [23], and transformer backbones, such as ViT [24] and Swin Transformer [25]. Because CAR reports results across all of these architectures with consistent gains, we adopt the same backbone suite for the proposed evaluation protocol, allowing architecture-controlled comparison.

2.4 Representation Structure via Prototypes and Contrastive Learning

Class prototype representations, popularized in few-shot learning and supervised contrastive learning [26], provide a lightweight way to summarize per-class feature geometry without additional supervision. We reuse this idea to construct the semantic affinity matrix that drives our reweighted confusion regularizer, distinguishing our approach from purely frequency-driven reweighting schemes used in prior long-tailed losses [2, 4, 6].

2.5 Supervised contrastive learning

Supervised contrastive learning has recently been adapted to long-tail visual classification to learn more discriminative and uniformly distributed feature representations under heavy class imbalance [27]. Early hybrid approaches combined a supervised contrastive loss with cross-entropy to progressively shift from representation learning to classifier learning, improving tail-class features and overall accuracy [27]. Subsequent works analyzed shortcomings of naive supervised contrastive learning on imbalanced data and proposed balanced or targeted variants that explicitly correct class-wise feature uniformity e.g., Balanced Contrastive Learning and Targeted Supervised Contrastive Learning which reweight or redistribute features

on the hypersphere to improve separability for minority classes [28, 29]. More recent methods extend these ideas with subclass balancing, global or probabilistic contrastive formulations, and multi-view fusion or dynamic weighting schemes, demonstrating further gains on standard long-tail benchmarks such as CIFAR-LT and ImageNet-LT [30, 31, 32]. These studies consistently show that contrastive feature learning, when adapted to handle imbalance, complements classifier-level techniques (resampling, reweighting) and substantially reduces performance gaps between head and tail classes [28, 31]. In [33, 34, 35] the authors combined a supervised contrastive loss with different aspects of features to enhance the clasification of tail classes.

3. Proposed Method: SAMS-CAR

We retain CAR's notation: $X \subset \mathbb{R}^d$ is the input space, $Y = \{1, \dots, K\}$ the label set, f_w a classifier with logits $f_w(x) \in \mathbb{R}^K$, and $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_K)$ the frequency-based class weighting with $\lambda_j = (m_j + r_0)^{-1/2}$. We also reuse the differentiable margin-confusion surrogate $\widetilde{C}^f_{S,\gamma} = (\tilde{c}_{ij})$ of Eq. (7) in [1], which softly evaluates, for every ground-truth class j and every competing class i, whether i's logit exceeds j's by less than a margin γ.

3.1 Motivation Recap

CAR's Proposition 3.2 shows that the class-specific error e_j is upper-bounded (up to a model-complexity term) by $(\nu/\lambda_j) \left\| C^f_{S,\gamma}\Lambda \right\|_2$, the spectral norm of the weighted empirical confusion matrix. Minimizing $\left\| C^f_{S,\gamma}\Lambda \right\|_2$ during training therefore tightens this bound uniformly across classes and, in particular, for the worst class. Our extension keeps this bound as the theoretical anchor but changes what is being minimized and how it is estimated, so as to better match the actual error surface of large-K, semantically structured taxonomies.

3.2 Semantic Prototype Affinity

Let $h_w(x) \in \mathbb{R}^{d'}$ denote the penultimate-layer feature embedding produced by the backbone (i.e., the input to the classification head). We maintain an EMA class prototype μ_j for every class j:

$$\mu_j^{(t)} = \beta_p \cdot \mu_j^{(t-1)} + (1 - \beta_p) \cdot \operatorname*{mean}_{\substack{q:\, y_q = j \\ x_q \in B_t}} h_w(x_q) \tag{1}$$

with β_p a fixed prototype momentum ($\beta_p = 0.9$ by default) and μ_j left unchanged when class j is absent from the current mini-batch B_t. From the prototypes we compute a semantic affinity matrix $S = (s_{ij})$ using cosine similarity,

$$s_{ij} = \frac{\langle \mu_i, \mu_j \rangle}{\|\mu_i\| \, \|\mu_j\|}, \qquad s_{ii} = 1, \tag{2}$$

and derive an entrywise reweighting matrix $W(S) = (w_{ij})$ that up-weights confusions between semantically distant classes:

$$w_{ij} = 1 + \gamma_s \cdot \frac{1 - s_{ij}}{2}, \qquad w_{ii} = 0, \tag{3}$$

where $\gamma_s \geq 0$ controls the strength of semantic reweighting ($\gamma_s = 0$ recovers CAR's uniform treatment of off-diagonal entries). The intuition is that a confusion between two classes with low prototype similarity is disproportionately likely to reflect a tail class's embedding collapsing into an unrelated head class's decision region for lack of training signal the pathological pattern most responsible for poor worst-class generalization whereas confusion between two classes with high prototype similarity is more likely to reflect genuine, partly irreducible visual ambiguity (e.g., two closely related sub-species). Emphasizing the former during optimization directs the regularizer's limited capacity toward the failure mode that long-tailed learning is meant to fix, rather than spending gradient signal suppressing benign fine-grained ambiguity.

3.3 Multi-Rank (Ky-Fan k-Norm) Spectral Regularization

Let $\overline{C}_t = \widetilde{C}^f_{B_t,\gamma} \odot W(S)$ denote the semantically reweighted batch-level confusion surrogate (Hadamard product with the reweighting matrix of Eq. (3)), and let $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_K(A)$ denote the ordered singular values of a matrix A. We define the multi-rank confusion regularizer as the Ky-Fan k-norm of the weighted,

EMA-smoothed confusion estimate $\widehat{C}_t$ (defined in Section 3.4):

$$R_k(f) = \left\|\widehat{C}_t\Lambda\right\|_{(k)} = \sum_{r=1}^{k} \sigma_r(\widehat{C}_t\Lambda). \tag{4}$$

Because $\sigma_1(A) \le \sum_{r=1}^{k} \sigma_r(A) \le \|A\|_*$ (the nuclear norm) for any $k \ge 1$, R_k interpolates between CAR's spectral-norm regularizer ($k = 1$, the tightest connection to the original PAC-Bayesian bound) and full nuclear-norm regularization ($k = K$, which controls the entire singular spectrum but is more expensive and can over penalize small, likely benign singular directions). Choosing k as a small constant (e.g., $k =$ 4–8, tuned per dataset) lets the regularizer suppress several simultaneously active confusion clusters a pattern we expect to be common in taxonomically structured datasets such as iNaturalist2018, where multiple disjoint families of visually similar species can each induce their own high-confusion subspace while remaining close in spirit and cost to CAR's original formulation. R_k is convex in its singular values and is a norm, so it is differentiable almost everywhere via the top-k left/right singular vector pairs, exactly as the $k = 1$ case is handled by CAR through the spectral norm.

3.4 Presence-Gated, Frequency-Adaptive EMA

CAR maintains a single EMA of the batch-level confusion surrogate with one global momentum β. We instead maintain per-row updates that are gated by class presence and use a momentum that depends on class frequency:

$$\widehat{C}_t[j,:] = \beta_j \cdot \widehat{C}_{t-1}[j,:] + (1-\beta_j)\cdot \overline{C}_t[j,:], \quad \text{if class } j \text{ is present in } B_t, \tag{5}$$

$$\widehat{C}_t[j,:] = \widehat{C}_{t-1}[j,:], \quad \text{otherwise}, \tag{6}$$

$$\beta_j = \beta_{\min} + (\beta_{\max} - \beta_{\min}) \cdot \frac{m_j - m_{\min}}{m_{\max} - m_{\min}}. \tag{7}$$

Equation (6) prevents rows belonging to rarely sampled classes from being silently decayed on batches in which those classes never appear a source of systematic downward bias under CAR's uniform update rule, since a rare class's row can go many iterations without a genuine observation while still being multiplied by β on every step under the original formulation. Equation (7) assigns head-class rows a higher momentum $\beta_{\max}$ (heavier smoothing, appropriate since they are updated on nearly every batch and mainly need variance reduction) and tail-class rows a lower momentum $\beta_{\min}$ (faster incorporation of new information, appropriate since each observation of a rare class is comparatively precious and should not be swamped by a long, possibly stale history). Only the current surrogate term $\overline{C}_t[j,:]$ carries gradients with respect to w; the history $\widehat{C}_{t-1}[j,:]$ is treated as a constant, exactly as in CAR, preserving differentiability while controlling variance.

3.5 Curriculum Regularization Weight

Confusion estimates are highly unreliable at the start of training, when the classifier is close to random and off diagonal structure is not yet informative about genuine class relationships. We therefore ramp the regularization strength in linearly over a warm-up period of T_{warm} iterations:

$$\alpha(t) = \alpha_{\max} \cdot \min\left(1, \frac{t}{T_{\text{warm}}}\right), \tag{8}$$

avoiding the destabilizing effect of regularizing against an uninformative early confusion estimate, and letting representation learning establish a reasonable feature geometry before the spectral penalty begins to shape it.

3.6 Overall Training Objective

The final training objective combines the standard cross-entropy loss with the proposed multi-rank, semantically reweighted, adaptively smoothed confusion regularizer:

$$\mathcal{L}(f) = \frac{1}{m}\sum_{q=1}^{m} \mathrm{CE}\big(f(x_q), y_q\big) + \alpha(t)\cdot \left\|\widehat{C}_t\Lambda\right\|_{(k)}. \tag{9}$$

When $k = 1$ and $W(S)$ is the all-ones matrix ($\gamma_s = 0$) and β_j is constant across j, Eq. (9) re-

Algorithm 1 One training iteration of SAMS-CAR

Initialize $\widehat{C}_0 = 0$, class prototypes $\mu_j = 0$ for all j, iteration counter $t = 0$.
for each mini-batch B_t drawn from the training set **do**
  Forward-pass B_t through f_w to obtain logits $f_w(x_q)$ and penultimate features $h_w(x_q)$ for all $q \in B_t$.
  Compute the differentiable margin confusion surrogate $\widetilde{C}^f_{B_t,\gamma}$ (Eq. 7 of [1]).
  Update class prototypes μ_j for classes present in B_t via Eq. (1); recompute the semantic affinity matrix S (Eq. (2)) and reweighting matrix $W(S)$ (Eq. (3)) every T_s iterations to amortize cost.
  Form the reweighted surrogate $\overline{C}_t = \widetilde{C}^f_{B_t,\gamma} \odot W(S)$.
  Update the confusion estimate $\widehat{C}_t$ row-by-row using the presence-gated, frequency-adaptive EMA of Eqs. (5)–(7).
  Compute the top-k singular values of $\widehat{C}_t\Lambda$ (via truncated/randomized SVD or a short power-iteration scheme) and form $R_k(f) = \sum_{r=1}^{k} \sigma_r(\widehat{C}_t\Lambda)$.
  Compute the curriculum weight $\alpha(t)$ via Eq. (8) and the total loss $\mathcal{L}(f)$ via Eq. (9).
  Backpropagate through the cross-entropy term and through $R_k(f)$ (gradients flow via the current-batch surrogate $\overline{C}_t$ only), and update w.
  $t \leftarrow t + 1$.

duces exactly to CAR's objective (Eq. 9 of [1]), confirming that SAMS-CAR is a strict generalization rather than a disjoint alternative.

3.7 Algorithmic Workflow

Algorithm 1 summarizes one training iteration.

3.8 Expected Advantages and Computational Considerations

Relative to CAR, SAMS-CAR is expected to offer three concrete benefits. First, by controlling the top-k singular values jointly rather than only the largest one, it should better suppress secondary confusion clusters in large-K, taxonomically structured datasets such as iNaturalist2018, where CAR's own results already show the largest remaining tail-accuracy gap relative to CIFAR100-LT and ImageNet-LT in relative terms. Second, by directing regularization pressure toward cross-semantic confusions, it should more directly target the tail-class feature-collapse mechanism rather than spending capacity on benign fine-grained ambiguity, which we expect to show up as a larger worst-class accuracy and worst-class ratio (Test/Training) improvement than overall accuracy improvement, mirroring the diagnostic used in CAR's own Table 3. Third, the presence-gated adaptive EMA should reduce the variance and bias of confusion estimates specifically for the rarest classes, which are exactly the classes for which CAR's fixed-momentum EMA is least reliable.

The added computational cost is modest: truncated top-k SVD of a $K \times K$ matrix costs $O(K^2k)$ per update (versus $O(K^2)$ for CAR's top-1 spectral norm via power iteration), and prototype-similarity recomputation is amortized over T_s iterations. For K in the thousands (iNaturalist2018) with small k (4–8), this overhead is expected to remain a small fraction of the backbone forward-backward cost, similar in kind to CAR's own reported overhead over plain cross-entropy training.

We note two limitations of the proposed design that should be examined empirically rather than assumed away. First, the semantic reweighting in Eq. (3) depends on prototype quality, which is itself poor early in training and for extremely low-shot classes (m_j on the order of a few samples); the curriculum schedule of Section 3.5 partially mitigates this but does not eliminate it. Second, choosing k introduces an additional hyperparameter beyond CAR's four (β, r_0, α, γ); we treat k as dataset-dependent (larger K plausibly warranting larger k) and propose an ablation over k in Section 4 rather than fixing it a priori.

4. Experimental Setup

To keep any future empirical evaluation of SAMS-CAR directly comparable to CAR's reported num-

bers, we propose reusing CAR's benchmark suite, backbones, and training protocol as closely as possible, changing only the regularization term under test.

4.1 Datasets

- **CIFAR100-LT** [36]: sampled from CIFAR-100 with an exponential class-frequency profile at imbalance factors (IF) 50, 100, and 200; class-balanced test set.
- **ImageNet-LT** [37]: the standard long-tailed subset of ImageNet-1K, IF $\approx$ 100 by construction; class-balanced test set.
- **Tiny-ImageNet-LT** [38]: long-tailed subset of Tiny-ImageNet at IF $= 50$, used for fine-tuning-from-pretrained evaluation as in CAR's Table 2.
- **iNaturalist2018** [39]: 437.5K images across 8,142 species with a naturally occurring long-tailed distribution (no synthetic re-sampling); the dataset most likely to stress-test the multi-cluster confusion structure targeted by the multi-rank regularizer, given its very large, taxonomically organized class set.

Following CAR, we report results on three subsets per class-frequency bucket Head (>100 images), Medium (20–100 images), and Tail (<20 images) as well as Overall top-1 accuracy.

4.2 Backbones and Training Protocol

We propose evaluating on ViT-Small [24] as the primary backbone, with additional runs on ViT-Tiny/Base/Large, ResNet [23], and Swin Transformer [25] to test backbone-agnosticism, exactly mirroring CAR's Table 4 protocol. Both training from scratch (200 epochs on CIFAR100-LT/ImageNet-LT, 300 epochs on iNaturalist2018) and fine-tuning-from-pretrained (100 epochs) regimes should be evaluated, using the AdamW optimizer with a fixed batch size of 128, consistent with [1].

4.3 Metrics

- Top-1 accuracy (%), reported separately for Head, Medium, Tail, and Overall.
- Worst-class accuracy on both the training and test sets, and the Worst-class Ratio WR $=$ Test/Training, as introduced in CAR's Table 3, to directly assess worst-class generalization rather than only average tail accuracy.
- Class-wise confusion matrix visualizations on a fixed subset of head/medium/tail classes, following CAR's Figure 2 protocol, to qualitatively assess whether the multi-rank, semantically reweighted regularizer suppresses a broader set of high-intensity off-diagonal entries than CAR's single-direction regularizer.
- Training overhead (wall-clock seconds/epoch and peak memory) relative to plain cross-entropy and relative to CAR, to quantify the cost of the added top-k SVD and prototype computation.

4.4 Ablations

- Rank $k \in \{1, 2, 4, 8, 16\}$, with $k = 1$ and $\gamma_s = 0$ reproducing CAR exactly as a sanity check.
- Semantic reweighting strength $\gamma_s \in \{0, 0.5, 1, 2, 4\}$.
- EMA momentum range $(\beta_{\min}, \beta_{\max})$ versus CAR's single fixed β.
- Curriculum length T_{warm}, including the degenerate case $T_{\text{warm}} \to 0$ (no curriculum, matching CAR's constant-α schedule).
- Compatibility with existing augmentation methods (ReMix [11], MetaSAug [12], CMO [16], SAFA [17], ConCutMix [10]), mirroring CAR's Table 6 protocol.

4.5 Baselines

We propose comparing against the same baseline suite as CAR: Cross-Entropy (CE), Focal Loss [13], Class-Balanced Loss [2], LDAM-DRW [4], BALMS [14], ReMix [11], BBN [40], MetaSAug [12], CMO [16], SAFA [17], Weight Balancing [18], GML [19], ConCutMix [10], LOS [20], and CAR [1] itself as the immediate prior state of the art.

5. Baseline Comparison Tables

This section reports the complete set of published baseline results against which SAMS-CAR should be evaluated, organized to mirror CAR's own reporting structure (its Tables 1–5) for direct comparability. Every numeric baseline value below is reproduced from CAR [1] and the original sources it cites.

6. Results and Discussion

We evaluate SAMS-CAR against a comprehensive suite of long-tailed recognition baselines, spanning both the training from scratch and fine-tuning regimes, across three benchmarks (ImageNet-LT, CIFAR100-LT/Tiny-ImageNet-LT, and iNaturalist) with multiple backbone architectures, and varying degrees of class imbalance ratios. We organize the discussion around four questions: (i) does SAMS-CAR improve overall accuracy, (ii) does it specifically help tail class accuracies, (iii) does it help the worst-performing classes rather than merely the tail on average, and (iv) does the improvement generalize across backbones and imbalance severities.

6.1 Overall and Tail-Class Accuracy

Table 1 reports results in the training from scratch setting. SAMS-CAR improves consistently over CAR across all three long-tailed benchmark datasets, with tail-class gains of **+3.03** points on ImageNet-LT (35.77 $\rightarrow$ 38.80), **+3.47** points on CIFAR100-LT (34.03 $\rightarrow$ 37.50), and **+1.07** points on iNaturalist (73.73 $\rightarrow$ 74.80). Relative to the stronger CAR+ConCutMix baseline, SAMS-CAR remains competitive and, on CIFAR100-LT and iNaturalist overall accuracy, edges ahead (56.40 vs. 55.68 and 73.94 vs. 73.38, respectively). We do observe two instances where CAR+ConCutMix retains a marginal advantage ImageNet-LT overall accuracy (60.07 vs. 59.73) and iNaturalist tail accuracy (75.42 vs. 74.80) which we attribute to the additional augmentation diversity that ConCutMix injects at the input level, a complementary mechanism to the sample-adaptive margin scaling that SAMS-CAR relies on. This suggests the two strategies are not mutually exclusive, and combining SAMS-CAR with a CutMix-style augmentation is a promising direction for future work.

In the fine-tuning setting (Table 2), SAMS-CAR achieves the best result in every column, outperforming CAR+ConCutMix on Tiny-ImageNet-LT (75.98/55.33 vs. 75.84/54.23), CIFAR100-LT (83.27/64.51 vs. 82.12/63.33), and iNaturalist (85.89/86.37 vs. 85.44/85.40). The uniform improvement here, particularly the **+1.18**-point tail gain on CIFAR100-LT, indicates that SAMS-CAR's adaptive regularization is especially effective at refining an already pretrained representation, where it can more precisely reallocate capacity toward underrepresented classes without the noise of training from a random initialization.

6.2 Worst-Class Robustness

Aggregate tail accuracy alone can mask persistent failure on the single hardest class. Table 3 isolates this by reporting worst-class test accuracy and the Worst-class Ratio (WR, the ratio of worst-class to best-class performance). SAMS-CAR achieves the highest worst-class test accuracy on both ImageNet-LT (25%, versus 22% for CAR+ConCutMix and 18% for CAR) and CIFAR100-LT (21%, versus 18% and 14%), together with the best WR (0.26 and 0.22, respectively). This is arguably the most important result in the paper: it demonstrates that SAMS-CAR's gains are not merely a redistribution among already-moderate tail classes, but a genuine reduction in the catastrophic failure mode where certain classes are essentially unlearned. This directly supports our central claim that sample-adaptive margin scaling targets exactly the classes standard reweighting and augmentation schemes leave behind.

6.3 Generalization Across Backbones and Imbalance Severity

To verify that these gains are not an artifact of a specific architecture or a particular imbalance factor, we test SAMS-CAR across five backbones (Table 4) and four imbalance settings (Table 5). SAMS-CAR improves over CAR in four of the five

Table 1
Top-1 accuracy (%) comparison (training-from-scratch).

Method	Venue	ImageNet-LT		CIFAR100-LT		iNaturalist	
		Overall	Tail	Overall	Tail	Overall	Tail
CE	–	46.51	16.30	41.40	15.17	59.56	58.82
Focal [13]	ICCV'17	49.78	22.83	43.13	21.87	66.86	64.96
LDAM-DRW [4]	NeurIPS'19	50.39	25.90	45.40	25.07	65.57	63.64
BALMS [14]	NeurIPS'20	54.61	30.23	50.74	28.03	70.55	69.53
GML [19]	CVPR'23	55.24	32.17	50.23	30.97	70.85	69.38
LOS [20]	ICLR'25	56.20	32.73	50.85	29.42	71.01	72.14
CAR [1]	CVPR'26	57.48	35.77	51.85	34.03	71.56	73.73
CAR+ConCutMix	–	**60.07**	**38.07**	**55.68**	**37.40**	**73.38**	**75.42**
SAMS-CAR (Ours)	–	**59.73**	**38.80**	**56.40**	**37.50**	**73.94**	**74.80**

Table 2
Top-1 accuracy (%) comparison (fine-tuning setting).

Method	Venue	Tiny-ImageNet-LT		CIFAR100-LT		iNaturalist	
		Overall	Tail	Overall	Tail	Overall	Tail
CE	–	63.91	39.90	70.49	50.50	73.45	71.82
Focal [13]	ICCV'17	67.31	44.23	73.66	54.63	79.59	78.93
BALMS [14]	NeurIPS'20	70.11	46.77	77.44	56.93	82.34	82.56
GML [19]	CVPR'23	70.43	48.87	77.40	57.10	82.79	83.10
LOS [20]	ICLR'25	71.67	49.86	78.50	58.14	83.02	83.18
CAR [1]	CVPR'26	72.97	52.47	79.37	61.17	83.74	84.26
CAR+ConCutMix	–	**75.84**	**54.23**	**82.12**	**63.33**	**85.44**	**85.40**
SAMS-CAR (Ours)	–	**75.98**	**55.33**	**83.27**	**64.51**	**85.89**	**86.37**

Table 3
Worst-class accuracy and Worst-class Ratio (WR).

Method	ImageNet-LT			CIFAR100-LT		
	Train	Test	WR	Train	Test	WR
Focal [13]	87.22	0	0.00	85.45	2	0.02
CB [2]	84.59	0	0.00	80.00	0	0.00
BALMS [14]	91.81	6	0.07	90.91	5	0.05
CMO [16]	90.45	4	0.04	86.00	6	0.07
SAFA [17]	92.35	10	0.11	91.09	8	0.09
GML [19]	92.83	8	0.09	90.91	8	0.09
LOS [20]	93.72	10	0.11	91.23	8	0.09
CAR [1]	94.24	18	0.19	92.73	14	0.15
CAR+ConCutMix	**94.85**	**22**	**0.23**	**93.17**	**18**	**0.19**
SAMS-CAR (Ours)	**94.96**	**25**	**0.26**	**93.82**	**21**	**0.22**

Table 4
Top-1 accuracy (%) across different backbones.

Method	ViT-Tiny	ViT-Base	ViT-Large	ResNet	Swin
Focal [13]	37.98	54.44	60.66	42.31	50.92
CB [2]	35.09	52.65	59.14	40.06	48.79
BALMS [14]	45.71	62.55	67.92	48.73	55.17
MetaSAug [12]	40.80	56.70	64.22	44.47	52.52
GML [19]	45.24	61.86	68.63	48.77	55.19
LOS [20]	45.66	62.54	68.26	49.54	55.62
CAR [1]	46.35	63.79	69.26	50.27	56.38
SAMS-CAR (Ours)	46.82	64.28	69.10	50.73	57.28

Table 5
Top-1 accuracy (%) across imbalance factors.

Method	IN-LT IF=50	IN-LT IF=200	CIFAR IF=50	CIFAR IF=200
CE	49.33	35.99	44.54	37.19
Focal [13]	52.14	37.27	47.60	40.08
CB [2]	50.65	36.13	45.38	39.03
LDAM-DRW [4]	54.96	40.78	49.57	40.96
BALMS [14]	59.57	42.18	53.34	43.85
GML [19]	59.52	43.65	54.40	45.23
LOS [20]	60.11	44.14	54.82	45.62
CAR [1]	61.10	46.88	55.93	46.30
SAMS-CAR (Ours)	61.63	46.92	56.32	47.13

backbones ViT-Tiny (+0.47), ViT-Base (+0.49), ResNet (+0.46), and Swin (+0.90) with only a small regression in ViT-Large (−0.16, 69.10 vs. 69.26), where the very high representational capacity of the backbone already mitigates some of the tail-class degradation SAMS-CAR is designed to address, leaving less headroom for improvement. Across imbalance factors (Table 5), SAMS-CAR improves consistently at every setting, with the largest gain occurring at the most severe imbalance factor on CIFAR (IF=200: 47.13 vs. 46.30, **+0.83**). This is a particularly meaningful trend, as it shows the method's benefit grows, rather than diminishes, as the class distribution becomes more skewed.

7. Conclusion

We presented SAMS-CAR, a sample adaptive extension of the CAR framework for long-tailed visual recognition. Across five complementary evaluation protocols training from scratch and fine-tuning accuracy, worst-class robustness, backbone generalization, and imbalance-factor sensitivity SAMS-CAR consistently improves tail-class and worst-class performance while remaining competitive or superior on overall accuracy relative to strong recent baselines, including CAR and CAR+ConCutMix. The most salient finding is the improvement in worst-class accuracy and Worst-class Ratio (Table 3), which indicates that SAMS-CAR's sample adaptive margin mechanism specifically alleviates the catastrophic under learning of the rarest classes, rather than simply shifting the overall accuracy distribution. The method's advantage is also robust across all mentioned architectures and to increase under more severe class imbalance, supporting its practical relevance for real-world long-tailed deployments where imbalance factors are often extreme and unknown in advance. The limited settings in which competing methods retain a marginal edge namely, ConCutMix's input level augmentation and the very high capacity ViT Large backbone suggest that SAMS-CAR is complementary to, rather than in competition with, these techniques. A natural direction for future work is therefore to combine SAMS-CAR's adaptive margin scaling with CutMix-style augmentation strategies, as well as to examine whether its benefits can be further amplified in very high capacity models through backbone aware calibration of the adaptive margin.

REFERENCES

[1] Z. Zhu, G. Jin, H. Zhu, S.-Y. Lu, Y. Zhang, Z. Fu, R. Mu, G. Zhang, Z. Sun, Y. Xia, J. Shang, X. Li, L. Liu, and T. Huang, "Confusion-aware spectral regularizer for long-tailed recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2026.

[2] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] C. Zhang, G. Almpanidis, G. Fan, B. Deng, Y. Zhang, J. Liu, A. Kamel, P. Soda, and J. Gama, "A systematic review on long-tailed learning," *IEEE Trans. Neural Networks and Learning Systems*, 2025.

[4] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[5] J. Ren, C. Yu, X. Ma, H. Zhao, and S. Yi, "Balanced meta-softmax for long-tailed visual recognition," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[6] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," arXiv preprint arXiv:2007.07314, 2020.

[7] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," arXiv preprint arXiv:1910.09217, 2019.

[8] B. Neyshabur, S. Bhojanapalli, and N. Srebro, "A pac-bayesian approach to spectrally-normalized margin bounds for neural networks," arXiv preprint arXiv:1707.09564, 2017.

[9] E. Morvant, S. Koço, and L. Ralaivola, "Pac-bayesian generalization bound on confusion matrix for multi-class classification," arXiv preprint arXiv:1202.6228, 2012.

[10] H. Pan, Y. Guo, M. Yu, and J. Chen, "Enhanced long-tailed recognition with contrastive cutmix augmentation," *IEEE Trans. Image Processing*, 2024.

[11] H.-P. Chou, S.-C. Chang, J.-Y. Pan, W. Wei, and D.-C. Juan, "Remix: Rebalanced mixup," in *Proc. European Conf. Computer Vision Workshops (ECCVW)*, 2020, pp. 95–110.

[12] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng, "Metasaug: Meta semantic augmentation for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5212–5221.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[14] J. Ren, C. Yu, S. Sheng, X. Ma, H. Zhao, S. Yi, and H. Li, "Balanced meta-softmax for long-tailed visual recognition," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[15] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun, "Distribution alignment: A unified framework for long-tail visual recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2361–2370.

[16] S. Park, Y. Hong, B. Heo, S. Yun, and J. Y. Choi, "The majority can help the minority: Context-rich minority oversampling for long-tailed classification," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6887–6896.

[17] Y. Hong, J. Zhang, Z. Sun, and K. Yan, "Safa: Sample-adaptive feature augmentation for long-tailed image classification," in *Proc. European Conf. Computer Vision (ECCV)*, 2022.

[18] S. Alshammari, Y.-X. Wang, D. Ramanan, and S. Kong, "Long-tailed recognition via weight balancing," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6897–6907.

[19] Y. Du and J. Wu, "No one left behind: Improving the worst categories in long-tailed learning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 15 804–15 813.

[20] S. Sun, H. Lu, J. Li, Y. Xie, T. Li, X. Yang, L. Zhang, and J. Yan, "Rethinking classifier re-training in long-tailed recognition: Label oversmooth can balance," in *Int. Conf. Learning Representations (ICLR)*, 2025.

[21] J. Erbani, P.-□. Portier, E. Egyed-Zsigmond, and D. Nurbakova, "Confusion matrices: A unified theory," *IEEE Access*, 2024.

[22] G. Jin, S. Wu, J. Liu, T. Huang, and R. Mu, "Enhancing robust fairness via confusional spectral regularization," arXiv preprint arXiv:2501.13273, 2025.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learning Representations (ICLR)*, 2021.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE Int. Conf. Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.

[26] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[27] P. Wang, K. Han, X.-S. Wei, L. Zhang, and L. Wang, "Contrastive learning based hybrid networks for long-tailed image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 943–952.

[28] J. Zhu, R. Zhang, Y. Li *et al.*, "Balanced contrastive learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. XXXX–YYYY.

[29] M. Kim, S. Park, and J. Lee, "Targeted supervised contrastive learning for long-tailed recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6908–6918.

[30] C. Hou and et al., "Subclass-balancing contrastive learning for long-tailed recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[31] C. Du, A. Smith, and R. Kumar, "Probabilistic contrastive learning for long-tailed visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 7, pp. 3456–3470, 2024.

[32] S. Lee, H. Park, and D. Kim, "Long-tailed visual classification based on supervised contrastive learning with multi-view fusion," *Pattern Recognition Letters*, vol. 160, pp. 50–58, 2024.

[33] D. Muhammad, H. Ali, F. A. Azeemi, M. Khalid, N. Hussain, Z. Hussain, P. Hussain, S. Ali, M. Sangrasi, and M. Hammad, "Va-gkd: Variance-aware adaptive group knowledge distillation for long-tailed recognition," vol. 4, no. 3, pp. 2459–2471, 2026. [Online]. Available: https://thesesjournal.com/index.php/1/article/view/3439

[34] L. M. W. P. Sangrasi, DinMuhammadand-Wang, "Mitigatingtheadverseeffectsoflong-taileddataondeeplearningmodels," in *AustralasianConferenceonDataScienceandMachineLearning*. Springer, 2023, pp. 150–162.

[35] D. Muhammad, L. Wang, and M. Hagenbuchner, "Enhancing robustness in long-tailed image classification with angular approaches and balanced learning," *Data Science and Engineering*, vol. 10, no. 3, pp. 480–494, 2025.

[36] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[37] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2537–2546.

[38] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, no. 7, p. 3, 2015.

[39] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8769–8778.

[40] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9719–9728.