Long-Tailed Classification with Multi-Granularity Semantics

Yuting Liu Liu Yang✉ Yu Wang
Tianjin University, Tianjin, China
{liuyuting, yangliuyl, wang.yu}@tju.edu.cn

Abstract

Real-world data often exhibit long-tailed distributions, which degrade data quality and pose challenges for deep learning. To address this issue, knowledge transfer from head classes to tail classes has been shown to effectively mitigate feature sparsity. However, existing methods often overlook class differences, leading to suboptimal knowledge transfer. While the class space exhibits a label hierarchy, similarity relationships beyond hierarchically related categories remain underexplored. Considering the human ability to process visual perception problems in a multi-granularity manner guided by semantics, this paper presents a novel semantic knowledge-driven contrastive learning method. Inspired by the implicit knowledge embedded in large language models, the proposed LLM-based label semantic generation method overcomes the limitations of the label hierarchy. Additionally, a semantic knowledge graph is constructed based on the extended label information to guide representation learning. This enables the model to dynamically identify relevant classes for learning and facilitates multi-granularity knowledge transfer between similar categories. Experiments on long-tail benchmark datasets, including CIFAR-10-LT, CIFAR-100-LT, and ImageNet-LT, demonstrate that the proposed method significantly improves the accuracy of tail classes and enhances overall performance without compromising the accuracy of head classes.

1. Introduction

In recent years, deep learning [18, 33] has achieved remarkable progress in computer vision tasks [27, 45, 49]. These tasks typically require large amounts of data and a balanced distribution across classes, and this heavy reliance on data limits the development of deep learning. However, real-world datasets are often imbalanced, with a few classes containing an abundance of samples (*i.e.* head classes), while other classes have only a limited number (*i.e.* tail classes). The long-tailed distribution of data often leads to deep learning models being dominated by the performance of head classes, while the learning of tail classes remains severely underdeveloped. As a result, data imbalance presents novel challenges in training deep models capable of making accurate and reliable decisions between different classes. To address the issue of unbalanced data, various methods have been explored, mainly focusing on re-sampling the training data [12, 31] and re-weighting the loss of different classes [8, 15] to allocate more attention to minority class samples.

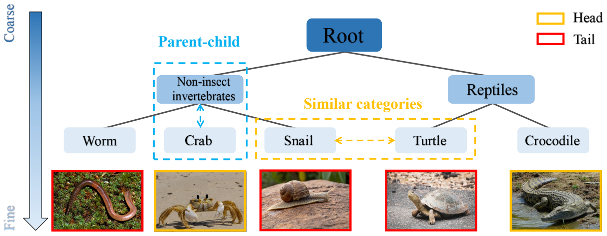


Figure 1. An example label hierarchy of CIFAR-100-LT. The coarse-grained class "Non-insect invertebrates" is the parent class of "Crab". Meanwhile, "Turtle" and "Snail" do not belong to the same subtree, but they all share semantic similarities.

However, compared to earlier approaches, research on knowledge transfer for long-tailed recognition remains relatively limited. The core objective of knowledge transfer is to transfer the knowledge learned from the substantial data of the head classes to the tail classes. Existing methods may fail to account for class differences, leading to ineffective knowledge transfer. For instance, in the case where the head class is "dog" and the tail class is "airplane", transferring knowledge from the head class to the tail class would be ineffective.

When addressing visual perception challenges, humans employ a hierarchical framework to associate each object with concepts of varying granularity, thereby enabling learning and reasoning based on multi-granularity semantic structures. In fact, the class space inherently contains multi-granularity relationships, where similar fine-grained

✉ Corresponding author.

categories can be organized into coarser-grained concepts based on semantic dependencies between classes in WordNet, thus forming a label hierarchy. Fig. 1 illustrates an example label hierarchy of CIFAR-100-LT dataset with an imbalance factor of 100. With the help of label hierarchy, Zhao *et al*. [52] proposed a hierarchical knowledge transfer method from coarse-grained to fine-grained classes.

However, the label hierarchy constructed based on WordNet is fixed. We observe that certain categories across different subtrees exhibit semantic similarities. For instance, "Snail" and "Turtle" share characteristics such as possessing a protective shell and moving slowly, as both can retract into their shells for defense. Therefore, this study introduces a novel perspective by focusing on the implicit similarity relationships mentioned above, and proposes a novel long-tailed classification method based on multi-granularity knowledge transfer. Humans can leverage prior knowledge to infer rare categories. For example, recognizing that airplanes and birds share similarities allows us to infer certain characteristics of airplanes.

Inspired by this, tail classes can borrow semantic transformations from other classes, even if they do not belong to the same coarse-grained category or are not at the same hierarchical level. In this paper, due to the limited semantic information contained in category labels, large language models (LLMs) (*e.g*., ChatGPT [47], DeepSeek [11]) are leveraged to enrich the label information as prior knowledge, thereby enhancing semantic knowledge generation and guiding the representation learning process. Due to their wealth of potential knowledge, LLMs serve as a teacher in our study, providing the model with effective prior knowledge. Our framework aims to leverage the capabilities of large language models to generate customized and diverse descriptions for each level. The generated descriptions are fed into a sentence-transformer model to construct a semantic similarity matrix, which captures relationships both within and across hierarchical levels, enabling the generation of a multi-granularity semantic knowledge graph.

The key to addressing the long-tailed problem lies in discovering and encoding the relationships between different classes. As shown in Fig. 2, we implement semantic knowledge-driven contrastive learning (SKCL) in a two-stage framework composed of a contrastive learning branch and a classification branch to better integrate the obtained semantic similarity relationships into the representation space. For each class, the Top-K classes are dynamically identified based on the multi-granularity semantic knowledge graph, which is constructed through leveraging LLMs. The proposed multi-granularity knowledge transfer constructs prototypes for each level, encouraging sample representations to be more closely with their corresponding similar category prototypes. This process facilitates both horizontal and vertical knowledge transfer, which is particularly crucial for representation learning.

Our main contributions are summarized as follows: (1) The proposed approach leverages the capabilities of large language models to enrich label semantics, which is utilized to construct a multi-granularity semantic knowledge graph as prior knowledge. (2) To encode category similarity, the proposed multi-granularity knowledge transfer constructs prototypes at different levels, enabling contrastive learning driven by semantic knowledge. (3) Extensive experiments on widely used long-tailed datasets demonstrate that our method improves recognition for both tail and head classes while achieving competitive overall performance.

2. Related Work

2.1. Long-tailed Recognition

Re-sampling is a fundamental approach in deep learning that balances data by over-sampling tail classes [31, 32] or under-sampling head classes [1, 12]. Re-weighting methods [8, 15] assign different weights to classes, enhancing the model's focus on tail samples. However, both approaches may lead to overfitting of tail classes [3]. Recently, decoupling methods [19, 53] have proven effective in long-tailed learning by separating representation learning from classifier learning, mitigating the negative impact of re-balancing on feature learning. To further enhance tail class diversity, data augmentation techniques have been explored [14, 29]. For instance, ECS-SC [14] leverages multi-granularity knowledge to identify semantic relationships, utilizing head classes to enrich tail classes. Unlike these approaches, our method guides representation learning by exploring latent class similarity relationships.

2.2. Knowledge Transfer

Due to the abundance of head class samples, recent studies [5, 26, 50, 52] have explored knowledge transfer from head to tail classes to mitigate the sparse representations of tail classes. Liu *et al*. [26] propose a feature cloud representation, where the intra-class distribution learned from head classes enhances the variability of tail class features. Chu *et al*. [5] leverage class activation maps to disentangle class-specific and class-generic features, augmenting tail-class features by integrating class-specific features with generic ones from head classes. However, these methods overlook relationships between classes, potentially introducing irrelevant knowledge. To address this, HCKC [50] incorporates hierarchical relationships as auxiliary information and transfers relevant knowledge through a hierarchical convolutional neural network. Furthermore, MGKT-MFF [52] employs a multi-scale feature fusion network to extract and utilize rich feature information, enhancing multi-granularity knowledge transfer. Unlike prior works, our approach

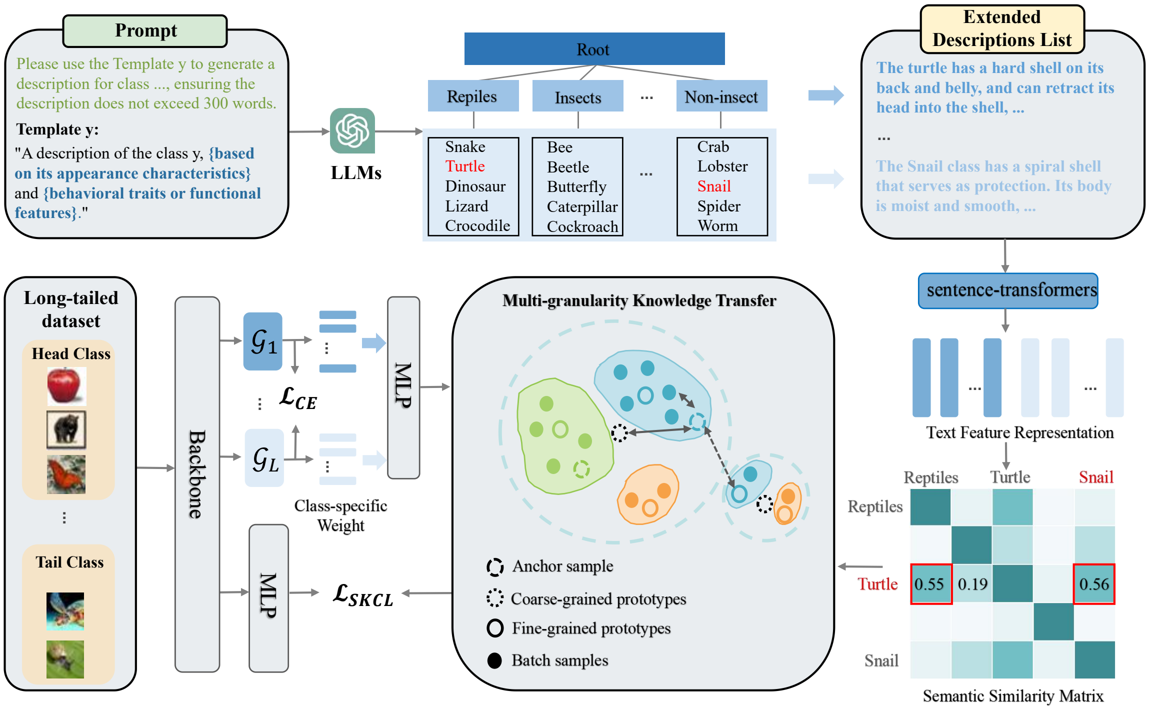

Prompt
Please use the Template y to generate a description for class ..., ensuring the description does not exceed 300 words.
Template y:
"A description of the class y, {based on its appearance characteristics} and {behavioral traits or functional features}."
LLMs
Root
Repiles
Insects
...
Non-insect
Snake
Turtle
Dinosaur
Lizard
Crocodile
Bee
Beetle
Butterfly
Caterpillar
Cockroach
Crab
Lobster
Snail
Spider
Worm
Extended Descriptions List
The turtle has a hard shell on its back and belly, and can retract its head into the shell, ...
...
The Snail class has a spiral shell that serves as protection. Its body is moist and smooth, ...
sentence-transformers
Text Feature Representation
Reptiles
Turtle
Snail
0.55
0.19
0.56
Semantic Similarity Matrix
Long-tailed dataset
Head Class
Tail Class
Backbone
$\mathcal{G}_1$
$\mathcal{G}_L$
$\mathcal{L}_{CE}$
MLP
Class-specific Weight
$\mathcal{L}_{SKCL}$
Multi-granularity Knowledge Transfer
Anchor sample
Coarse-grained prototypes
Fine-grained prototypes
Batch samples

Figure 2. An overview of the proposed framework. We first utilize LLMs to generate a unified extended description list, which enables the model uncover implicit category similarities beyond the label hierarchy and construct a semantic similarity matrix. Semantic knowledge-driven contrastive learning is employed to establish a semantic space for multi-granularity knowledge transfer, while the weights of linear classifiers at different levels are transformed non-linearly through an MLP to obtain representations of multi-granularity prototypes.

aims to improve tail class performance while ensuring a balanced enhancement of head class performance through multi-granularity transfer learning.

2.3. Contrastive Learning

Contrastive learning, which aggregates similar samples while excluding dissimilar ones, has been widely applied to feature representation learning across various tasks [23, 34, 46]. In long-tailed recognition, SSP [46] demonstrates that class imbalance can be alleviated through both semi-supervised and self-supervised approaches. Recently, supervised contrastive learning (SCL) [21] has been applied to image classification and long-tailed recognition, effectively combining the strengths of supervised and contrastive learning methods. For instance, Hybrid-SC and Hybrid-PSC [40] employ a two-branch network, where one branch utilizes supervised contrastive learning to enhance feature representations, while the other reduces classifier bias. However, due to the imbalanced data distribution, minority class samples often exhibit poor separability in feature space. To address this, TSC [23] enforces a uniform distribution of class features, while BCL [55] leverages class prototypes as additional samples to enhance long-tailed recognition.

2.4. Large Language Models

Recently, large language models (*e.g.*, BERT [9], ChatGPT [47], DeepSeek [11]) have been explored to address long-tailed challenges from a semantic perspective. LLMs serve as vast knowledge repositories [10]. Researchers have investigated their generative capabilities to mitigate long-tailed issues. For instance, LLM-AutoDA [41] discovers effective data augmentation strategies for imbalanced distributions, while LTGC [51] generates diverse and accurate tail data. Despite these advancements, practical deployment remains challenging due to the high computational cost and extensive training time. In this work, we propose a lightweight approach to harness LLMs' latent knowledge, enabling compact visual models to better understand category relationships and addressing long-tailed challenges.

3. Methodology

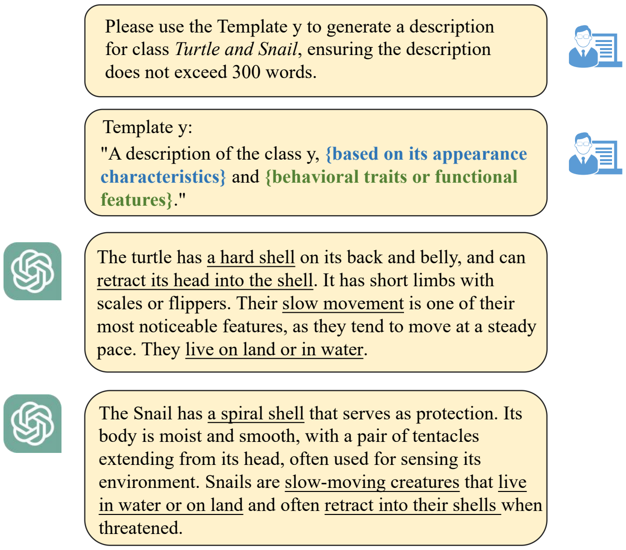
3.1. Problem Setting

In the long-tailed recognition task, our goal is to capture semantic similarities between categories to facilitate multi-granularity knowledge transfer and enrich the semantic knowledge of both head and tail classes through representation learning. Given a long-tailed training dataset $\mathbf{D} = \{(\mathbf{x_i}, y_i)\}_{i=1}^{N}$, where $\mathbf{x_i}$ denotes a sample and $y_i \in \mathcal{Y}$ represents its corresponding label. When the input image $\mathbf{x_i}$ is processed by a deep feature extractor $f_\theta(\cdot)$, the corresponding representation $\mathbf{z_i}$ is extracted as $\mathbf{z_i} = f_\theta(\mathbf{x_i})$. Consequently, the sample $\mathbf{x_i}$ is mapped to a feature space and the final prediction is given by the linear classifier. The quality of these representations plays a crucial role in classification accuracy. Therefore, our objective is to learn an effective encoder $f_\theta(\cdot)$ to enhance long-tailed learning.

3.2. LLM-based Label Semantic Generation

Real-world data often exhibit long-tailed distributions and inherently possess a structure known as label hierarchy. However, a significant limitation is that label hierarchy derived from WordNet is static and contains restricted semantic information. Intuitively, the different categories exhibit varying degrees of similarity. We observe that fine-grained categories under different coarse-grained categories may have latent similarities, while fine-grained categories under the same coarse-grained category also exhibit varying levels of similarity. For example, although "turtle" and "crocodile" belong to the category "reptiles", the class "turtle" actually exhibits greater semantic similarity with "snail", which belongs to the category "non-insect invertebrates". Both share characteristics such as a hard shell and a slow-moving, ground-dwelling nature, which contribute to their greater similarity in appearance and behavior.

To capture the implicit similarity relationships among categories in the label hierarchy, semantic knowledge of category labels can be utilized. However, the semantic knowledge of category labels is inherently limited. Therefore, we leverage the extensive common-sense knowledge of large language models (LLMs), such as ChatGPT [47], to generate textual descriptions based on existing labels. In this process, the responses generated by LLMs may vary, and at times, contain redundant information. This variability and redundancy may hinder the subsequent extraction of similarity, as it introduces unnecessary complexity. To unify the format of label expansion descriptions, we employ predefined text templates to constrain the responses of LLMs. Specifically, for a given class y, the template is defined as follows: `"A description of the class [y], {based on its appearance characteristics} and {behavioral traits or functional features}."`, which includes the given class, its appearance characteristics, and either its behavioral traits (for animals) or functional features (for other objects). The response template and Prompt are input together into the LLMs to obtain the final description: `"Please use the Template y to generate a description for class [y], ensuring the description does not exceed 300 words."` To ensure fairness in the subsequent similarity comparison, we also impose a word limit on the descriptions. As shown in Fig. 3, y represents the given class label, and the LLMs automatically generate descriptions by incorporating the class name, appearance characteristics, and either functional features or behavioral traits according to Template y. This standardization ensures that the generated descriptions are consistent and comparable across all categories.



Please use the Template y to generate a description for class *Turtle and Snail*, ensuring the description does not exceed 300 words.

Template y:
"A description of the class y, **{based on its appearance characteristics}** and **{behavioral traits or functional features}**."

The turtle has a hard shell on its back and belly, and can retract its head into the shell. It has short limbs with scales or flippers. Their slow movement is one of their most noticeable features, as they tend to move at a steady pace. They live on land or in water.

The Snail has a spiral shell that serves as protection. Its body is moist and smooth, with a pair of tentacles extending from its head, often used for sensing its environment. Snails are slow-moving creatures that live in water or on land and often retract into their shells when threatened.

Figure 3. An example instruction for LLMs. When text templates and multi-granularity labels are input into LLMs, semantically rich descriptions can be obtained, revealing similarities between categories, such as the common traits of "turtle" and "snail".

3.3. Semantic Knowledge Graph

Transfer learning is based on the continuous and iterative assumption that the processing mechanism of deep neural networks is similar to that of the human brain, as both rely on existing knowledge to recognize new concepts [17]. Inspired by the human cognitive process when encountering new concepts, we construct a semantic knowledge graph as prior knowledge using an extended description list, which incorporates rich common sense knowledge from LLMs to assist the model in recognizing rare categories. Let us define the semantic knowledge graph $G = (V, E)$ for the class space, where $V = \{v_1, v_2, \ldots, v_n\}$ represents the set of n nodes, and $E \subseteq V \times V$ denotes the set of edges. Each node $v \in V$ represents a distinct class label, and $(v_i, v_j) \in E$ indicates the semantic similarity relationship between nodes v_i and v_j.

To construct the semantic knowledge graph, we first employ the pre-trained model all-MiniLM-L6-v2 [42] to generate sentence embeddings $\mathbf{e_i}$ for class v_i, where $\mathbf{e_i}$ is a 384-dimensional vector representing the embedding of each description. Then, the obtained sentence embeddings are ℓ_2-normalized to ensure that all sentence embeddings lie on the unit sphere, which guarantees fairness in similarity computation. This normalization process is expressed as follows:

$$\mathbf{\hat{e}_i} = \frac{\mathbf{e_i}}{\|\mathbf{e_i}\|_2}. \tag{1}$$

In the resulting sentence embedding space, the similarity between v_i and v_j in the semantic knowledge graph is computed using cosine similarity:

$$\mathbf{S}(i,j) = \mathbf{\hat{e}_i} \cdot \mathbf{\hat{e}_j}, \tag{2}$$

where $\mathbf{S} \in \mathbb{R}^{n\times n}$ represents the semantic similarity matrix.

3.4. Multi-granularity Knowledge Transfer

To effectively incorporate semantic similarity knowledge into the learned embedding space, we construct a multi-granularity prototype space. In this space, class prototypes can serve as learnable class centers without introducing additional computational overhead. Specifically, we apply nonlinear mappings to the weights of linear classifiers at different levels, and treat the resulting outputs as prototypes for each class. Through experimental investigations within long-tail recognition tasks, we have observed that an excessive amount of prior knowledge may introduce misleading signals, thereby hindering the training process. To mitigate this issue, we propose transferring the existing semantic relationships between categories into the model by considering only the Top-K most similar classes. Formally, for the i-th row of the similarity matrix $\mathbf{S}$, the Top-K most similar categories are found as follows:

$$I_i = TopK\left\{\mathbf{S}(i,j), K\right\}_{j=1}^{n}, i \neq j, \tag{3}$$

where n represents the total number of nodes in the semantic knowledge graph G. Thus, $(v_i, v_j) \in E$ if and only if $j \in I_i$.

To facilitate knowledge transfer between similar categories, we introduce multi-granularity class center representations, *i.e*., multi-granularity prototypes for semantic knowledge-driven contrastive learning. The representation $\mathbf{z_i}$ in semantic knowledge-driven contrastive learning is obtained using a multi-layer perceptron with one hidden layer, followed by ℓ_2-normalization of $\mathbf{z_i}$ to produce $\mathbf{\bar{z}_i}$. For an instance $\mathbf{x_i}$ with representation $\mathbf{\bar{z}_i}$ in batch B, the semantic knowledge-driven contrastive loss $\mathcal{L}_{SKCL}$ is defined by the following expression:

$$\mathcal{L}_{SKCL}(\mathbf{x_i}) = -\Big(\frac{1}{|M_i|}\sum_{\mathbf{c_q}\in M_i}\log\frac{\exp(\mathbf{\bar{z}_i}\cdot\mathbf{c_q}/\tau')}{\sum_{j=1}^{n}\exp(\mathbf{\bar{z}_i}\cdot\mathbf{c_j}/\tau')}+ \frac{1}{|P_i|}\times\sum_{\mathbf{\bar{z}_p}\in P_i}\log\frac{\exp(\mathbf{\bar{z}_i}\cdot\mathbf{\bar{z}_p}/\tau)}{\sum\limits_{l\in\mathcal{Y}}\frac{1}{|A_l|}\sum\limits_{\mathbf{\bar{z}_k}\in A_l}\exp(\mathbf{\bar{z}_i}\cdot\mathbf{\bar{z}_k}/\tau)}\Big), \tag{4}$$

where M_i represents the set consisting of the class prototype corresponding to y_i and the prototypes of its Top-K most similar classes, in other words, $q \in I_{y_i} \cup \{y_i\}$. Specifically, A_l denotes the set of representations $\mathbf{\bar{z}_l}$ belonging to the class l, and P_i is a subset of B that contains elements from A_{y_i} excluding $\mathbf{\bar{z}_i}$. $|\cdot|$ denotes the size of the set, representing the number of samples.τ' and τ ($\tau' > \tau > 0$) are scalar temperature hyperparameters that control the model's sensitivity to similar samples and class prototypes. A smaller temperature value results in a lower tolerance for similar samples [39].

By pulling the samples of the target class closer to the prototypes of similar classes in the representation space, our method achieves multi-granularity knowledge transfer. This enables knowledge to be transferred not only within the same label level but also across different granularity levels.

3.5. Framework

The overview of the proposed framework is shown in Fig. 2. We implement SKCL in a two-stage framework, consisting of two main components: a contrastive learning branch and a classification branch. To build multi-granularity prototypes, for each level $i \in \{1, 2, \ldots, L\}$ in the label hierarchy, we train a linear classifier on the learned representations along with the corresponding labels to obtain the classifier $\mathcal{G}_i$. Then, the cross-entropy loss function $\mathcal{L}_{ce,i}$ is computed based on the classifier's predictions. Accordingly, the total classification loss function can be calculated as follows:

$$\mathcal{L}_{CE} = \textstyle\sum_{i=1}^{L} w_i \mathcal{L}_{ce\_i} \tag{5}$$

where w_i denotes the weights assigned to learning features at different levels. The weights of the linear classifiers are transformed non-linearly through an MLP to obtain multi-granularity prototype representations. To unify the format of an extended description list, we use text templates to constrain the responses of LLMs. Then, the extended description list enriched with commonsense knowledge from LLMs is input into a sentence-transformer model, and the semantic similarity matrix is computed using cosine similarity to construct the multi-granularity semantic knowledge graph.

In the constructed semantic feature space, SKCL performs multi-granularity knowledge transfer by encoding the implicit similarity relationships between categories. Overall, we use a training objective composed of cross-entropy loss $\mathcal{L}_{CE}$ and contrastive loss $\mathcal{L}_{\text{SKCL}}$:

$$\mathcal{L} = \lambda\mathcal{L}_{\text{SKCL}} + \mathcal{L}_{\text{CE}}, \tag{6}$$

where λ controls the impact of $\mathcal{L}_{\text{SKCL}}$. In the representation space, SKCL enables both vertical and horizontal knowledge transfer through the construction of multi-granularity prototypes, while simultaneously enriching the semantic knowledge of both head and tail classes.

Table 1. Comparison results with state-of-the-art methods using ResNet-32 on the CIFAR-100-LT and CIFAR-10-LT datasets under different imbalance factors, focusing on Top-1 accuracy(%). The best results are highlighted in bold.

Methods	CIFAR-100-LT			CIFAR-10-LT		
	100	50	10	100	50	10
Focal Loss [25]	38.41	44.32	55.78	70.38	76.72	86.66
CB-Focal [8]	39.60	45.17	57.99	74.57	79.27	87.10
LDAM-DRW [2]	42.04	46.62	58.71	77.03	81.03	88.16
SSP [46]	43.43	47.11	58.91	77.83	82.13	88.53
BBN [54]	42.56	47.02	59.12	79.82	81.18	88.32
Casual Model [38]	44.10	50.30	59.60	80.60	83.60	88.50
KCL [20]	42.80	46.30	57.60	77.60	81.70	88.00
Hybrid-SC [40]	46.72	51.87	63.05	81.40	85.36	91.12
MetaSAug-LDAM [22]	48.01	52.27	61.28	80.66	84.34	89.68
TSC [24]	43.80	47.40	59.00	79.70	82.90	88.70
ResLT [7]	48.21	52.71	62.01	82.40	85.17	89.70
Remix [4]	41.94	49.50	59.36	75.36	-	88.15
UniMix [44]	45.45	51.11	61.25	82.75	84.32	89.66
SMC [16]	48.90	52.30	62.50	-	-	-
ECS-SC [14]	43.16	47.32	59.68	-	-	-
HCKC [50]	39.00	45.31	57.56	77.05	-	87.81
MGKT-MFF [52]	46.36	52.34	64.18	-	-	-
BCL [55]	52.01	56.32	64.01	84.31	87.26	90.91
ConCutMix [30]	53.16	57.40	64.53	86.07	88.00	91.42
Ours	**54.02**	**58.13**	**65.86**	**87.50**	**88.16**	**92.37**

Table 2. Comparison results with state-of-the-art methods using ResNet-32 on the CIFAR-100-LT dataset with an imbalance factor of 100, focusing on Top-1 accuracy(%).

Methods	Many	Medium	Few	All
τ-norm [19]	61.4	42.5	15.7	41.4
Hybrid-SC [40]	-	-	-	46.7
MetaSAug-LDAM [22]	-	-	-	48.0
DRO-LT [37]	64.7	50.0	23.8	47.3
RIDE(3 experts) [43]	68.1	49.2	23.9	48.0
BCL [55]	67.2	53.1	32.9	51.9
ConCutMix [30]	67.4	53.9	35.8	53.2
Ours	**68.3**	**54.1**	**37.2**	**54.0**

4. Experiments

4.1. Datasets

Long-Tailed CIFAR. CIFAR-10-LT and CIFAR-100-LT [2] are long-tailed versions of the CIFAR-10 and CIFAR-100 datasets, respectively. Both datasets contain 60,000 images of size 32×32, covering 10 and 100 classes, respectively. CIFAR-10/100-LT is constructed by exponentially decreasing the number of training samples per class. The imbalance ratio β is typically defined as $\beta = N_{max}/N_{min}$, which quantifies the degree of class imbalance in the dataset. In our experiments, we set β to 100, 50, and 10 to analyze different levels of class imbalance.

ImageNet-LT. ImageNet-LT [28] is a long-tailed subset of the large-scale ImageNet dataset [36], following a Pareto distribution with a shape parameter of $\alpha = 0.6$. It contains 115.8K images across 1,000 categories, with a maximum of 1,280 images per class and a minimum of 5 images per class.

4.2. Experimental Setup

Implementation Details. For both CIFAR-10-LT and CIFAR-100-LT, following [30], we adopt ResNet-32 [13] with AutoAugment [6] as the network backbone, with a weight decay of $5e-4$. The batch size is set to 256, and the temperatures τ and τ' are set to 0.1 and 0.2, respectively. Each experiment is trained for 300 epochs, and K is set to 2. For ImageNet-LT, we use ResNet-50 [13] as our backbone. During training, we apply the same augmentation strategy as in [30] for the two-branch framework, with an initial learning rate of 0.1 and a weight decay of $5e-4$. The model is optimized using stochastic gradient descent (SGD) with a momentum value of 0.9 and K set to 4. Since the ImageNet-LT dataset is inherently structured based on the WordNet hierarchy with 12 subtrees, it does not require classifiers at different levels. We utilize gpt-4.0-

turbo for label semantic generation. CIFAR-100-LT consists of 20 coarse-grained classes, each of which comprises 5 fine-grained classes. Meanwhile, we group the 10 classes of CIFAR-10-LT into two coarse-grained categories, “vehicles” and “animals”, based on semantic similarity.

Evaluation Metrics. To evaluate the performance of our method more reasonably, we divide these classes into three groups based on the number of training samples, following [55]: classes with more than 100 training samples are considered head classes, those with 20 to 100 training samples are classified as medium classes, and classes with fewer than 20 training samples are categorized as tail classes. Additionally, we compute the Top-1 accuracy across all test samples under different imbalance factors.

Compared Methods To validate the effectiveness of the proposed method, we compare it with several state-of-the-art methods from six different categories: (1) Class-blalanced classifiers: τ-norm [19], LWS [19], and DisAlign [48]. (2) Loss functions: Focal Loss [25], CB-Focal [8], LDAM-DRW [2], DRO-LT [37], BALMS [35], and LADE [15]. (3) Contrastive learning methods: SSP [46], Hybrid-SC [40], KCL [20], BCL [55], TSC [24], and SMC [16]. (4) Transfer learning: BBN [54], HCKC [50], and MGKT-MFF [52]. (5) Data augmentation: Remix [4], UniMix [44], MetaSAug-LDAM [22], ECS-SC [14], and ConCutMix [30]. (6) Other methods: [38], ResLT [7], and RIDE [43].

4.3. Comparison with State-of-the-art Methods

Long-tailed CIFAR. As illustrated in Tab. 1, our method outperforms existing advanced approaches across all imbalance ratios on CIFAR-100-LT and CIFAR-10-LT, demonstrating its effectiveness in addressing long-tailed challenges. Recent contrastive learning methods [21, 23, 55] exhibit limited performance due to their disregard for semantic relationships between classes, as they often treat all classes equally. This limitation hinders the model’s ability to learn discriminative and meaningful representations, particularly in imbalanced settings.

We observe that SKCL, which integrates multi-granularity knowledge for representation learning, surpasses ECS-SC [14], which focuses on selecting and combining easily confused tail samples. This advantage stems from leveraging the common-sense knowledge of large models to capture implicit semantic relationships between categories. In addition, we compare the performance of BCL [55] and SKCL at different imbalance factors in Fig. 4, and the results demonstrate that the proposed method achieves significant performance improvements across various imbalance factors. By constructing a multi-granularity prototype space, our method enables the model to learn a more semantically discriminative feature space, mitigating the representation learning deficiency caused by limited samples.

Table 3. Comparison results with state-of-the-art methods using ResNet-50 on the ImageNet-LT dataset, focusing on Top-1 accuracy(%).

Methods	Many	Medium	Few	All
Focal Loss [25]	64.3	37.1	8.2	43.7
τ-norm [19]	59.1	46.9	30.7	49.4
BALMS [35]	62.2	48.8	29.8	51.4
LWS [19]	60.2	47.2	30.3	49.9
Casual Model [38]	62.7	48.8	31.6	51.8
LADE [15]	62.3	49.3	31.2	51.9
DisAlign [48]	62.7	52.1	31.4	53.4
RIDE(2 experts) [43]	-	-	-	55.9
BCL [55]	67.2	53.9	36.5	56.7
ConCutMix [30]	70.7	56.6	39.8	59.7
Ours	**71.5**	**56.8**	**41.5**	**60.4**

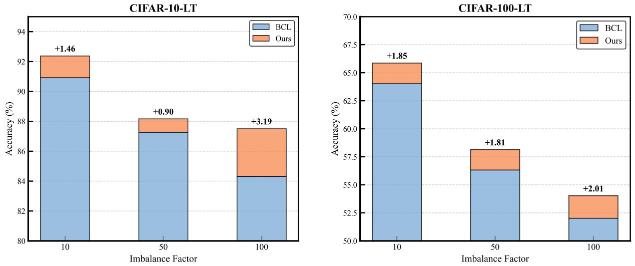
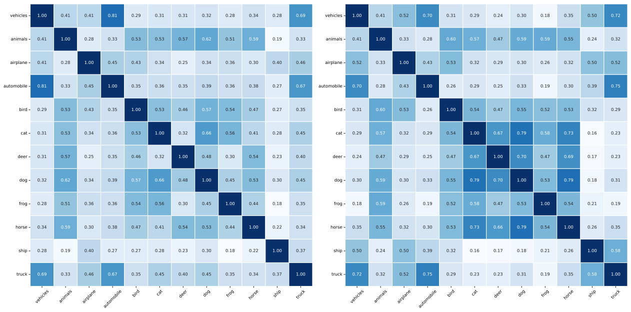


Figure 4. Comparison of Top-1 accuracy(%) with the baseline BCL on CIFAR-100-LT and CIFAR-10-LT datasets under different imbalance factors.

Furthermore, we report the accuracy of three groups of classes on CIFAR-100-LT with an imbalance factor of 100 to further verify the effectiveness of the proposed method. As shown in Tab. 2, our method achieves the highest Top-1 accuracy across all class groups. Notably, SKCL not only benefits tail classes by effectively utilizing semantic knowledge but also enhances feature learning for head classes, in contrast to methods that improve tail-class performance at the expense of head-class accuracy. Our method achieves state-of-the-art performance, improving accuracy for many-shot, medium-shot, and few-shot classes by 0.9%, 0.2%, and 1.4%, respectively.

ImageNet-LT. Experiments are also conducted with ResNet-50 on ImageNet-LT as shown in Tab. 3. The methods LWS [19], τ-norm [19], and DisAlign [48] follow a two-stage learning strategy, focusing on fine-tuning the classifier in the second stage, while overlooking the impact of tail-class data scarcity during the representation learning stage. Compared to the recent strong baseline BCL [55], SKCL achieves a performance gain of 4.3% for many-shot classes, 2.9% for medium-shot classes, and a notable 5.0% for few-shot classes, leading to an overall performance im-



(a) Label Semantics (b) LLM-based Label Semantic

Figure 5. Comparisons of similarity matrices constructed using original label semantics and LLM-based label semantics on CIFAR-10-LT.

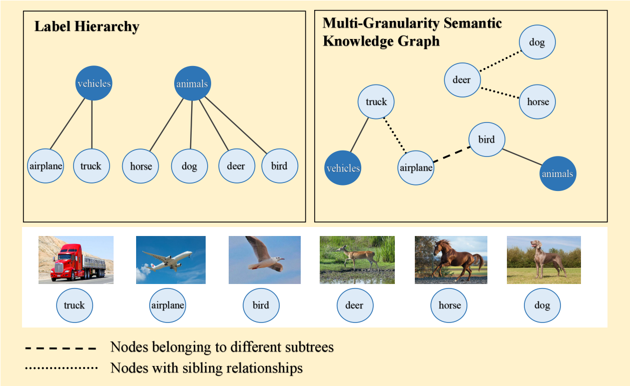


Figure 6. Comparison between a portion of the multi-granularity semantic knowledge graph and the label hierarchy on CIFAR-10-LT.

provement of 3.7%. The recently proposed data augmentation method, ConCutMix [30], constructs augmented samples with semantically consistent labels to improve long-tailed recognition. In contrast, SKCL further extracts semantic knowledge without requiring significant additional computation, leading to an overall performance improvement of 0.7%. This demonstrates that leveraging the rich common-sense knowledge of LLMs to expand semantic labels is effective. The constructed multi-granularity prototype space enables knowledge transfer across levels and subtrees between similar categories.

4.4. Ablation Study

To demonstrate the effectiveness of multi-granularity knowledge transfer, we conducted an ablation study, with all experiments performed on CIFAR-100-LT at an imbalance ratio of 100. We use a strategy that transfers knowledge exclusively between coarse-grained and fine-grained levels (C-F) as a baseline. Meanwhile, F-F serves as a baseline that focuses on the similarity within the same level for knowledge transfer. As shown in Tab. 4, limiting knowledge transfer to only the corresponding coarse-grained category does not effectively enhance the performance of tail classes, while constraining transfer learning within the same hierarchical level prevents head classes from benefiting. In contrast, incorporating both horizontal and vertical knowledge transfer leads to significant performance improvements, simultaneously enriching the representations of both head and tail classes.

Table 4. Ablation study on the CIFAR-100-LT dataset with an imbalance factor of 100, using different knowledge transfer strategies.

C-F	F-F	Many	Medium	Few	All
		67.2	53.1	32.9	51.9
✓		68.1	53.9	34.6	53.1
	✓	67.5	53.3	36.5	53.2
✓	✓	**68.3**	**54.1**	**37.2**	**54.0**

4.5. Semantic Similarity Analysis

To verify the effectiveness of LLMs in semantic expansion, we visualize the semantic similarity matrices in Fig. 5, which are generated from original and LLM-based label semantics. In the LLM-based matrix, the most similar category to "airplane" is "bird," which is not in the same subtree. This highlights that the original label semantics matrix fails to reveal such relationships, despite the fact that both "airplane" and "bird" share the characteristic of flight. Fig. 6 further illustrates that the multi-granularity semantic knowledge graph can effectively uncover latent class relationships (*i.e.* "bird" and "airplane"). Even though the categories share a sibling relationship, they are still connected based on semantic similarity to the most similar ones (*i.e.* "deer" and "horse"). Unlike the fixed label hierarchy, the multi-granularity semantic knowledge graph can adapt based on the semantic knowledge of categories, which is crucial for representation learning.

5. Conclusion

In this paper, we introduce semantic knowledge-driven contrastive learning (SKCL) to address the challenges of long-tailed classification. Inspired by the implicit knowledge in large language models, SKCL overcomes the limitations of label hierarchy by utilizing an extended semantic list to capture category similarities, thereby constructing a multi-granularity semantic knowledge graph. Furthermore, a prototype space is constructed to encourage the model to dynamically identify relevant classes for learning, thereby facilitating multi-granularity knowledge transfer. Extensive experiments on long-tailed benchmark datasets validate the effectiveness of the proposed method.

Acknowledgement This research was supported in part by the National Natural Science Foundation of China under grant No. 62476194, No. U23B2049 and No. 62476195, the Tianjin Young Scientific and Technological Talents Project under grant QN20230305, and the Tianjin Science and Technology Plan Project under grants 24YDT-PJC00150 and 24JCYBJC00950.

References

[1] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 2

[2] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 32, 2019. 6, 7

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. 2

[4] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *ECCV*, pages 95–110. Springer, 2020. 6, 7

[5] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *ECCV*, pages 694–710. Springer, 2020. 2

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 6

[7] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3695–3706, 2022. 6, 7

[8] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277, 2019. 1, 2, 6, 7

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. 3

[10] Lin Geng Foo, Hossein Rahmani, and Jun Liu. Ai-generated content (aigc) for various data modalities: A survey. *arXiv preprint arXiv:2308.14177*, 2023. 3

[11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 3

[12] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017. 1, 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[14] Wenwei He, Junyan Xu, Jie Shi, and Hong Zhao. Ecs-sc: Long-tailed classification via data augmentation based on easily confused sample selection and combination. *Expert Systems with Applications*, 246:123138, 2024. 2, 6, 7

[15] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, pages 6626–6636, 2021. 1, 2, 7

[16] Minki Jeong and Changick Kim. Supervised contrastive learning on blended images for long-tailed recognition. *arXiv preprint arXiv:2211.11938*, 2022. 6, 7

[17] Junjie Jiang, Zaixing He, Shuyou Zhang, Xinyue Zhao, and Jianrong Tan. Learning to transfer focus of graph neural network for scene graph parsing. *Pattern Recognition*, 112: 107707, 2021. 4

[18] Di Jin, Jingyi Cao, Xiaobao Wang, Bingdao Feng, Dongxiao He, Longbiao Wang, and Jianwu Dang. Rethinking contrastive learning in graph anomaly detection: A clean-view perspective. *arXiv preprint arXiv:2505.18002*, 2025. 1

[19] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 2, 6, 7

[20] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *ICLR*, 2020. 6, 7

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 33:18661–18673, 2020. 3, 7

[22] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *CVPR*, pages 5212–5221, 2021. 6, 7

[23] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *CVPR*, pages 6918–6928, 2022. 3, 7

[24] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *CVPR*, pages 6918–6928, 2022. 6, 7

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 6, 7

[26] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *CVPR*, 2020. 2

[27] Yuting Liu, Liu Yang, and Yu Wang. Hierarchical fine-grained visual classification leveraging consistent hierarchical knowledge. In *ECML-PKDD*, pages 279–295. Springer, 2024. 1

[28] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pages 2537–2546, 2019. 6

[29] Haolin Pan, Yong Guo, Mianjie Yu, and Jian Chen. Enhanced long-tailed recognition with contrastive cutmix augmentation. *IEEE Transactions on Image Processing*, 2024. 2

[30] Haolin Pan, Yong Guo, Mianjie Yu, and Jian Chen. Enhanced long-tailed recognition with contrastive cutmix augmentation. *IEEE Transactions on Image Processing*, 2024. 6, 7, 8

[31] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *CVPR*, pages 6887–6896, 2022. 1, 2

[32] Junran Peng, Xingyuan Bu, Ming Sun, Zhaoxiang Zhang, Tieniu Tan, and Junjie Yan. Large-scale object detection in the wild from imbalanced multi-labels. In *CVPR*, pages 9709–9718, 2020. 2

[33] Ziqi Qiu, Jianxing Yu, Yufeng Zhang, Hanjiang Lai, Yanghui Rao, Qinliang Su, and Jian Yin. Detecting emotional incongruity of sarcasm by commonsense reasoning. In *COLING*, pages 9062–9073, 2025. 1

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmLR, 2021. 3

[35] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *NeurIPS*, 33:4175–4186, 2020. 7

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 6

[37] Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *ICCV*, pages 9495–9504, 2021. 6, 7

[38] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *NeurIPS*, 33:1513–1524, 2020. 6, 7

[39] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, pages 2495–2504, 2021. 5

[40] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *CVPR*, pages 943–952, 2021. 3, 6, 7

[41] Pengkun Wang, Zhe Zhao, HaiBin Wen, Fanfu Wang, Binwu Wang, Qingfu Zhang, and Yang Wang. Llm-autoda: Large language model-driven automatic data augmentation for long-tailed problems. In *NeurIPS*, pages 64915–64941. Curran Associates, Inc., 2024. 3

[42] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *NeurIPS*, 33:5776–5788, 2020. 4

[43] 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. 6, 7

[44] Zhengzhuo Xu, Zenghao Chai, and Chun Yuan. Towards calibrated model for long-tailed visual recognition from prior perspective. *NeurIPS*, 34:7139–7152, 2021. 6, 7

[45] Zhikang Xu, Xiaodong Yue, Ying Lv, Wei Liu, and Zihao Li. Trusted fine-grained image classification through hierarchical evidence fusion. In *AAAI*, pages 10657–10665, 2023. 1

[46] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *NeurIPS*, 33:19290–19301, 2020. 3, 6, 7

[47] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 2, 3, 4

[48] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, pages 2361–2370, 2021. 7

[49] Wenqiao Zhang, Changshuo Liu, Lingze Zeng, Bengchin Ooi, Siliang Tang, and Yueting Zhuang. Learning in imperfect environment: Multi-label classification with long-tailed distribution and partial labels. In *ICCV*, pages 1423–1432, 2023. 1

[50] Hong Zhao, Zhengyu Li, Wenwei He, and Yan Zhao. Hierarchical convolutional neural network with knowledge complementation for long-tailed classification. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–22, 2024. 2, 6, 7

[51] Qihao Zhao, Yalun Dai, Hao Li, Wei Hu, Fan Zhang, and Jun Liu. Ltgc: Long-tail recognition via leveraging llms-driven generated content. In *CVPR*, pages 19510–19520, 2024. 3

[52] Wei Zhao and Hong Zhao. Hierarchical long-tailed classification based on multi-granularity knowledge transfer driven by multi-scale feature fusion. *Pattern Recognition*, 145: 109842, 2024. 2, 6, 7

[53] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *CVPR*, pages 16489–16498, 2021. 2

[54] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, pages 9719–9728, 2020. 6, 7

[55] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *CVPR*, pages 6908–6917, 2022. 3, 6, 7