MemTailor: Hierarchical Memory-Augmented Multi-Expert Learning for Long-Tailed Recognition

Yuyang Sun[1,†] **Senyang Su**[2,†] **Xiaotian Wang**[3,4] **Pengkun Wang**[1,5,*] **Yang Wang**[1,5,*]

[1]University of Science and Technology of China
[2]Huazhong University of Science and Technology
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
[4]Institute of Automation, Chinese Academy of Sciences
[5]Suzhou Institute for Advanced Research, USTC

frontsea@mail.ustc.edu.cn susenyang@hust.edu.cn
wangxiaotian26@mails.ucas.ac.cn {pengkun, angyan}@ustc.edu.cn

Abstract

Long-tailed recognition remains difficult because rare-class evidence is not only weakly supervised during training, but is also poorly retained in parameters dominated by head-class gradients. We present **MemTailor**, a hierarchical memory-augmented multi-expert framework that addresses this problem by coupling frequency-aware expert learning with class-aware evidence preservation. MemTailor maintains a hierarchical memory for each expert, composed of class prototypes and bounded instance slots over shallow and deep features, with class-dependent capacity that allocates richer storage to rare classes. During training, the memory performs same-class feature augmentation to reinforce under-represented categories without introducing cross-class drift. During inference, memory quality and prediction uncertainty jointly control when stored evidence is allowed to refine expert logits, and the refined predictions are further fused across multiple input scales. Extensive experiments on CIFAR-100-LT, ImageNet-LT, iNaturalist 2018, and Places-LT show that MemTailor consistently improves long-tailed recognition, with especially strong gains on few-shot classes and under shifted test distributions.

1 Introduction

Representation learning under imbalanced data distributions is common in real-world settings but is often overlooked by standard learning paradigms. Long-tailed distributions are a particularly representative and severe form of imbalance. In computer vision, long-tailed visual recognition studies classification with highly imbalanced class frequencies: a small number of frequent head classes account for most training samples, while many rare tail classes provide only limited visual evidence. Standard empirical risk minimization (ERM) therefore learns decision boundaries biased toward head classes. Existing methods mitigate this bias at the optimization level, for example through re-sampling, re-weighting, margin adjustment, classifier calibration, or decoupled training [Cui et al., 2019, Cao et al., 2019a, Kang et al., 2020, Ren et al., 2020b, Alshammari et al., 2022]. These methods have substantially advanced long-tailed learning, but share a deeper limitation: in a single model dominated by head-class gradients, the visual evidence contributed by rare classes is often systematically under-retained in the model parameters. Improving the loss alone cannot recover information that the network has failed to preserve.

*Corresponding authors. †Equal contribution.

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

To move beyond single-model optimization, multi-expert architectures assign specialized branches to different regions of the data distribution, thereby increasing prediction diversity [Wang et al., 2020, Li et al., 2022b]. This is a promising architectural direction, but both optimization-level loss correction and expert specialization overlook a deeper evidence-preservation problem: when head-class gradients dominate, visual evidence from rare classes may already be systematically discarded during feature extraction. Neither downstream loss adjustment nor branch specialization can restore information that the network has not retained. In a multi-expert model, each expert still depends on features from a backbone shaped by head-class dominance. Rare-class representations are therefore affected in two compounding ways: gradient suppression during training weakens their feature regions, while the resulting low-confidence predictions at inference leave the model without reliable evidence to support rare-class decisions.

These two failure modes point to the same missing component and raise a natural question: **How can we preserve rare-class information independently of imbalanced parameter learning and make it available when needed?** A solution must address two challenges:

(1) **During training**, it should record class prototypes and diverse boundary instances, providing rare classes with a retrieval-based evidence channel that is not eroded by head-class gradients.

(2) **At inference**, it should use uncertainty to control evidence consultation, refining low-confidence predictions while avoiding disruption to decisions the model can already make reliably.

Effective external evidence must also span multiple levels of feature abstraction. Shallow features capture low-level discriminative cues that can be lost in deeper layers, whereas deep features encode the high-level semantic structure needed for prototype matching. We therefore propose **MemTailor**, a memory-guided framework for long-tailed visual recognition built on a multi-expert backbone. MemTailor maintains a hierarchical, class-aware memory over shallow and deep feature spaces, with capacity allocated according to class frequency so that rare classes receive relatively richer storage. During training, same-class retrieval augments tail-class features and provides a training signal less diluted by head-class gradients. At inference, quality-gated deep-memory scores first refine expert logits; prediction uncertainty then determines whether an additional shallow–deep correction is applied. Finally, MemTailor fuses the memory-corrected logits from multiple input scales to form the prediction.

Our contributions are as follows:

- We identify gradient-driven evidence under-retention as a shared limitation of existing long-tailed recognition methods and show how an external class-aware memory addresses its manifestations during both training and inference.
- We propose a hierarchical class-aware memory and its associated training and inference mechanisms. The memory preserves prototype- and instance-level evidence across shallow and deep feature spaces, allocates capacity according to class frequency, augments tail-class features via same-class retrieval during training, and improves rare-class predictions through uncertainty- and quality-gated correction at test time.
- We combine memory-guided expert inference with multi-scale logit fusion, aggregating memory-corrected predictions across input scales rather than relying on raw logits from a single scale.

2 Related Work

Learning with limited data. Learning with limited or imperfectly distributed data has also been studied through training dynamics, distribution geometry, and objective design. Mentored learning examines teacher-guided student learning to improve generalization and convergence [Cao et al., 2024]; related work studies distribution disagreement with a Lorentzian focal representation [Cao and Tsang, 2022] and data-efficient learning by minimizing hyperspherical energy [Cao et al., 2023]. A recent survey organizes methods for low-resource data learning [Cao et al., 2025]. These works address complementary aspects of learning under data constraints; MemTailor focuses specifically on preserving class-conditioned visual evidence for tail classes and consulting it during training and inference.

Long-tailed visual recognition. Long-tailed recognition methods primarily reduce the dominance of frequent classes by modifying data sampling, loss functions, classifier calibration, or training schedules. Re-sampling and augmentation methods rebalance the training distribution or synthesize more informative minority-class examples [Chou et al., 2020, Zhang et al., 2021], while loss-based methods reshape the optimization objective through class-balanced weights, margins, or distribution-aware normalization [Cui et al., 2019, Cao et al., 2019a, Wu et al., 2020]. Another line decouples representation learning from classifier learning or calibrates logits according to class priors [Kang et al., 2020, Ren et al., 2020b, Alshammari et al., 2022]. These methods improve the training signal seen by rare classes, but they still rely on parameters learned from an imbalanced dataset to preserve rare-class evidence. MemTailor instead keeps a class-aware memory alongside the classifier, so rare-class features can be stored, retrieved, and consulted after the backbone has produced its prediction.

Multi-expert long-tailed learning. Multi-expert models improve long-tailed recognition by assigning different branches to different regions of the class distribution. Routing-based or distribution-aware expert methods encourage experts to specialize at the decision level [Wang et al., 2021], and later methods further strengthen complementarity through self-distillation, aggregation, calibration, or collaborative training [Li et al., 2022b, Zhang et al., 2022, Sanchez Aimar et al., 2023, Zhao et al., 2023]. More broadly, mixture-of-experts architectures show that conditional or sparse expert computation can increase model capacity without forcing all samples through the same predictor [Jacobs et al., 1991, Shazeer et al., 2017, Fedus et al., 2022, Zhou et al., 2022]. MemTailor follows this expert-specialization principle but adds a memory path that is class-conditioned and feature-aware. The experts provide diverse logits and uncertainty estimates, while the memory supplies stored visual evidence that is not limited to the current parametric representation.

Memory and retrieval for rare-class evidence. Memory-based recognition methods are useful when a parametric classifier may not retain all instance-level evidence needed for later decisions. In long-tailed recognition, this issue is most visible for rare classes, where few examples can be overwhelmed by head-class gradients during representation learning. MemTailor addresses this gap with a hierarchical memory composed of class prototypes and instance slots. The prototype captures stable class-level evidence, while the slots preserve hard or diverse samples under a bounded capacity. Unlike generic feature caching, the memory is written with class-aware gates and is read only from same-class memory during training augmentation, reducing the risk of moving rare-class features toward unrelated classes.

Uncertainty-guided prediction. Uncertainty estimation provides a mechanism for deciding when a model prediction should be trusted or refined. Bayesian dropout, deep ensembles, and calibration studies provide common tools for estimating predictive uncertainty and reliability [Gal and Ghahramani, 2016, Kendall and Gal, 2017, Lakshminarayanan et al., 2017, Guo et al., 2017]. In long-tailed recognition, uncertainty is especially relevant because a low-confidence prediction may come from rare-class under-representation rather than only image ambiguity. MemTailor uses expert uncertainty as a gate for memory consultation: confident predictions remain close to the expert ensemble, while uncertain predictions receive additional shallow–deep memory evidence. This design treats uncertainty as a control signal for selective retrieval rather than as an auxiliary diagnostic.

3 Method

3.1 Problem setup and overview

We study long-tailed image classification on $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{1, \dots, C\}$ and class c contains n_c training samples. The goal is to learn a classifier that remains reliable for rare classes without losing accuracy on frequent classes. MemTailor addresses this goal with a frequency-aware expert backbone and a class-aware memory that stores shallow and deep visual evidence for selective use.

For an input image x, expert $e \in \{1, \dots, E\}$ produces a deep feature $h^{(e)} \in \mathbb{R}^d$ and logits $z^{(e)} \in \mathbb{R}^C$. Before memory correction, the ensemble prediction is $\bar{z} = E^{-1} \sum_{e=1}^{E} z^{(e)}$. The memory for class c and expert e contains a prototype $\mu_c^{(e)}$ and bounded instance slots $\mathcal{M}_c^{(e)} = \{m_{c,k}^{(e)}\}_{k=1}^{K_c}$, built from

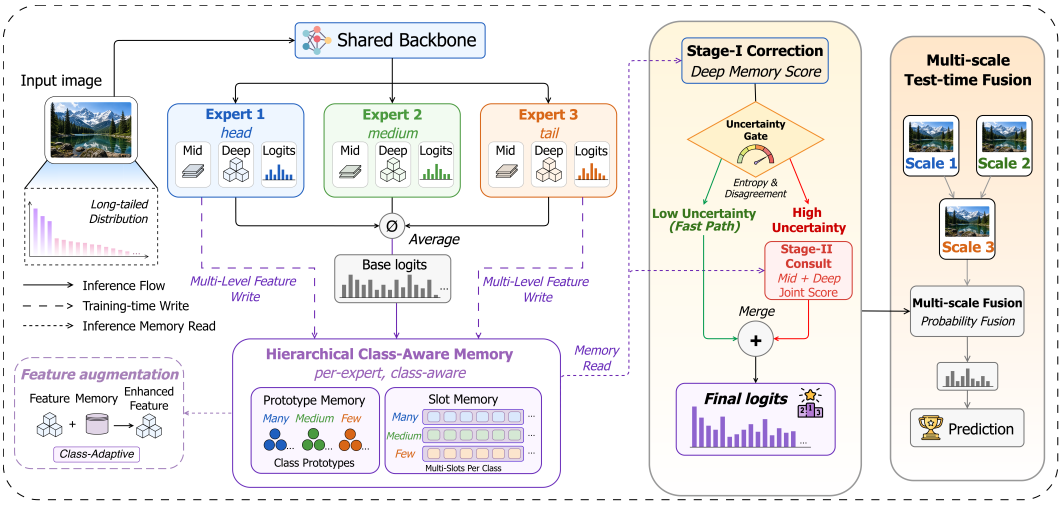


Figure 1: Overview of MemTailor. A shared backbone feeds multiple frequency-aware experts, while shallow and deep features update class-aware prototype and slot memories. At inference time, memory evidence is quality-gated and used for uncertainty-guided prediction correction before multi-scale logit fusion.

both shallow features $h_{\mathrm{s}}^{(e)}$ and deep features $h_{\mathrm{d}}^{(e)}$. Figure 1 summarizes the resulting flow: expert predictions provide the base decision, memory records class-conditioned evidence during training, and uncertain test samples receive additional memory-guided correction.

3.2 Frequency-aware expert backbone

MemTailor uses a shared-to-specialized backbone. Early layers extract a common representation, and each expert branch maps it to an expert-specific feature and classifier output:

$$u = \phi_{\mathrm{shared}}(x), \qquad h^{(e)} = \phi_e(u), \qquad z^{(e)} = s \cdot \mathrm{NormLinear}^{(e)}(h^{(e)}). \tag{1}$$

Here s is a fixed logit scale, and the branch-specific mapping allows experts to specialize after sharing low-level visual extraction. To make the experts complementary under class imbalance, their supervised losses use different class-prior biases. Let $\pi = [n_1, \ldots, n_C]$ be the frequency vector after classes are ordered by frequency. For expert-specific powers $a_e \in \{-0.5, 1.0, 2.5\}$, the biased logits and supervised loss are

$$\tilde{z}^{(e)} = z^{(e)} + \log(\pi^{a_e}), \qquad \mathcal{L}_{\mathrm{ce}} = \sum_{e=1}^{E} \mathrm{CE}(\tilde{z}^{(e)}, y). \tag{2}$$

The different powers bias experts toward different frequency regions while preserving a shared input representation.

The consistency term is added only after checking the teacher prediction, and its weight is increased gradually during warm-up:

$$\mathcal{L}_{\mathrm{mdcs}} = \mathcal{L}_{\mathrm{ce}} + \lambda_{\mathrm{kl}}(t) \sum_{e=1}^{E} \mathrm{KL}\left(\sigma(\tilde{z}_{\mathrm{tea}}^{(e)}) \middle\| \sigma(\tilde{z}_{\mathrm{stu}}^{(e)})\right), \qquad \lambda_{\mathrm{kl}}(t) = \lambda_{\max} \min\left(\frac{t}{T_w}, 1\right). \tag{3}$$

This loss encourages augmentation-stable expert predictions without propagating incorrect teacher labels.

3.3 Hierarchical class-aware memory

The memory stores compact class evidence rather than all training features. For each expert and class, a prototype tracks the class mean by momentum update,

$$\mu_c^{(e)} \leftarrow \rho \mu_c^{(e)} + (1 - \rho)\bar{h}_c^{(e)}, \tag{4}$$

where $\bar{h}_c^{(e)}$ is the current mini-batch class mean. Instance slots preserve harder or more diverse samples, with class capacity inversely related to frequency, $K_c \propto 1/n_c$, so tail classes can retain more evidence.

A candidate feature is written only if it is difficult or adds diversity. With classification loss ℓ_i and nearest same-class memory distance d_i, the memory first computes a shallow–deep distance and then applies the write gate:

$$d(h, m) = \frac{\lambda_{\mathrm{s}}(1 - \cos(h_{\mathrm{s}}, m_{\mathrm{s}})) + \lambda_{\mathrm{d}}(1 - \cos(h_{\mathrm{d}}, m_{\mathrm{d}}))}{\lambda_{\mathrm{s}} + \lambda_{\mathrm{d}}}, \qquad \ell_i \geq \tau_{\mathrm{hard}} \quad \text{or} \quad d_i \geq \tau_{\mathrm{div}}. \tag{5}$$

The distance term measures whether a feature is redundant with existing same-class evidence, while the loss term preserves examples near difficult decision regions. When a class memory is full, MemTailor replaces the slot with the lowest retention score, computed from discriminability, boundary usefulness, diversity, and freshness. This bounded policy keeps the memory class-conditioned and informative instead of turning it into a recent-sample cache.

3.4 Memory-guided training augmentation

Once memory is populated, MemTailor retrieves only same-class evidence to enhance rare-class training features. For sample (x_i, y_i) and expert e, slots in $\mathcal{M}_{y_i}^{(e)}$ are ranked by an uncertainty-aware attention score, and the top-ranked slots are mixed with the class prototype:

$$a_{ij}^{(e)} = \frac{(h_i^{(e)})^\top m_j^{(e)}}{\sqrt{d}}(1 - u_j), \qquad r_i^{(e)} = \lambda_{\mathrm{p}} \mu_{y_i}^{(e)} + \lambda_{\mathrm{slot}} \sum_{j \in \mathrm{TopK}} \omega_{ij}^{(e)} m_j^{(e)}. \tag{6}$$

Here u_j downweights uncertain slots, and $\omega_{ij}^{(e)}$ is the softmax-normalized attention weight over the retrieved slots. The resulting memory feature is added with a class-adaptive strength,

$$\hat{h}_i^{(e)} = h_i^{(e)} + \eta_{y_i} r_i^{(e)}, \qquad \mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{mdcs}} + \lambda_{\mathrm{aug}} \mathcal{L}_{\mathrm{aug}}. \tag{7}$$

Here η_{y_i} is larger for tail classes and is zero for classes above the median frequency. Because retrieval is restricted to the ground-truth class, the augmentation strengthens rare-class evidence without mixing labels across classes.

3.5 Uncertainty-guided inference and test-time fusion

At inference time, memory acts as a quality-gated evidence source. For query feature h from expert e, the memory score for class c combines prototype and slot similarity. This raw score is then rescaled by memory quality before it can influence the prediction:

$$s_c^{(e)}(h) = \lambda_{\mathrm{p}} s_{c,\mathrm{proto}}^{(e)}(h) + \lambda_{\mathrm{slot}} s_{c,\mathrm{slot}}^{(e)}(h), \qquad \tilde{s}_c^{(e)}(h) = q_c^{(e)} s_c^{(e)}(h), \tag{8}$$

where $q_c^{(e)}$ summarizes fill ratio, average uncertainty, and prototype validity. This gate suppresses invalid or low-quality class memories before they affect the logits.

The inference rule has three parts. A deep-memory score first corrects each expert logit, entropy of the corrected ensemble determines whether the sample is uncertain, and only uncertain samples receive an additional shallow–deep correction:

$$\hat{z}^{(e)} = z^{(e)} + \gamma^{(e)} \tilde{s}_{\mathrm{d}}^{(e)}, \tag{9}$$

$$u_{\mathrm{guide}} = H\left(\frac{1}{E}\sum_e p^{(e)}\right), \tag{10}$$

$$z_{\mathrm{final}}^{(e)} = \hat{z}^{(e)} + \mathbf{1}[u_{\mathrm{guide}} > \tau_u] \gamma_2^{(e)} \left(\beta_{\mathrm{s}} \tilde{s}_{\mathrm{s}}^{(e)} + \beta_{\mathrm{d}} \tilde{s}_{\mathrm{d}}^{(e)}\right), \tag{11}$$

where $p^{(e)} = \sigma(\hat{z}^{(e)})$. The final logits for one view are averaged across experts, $z_{\mathrm{final}} = E^{-1} \sum_e z_{\mathrm{final}}^{(e)}$. This rule keeps confident examples close to the backbone prediction and reserves the richer memory consultation for uncertain examples.

MemTailor makes the final prediction from multiple input scales rather than from a single resized view. Each scale s_m is passed through the same memory-guided inference rule to obtain post-memory logits z_{s_m}, and the logits are fused before the final softmax decision:

$$z^* = \sum_{m=1}^{M} w_m z_{s_m}, \qquad \sum_{m=1}^{M} w_m = 1, \qquad \hat{y} = \arg\max_c \sigma(z^*)_c. \tag{12}$$

This multi-scale inference step is part of the prediction pipeline: memory correction is first applied at each scale, and the resulting logits are then combined to make the final decision.

4 Experiments

4.1 Experimental setup

Datasets. We evaluate MemTailor on four standard long-tailed recognition benchmarks that cover both manually constructed and naturally imbalanced class distributions. CIFAR-100-LT [Cui et al., 2019] is built from CIFAR-100 with exponential imbalance, and we use imbalance factors 100, 50, and 10. ImageNet-LT and Places-LT [Liu et al., 2019] provide large-scale long-tailed recognition settings with 1,000 and 365 classes, respectively. iNaturalist 2018 [Van Horn et al., 2018] is a fine-grained natural long-tailed dataset with 8,142 species categories. Across datasets, the official validation or test split is used for evaluation, while the training split provides the class-frequency statistics used by the expert loss, memory allocation, and shot-based analysis.

Baselines. We compare MemTailor with representative long-tailed recognition methods spanning re-balancing, margin-based learning, classifier calibration, decoupled representation learning, contrastive learning, and multi-expert learning. The compared methods include LDAM-DRW [Cao et al., 2019b], Balanced Softmax [Ren et al., 2020a], LADE [Hong et al., 2021], MiSLAS [Zhong et al., 2021], RIDE [Wang et al., 2021], SADE [Zhang et al., 2022], BalPoE [Sanchez Aimar et al., 2023], MDCS [Zhao et al., 2023], BCL [Zhu et al., 2022], PaCo [Cui et al., 2021], NCL [Li et al., 2022a], LSC [Wei et al., 2024], PRL [Zhao et al., 2024], ProCo [Du et al., 2024], ConCutMix [Pan et al., 2024], LOS [Sun et al., 2025], and FeatRecon [Yi et al., 2025]. The comparison is centered on methods that report results under the same benchmark protocols and backbone families. When available, we use numbers from the original papers; otherwise, we reproduce results with public implementations under the corresponding dataset and backbone setting. Reproduced entries are marked separately in the result tables to distinguish them from reported results.

Implementation and evaluation details. We use standard backbones for each benchmark: ResNet-32 for CIFAR-100-LT, ResNet-50 for ImageNet-LT and iNaturalist 2018, and ResNet-152 for Places-LT. MemTailor uses three experts; multi-expert baselines follow their corresponding expert configurations, while the ablation includes a separate single-model CE baseline. All models are trained with SGD under the training schedules commonly used for the corresponding datasets. We report top-1 accuracy on the official evaluation split of each benchmark, together with many-, medium-, and few-shot accuracy based on the number of training samples per class. The shot groups follow the standard thresholds: many-shot classes have more than 100 samples, medium-shot classes have 20 to 100 samples, and few-shot classes have fewer than 20 samples.

4.2 Main results

Overall comparison across benchmarks. Table 1 shows that MemTailor improves over strong long-tailed baselines across all four benchmarks. Under the standard training schedule, MemTailor reaches the best overall accuracy on CIFAR-100-LT with IR=50 and IR=100, on ImageNet-LT, on iNaturalist 2018, and on Places-LT, while remaining competitive on the milder CIFAR-100-LT IR=10 setting. The gains are especially consistent against recent multi-expert competitors such as GUIDE, indicating that external class-aware memory contributes information that is not recovered by expert specialization alone.

The longer-training setting shows the same trend. MemTailor achieves the best performance on CIFAR-100-LT with IR=10 and IR=100, ImageNet-LT, iNaturalist 2018, and Places-LT, while remaining within 0.4 points of the best method on CIFAR-100-LT IR=50. This pattern is important

Table 1: Comparison of top-1 accuracy (%) with state-of-the-art methods. Best results are in **bold**, and second-best results are underlined.

Method	CIFAR-100-LT			ImageNet-LT	iNaturalist 2018	Places-LT
	IR=10	IR=50	IR=100			
Softmax	59.1	45.6	41.4	48.0	64.7	31.4
LDAM-DRW [Cao et al., 2019b]	58.7	46.6	42.0	48.8	64.6	36.8
Balanced Softmax [Ren et al., 2020a]	61.0	50.9	46.1	52.3	70.6	39.4
LADE [Hong et al., 2021]	61.6	50.1	45.6	52.3	69.3	39.2
MiSLAS [Zhong et al., 2021]	62.5	51.5	46.8	51.4	70.7	38.3
RIDE (3 experts) [Wang et al., 2021]	61.8	51.7	48.0	56.3	71.8	40.3
SADE [Zhang et al., 2022]	63.6	53.8	48.8	58.8	72.7	40.9
BalPoE [Sanchez Aimar et al., 2023]	64.8	56.3	52.0	59.3	75.0	40.8
MDCS [Zhao et al., 2023]	–	57.2	53.2	59.3	72.5	42.4
LSC [Wei et al., 2024]	65.0	56.5	51.8	60.2	73.9	41.3
BCL [Zhu et al., 2022]	64.9	56.6	51.9	56.0	71.8	–
PRL [Zhao et al., 2024]	65.6	57.3	52.8	60.8	75.1	41.6
ProCo [Du et al., 2024]	65.5	57.1	52.8	57.3	73.5	–
ConCutMix [Pan et al., 2024]	64.5	57.4	53.2	58.5	72.1	–
LOS [Sun et al., 2025]	**69.7**	58.8	54.9	54.4	70.8	–
FeatRecon [Yi et al., 2025]	65.3	57.0	52.5	56.8	72.9	41.3
GUIDE	69.2	60.3	56.4	62.5	76.1	42.2
MemTailor	69.5	**61.2**	**57.5**	**64.0**	**78.6**	**43.4**
Longer training						
PaCo [Cui et al., 2021]	64.2	56.0	52.0	57.0	73.2	–
SADE [Zhang et al., 2022]	65.3	57.3	52.2	–	74.0	39.5
NCL [Li et al., 2022a]	–	58.2	54.2	59.5	76.9	–
BCL [Zhu et al., 2022]	66.9	58.6	53.9	59.0	74.8	–
BalPoE [Sanchez Aimar et al., 2023]	68.1	60.1	55.9	59.7	76.9	41.1
MDCS [Zhao et al., 2023]	–	60.1	56.1	60.7	75.6	–
ProCo [Du et al., 2024]	67.8	58.9	54.2	58.5	75.6	–
ConCutMix [Pan et al., 2024]	68.3	59.5	55.1	58.9	75.7	–
FeatRecon [Yi et al., 2025]	67.2	58.5	54.0	59.3	75.9	–
GUIDE	70.2	**62.3**	57.7	63.4	77.8	43.1
MemTailor	**70.8**	61.9	**58.4**	**64.6**	**78.9**	**44.0**

Table 2: Detailed performance breakdown (%) on CIFAR-100-LT with ResNet-32.

Method	CIFAR-100-LT (IR=100)				CIFAR-100-LT (IR=50)			
	Many	Medium	Few	Overall	Many	Medium	Few	Overall
LDAM-DRW [Cao et al., 2019b]	61.2	39.5	14.4	42.0	64.9	45.1	28.7	46.6
RIDE (3 experts) [Wang et al., 2021]	64.7	46.5	22.3	48.0	67.5	50.8	27.9	51.7
SADE [Zhang et al., 2022]	63.4	49.2	25.1	49.8	67.2	53.5	31.8	53.9
BalPoE [Sanchez Aimar et al., 2023]	65.3	51.1	28.0	52.0	69.2	55.9	34.5	56.3
PRL [Zhao et al., 2024]	68.7	55.3	31.2	52.8	67.5	53.5	39.0	56.6
GUIDE	71.3	**59.1**	36.0	56.4	70.2	56.1	47.4	60.3
MemTailor	**72.1**	58.8	**38.9**	**57.5**	**71.0**	**56.3**	**50.2**	**61.2**
Longer training								
GUIDE (longer)	**74.4**	60.0	36.1	57.7	**72.4**	57.5	50.1	62.3
MemTailor (longer)	74.1	**60.2**	**39.5**	**58.8**	71.6	**58.5**	**51.2**	**62.6**

because it suggests that the improvement is not only a short-schedule optimization effect: even when strong baselines are trained longer, the memory-guided mechanism continues to add value. The longer-training CIFAR-100-LT entries in Tables 1 and 2 come from independent runs, so their overall accuracies are not expected to match exactly.

Shot-wise behavior on CIFAR-100-LT. Table 2 shows where the gains come from. On CIFAR-100-LT with IR=100, MemTailor improves the overall accuracy from GUIDE's 56.4 to 57.5 under the standard schedule, and the largest absolute gain appears on few-shot classes, where the accuracy

Table 3: Top-1 accuracy (%) on CIFAR-100-LT (IR=100) under unknown test class distributions. "Prior" indicates whether the method uses the test class prior for post-hoc adjustment.

Method	Prior	Forward-LT					Uni.	Backward-LT				
		50	25	10	5	2	1	2	5	10	25	50
Softmax	×	63.3	62.0	56.2	52.5	46.4	41.4	36.5	30.5	25.8	21.7	17.5
MiSLAS [Zhong et al., 2021]	×	58.8	57.2	55.2	53.0	49.6	46.8	43.6	40.1	37.7	33.9	32.1
LADE [Hong et al., 2021]	×	56.0	55.5	52.8	51.0	48.0	45.6	43.2	40.0	38.3	35.5	34.0
LADE [Hong et al., 2021]	✓	62.6	60.2	55.6	52.7	48.2	45.6	43.8	41.1	41.5	40.7	41.6
RIDE [Wang et al., 2021]	×	63.0	59.9	57.0	53.6	49.4	48.0	42.5	38.1	35.4	31.6	29.2
SADE [Zhang et al., 2022]	×	65.2	62.5	58.8	55.4	51.2	48.8	43.0	43.9	42.4	42.2	42.0
BalPoE [Sanchez Aimar et al., 2023]	×	69.0	65.2	61.2	59.0	54.2	51.7	45.7	46.6	45.2	45.2	45.8
LSC [Wei et al., 2024]	×	67.8	64.2	60.2	58.1	53.2	51.6	44.7	45.7	44.2	44.7	48.0
PRL [Zhao et al., 2024]	×	69.5	65.7	61.7	59.5	54.7	52.2	46.2	47.1	45.7	**45.7**	**48.5**
GUIDE	×	68.5	66.7	64.4	62.6	59.2	56.4	53.3	49.7	46.7	43.2	40.7
GUIDE (longer)	×	71.9	**70.0**	**67.3**	**64.7**	60.8	57.7	54.4	**50.6**	47.7	43.6	41.5
MemTailor	×	71.2	67.8	65.1	63.3	60.5	57.5	55.4	49.5	47.7	45.2	43.0
MemTailor (longer)	×	**72.2**	68.3	67.0	**64.7**	**61.6**	**58.8**	**55.6**	49.7	**48.0**	44.9	44.2

rises from 36.0 to 38.9. Under longer training, MemTailor further improves the few-shot score to 39.5 and the overall score to 58.8. A similar trend holds for IR=50, where MemTailor improves both few-shot and overall performance in both the standard and longer schedules. These results support the central claim of the paper: preserving class-specific evidence outside the backbone is especially useful when the classifier must retain decision support for under-represented classes.

Robustness to test-distribution shift. Table 3 evaluates a stricter setting in which the test class distribution changes from forward long-tail to uniform and backward long-tail regimes. MemTailor remains strong across this spectrum and obtains the best result in six of the eleven evaluation columns in the longer-training setting. The gains are particularly clear around the balanced and mildly backward-shifted regimes, where memory-guided correction can help uncertain predictions without requiring access to the test prior. This behavior is consistent with the intended role of the memory module as an evidence source that is conditioned on class content rather than tied to a single fixed deployment prior.

4.3 Ablation study

Each MemTailor component contributes complementary gains. Table 4 reports the corrected component-wise ablation on CIFAR-100-LT (IR=100) with ResNet-32. The original caption incorrectly identified Places-LT with ResNet-152. The first row is a single-model CE baseline; the second is plain multi-expert CE and does not include DL or CS. Adding DL and then CS raises overall accuracy from 47.3 to 53.2. Comparing +CS with +Memory and Feat-Aug isolates training-time memory augmentation: Many/Medium/Few/Overall accuracy changes by $+0.3/+1.7/+1.3/+1.1$ points, preserving many-shot accuracy while improving medium- and few-shot accuracy.

Memory-guided inference raises the single-scale overall accuracy from 54.3 to 56.3. Multi-scale fusion without memory inference reaches 54.8, while combining memory inference and multi-scale fusion reaches 57.5. Thus, memory correction provides the larger inference gain and multi-scale fusion adds a complementary improvement.

Visualization of memory allocation. Figure 2 visualizes the allocation rule induced by the hierarchical memory. Although head classes dominate the original sample count, the bounded memory budget is intentionally shifted toward medium- and few-shot groups, so the stored evidence does not simply mirror the training imbalance. This behavior is important for the full method: the external memory is designed to over-preserve classes that are most vulnerable to being washed out by head-dominated optimization.

Retrieved-memory diagnostic. We evaluate retrieved-memory quality using the final seed-0 CIFAR-100-LT IR=100 checkpoint. Cross-class memory retrieval reaches 39.07% Top-1 and 74.88% Top-5 overall, and 34.73% Top-1 and 72.80% Top-5 on few-shot classes. Shuffling class assignments

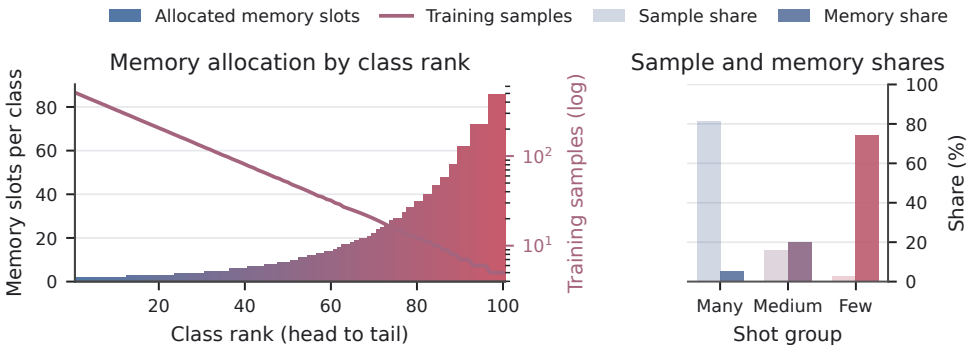


Figure 2: Inverse-frequency memory allocation in MemTailor. Rare classes receive a larger share of bounded memory slots, so the external memory emphasizes evidence preservation where the training distribution is most sparse.

Table 4: Component-wise ablation on CIFAR-100-LT (IR=100) with ResNet-32. ME: multi-expert; DL: decoupled learning; CS: class-specialized objective; Feat-Aug: training-time memory feature augmentation.

Method	Components							Accuracy (%)			
	ME	DL	CS	Memory	Feat-Aug	Memory Inf.	Multi-scale	Many	Med.	Few	All
Single-model CE	–	–	–	–	–	–	–	65.2	40.4	24.6	44.3
Plain multi-expert CE	✓	–	–	–	–	–	–	67.2	44.1	27.8	47.3
+ DL	✓	✓	–	–	–	–	–	68.8	48.4	32.2	50.7
+ CS	✓	✓	✓	–	–	–	–	69.7	52.0	35.5	53.2
+ Memory and Feat-Aug	✓	✓	✓	✓	✓	–	–	70.0	53.7	36.8	54.3
+ Memory Inf., single-scale	✓	✓	✓	✓	✓	✓	–	71.4	57.4	37.4	56.3
+ Multi-scale, no Memory Inf.	✓	✓	✓	✓	✓	–	✓	70.4	54.0	37.5	54.8
MemTailor	✓	✓	✓	✓	✓	✓	✓	**72.1**	**58.8**	**38.9**	**57.5**

in memory reduces few-shot Top-1 retrieval accuracy to 0.17%, indicating that the stored items carry class-discriminative evidence. Relative to the deployable proxy-routing correction path, a mechanism-only ground-truth routing oracle improves medium- and few-shot accuracy by $+4.77$ and $+6.60$ points, respectively. This oracle is not deployable; it indicates the correction potential of stored evidence and points to proxy routing and fusion calibration as remaining limitations.

Discussion and limitations. Taken together, the experiments support a consistent picture. The cross-benchmark results show that MemTailor is competitive at scale, the shot-based analysis shows that the gains are concentrated where long-tailed recognition is most difficult, and the ablation confirms that these improvements are produced by the intended interaction between expert diversity, class-aware storage, training-time retrieval, and selective inference-time correction. This alignment between design and empirical behavior strengthens the claim that external memory addresses a failure mode not solved by decision-level expert diversification alone.

The current study also has a limited scope, although the remaining caveats are moderate. MemTailor introduces extra memory states and a small set of retrieval-related hyperparameters beyond a plain multi-expert baseline, and the experiments focus on standard closed-set long-tailed recognition benchmarks with conventional CNN backbones. We therefore view extension to other backbone families, more aggressive efficiency optimization, and broader deployment settings as natural follow-up directions rather than prerequisites for the core claim validated here.

Figure 3 provides a complementary view of the same behavior. In the early stage, MemTailor tracks the baseline closely while the memory is being populated. Once retrieval-based augmentation and inference-time consultation become useful, the overall gain widens and the few-shot curve improves more sharply than the baseline, matching the shot-wise results in Table 2. The middle panel shows that memory utilization rises quickly while the consult rate remains moderate, consistent with selective memory use. The right panel reports the retrieval diagnostic described above: class-shuffled few-shot Top-1 retrieval falls to 0.17%, compared with 34.73% using the true class assignments.

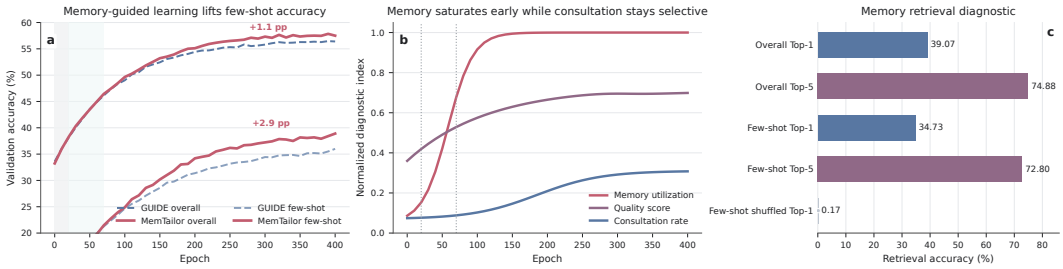


Figure 3: Training dynamics and retrieval diagnostics on CIFAR-100-LT (IR=100). (a) GUIDE and MemTailor validation accuracy for overall and few-shot groups; endpoint annotations show the accuracy gains. (b) Normalized memory utilization, quality score, and consultation rate during training. (c) Top-1 and Top-5 accuracy of cross-class memory retrieval overall and on few-shot classes, with few-shot Top-1 under shuffled class assignments.

5 Conclusion

MemTailor improves long-tailed visual recognition by combining a frequency-aware multi-expert classifier with hierarchical class-aware memory and selective memory-guided refinement. Across CIFAR-100-LT, ImageNet-LT, iNaturalist 2018, and Places-LT, it delivers consistent gains, especially on under-represented classes. These results indicate that explicit evidence retention is a useful complement to existing expert-based long-tailed learning strategies.

References

Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6907, 2022.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/621461af90cadfdaf0e8d4cc25129f91-Paper.pdf`.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019b.

X. Cao and I. W. Tsang. Distribution disagreement via lorentzian focal representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6872–6889, 2022. doi: 10.1109/TPAMI.2021.3093590.

X. Cao, W. Liu, and I. W. Tsang. Data-efficient learning via minimizing hyperspherical energy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13422–13437, 2023. doi: 10.1109/TPAMI.2023.3290544.

X. Cao, Y. Guo, H. T. Shen, I. W. Tsang, and J. T. Kwok. Mentored learning: Improving generalization and convergence of student learner. *Journal of Machine Learning Research*, 25(325):1–45, 2024.

Xiaofeng Cao, Mingwei Xu, Xin Yu, Jiangchao Yao, Wei Ye, Shengjun Huang, Minling Zhang, et al. Analytical survey of learning with low-resource data: From analysis to investigation. *ACM Computing Surveys*, 58(6):1–47, 2025.

Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *European conference on computer vision*, pages 95–110. Springer, 2020.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 715–724, 2021.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6626–6636, 2021.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition, 2020. URL `https://arxiv.org/abs/1910.09217`.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6949–6958, 2022a.

Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6949–6958, 2022b.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019.

Haolin Pan, Yong Guo, Mianjie Yu, and Jian Chen. Enhanced long-tailed recognition with contrastive CutMix augmentation. *IEEE Transactions on Image Processing*, 2024.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a.

Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4175–4186. Curran Associates, Inc., 2020b. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/2ba61cc3a8f44143e1f2f13b2b729ab3-Paper.pdf`.

Emanuel Sanchez Aimar, Arvi Jonnarth, Michael Felsberg, and Marco Kuhlmann. Balanced product of calibrated experts for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19967–19977, 2023.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

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 (ICLR)*, 2025.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Tong Wei, Zhen Mao, Zi-Hao Zhou, Yuanyu Wan, and Min-Ling Zhang. Learning label shift correction for test-agnostic long-tailed recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 162–178, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58548-8.

Lingjie Yi, Michael Yao, Weimin Lyu, Haibin Ling, Raphael Douady, and Chao Chen. Geometry of long-tailed representation learning: Rebalancing features for skewed distributions. In *International Conference on Learning Representations (ICLR)*, 2025.

Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yongshun Zhang, Xiu-Shen Wei, Boyan Zhou, and Jianxin Wu. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3447–3455, May 2021. doi: 10.1609/aaai.v35i4.16458. URL `https://ojs.aaai.org/index.php/AAAI/article/view/16458`.

Qihao Zhao, Chen Jiang, Wei Hu, Fan Zhang, and Jun Liu. MDCS: More diverse experts with consistency self-distillation for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11597–11608, 2023.

Zhe Zhao, Haibin Wen, Zikang Wang, Pengkun Wang, Fanfu Wang, Song Lai, Qingfu Zhang, and Yang Wang. Breaking long-tailed learning bottlenecks: A controllable paradigm with hypernetwork-generated diverse experts. 2024.

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

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6908–6917, 2022.

A Technical appendices and supplementary material

This appendix summarizes the implementation details that support the main results and clarifies the scope of the current submission.

A.1 Reproducibility and implementation details

MemTailor uses three experts, while multi-expert baselines follow their corresponding expert configurations; the ablation includes a separate single-model CE baseline. All experiments use the standard backbone family for each benchmark: ResNet-32 for CIFAR-100-LT, ResNet-50 for ImageNet-LT and iNaturalist 2018, and ResNet-152 for Places-LT. Training uses SGD with momentum 0.9, Nesterov acceleration, mixed precision, and validation-based model selection. The default memory design uses a class prototype plus bounded instance slots per expert, prototype momentum 0.9, top-k retrieval with $k = 4$ on CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018 and $k = 6$ on Places-LT, prototype weight 0.3, slot weight 0.7, and training-time auxiliary memory augmentation weight 0.5 after warm-up.

For CIFAR-100-LT, we use one GPU, batch size 256, initial learning rate 0.1, weight decay 5×10^{-4}, and a 400-epoch schedule with decay milestones at epochs 150, 320, and 380 after a 20-epoch warm-up. For ImageNet-LT, we use four GPUs, batch size 512, initial learning rate 0.2, weight decay 5×10^{-4}, and cosine decay for 180 epochs. For iNaturalist 2018, we use four GPUs, batch size 512, initial learning rate 0.2, weight decay 2×10^{-4}, and a 400-epoch schedule with warm-up and staged decay. For Places-LT, we use four GPUs, batch size 64, initial learning rate 0.01 for expert-specific layers and 0.001 for shared layers, weight decay 4×10^{-4}, and a 180-epoch schedule. The exact configuration files and run commands corresponding to the reported results are intended to be included in the anonymized supplementary package.

A.2 Limitations

MemTailor improves rare-class retention by introducing an external memory, which adds moderate storage and tuning overhead beyond a plain multi-expert backbone. In practice the method remains compact enough for the benchmark settings studied here, but streamlining the retrieval path further would still be valuable for highly constrained deployments.

Our experiments focus on standard closed-set long-tailed recognition benchmarks and conventional CNN backbones. The reported results therefore validate the main mechanism in a strong benchmark setting, while leaving broader studies on alternative backbones, larger-scale efficiency optimization, and additional deployment scenarios to future work.

A.3 Broader impacts

Positive impacts include improved recognition of under-represented classes in scientific categorization, biodiversity monitoring, and other applications where rare categories are important and under-served by standard imbalanced learning. Better long-tailed recognition can also reduce the tendency of models to over-prioritize frequent categories in downstream decision pipelines.

Potential negative impacts include over-trust in automated predictions on rare categories, especially when a benchmark gain is mistaken for deployment-level reliability. Like other visual recognition systems, long-tailed classifiers may also inherit dataset bias and may be used in settings where mistakes disproportionately affect minority categories. For this reason, MemTailor should be used with dataset-appropriate validation, uncertainty-aware monitoring, and human oversight in high-stakes applications.

A.4 Assets and release scope

The paper uses public long-tailed recognition benchmarks and cites the original dataset papers for CIFAR-100-LT, ImageNet-LT, iNaturalist 2018, and Places-LT, together with the original papers for the compared baseline methods. The paper does not release a new public dataset. Code and configuration files will be made available to support reproducibility.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope, method, and intended gains, and these claims are matched to the method description and benchmark results in Sections 1, 3, and 4.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses computational overhead, memory-quality dependence, benchmark scope, and robustness limits in Appendix A.2.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: The paper is empirical and does not present theorem-level theoretical results requiring formal proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The architecture, training objective, evaluation protocol, backbone choices, and dataset-specific optimization settings are described in Sections 3–4 and Appendix A.1, which provides the key details needed to reproduce the reported results.

Guidelines:

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

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

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

   Answer: [No]

   Justification: The paper does not currently provide an open public code or data package, although the experiments use public benchmarks and Appendix A.1 states the intended supplementary release scope for configurations and commands.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Sections 3–4 describe the model, losses, datasets, and evaluation protocol, and Appendix A.1 summarizes dataset-specific schedules, batch sizes, optimizers, and memory-related settings.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: The current submission reports single-run benchmark accuracies on standard splits and does not include error bars or formal significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: Appendix A.1 reports the GPU counts and core training settings, but the current submission does not yet provide a complete accounting of worker type, memory size, and per-run wall-clock cost for every experiment.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work uses standard benchmark datasets, does not involve human subjects, and the submission preserves anonymity while discussing limitations and responsible use considerations in Appendix A.2–A.4.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

Justification: Appendix A.3 discusses both beneficial uses of improved rare-class recognition and possible risks from deployment, dataset bias, and over-trust in automated predictions.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper does not release a high-risk generative model, scraped dataset, or similarly misuse-sensitive asset.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [No]

Justification: The paper cites the original datasets and baseline methods, but the current submission does not yet provide a complete per-asset list of license names or URLs for every external dataset and codebase.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

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

13. **New assets**

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

    Answer: [N/A]

    Justification: The current submission does not release a new public dataset, model, or code package as a standalone asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing experiments or human-subject data collection.

    Guidelines:

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

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

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

    Answer: [N/A]

Justification: The work does not involve human subjects or crowdsourcing and therefore does not require IRB-style review in the scope of this submission.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs are not part of the core method, experiments, or scientific claims of this work.

Guidelines:

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