HGLTR: Hierarchical Knowledge Injection for Calibrating Pre-trained Models in Long-Tail Recognition

Jinpeng Zheng[1,2], Shao-Yuan Li[1,2 3*], Gan Xu[4], Wenhai Wan[5], Zijian Tao[1,2], Songcan Chen[1,2], Kangkan Wang[6*]

[1]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
[2]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
[3]State Key Lab. for Novel Software Technology, Nanjing University
[4]College of Information Engineering, Zhejiang University of Technology
[5]School of Computer Science and Technology, Huazhong University of Science and Technology
[6]School of Computer Science and Engineering, Nanjing University of Science and Technology

Abstract

Long-tail recognition remains challenging for pre-trained foundation models like CLIP, which often suffer from performance degradation under imbalanced data. This stems not only from the overfitting/underfitting issues during fine-tuning but, more fundamentally, from the inherent bias inherited from the long-tail distribution of their massive pre-training datasets. To address this, we propose **HGLTR** (Hierarchy-Guided Long-Tail Recognition), a novel framework that calibrates pre-trained models by injecting objective class hierarchy knowledge. We argue that the semantic proximity defined by a hierarchy provides a robust, data-independent prior to counteract model bias. Our method is specifically designed for vision-language models' dual-modality architecture. At the feature level, we align image embeddings with a hierarchy-guided text similarity structure. At the classifier level, we employ a distillation loss to regularize predictions using soft labels derived from the hierarchy. This dual-level injection effectively transfers knowledge from head to tail classes. Experiments on ImageNet-LT, Places-LT, and iNaturalist 2018 demonstrate that HGLTR achieves state-of-the-art performance, particularly in tail-classes accuracy, highlighting the importance of leveraging structural priors to calibrate foundation models for real-world data.

Introduction

Recent years have witnessed revolutionary progress in large-scale self-supervised pre-trained models, particularly vision-language foundation models. Among these, CLIP (Radford et al. 2021) has emerged as a dominant framework by learning powerful cross-modal semantic alignment through training on massive internet-scale image-text pairs. Its core approach maps images and text into a shared embedding space where semantically related pairs are pulled closer, endowing CLIP with remarkable zero-shot classification capabilities and strong generalization that enables competitive results on various downstream tasks without fine-tuning. To adapt these powerful pre-trained models to specific downstream tasks while balancing performance improvement with computational efficiency, Parameter-Efficient Fine-Tuning (PEFT) techniques (Houlsby et al. 2019; Yu et al. 2023) have become mainstream and effectively enhanced model performance across diverse applications.

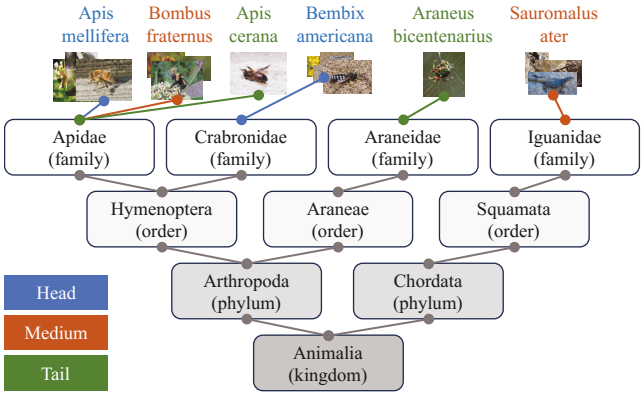


Figure 1: Using the hierarchical structure from the iNaturalist 2018, we illustrate the hierarchical semantic relationships among different distribution categories.

Despite the impressive performance, pre-trained models often degrade significantly when applied to long-tail recognition tasks, where head classes dominate training samples while tail classes suffer from severe scarcity. Recent studies reveal critical insights: Wang et al. (2024) explored various tuning and distribution alignment strategies, Shi et al. (2024b) showed that aggressive fine-tuning harms generalization in long-tail settings, and even parameter-efficient methods struggle with extreme imbalance (Shi et al. 2024c).

We argue that the poor performance of pre-trained models in long-tail recognition stems from the interplay between inherent data bias from pre-training and downstream fine-tuning inadequacy. First, from the fine-tuning perspective, the observations in literature (Shi et al. 2024b,c; Wang et al. 2024) indicate that, models inevitably overfit to head classes and underfit tail classes—a problem analogous to training from scratch (Kang et al. 2020; Cao et al. 2019; Menon et al. 2021). More critically, as examined in Zhu et al. (2023),

*Corresponding author.

Methods	Learnable Params.	Overall	Tail
CLIP (Zero-Shot)	0.00M	67.2	66.6
ProCo(Du et al. 2024)	23.51M	57.8	38.1
BALLAD (Ma et al. 2021)	149.62M	75.2	69.8
HGLTR (Orus)	1.61M	**77.4**	**72.1**

Table 1: Performance (Top-1 Accuracy%) on ImageNet-LT.

a fundamental and intrinsic cause lies in the inherent bias of the pre-trained model itself. The datasets used to train models like CLIP (e.g., web-scraped image-text pairs) intrinsically follow a long-tail distribution. During the self-supervised learning process, the model learns the statistical regularities of this data distribution, causing its feature extractor and zero-shot classifier to implicitly favor categories that were frequent during pre-training. This model bias, induced by the pre-training data distribution, is transferred to downstream tasks. When the downstream task also exhibits a long-tail distribution, this bias is amplified, making the model more prone to predicting head classes and thereby exacerbating the difficulty of recognizing tail classes.

To counteract this interplay bias, we propose leveraging semantic class hierarchies as an objective, data-independent prior. As shown in Fig. 1, hierarchical relationships (e.g., biological taxonomy) define semantic proximity between categories (e.g., "Apis cerana" is closer to "Apis mellifera" than to "Sauromalus ater"), providing robust structural knowledge unaffected by sample distribution. Crucially, this hierarchy enables knowledge in the pre-trained model to transfer from well-represented head classes to semantically related tail classes, effectively mitigating the representation collapse that typically plagues tail categories.

To this end, we introduce Hierarchy-Guided Long-Tail Recognition (HGLTR), a novel framework that calibrates pre-trained models through dual-level hierarchical injection. At the feature level, we align image embeddings with a hierarchy-guided similarity structure to organize the visual feature space according to semantic relationships. At the classifier level, we employ knowledge distillation with soft labels derived from the hierarchy to regularize predictions, ensuring that classification decisions adhere to the class taxonomy. Built upon PEFT techniques, HGLTR effectively injects objective hierarchical knowledge into the pre-trained model, calibrating it for better performance on long-tail data.

Notably, while prior works (Landrieu and Garnot 2021; Bertinetto et al. 2020; Shi et al. 2024a) were explored using hierarchies, they were typically designed not for imbalance learning. Besides, optimized for small single-modal networks trained from scratch, their implementation fail to fully exploit the rich cross-modal information already present in pre-trained models like CLIP. In contrast, the proposed HGLTR approach is tailored for a pre-trained dual-branch vision/text architecture. Table 1 indicates that HGLTR attains superior performance at a low computational cost. Specifically, it requires only 1.61M learnable parameters in the backbone and 10 epochs to converge, without sacrificing the original model's proficiency.

In summary, our contributions are threefold:

• We establish that pre-training data bias, not just fine-tuning limitations, is the fundamental cause of long-tail performance degradation.

• We design HGLTR, the first framework specifically tailored for injecting hierarchical knowledge into CLIP-like models through dual-level calibration.

• Extensive experiments demonstrate state-of-the-art (SOTA) results, with particularly significant gains in tail-class accuracy across multiple benchmarks, validating our approach's effectiveness in calibrating foundation models for real-world imbalanced data.

Related Works

Vision-Language Models (VLMs) VLMs, such as CLIP (Radford et al. 2021), BLIP (Li et al. 2022a), and CoCa (Yu et al. 2022), align visual and textual modalities by learning a shared representation space through joint architectures. Pre-trained on large-scale image-text pairs, they exhibit strong generalization and have been applied to tasks like noisy-label detection (Feng, Tzimiropoulos, and Patras 2024), anomaly analysis (Gu et al. 2024), and long-tail recognition (Shi et al. 2024c). In this work, we adopt CLIP as the backbone for long-tail learning. Directly using CLIP often suffers from domain gaps, so PEFT is commonly employed. Typical PEFT strategies include inserting low-rank adapters (LoRA) into Transformer blocks (Yu et al. 2023), attaching lightweight adapters after linear layers (Houlsby et al. 2019), or employing learnable prompts (Jia et al. 2022). These methods adapt CLIP effectively with minimal trainable parameters.

Long-Tail Recognition For models trained from scratch, long-tail recognition methods mainly fall into two groups: (1) data re-sampling, which balances the distribution by oversampling tail classes with strong augmentations to reduce overfitting (Kang et al. 2020; Shi et al. 2023; Zhong et al. 2021). (2) loss function engineering, which re-weights or redesigns objectives to emphasize tail classes (Cao et al. 2019; Menon et al. 2021). More advanced techniques, such as multi-expert learning, employ multiple classifiers or tailored losses to further improve robustness on imbalanced data (Zhao et al. 2024; Zhang et al. 2022; Wang et al. 2021).

Long-Tail Recognition via Pre-trained Models Recent works adapt pre-trained models to long-tail distributions by designing lightweight adaptation modules. BALLAD (Ma et al. 2021) fine-tunes on the target data and applies a linear adapter on balanced samples to strengthen tail-class. LIFT (Shi et al. 2024b) constrains adapter parameters during fine-tuning to mitigate overfitting. Candle (Shi et al. 2024c) enhances VLMs by combining logit-adjusted loss with cross-modal feature fusion to generalize new classes. Wang et al. (2024) appends a lightweight decoder to frozen visual encoders and integrates imbalanced losses. LPT (Dong et al. 2023) employs a shared prompt for domain alignment alongside group-specific prompts for fine-grained discrimination. However, they mostly overlook biases inherent in the pre-training data, which can degrade downstream performance. To address this issue, we introduce hierarchical information to calibrate pre-trained models.

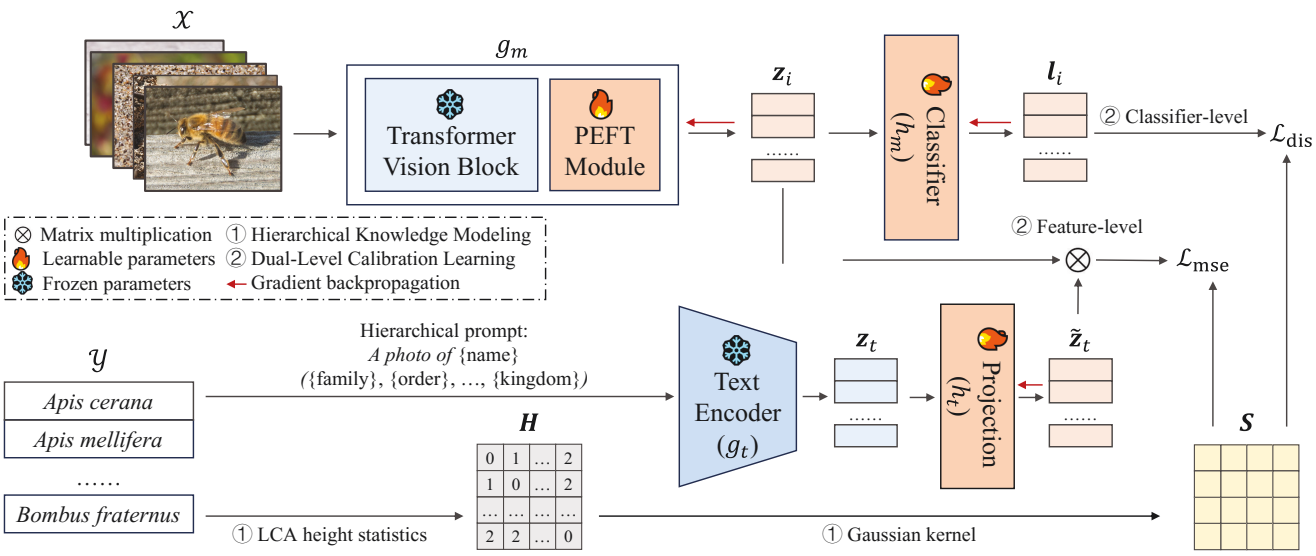


Figure 2: HGLTR's two-stage framework: ① Hierarchical Knowledge Modeling constructs the semantic blueprint by computing LCA-based distance matrix $\boldsymbol{H}$ from class hierarchy and converting it to hierarchical similarity matrix $\boldsymbol{S}$; ② Dual-Level Calibration Learning fine-tunes the model: at the feature level, image features $\boldsymbol{z}_i$ from frozen CLIP backbone with PEFT modules are aligned with $\boldsymbol{S}$ using $\mathcal{L}_{\text{mse}}$; at the classifier level, predictions are regularized with soft labels from $\boldsymbol{S}$ via $\mathcal{L}_{\text{dis}}$.

Preliminary

We address the problem of long-tail image recognition using pre-trained VLMs such as CLIP. Given a training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ denotes an image and $y_i \in \mathcal{Y} = \{1, 2, \ldots, K\}$ is its class label, the samples size of class k is N_k. The dataset exhibits a long-tail distribution: a small number of "head" classes contain the majority of samples, while the vast majority of "tail" classes are severely underrepresented. Our goal is to fine-tune a pre-trained VLM to achieve high accuracy, particularly on the tail classes, while maintaining overall performance.

As discussed in Introduction, directly fine-tuning VLMs on long-tail data is suboptimal due to dual challenges: (1) the fine-tuning process itself is prone to overfitting head classes and underfitting tail classes, and (2) more fundamentally, the model inherits a bias from the long-tail distribution of its pre-training data, which is amplified under downstream long-tail conditions. To overcome this, we propose to leverage an objective class hierarchy $\boldsymbol{H}$ (e.g., a biological taxonomy) as a structural prior. This hierarchy defines the semantic relatedness between classes (e.g., the Lowest Common Ancestor (LCA) height), providing a data-independent source of knowledge. By using $\boldsymbol{H}$, we can guide the model to transfer knowledge from well-represented head classes to their semantically related tail classes, thereby mitigating the negative effects of data imbalance and pre-training bias.

Methodology

Overall Framework

Based on the above insights, we propose HGLTR, a novel framework that calibrates a pre-trained VLM by injecting hierarchical knowledge at two levels. As illustrated in Fig. 2, HGLTR consists of two key stages:

1. Hierarchical Knowledge Modeling We first construct an "ideal" semantic blueprint from the class hierarchy $\boldsymbol{H}$. This is done by computing a hierarchical similarity matrix $\boldsymbol{S}$ based solely on the predefined hierarchy structure $\boldsymbol{H}$, independent of any pre-trained model's feature space.

2. Dual-Level Calibration Learning We fine-tune the VLM by optimizing a combined objective that calibrates the model towards the ideal blueprint $\boldsymbol{S}$ at two levels:

- *Feature-Level Calibration* aligns visual features extracted from images with the hierarchical similarity structure $\boldsymbol{S}$.
- *Classifier-Level Calibration* regularizes the model's predictions to conform to soft labels derived from $\boldsymbol{S}$.

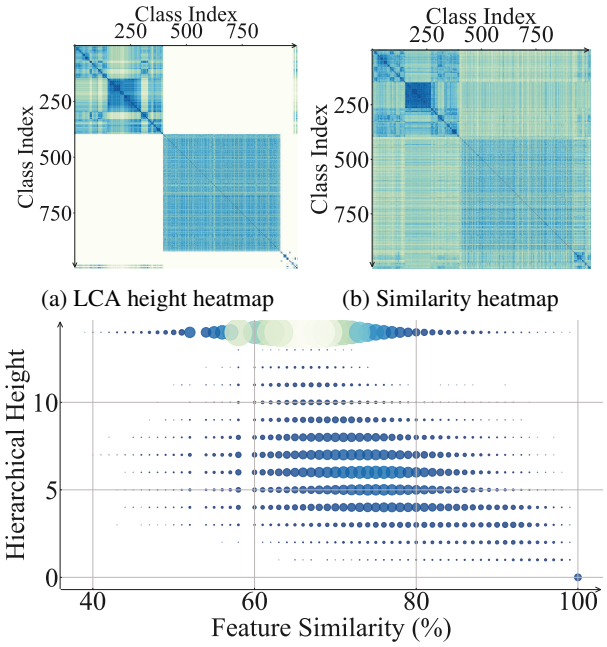
The dual-level approach ensures that both the learned representations and the final predictions are consistent with the objective semantic structure, effectively adapting the pre-trained model for long-tail recognition.

Hierarchical Knowledge Modeling

We begin by constructing the hierarchical similarity matrix $\boldsymbol{S} \in \mathbb{R}^{K \times K}$ based on the predefined class hierarchy $\boldsymbol{H} \in \mathbb{R}^{K \times K}$. We adopt the LCA height (Liang and Davis 2023) as the distance for semantic proximity. The distance between two classes k_1 and k_2, denoted as H_{k_1,k_2}, is defined as the height of their LCA node in the tree, with leaf nodes assigned a height of zero. To transform these distances into a similarity matrix, we apply a Gaussian kernel:

$$S_{k_1,k_2} = \frac{\exp(-H_{k_1,k_2} \cdot \tau_s)}{\sum_{k=1}^{K} \exp(-H_{k_1,k} \cdot \tau_s)}, \tag{1}$$

where τ_s is a hyperparameter that controls the sensitivity of similarity scores to hierarchical distances. A smaller τ_s



(a) LCA height heatmap (b) Similarity heatmap

(c) Relationship between LCA height and similarity

Figure 3: Hierarchical structure and feature space analysis on ImageNet. (a) LCA height heatmap: darker colors indicate smaller semantic distance. (b) Feature similarity heatmap: darker colors indicate higher cosine similarity between class prototypes. (c) Scatter plot showing a negative correlation between LCA height (y-axis) and feature similarity (x-axis); point size denotes sample pair counts.

yields a flatter similarity distribution, while larger values lead to sharper distinctions. In the limit as $\tau_s \to \infty$, $\boldsymbol{S}_k$ approaches a one-hot vector that highlights the closest class; as $\tau_s \to 0$, the distribution becomes uniform.

Given that a significant number of class pairs in the dataset are semantically unrelated—having the root node as their LCA — it is reasonable to define their hierarchical distance as ∞. This assignment, in accordance with Eq. 1, results in a similarity score of zero, correctly preventing any attribution of semantic similarity.

To validate the rationale of our way of constructing $\boldsymbol{S}$, we conduct a preliminary analysis to investigate the relationship between the objective class hierarchy and the feature space learned by a pre-trained model. Specifically, we utilize the frozen, pre-trained CLIP model (ViT-B/16) and extract image features for all training samples from the ImageNet-1K dataset. We then compute the class prototype (i.e., the mean feature vector) for each of the 1,000 ImageNet classes. The pairwise cosine similarity between all class prototypes is calculated to form the feature similarity matrix.

The class hierarchy $\boldsymbol{H}$ is derived from WordNet (Miller 1995), which provides the taxonomic structure for ImageNet classes. As shown in Fig. 3, we compare the LCA-based hierarchical distance (Fig. 3a) and the feature similarity from CLIP prototypes (Fig. 3b). A striking visual correspondence is observed: classes that are semantically closer in the hierarchy (lower LCA height, darker in Fig. 3a) exhibit higher feature similarity (darker in Fig. 3b). To quantify this relationship, we plot the LCA height against the feature similarity for all class pairs in Fig. 3c. The clear negative correlation in the scatter plot demonstrates that classes with smaller semantic distances in the hierarchy tend to have more similar representations in the pre-trained model's embedding space.

This key observation provides strong empirical support for our approach. It indicates that the embedding space learned by pre-trained VLMs like CLIP is inherently organized according to the underlying semantic structure defined by the class hierarchy. Therefore, leveraging this objective structural prior $\boldsymbol{S}$ to calibrate the model is a principled strategy. It guides the model to maintain and refine this desirable semantic organization during fine-tuning on long-tail data, effectively counteracting the performance degradation caused by data imbalance.

Dual-Level Calibration Learning

Building on the constructed hierarchical similarity matrix $\boldsymbol{S}$, we introduce a dual-level calibration that leverages $\boldsymbol{S}$ to counteract model bias from long-tail pre-training. At the feature level, we align image embeddings with $\boldsymbol{S}$ to reorganize the visual space according to true semantic relationships, preventing tail classes from being pulled toward head classes. At the classifier level, we use knowledge distillation with soft labels derived from $\boldsymbol{S}$ to guide predictions. These two levels work synergistically, effectively transferring knowledge from head to tail classes while maintaining balanced performance across all categories.

Feature-Level Calibration Building upon the hierarchical similarity matrix $\boldsymbol{S}$, our feature-level calibration reorganizes the visual feature space according to the objective semantic structure. This is critical because pre-trained models often exhibit feature representations biased by the long-tail distribution of their pre-training data.

For a direct and fair comparison with SOTA methods (Shi et al. 2024b; Wang et al. 2024; Ma et al. 2021), we adopt the pre-trained CLIP model (Radford et al. 2021) as our feature backbone, keeping the main architecture frozen while integrating Adapter modules (Houlsby et al. 2019) into each Transformer block for parameter-efficient adaptation. Preliminary experiments in Shi et al. (2024b) confirmed that a simple linear classifier on frozen features yields suboptimal performance, as the limited adaptability fails to address domain-specific nuances in long-tail distributions.

For each class, we construct hierarchy-guided textual prompts in the format "*a photo of a* {*class name*}({*family*}, {*order*},..., {*kingdom*})". These prompts are fed into CLIP's frozen text encoder g_t, producing text embeddings $\boldsymbol{z}_t \in \mathbb{R}^{K \times d_t}$, where each row corresponds to one class prototype with feature dimension d_t. Importantly, these prototypes naturally capture the semantic structure imposed by our hierarchical design. For an input image $\boldsymbol{x}_i$, its visual representation is obtained via the CLIP image encoder g_m, yielding an image feature $\boldsymbol{z}_i \in \mathbb{R}^{d_m}$, where d_m denotes the dimension of the image embedding space.

To bridge potential distribution shifts between the text and image feature spaces introduced during fine-tuning, we in-

troduce a learnable text projection head h_t implemented as a linear adapter parameterized by a weight matrix $\boldsymbol{W}_t \in \mathbb{R}^{d_m \times d_t}$ and bias $\boldsymbol{b}_t \in \mathbb{R}^{d_m}$:

$$\tilde{\boldsymbol{z}}_t = h_t(\boldsymbol{z}_t) = \boldsymbol{z}_t \boldsymbol{W}_t^\top + \boldsymbol{b}_t. \tag{2}$$

This projection ensures that text embeddings reside in the same space as image features $\boldsymbol{z}_i$, facilitating meaningful cross-modal comparisons.

We formulate our calibration objective by minimizing the mean squared error (MSE) between the observed image-text similarity and the hierarchical similarity matrix $\boldsymbol{S}$:

$$\mathcal{L}_{\text{mse}} = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \left(\text{sim}(\boldsymbol{z}_i, \tilde{\boldsymbol{z}}_t) - \boldsymbol{S}_{y_i}\right)^2, \tag{3}$$

where $\text{sim}(\boldsymbol{z}_i, \tilde{\boldsymbol{z}}_t) = \frac{\boldsymbol{z}_i \cdot \tilde{\boldsymbol{z}}_t^\top}{\|\boldsymbol{z}_i\| \|\tilde{\boldsymbol{z}}_t\|}$ denotes the cosine similarity between the image feature $\boldsymbol{z}_i$ and the transformed text embedding $\tilde{\boldsymbol{z}}_t$. $\tilde{N}$ is the number of non-tail samples, and $\boldsymbol{S}_{y_i}$ corresponds to the y_i-th row of the hierarchical similarity matrix, representing the ground-truth similarity.

Critically, we apply this calibration *only* to non-tail samples to prevent confusion in tail classes with scarce data. This directional constraint ensures that knowledge flows from head to tail classes, leveraging the semantic proximity defined by the hierarchy to enhance tail class representations. The resulting feature alignment forms a more structured visual space, providing a solid basis for the subsequent classifier-level calibration, establishing a coherent pipeline from structural prior to final predictions.

Classifier-Level Calibration While feature-level calibration organizes visual representations according to the hierarchy, it doesn't directly constrain classifier decision boundaries. This is critical in long-tail recognition, where tail classes suffer from ambiguous boundaries due to insufficient samples. To address this, we introduce classifier-level calibration that leverages hierarchical structure to guide predictions toward semantically plausible decisions.

Formally, we treat the hierarchical similarity matrix $\boldsymbol{S}$ as a structured "teacher" for knowledge distillation. Instead of conventional one-hot targets, we use rows of $\boldsymbol{S}$ as soft labels that encode semantic relatedness between classes. These soft labels provide richer supervision signals, ensuring predictions respect semantic proximity—classes that are hierarchically closer receive higher probabilities for related samples. This directly counters the long-tail bias inherited from pre-training that causes standard classifiers to violate semantic relationships, particularly for tail classes.

Given an input $\boldsymbol{x}_i$ with feature $\boldsymbol{z}_i$, the classifier h_m produces logits $\boldsymbol{l}_i = h_m(\boldsymbol{z}_i)$, which are converted to probabilities via a temperature-scaled softmax: $\boldsymbol{p}_i = \text{softmax}(\boldsymbol{l}_i/\tau_l)$, where τ_l is a hyperparameter that controls the sharpness of the output distribution. The distillation loss $\mathcal{L}_{\text{dis}}$ is then formulated as the Kullback-Leibler (KL) divergence between $\boldsymbol{p}_i$ and the corresponding hierarchical soft label $\boldsymbol{S}_{y_i}$:

$$\mathcal{L}_{\text{dis}} = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \text{KL}\left(\boldsymbol{p}_i \,\|\, \boldsymbol{S}_{y_i}\right). \tag{4}$$

This loss effectively regularizes the model's output distribution to align with the semantic structure defined by the hierarchy, ensuring that predictions for tail classes are informed by their semantic relationships with better-represented head classes. This classifier-level calibration complements the feature-level one: while feature-level alignment constructs a semantically organized representation space as a robust foundation, classifier-level constraints directly shape decision boundaries to respect hierarchical relationships.

Overall Learning Objective

Our complete training objective integrates three complementary components that address different aspects of the long-tail recognition challenge. The primary component is the classification loss $\mathcal{L}_{\text{cls}}$, for which we adopt the logit adjustment (LA) strategy (Menon et al. 2021), a technique widely validated for long-tail recognition. This loss adjusts the model's output logits based on class distribution priors to counteract the tendency to over-predict head classes:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(l_{i,y_i} + \log q_{y_i})}{\sum_{k=1}^{K} \exp(l_{i,k} + \log q_k)}, \tag{5}$$

where l_{i,y_i} is the logits for the ground-truth class y_i, and q_k is the prior probability of class k, estimated from its empirical frequency (N_k/N) in the training set.

To this foundation, we add our two hierarchical calibration losses: the feature-level MSE loss $\mathcal{L}_{\text{mse}}$ that aligns visual features with the hierarchical structure, and the classifier-level distillation loss $\mathcal{L}_{\text{dis}}$ that regularizes predictions. The final training objective is a weighted combination:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha \mathcal{L}_{\text{mse}} + \beta \mathcal{L}_{\text{dis}}, \tag{6}$$

where α and β are scalar hyperparameters that balance the contribution of each component. The effectiveness of this combined objective and an analysis of alternative classification losses are validated in our experimental section. The overall training procedure can be found in Appendix A.

Experiments

Experimental Setup

Datasets We evaluate our method on three widely used long-tail recognition benchmarks: ImageNet-LT (Liu et al. 2022), Places-LT (Liu et al. 2022), and iNaturalist 2018 (Van Horn et al. 2018). The hierarchical structure of ImageNet is derived from WordNet (Miller 1995); Places provides a two-level scene hierarchy; and iNaturalist 2018 includes a seven-level biological taxonomy, ranging from kingdom to species. Dataset details are provided in Appendix B.

Baseline Methods Following Shi et al. (2024b), we compare against three categories of approaches: (1) conventional methods that train models from scratch, (2) methods based on pre-trained foundation models, and (3) methods that utilize additional external data. For fairness, we adopt the same training settings as the competing methods. Details of the compared baselines are provided in Appendix C.

Implementation Details The image encoder g_m is built on a frozen ViT-B/16 from CLIP (Radford et al. 2021) with learnable Adapter modules (Houlsby et al. 2019) inserted into

Methods	Backbone	Learnable Params.	Epochs	Overall	Head	Medium	Tail
Training from scratch							
MiSLAS (Zhong et al. 2021)	ResNet-50	23.51M	180+10	52.7	62.9	50.7	34.3
LA (Menon et al. 2021)	ResNet-50	23.51M	90	51.1	-	-	-
DisAlign (Zhang et al. 2021)	ResNet-50	23.51M	90	52.9	61.3	52.2	31.4
NCL (Li et al. 2022b)	ResNet-50	23.51M	400	57.4	-	-	-
ProCo (Du et al. 2024)	ResNet-50	23.51M	180	57.8	68.2	55.1	38.1
CBTS (Zhao et al. 2025)	ResNet-50	23.51M	90	59.8	70.0	57.3	39.1
LiVT (Xu et al. 2023)	ViT-B/16	85.80M	100	60.9	73.6	56.4	41.0
Fine-tuning foundation model							
BALLAD (Ma et al. 2021)	ViT-B/16	149.62M	50+10	75.7	79.1	74.5	69.8
Decoder (Wang et al. 2024)	ViT-B/16	21.26M	$\sim$18	73.2	-	-	-
HGLTR(Ours)	ViT-B/16	**1.61M**	**10**	**77.4**	**80.4**	**76.5**	**72.1**
Fine-tuning with extra data							
VL-LTR (Tian et al. 2022)	ViT-B/16	149.62M	100	77.2	**84.5**	74.6	59.3
GML (Suh and Seo 2023)	ViT-B/16	149.62M	100	**78.0**	-	-	-

Table 2: Comparison performance (Top-1 Accuracy%) on ImageNet-LT.

Methods	Backbone	Learnable Params.	Epochs	Overall	Head	Medium	Tail
Training from scratch (with an ImageNet-1K pre-trained backbone)							
OLTR (Liu et al. 2022)	ResNet-152	58.14M	30	35.9	44.7	37.0	25.3
MiSLAS (Zhong et al. 2021)	ResNet-152	58.14M	90+10	40.4	39.6	43.3	36.1
DisAlign (Zhang et al. 2021)	ResNet-152	58.14M	30	39.3	40.4	42.4	30.1
ALA (Zhao et al. 2022)	ResNet-152	58.14M	30	40.1	43.9	40.1	32.9
LiVT (Xu et al. 2023)	ViT-B/16	85.80M	100	40.8	48.1	40.6	27.5
Fine-tuning foundation model							
BALLAD (Ma et al. 2021)	ViT-B/16	149.62M	50+10	49.5	49.3	50.2	48.4
Decoder (Wang et al. 2024)	ViT-B/16	21.26M	$\sim$34	46.8	-	-	-
LPT (Dong et al. 2023)	ViT-B/16	**1.01M**	40+40	50.1	49.3	52.3	46.9
GNM-PT (Li et al. 2024)	ViT-B/16	**1.01M**	100	50.1	46.6	**53.3**	49.4
HGLTR(Ours)	ViT-B/16	1.61M	**10**	**51.4**	**51.1**	51.6	**51.2**
Fine-tuning with extra data							
VL-LTR (Tian et al. 2022)	ViT-B/16	149.62M	100	50.1	**54.2**	48.5	42.0
RAC (Long et al. 2022)	ViT-B/16	85.80M	30	47.2	48.7	48.3	41.8

Table 3: Comparison performance (Top-1 Accuracy%) on Places-LT.

each Transformer block. The linear classifier h_m has a feature dimension $d_m = 768$ that matches the backbone output, and the text embeddings have a dimension $d_t = 512$. Both dimensions are determined by the pre-trained model. Training details are provided in Appendix D.

Result Analysis

Tables 2–4 present the Top-1 accuracy, learnable backbone parameters, and training epochs on three benchmarks, with the best scores **bolded**. The results reveal consistent and notable advantages of HGLTR over both traditional and foundation model-based baselines, confirming HGLTR's universal applicability. Leveraging domain knowledge of class semantic hierarchies, our dual-level calibration effectively transfers knowledge from head to tail classes while preserving the model's original capabilities.

ImageNet-LT HGLTR achieves the highest overall accuracy of 77.4%, outperforming all baselines in tailed classes, including those using external data. It significantly improves tail-class accuracy (72.1%), highlighting its strength in addressing class imbalance. With only 1.61M learnable parameters and 10 training epochs, HGLTR is not only accurate but also highly efficient. Compared to BALLAD and Decoder, HGLTR delivers gains of +1.7% and +4.2% overall. While VL-LTP (Tian et al. 2022) and GML (Suh and Seo 2023) incorporate additional external data sources, GML achieves a 0.6% higher accuracy than our method overall. Importantly, our approach does not rely on any external data; all hierar-

Methods	Overall	Head	Medium	Tail
ALA (Zhao et al. 2022)	70.7	71.3	70.8	70.4
ProCo (Du et al. 2024)	75.8	74.0	76.0	76.0
CBTS (Zhao et al. 2025)	72.8	72.1	73.2	72.4
LPT (Dong et al. 2023)	76.1	-	-	**79.3**
LiVT (Xu et al. 2023)	76.1	**78.9**	76.5	74.8
GNM-PT (Li et al. 2024)	76.3	76.3	**77.6**	75.0
HGLTR (Ours)	**76.8**	70.6	76.4	78.8

Table 4: Comparison performance (Top-1 Accuracy%) on iNaturalist 2018.

chical information is extracted solely based on classes, making it more practical for real-world applications where additional data may not be available.

Places-LT HGLTR again sets a new SOTA with 51.4% overall accuracy. It achieves balanced improvements across head (51.1%), medium (51.6%), and tail (51.2%) classes, demonstrating robust generalization across the long-tail distribution. The near-identical performance across distribution categories reveals HGLTR's unique ability to mitigate representation collapse in tail classes, which typically suffer from ambiguous decision boundaries in scene recognition tasks.

iNaturalist 2018 HGLTR achieves the highest overall accuracy (76.8%) and excels on tail classes (78.8%), demonstrating strong scalability to large-scale, fine-grained datasets with a seven-level taxonomy. Its notable tail-class advantage (+8.2% over head classes) arises from effective knowledge transfer across taxonomic levels, leveraging hierarchical structure to compensate for limited training data on rare species. Full results are provided in Appendix E.

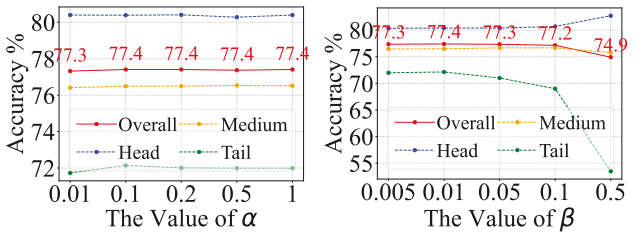
Ablation Study

Effectiveness of Components We conduct an ablation study on ImageNet-LT to evaluate HGLTR's components. Table 5 shows that both loss terms: $\mathcal{L}_{\text{mse}}$ and $\mathcal{L}_{\text{dis}}$ contribute positively to model performance. Their combination leads to the highest gains, pushing tail accuracy to 72.1%, which is critical for long-tail recognition. Notably, $\mathcal{L}_{\text{dis}}$ provides a +3.6% boost in tail accuracy over using $\mathcal{L}_{\text{mse}}$ alone, demonstrating that hierarchical distance-based supervision is highly effective for enhancing generalization to rare classes. Additional results on hierarchical prompts and head-to-tail transfer constraints are presented in the Appendix F.

Sensitivity to Hyperparameters We evaluate HGLTR's sensitivity to key hyperparameters α and β on the ImageNet-LT dataset. As shown in Fig. 4, HGLTR is robust to variations in the weight α, while performance is most sensitive to β, which governs the influence of hierarchy-aware regularization. A moderate β (e.g., 0.1) yields optimal results, confirming that softly guided transfer via hierarchical priors is both stable and impactful.

Component		Overall	Head	Medium	Tail
$\mathcal{L}_{\text{mse}}$	$\mathcal{L}_{\text{dis}}$				
✓	-	76.3	80.6	75.3	68.5
-	✓	76.4	80.3	75.4	69.5
✓	✓	**77.4**	**80.4**	**76.5**	**72.1**

Table 5: Effectiveness (Top-1 Accuracy%) of HGLTR components on ImageNet-LT.



(a) Sensitivity effect of α. (b) Sensitivity effect of β.

Figure 4: Hyperparameter sensitivity analysis (Top-1 Accuracy%) on ImageNet-LT. (a) Performance with α and β fixed at 0.01. (b) Performance with β and α fixed at 0.1. Overall accuracy is highlighted in the plot.

PEFT	Learnable Params.	Overall	Head	Tail
Zero-Shot	0.00M	67.2	68.2	66.6
LoRA (Yu et al. 2023)	2.76M	75.7	79.9	64.8
LIFT (Shi et al. 2024b)	1.25M	76.7	80.2	70.6
VPT-Shallow (Jia et al. 2022)	0.41M	74.2	79.3	60.0
VPT-Deep (Jia et al. 2022)	0.49M	76.3	79.7	69.2
Adapter (Houlsby et al. 2019)	1.61M	**77.4**	**80.4**	**72.1**

Table 6: Performance (Top-1 Accuracy%) of HGLTR using different PEFT on ImageNet-LT.

Analysis of PEFT Table 6 compares HGLTR with other PEFT strategies. The Adapter achieves the best overall accuracy (77.4%), outperforming others. While all PEFT modules improve upon the CLIP zero-shot baseline, Adapters strike the best balance between adaptation capacity and parameter efficiency, confirming their suitability for long-tail transfer. Additional comparisons with different loss functions are reported in the Appendix G.

Conclusion

This work addresses performance degradation of pre-trained VLMs in long-tail recognition, caused not only by fine-tuning limitations but fundamentally by inherent bias from the long-tail distribution. We propose HGLTR, the first framework injecting hierarchical knowledge into CLIP-like models through dual-level calibration, aligning visual features with hierarchy-guided structures at the feature level and employing distillation with hierarchical soft labels at the classifier level. Extensive experiments show HGLTR achieves SOTA results with significant tail-class accuracy gains. In the future, we plan to explore automatically constructing hierarchies for domains lacking taxonomic structures and applying hierarchical calibration to broader distribution-shift scenarios.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62576168, 62376126), Fundamental Research Funds for the Central Universities (NS2024059), Open Project Funds for the Joint Laboratory of Spatial Intelligent Perception and Large Model Application (SIPLMA-2024-YB-05).

References

Bertinetto, L.; Mueller, R.; Tertikas, K.; Samangooei, S.; and Lord, N. A. 2020. Making better mistakes: leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12506–12515.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, 32.

Dong, B.; Zhou, P.; YAN, S.; and Zuo, W. 2023. Lpt: long-tailed prompt tuning for image classification. In *International Conference on Learning Representations*.

Du, C.; Wang, Y.; Song, S.; and Huang, G. 2024. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9): 5890–5904.

Feng, C.; Tzimiropoulos, G.; and Patras, I. 2024. Clipcleaner: cleaning noisy labels with clip. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 876–885.

Gu, Z.; Zhu, B.; Zhu, G.; Chen, Y.; Tang, M.; and Wang, J. 2024. Anomalygpt: detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 1932–1940.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, 2790–2799. PMLR.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727. Springer.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*.

Landrieu, L.; and Garnot, V. S. F. 2021. Leveraging class hierarchies with metric-guided prototype learning. In *British Machine Vision Conference*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, J.; Tan, Z.; Wan, J.; Lei, Z.; and Guo, G. 2022b. Nested collaborative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6949–6958.

Li, M.; Liu, Y.; Lu, Y.; Zhang, Y.; Cheung, Y.-m.; and Huang, H. 2024. Improving visual prompt tuning by gaussian neighborhood minimization for long-tailed visual recognition. *Advances in Neural Information Processing Systems*, 37: 103985–104009.

Liang, T.; and Davis, J. 2023. Inducing neural collapse to a fixed hierarchy-aware frame for reducing mistake severity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1443–1452.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2022. Open long-tailed recognition in a dynamic world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3): 1836–1851.

Long, A.; Yin, W.; Ajanthan, T.; Nguyen, V.; Purkait, P.; Garg, R.; Blair, A.; Shen, C.; and Van den Hengel, A. 2022. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6959–6969.

Ma, T.; Geng, S.; Wang, M.; Shao, J.; Lu, J.; Li, H.; Gao, P.; and Qiao, Y. 2021. A simple long-tailed recognition baseline via vision-language model. *arXiv preprint arXiv:2111.14745*.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Shi, J.; Gare, G.; Tian, J.; Chai, S.; Lin, Z.; Vasudevan, A.; Feng, D.; Ferroni, F.; and Kong, S. 2024a. Lca-on-the-line: benchmarking out-of-distribution generalization with class taxonomies. In *International Conference on Machine Learning*, 44887–44908.

Shi, J.-X.; Wei, T.; Xiang, Y.; and Li, Y.-F. 2023. How re-sampling helps for long-tail learning? *Advances in Neural Information Processing Systems*, 36: 75669–75687.

Shi, J.-X.; Wei, T.; Zhou, Z.; Shao, J.-J.; Han, X.-Y.; and Li, Y.-F. 2024b. Long-tail learning with foundation model: heavy fine-tuning hurts. In *International Conference on Machine Learning*, 45014–45039. PMLR.

Shi, J.-X.; Zhang, C.; Wei, T.; and Li, Y.-F. 2024c. Efficient and long-tailed generalization for pre-trained vision-language model. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, 2663–2673.

Suh, M.-K.; and Seo, S.-W. 2023. Long-tailed recognition by mutual information maximization between latent features and ground-truth labels. In *International Conference on Machine Learning*, 32770–32782. PMLR.

Tian, C.; Wang, W.; Zhu, X.; Dai, J.; and Qiao, Y. 2022. Vl-ltr: learning class-wise visual-linguistic representation for

long-tailed visual recognition. In *European Conference on Computer Vision*, 73–91. Springer.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8769–8778.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. 2021. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*.

Wang, Y.; Yu, Z.; Wang, J.; Heng, Q.; Chen, H.; Ye, W.; Xie, R.; Xie, X.; and Zhang, S. 2024. Exploring vision-language models for imbalanced learning. *International Journal of Computer Vision*, 132(1): 224–237.

Xu, Z.; Liu, R.; Yang, S.; Chai, Z.; and Yuan, C. 2023. Learning imbalanced data with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15793–15803.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.

Yu, Y.; Yang, C.-H. H.; Kolehmainen, J.; Shivakumar, P. G.; Gu, Y.; Ren, S. R. R.; Luo, Q.; Gourav, A.; Chen, I.-F.; Liu, Y.-C.; et al. 2023. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 1–8. IEEE.

Zhang, S.; Li, Z.; Yan, S.; He, X.; and Sun, J. 2021. Distribution alignment: a unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2361–2370.

Zhang, Y.; Hooi, B.; Hong, L.; and Feng, J. 2022. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. *Advances in Neural Information Processing Systems*, 35: 34077–34090.

Zhao, W.; Li, W.; Li, Y.; Yang, L.; Liang, Z.; Hu, E.; Zhang, W.; and Yang, H. 2025. Constructing balanced training samples: a new perspective on long-tailed classification. *IEEE Transactions on Multimedia*, 27: 5130–5143.

Zhao, Y.; Chen, W.; Tan, X.; Huang, K.; and Zhu, J. 2022. Adaptive logit adjustment loss for long-tailed visual recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 3472–3480.

Zhao, Z.; Wen, H.; Wang, Z.; Wang, P.; Wang, F.; Lai, S.; Zhang, Q.; and Wang, Y. 2024. Breaking long-tailed learning bottlenecks: a controllable paradigm with hypernetwork-generated diverse experts. *Advances in Neural Information Processing Systems*, 37: 7493–7520.

Zhong, Z.; Cui, J.; Liu, S.; and Jia, J. 2021. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16489–16498.

Zhu, B.; Tang, K.; Sun, Q.; and Zhang, H. 2023. Generalized logit adjustment: calibrating fine-tuned models by removing label bias in foundation models. *Advances in Neural Information Processing Systems*, 36: 64663–64680.