Feature Manifold-Aware Margin for Logits Calibration in Long-Tailed Recognition

Jiaheng Chen
Software College
Northeastern University
Shenyang, China
20236778@stu.neu.edu.cn

Chaopeng Guo *
Software College
Northeastern University
Shenyang, China
guochaopeng@swc.neu.edu.cn

***Abstract*—Decoupled training has become a standard paradigm for long-tailed recognition by separating representation learning from classifier retraining. However, most existing second-stage retraining methods characterize class-wise learning difficulty primarily based on sample quantity. As a result, differences in how classes are structured in the learned feature space are often overlooked. In this work, we identify representation-level characteristics as an additional factor influencing classifier difficulty, and propose Feature Manifold-Aware Margin, a lightweight plug-and-play module for classifier retraining. FMAM estimates class-wise geometric complexity using intra-class feature dispersion. Based on this estimate, it applies adaptive margins to calibrate class logits during training. Experiments on CIFAR-100-LT and ImageNet-LT demonstrate that FMAM consistently improves strong second-stage baselines, including cRT, LOS, and ViT-based models, with notable gains on few-shot categories of up to 3.87%. In particular, FMAM achieves a 3.79% improvement in overall accuracy on ImageNet-LT with a ViT backbone.**

***Index Terms*—Long-Tailed Recognition, Decoupled Learning, Classifier Retraining, Feature Manifold, Loss Function Design.**

I. Introduction

Real-world recognition datasets are often long-tailed, with a small number of classes containing most of the samples, while many others are represented by only a handful of instances. Training deep models under such imbalance tends to favor majority classes, which in turn leads to degraded performance on rare categories.

To address this issue, recent works have adopted the decoupled training paradigm, where representation learning and classifier retraining are performed in separate stages. In practice, this strategy has proven effective at reducing representation bias caused by imbalanced data, especially during the classifier retraining stage. [1]

Despite these advances, most second-stage classifier retraining methods implicitly rely on a simplifying assumption: once the feature representation is frozen, class-wise learning difficulty can be largely characterized by sample quantity. Under this assumption, existing approaches mainly focus on resampling, re-weighting, or applying uniform calibration strategies to balance the learning strength across classes. [2] [3] However, such treatments overlook an important aspect of the frozen feature space—different classes may exhibit substantially different geometric structures, even when trained under identical sampling conditions. [4] [5]

In deep models, feature representations are not distributed uniformly across classes. Some categories form compact clusters that are easy to separate, while others occupy much broader regions of the feature space, often with irregular or overlapping structures. Importantly, these differences do not disappear simply because the data are rebalanced. Even under class-balanced sampling, certain categories remain inherently more difficult to model due to variations in visual appearance or semantic definition. [6]

This observation becomes particularly relevant in the second stage of decoupled training, where the feature extractor is fixed and the classifier operates within a shared representation space. At this stage, performance differences across classes cannot be attributed to sample quantity alone. In our view, they are also shaped by how reliably class distributions can be separated in the frozen feature space.

Classes with dispersed or unstable feature distributions often require stronger constraints to obtain consistent decision boundaries. Applying identical assumptions or regularization strategies to all classes, however, may be unnecessarily restrictive for some and insufficient for others. [7]

Based on this observation, we propose Feature Manifold-Aware Margin (FMAM) for second-stage classifier retraining. Rather than introducing additional architectural components or relying on prediction uncertainty [4], FMAM leverages a simple statistic—the dispersion of class features in the frozen representation space—to approximate class-wise geometric complexity. This quantity is then used to modulate class-specific logit margins during training. The resulting formulation is lightweight and easy to integrate, and it does not alter inference-time computation.

The main contributions of this work can be summarized as follows:

- We highlight that, in the classifier retraining stage of long-tailed recognition, sample quantity alone is insufficient to fully characterize class-wise learning difficulty, and that geometric differences in the representation space constitute an important complementary factor.
- We use intra-class feature dispersion as a stable proxy for class-wise geometric complexity, and employ it to guide

a feature-aware margin design for classifier calibration.
- Experiments on CIFAR-100-LT and ImageNet-LT show that FMAM integrates naturally with strong second-stage baselines and improves performance across architectures, especially on medium- and few-shot categories.

Our code will be made publicly available upon acceptance.

II. Related Work

A. Long-Tailed Recognition via Class Frequency Modeling

Early approaches to long-tailed recognition primarily attribute class-wise learning difficulty to sample frequency imbalance, and address this issue through re-sampling, re-weighting, or frequency-aware loss design. Representative methods include LDAM [2], which introduces class-dependent margins derived from class sample counts, and Balanced Softmax, which corrects biased class priors by modifying the softmax normalization term.

With the development of the decoupled training paradigm, methods such as cRT [1] separate representation learning from classifier retraining, allowing the classifier to be optimized under more balanced sampling within a frozen feature space. This strategy has become a standard baseline in long-tailed recognition. However, during second-stage classifier retraining, most decoupled methods still rely on class frequency or uniform regularization schemes, applying identical constraints across classes without explicitly modeling structural differences in their feature distributions.

B. Optimization-Driven and Calibration-Based Methods

More recent studies recognize that learning difficulty cannot be fully characterized by sample quantity alone, and introduce adaptive mechanisms based on optimization dynamics or prediction-level signals. For instance, ALA [4] adjusts logit biases according to prediction uncertainty, while DBM [5] imposes instance-specific margin constraints on hard or misclassified samples. Another line of work focuses on logit calibration and regularization under the decoupled training framework, including methods such as MiSLAS [3], MARC [8], and the recently proposed LOS [9].

Although effective in practice, these approaches primarily operate at the optimization or calibration level, and typically adopt uniform or empirically designed adjustment schemes. In the second stage of decoupled training, as feature representations are frozen, the classifier increasingly faces the challenge of separating different classes within a fixed representation space. Difficulty modeling based solely on optimization dynamics or calibration mechanisms is therefore insufficient to capture class-level structural differences in the learned feature space.

In contrast to these approaches, our method focuses on modeling class-level representation difficulty in the frozen feature space, and introduces a geometry-aware margin mechanism that complements existing frequency-based and calibration strategies.

C. Architecture- and Contrastive-Learning-Based Approaches

Another line of research addresses long-tailed recognition by modifying the training paradigm or model structure, rather than focusing on classifier retraining under a fixed representation. For example, RIDE [10] adopts a multi-expert architecture and routes samples to different experts to handle distribution imbalance. Such approaches mainly improve performance by enhancing representation capacity during training, and therefore operate in a setting where feature representations are still being actively optimized.

Contrastive learning based methods follow a similar line of thought. Approaches such as BCL [11] and PaCo [12] introduce class-aware contrastive objectives to mitigate representation bias caused by long-tailed data, typically together with carefully designed sampling strategies. These methods explicitly reshape the feature space during representation learning, and their effectiveness relies on the ability to modify representations through the loss.

In contrast, this work focuses on the classifier retraining problem after representations are fixed, rather than on representation learning itself. Under this premise, contrastive learning methods and multi-expert architectures address a different stage of the training pipeline, and thus are not situated in the same problem setting.

III. Method

A. Problem Setup and Decoupled Training

We study long-tailed recognition in the decoupled training setting, where model optimization is performed in two stages. In the first stage, a feature extractor is trained on the original long-tailed dataset to learn discriminative visual representations. In the second stage, the feature extractor is frozen and only the classifier is retrained, with the goal of mitigating decision bias caused by class imbalance. Our method is strictly applied to the second-stage classifier retraining.

Let the frozen feature extractor be denoted as $f(\cdot)$. Given an input sample x, the extracted feature is

$$\mathbf{z} = f(x). \tag{1}$$

The classifier is linear, and the logit corresponding to class c is given by

$$\ell_c = \mathbf{w}_c^\top \mathbf{z} + b_c, \tag{2}$$

where $\mathbf{w}_c$ and b_c represent the classifier weight and bias, respectively.

As all classes share a frozen representation space in the second stage, learning reliable and well-separated decision boundaries becomes non-trivial.

Conventional second-stage paradigms typically operate under the assumption that class-wise learning difficulty is a direct function of sample frequency alone. We contend, however, that within the shared frozen feature space, the geometric structure of class distributions—specifically their compactness and stability—plays a decisive role in establishing robust decision boundaries. Dispersed categories remain inherently more susceptible to boundary perturbations even when sample sizes

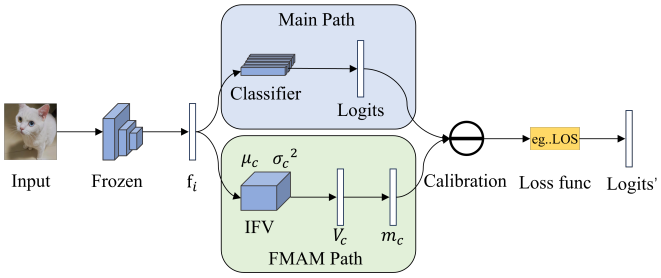

Fig. 1. FMAM Overview: FMAM acts as a plug-and-play module in the second stage of decoupled training. For a feature $\mathbf{z}_i$ from the frozen feature extractor, our framework uses a dual-path approach: (Top) The main path feeds $\mathbf{z}_i$ into the baseline classifier to obtain the original logits. (Bottom) The parallel FMAM path uses $\mathbf{z}_i$ to online-estimate the intra-class feature variance, thereby calculating the manifold complexity V_c and the final adaptive margin m_c. Finally, m_c refines the original logits via the logits calibration module. This design highlights FMAM's lightweight and universal nature, allowing seamless integration with any baseline method (e.g., cRT [1], LOS [9]).

are comparable. Consequently, frequency-based calibration proves insufficient, necessitating a shift toward remodeling class difficulty through the lens of representation geometry.

B. Estimating Class-wise Feature Manifold Complexity

Directly modeling the geometric complexity of high-dimensional feature manifolds is computationally expensive and often unstable in practice. Instead, we adopt a lightweight and robust proxy by characterizing the intra-class feature dispersion in the frozen representation space.

For each class c, we maintain a running estimate of its feature mean $\boldsymbol{\mu}_c$ and intra-class variance σ_c^2. During second-stage training, these statistics are updated online using an exponential moving average.

Given a feature vector $\mathbf{z}_i$ belonging to class c, the update rules are

$$\boldsymbol{\mu}_c \leftarrow (1-\eta)\boldsymbol{\mu}_c + \eta \mathbf{z}_i, \tag{3}$$

$$\sigma_c^2 \leftarrow (1-\eta)\sigma_c^2 + \eta \|\mathbf{z}_i - \boldsymbol{\mu}_c\|^2, \tag{4}$$

where η is the momentum coefficient.

To reduce scale discrepancies across classes, we apply a logarithmic transformation followed by normalization, yielding the class-wise feature manifold complexity score

$$V_c = \mathrm{Norm}(\log \sigma_c^2), \quad V_c \in [0,1]. \tag{5}$$

We emphasize that our goal is not to precisely model the full geometry of the feature manifold. Instead, intra-class feature variance serves as a computable, interpretable, and training-dynamics-independent proxy that captures relative structural complexity across classes in the frozen representation space.

C. Feature Manifold-Aware Margin

Based on the estimated class-wise geometric complexity, we introduce Feature Manifold-Aware Margin (FMAM) to impose representation-aware constraints during second-stage classifier retraining, as illustrated in Fig. 1.

For each class c, the margin is defined as

$$m_c = \alpha V_c + \beta S_c, \tag{6}$$

Specifically, the data scarcity term S_c is formulated as $(\frac{1}{N_c})^\gamma$, where N_c denotes the number of training samples for class c. The exponent γ serves as a scaling factor to control the sensitivity of the margin to sample frequency. While V_c captures the intrinsic representation difficulty from the feature space, S_c addresses the external imbalance in data distribution. By incorporating both terms, the total margin m_c provides a multi-dimensional characterization of class-wise learning difficulty. This hybrid design ensures that classes with either limited samples or dispersed feature distributions are prioritized through stronger logit constraints. In scenarios with strong baseline regularizers like LOS [9], we typically deactivate the scarcity term by setting $\beta = 0$, allowing the module to isolate and address the orthogonal dimension of representation geometry.

In strong baseline settings such as LOS [9], we set $\beta = 0$ to isolate the effect of representation geometry and highlight its orthogonality to sample-frequency-based modeling.

During training, FMAM calibrates the logits by subtracting the class-specific margin from the ground-truth class:

$$\ell'_j = \begin{cases} \ell_j - m_j, & j = y, \\ \ell_j, & j \neq y, \end{cases} \tag{7}$$

where y denotes the ground-truth label.

This formulation enforces stronger discriminative constraints on classes with higher geometric complexity, encouraging the classifier to learn more stable decision boundaries. Conversely, classes with simpler and more compact feature distributions are not over-regularized, avoiding unnecessary performance degradation.

FMAM operates exclusively during second-stage training and does not modify the classifier architecture or loss function. To ensure training stability, we adopt a warm-up strategy during which the FMAM margins are temporarily disabled. Specifically, during the first T_{warm} epochs, we set $\alpha = \beta = 0$, allowing the classifier to converge under the baseline optimization dynamics. After warm-up, FMAM is gradually activated together with a cosine annealing learning rate schedule with warm restarts, enabling stable adaptation to the geometry-aware margins. The overall training schedule with warm-up is summarized in Fig. 2.

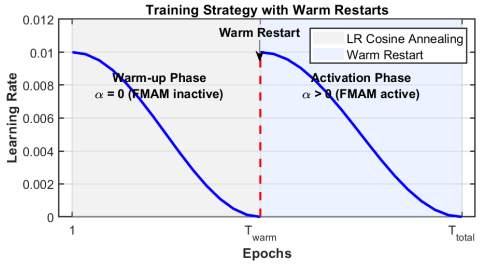

Fig. 2. Schematic of the two-stage training strategy with warm restarts

After training, the learned margins are absorbed into the classifier bias terms, leaving inference-time computation unchanged.

TABLE I
MAIN RESULTS ON LONG-TAILED RECOGNITION BENCHMARKS (TOP-1 ACCURACY %). WE SHOWCASE FMAM'S PLUG-AND-PLAY EFFECTIVENESS ON BOTH CRT [1] AND THE STRONG LOS [9] BASELINE ACROSS DIFFERENT BACKBONES.

Method	Backbone	CIFAR-100-LT			ImageNet-LT			
		IR=100	IR=50	IR=10	All	Many	Medium	Few
Baselines on ResNet-34								
cRT [1]	ResNet-34	41.20	46.80	57.90	49.60	61.80	46.20	27.40
cRT [1] + FMAM (ours)	ResNet-34	**43.87**	49.40	59.78	51.43	63.16	48.38	30.39
LOS [9]	ResNet-34	52.47	56.67	64.88	54.40	63.20	50.70	42.30
LOS [9] + FMAM (ours)	ResNet-34	**53.14**	57.07	65.06	54.88	63.57	51.17	43.08
Baselines on ViT-B/16 [13]								
cRT [1]	ViT-B/16 [13]	—	—	—	77.54	79.60	77.69	73.47
cRT [1] + FMAM (ours)	ViT-B/16 [13]	—	—	—	**81.33**	**84.87**	**82.69**	**74.42**

IV. EXPERIMENTS

A. Experimental Setup

a) Datasets: We evaluate FMAM on two representative long-tailed recognition benchmarks: CIFAR-100-LT and ImageNet-LT.

CIFAR-100-LT is constructed by applying exponential downsampling to CIFAR-100, resulting in controlled imbalance ratios (IR = 10, 50, 100). This dataset allows systematic evaluation under varying degrees of long-tailed severity.

ImageNet-LT is a real-world long-tailed dataset with a much larger number of classes and a highly skewed sample distribution, serving as a challenging benchmark for large-scale evaluation.

b) Backbone architectures and baselines: To demonstrate the generality of FMAM across different inductive biases, we adopt two representative architectures: ResNet-34 and ViT-B/16 [13].

For second-stage classifier retraining, we adopt cRT [1] and LOS [9] as baseline methods. cRT [1] is a widely used and stable decoupled-training baseline, while LOS [9] represents one of the strongest recent approaches for second-stage classifier optimization. Our goal is to verify whether FMAM can consistently improve performance on top of such strong baselines.

c) Implementation details: All experiments follow the standard decoupled training protocol. The feature extractor is trained in the first stage and frozen thereafter. In the second stage, only the classifier is retrained using class-balanced sampling. The retraining stage lasts for 30 epochs, with the first 15 epochs serving as a warm-up period. We adopt a cosine annealing learning rate schedule with warm restarts. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

B. Main Results

Table I summarizes the overall performance of FMAM on CIFAR-100-LT and ImageNet-LT. To highlight the plug-and-play nature of FMAM, we report results when integrating it into both cRT [1] and LOS [9] baselines.

Several observations can be drawn from the results:

TABLE II
PERFORMANCE GAIN OF THE FMAM MODULE ON DIFFERENT BASELINES

Baseline Method	Overall Acc. (%)	Few-shot Acc. (%)
cRT [1]	41.2	26.33
cRT [1] + FMAM (ours)	43.87	27.80
Gain (Δ)	+2.67	+1.47
LOS [9]	52.47	27.60
LOS [9] + FMAM (ours)	53.14	31.47
Gain (Δ)	+0.67	**+3.87**

TABLE III
COMPARISON WITH REPRESENTATIVE SOTA METHODS ON IMAGENET-LT.

Method	All (%)	Few (%)
cRT [1]	49.60	27.40
EWB-FDR [14]	53.20	35.10
LTWB [15]	53.90	41.50
LOS [9]	54.40	42.30
LOS + FMAM(ours)	**54.88**	**43.08**

a) Consistent performance gains: FMAM yields consistent improvements over both cRT [1] and LOS [9] across different imbalance ratios. The gains become more pronounced as the long-tailed severity increases (e.g., IR = 100), indicating that FMAM effectively complements existing second-stage strategies under challenging imbalance conditions. These results suggest that FMAM does not rely on a specific loss function or architecture and can serve as a general enhancement module.

b) Improvements on medium- and few-shot classes: The overall accuracy gains mainly come from medium-shot and few-shot categories. For example, on CIFAR-100-LT with IR = 100, FMAM improves the few-shot accuracy of LOS [9] by 3.87%. The detailed performance gains across different baselines are summarized in Table II.This suggests that incorporating representation-level difficulty is beneficial for classes with more complex feature distributions.

c) Cross-architecture generalization: Results on ImageNet-LT show that FMAM also benefits ViT-based [13] architectures. Compared with convolutional models, Vision Transformers tend to be more sensitive to variations in

TABLE IV
SENSITIVITY ANALYSIS OF COMPLEXITY WEIGHT α AND SCARCITY WEIGHT β (BASED ON LOS [9] BASELINE)

Case	β	α	Overall Acc. (%)	Few-shot Acc. (%)
LOS [9] (Baseline)	0	0	52.47	27.60
+ Scarcity Margin Only	1.0	0	52.40 (-0.07)	26.90
+ Hybrid Margin	0.5	0.5	52.87 (+0.40)	29.97
+ Complexity Margin Only (ours)	0	0.3	**53.14 (+0.67)**	**31.47**

TABLE V
ABLATION STUDY ON VIT [13] BACKBONE

Case	β	α	Overall Acc. (%)	Few-shot Acc. (%)
ViT [13] (Baseline)	0	0	77.54	73.47
+ Scarcity Margin Only	0.2	0	81.23	74.13
+ Hybrid Margin (ours)	0.2	0.5	**81.33(+3.79)**	**74.42(+0.95)**
+ Complexity Margin Only	0	0.5	81.22	74.11

class-level feature distributions. In this setting, FMAM yields more noticeable overall improvements, indicating that it remains effective across different types of representation spaces.

C. Comparison with Related Margin and Calibration Methods

To contextualize FMAM within existing approaches, we further compare it with representative margin- and calibration-based methods, including LDAM [2] and MiSLAS [3].

The results show that FMAM achieves competitive or superior performance across multiple settings, particularly on medium- and few-shot categories. It is worth noting that these competing methods primarily modulate classifier behavior based on sample frequency or prediction-level calibration. In contrast, FMAM focuses on class-wise geometric structure in the frozen feature space. This difference in modeling perspective explains why FMAM can complement, rather than replace, existing methods and still provide additional gains without introducing complex architectural changes.

D. Ablation Studies and Analysis

a) Plug-and-Play Property of FMAM: We first examine the effectiveness of FMAM as an independent module under different baseline frameworks. As shown in Table IV, FMAM consistently improves performance when integrated into both cRT [1] and LOS [9], without requiring any modification to the original model architecture or training pipeline. This confirms the plug-and-play nature of the proposed method.

b) Effects of Different Difficulty Modeling Factors: To analyze the respective roles of geometric complexity and data scarcity, we conduct ablation studies by varying the coefficients α and β in Eq. (4).

Under strong regularization baselines such as LOS [9], relying solely on sample-frequency-based margins can be sub-optimal. In contrast, incorporating geometry-aware calibration results in more stable performance gains across settings.

On ViT [13] backbones, jointly considering geometric complexity and data scarcity yields the best performance, suggesting that different architectures may emphasize different aspects of class difficulty.Quantitative results on ViT backbones are reported in Table V.

c) Relationship Between Geometric Complexity and Performance Gain: We analyze the relationship between class-wise geometric complexity and the performance gains introduced by FMAM.

Classes with higher complexity scores generally show larger improvements, while those with more compact feature distributions exhibit smaller changes. This pattern is consistent with the role of intra-class feature dispersion as an indicator of representation-level complexity.

In addition, FMAM introduces no additional learnable parameters, and its computational overhead during training is negligible. At inference time, the learned margins are absorbed into the classifier bias, leaving the test-time cost unchanged.

E. Qualitative Analysis of Feature Manifold Complexity

To further interpret the estimated feature manifold complexity, we conduct a qualitative analysis on ImageNet-100, as shown in Fig. 3.

Classes with higher estimated complexity generally show larger appearance variation, more diverse materials, and more heterogeneous backgrounds, whereas low-complexity classes tend to exhibit more stable visual structures and consistent discriminative patterns.

This observation supports our motivation that the proposed complexity estimate captures meaningful intra-class variability in the learned feature space.

We also observe a counter-intuitive case for small objects such as *safety pin*, whose high dispersion is driven primarily by background variation rather than object geometry.

This is likely because ImageNet contains complex backgrounds, and classification-oriented training may be less effective when the object occupies only a small image region.

V. CONCLUSION

In this work, we revisit the second-stage classifier retraining problem in long-tailed recognition and propose Feature Manifold-Aware Margin, a representation-aware calibration strategy grounded in the geometry of the frozen feature space. Compared with approaches that rely on sample quantity or optimization dynamics, FMAM focuses on modeling class-wise difficulty in the frozen representation space using intra-class feature dispersion. Experiments on CIFAR-100-LT and ImageNet-LT show that FMAM integrates well with a range of second-stage baselines and improves performance across different architectures. For example, on CIFAR-100-LT with IR = 100, FMAM improves the few-shot accuracy of LOS by 3.87%. On ImageNet-LT with a ViT backbone, it achieves a 3.79% gain in overall accuracy, indicating its effectiveness in representation spaces with weaker inductive bias.

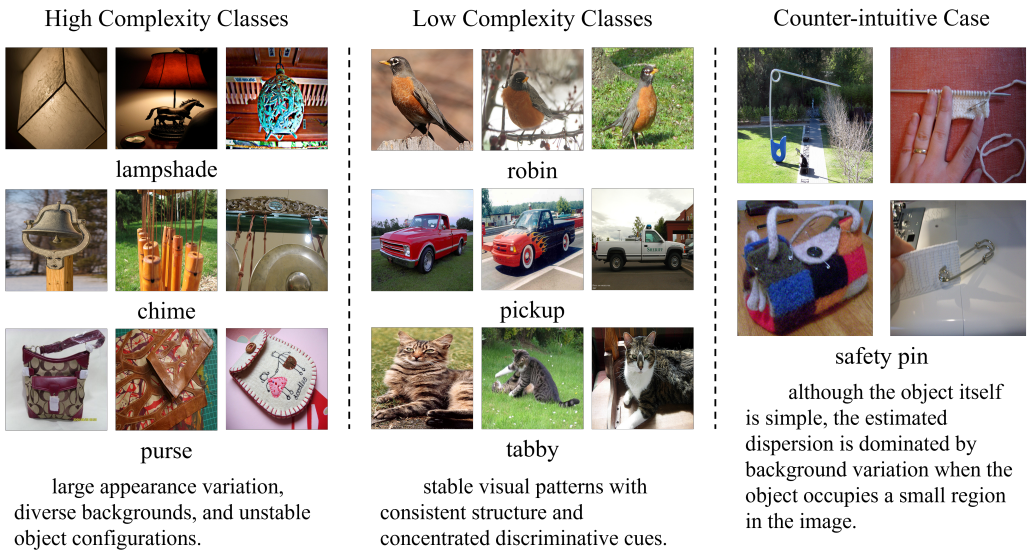

Fig. 3. Qualitative analysis of class-wise feature manifold dispersion on ImageNet-100. **Left (high complexity):** classes such as *lampshade*, *purse*, and *chime* exhibit large intra-class variation due to diverse appearances, materials, and backgrounds. In particular, the category of *chime* includes visually distinct objects (e.g., wind chimes and metal chimes), resulting in substantial feature dispersion. **Middle (low complexity):** classes such as *robin*, *pickup*, and *tabby cat* show relatively stable visual patterns. For example, robins share consistent color distributions on the chest and head, pickup trucks have similar geometric structures with characteristic cargo beds, and tabby cats exhibit comparable stripe patterns across images. **Right (counter-intuitive case):** although the structure of *safety pin* is visually simple and consistent, its estimated dispersion is high, mainly because the object occupies a small region in the image and the surrounding background varies significantly.

ACKNOWLEDGMENT

This paper is supported by the National Natural Science Foundation of China (Grant No. 62302086).

REFERENCES

[1] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia, April 2020. OpenReview.net.

[2] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett, editors, *Annual Conference on Neural Information Processing Systems 2019*, pages 1565–1576, Vancouver, BC, Canada, December 2019.

[3] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving Calibration for Long-Tailed Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16489–16498, virtual, June 2021. Computer Vision Foundation / IEEE.

[4] Yan Zhao, Weicong Chen, Xu Tan, Kai Huang, and Jihong Zhu. Adaptive Logit Adjustment Loss for Long-Tailed Visual Recognition. In *AAAI Conference on Artificial Intelligence*, pages 3472–3480, Virtual Event, February 2022. AAAI Press.

[5] Minseok Son, Inyong Koo, Jinyoung Park, and Changick Kim. Difficulty-aware Balancing Margin Loss for Long-tailed Recognition. In Toby Walsh, Julie Shah, and Zico Kolter, editors, *the Association for the Advancement of Artificial Intelligence*, pages 20522–20530, Philadelphia, PA, USA, February 2025. AAAI Press.

[6] Pietro Boni, Mirko Mazzoleni, Matteo Scandella, and Fabio Previdi. A graph learning approach for kernel-based system identification with manifold regularization. *Journal of the Franklin Institute*, 362(12):107793, 2025.

[7] Kyle Erwin and Andries P. Engelbrecht. Feature-based complexity measure for multinomial classification datasets. *Entropy*, 25(7):1000, 2023.

[8] Yidong Wang, Bowen Zhang, Wenxin Hou, Zhen Wu, Jindong Wang, and Takahiro Shinozaki. Margin Calibration for Long-Tailed Visual Recognition. In Vineeth N. Balasubramanian and Ivor W. Tsang, editors, *Asian Conference on Machine Learning*, volume 189 of *Proceedings of Machine Learning Research*, pages 1101–1116, Hyderabad, India, December 2022. PMLR.

[9] Siyu Sun, Han Lu, Jiangtong Li, Yichen Xie, Tianjiao Li, Xiaokang Yang, Liqing Zhang, and Junchi Yan. Rethinking Classifier Re-Training in Long-Tailed Recognition: Label Over-Smooth Can Balance. In *International Conference on Learning Representations*, Singapore, April 2025. OpenReview.net.

[10] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. Long-tailed Recognition by Routing Diverse Distribution-Aware Experts. In *International Conference on Learning Representations*, Austria, May 2021. OpenReview.net.

[11] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced Contrastive Learning for Long-Tailed Visual Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6898–6907, New Orleans, LA, USA, June 2022. IEEE.

[12] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric Contrastive Learning. In *IEEE/CVF International Conference on Computer Vision*, pages 695–704, Montreal, QC, Canada, October 2021. IEEE.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In Marina Meila and Tong Zhang, editors, *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, Virtual Event, July 2021. PMLR.

[14] Naoya Hasegawa and Issei Sato. Exploring Weight Balancing on Long-Tailed Recognition Problem. In *International Conference on Learning Representations*, Vienna, Austria, May 2024. OpenReview.net.

[15] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long- Tailed Recognition via Weight Balancing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6887–6897, New Orleans, LA, USA, June 2022. IEEE.