

# Hierarchical Label Auto-Labeling and Relationship Constraints for Multi-Granularity Image Classification

Changwang Mei<sup>[0009-0008-4812-7293]</sup>, Xindong You, Shangzhi Teng<sup>\*[0000-0001-7098-9932]</sup>, and Xueqiang LYU

Beijing Key Laboratory of Internet Culture Digital Dissemination, Beijing Information Science and Technology University, Beijing 100101, China  
{2021020593, xindongyou, tengshangzhi, lxq}@bistu.edu.cn

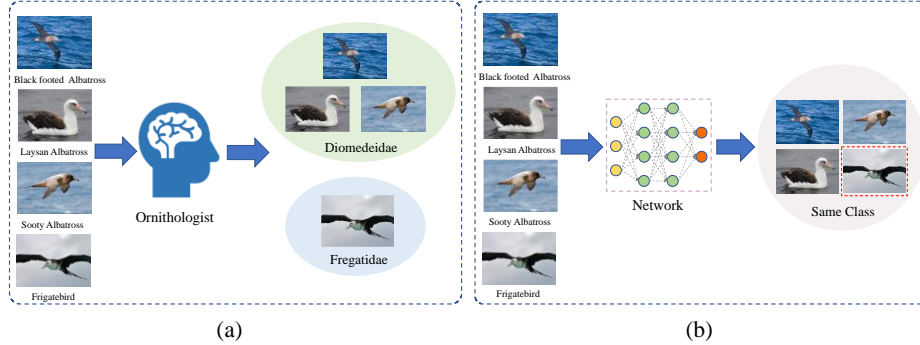
**Abstract.** Hierarchical multi-Granularity classification (HMC) aims to assign each object a label with multiple granularities from coarse to fine, focusing on the hierarchical structure of the label encoding. However, obtaining multi-granularity image labels through extensive manual labeling by experts is both costly and impractical for large-scale Fine-grained visual classification (FGVC) datasets and new scenarios. In this paper, we propose a hierarchical label auto-labeling clustering algorithm (HLA) to automatically generate hierarchical multi-granularity image labels. Additionally, we introduce a hierarchical constraint loss (HCL) and propose a hierarchical prediction constraint loss (HPCL) to constrain the relationship between different hierarchies. Extensive experiments on three commonly used FGVC datasets demonstrate that the proposed HLA can obtain similar performance with manual label method on CUB-200-2011, FGVC-Aircraft and Stanford Cars datasets. The introduced HCL and HPCL achieves promising performance on multi-granularity image classification datasets. Meanwhile, the consistent improvement on all object re-identification tasks demonstrates the effectiveness of our method.

**Keywords:** Hierarchical Multi-Granularity Classification, Fine-Grained Visual Classification, Automatic Labeling.

## 1 Introduction

Fine-grained visual classification (FGVC) aims to retrieve and recognize images belonging to multiple subordinate classes of a superclass. Hierarchical multi-granularity classification (HMC) is a classification task for hierarchical structures. Multi-granularity image classification (MIC) is an HMC task in the image area. Previous studies [5, 29] have demonstrated that HMC can improve the performance of FGVC tasks. However, these studies necessitate domain-specific experts to manually assign hierarchical labels, which is expensive and impractical for large-scale FGVC and MIC datasets and new scenarios. The clustering algorithm emerges as the most used auto-labeling approach. Both the K-means algorithm [14] and Hierarchical clustering [1] demand prior knowledge of the number of clusters in each hierarchy, and Hierarchical clustering has high computational complexity.

\* Corresponding author



**Fig. 1.** (a) Illustration of ornithologists grouping fine-grained labels. (b) Illustration of network grouping fine-grained labels.

Compared with traditional image classification tasks, FGVC is more challenging due to the high similarity in appearance among subordinate classes. Current FGVC methods often utilize attention mechanisms or design specific loss functions to optimize feature representation [3, 30, 32]. Although these works have achieved remarkable success, they rely on discrete labels that fail to adequately exploit the similarity relationships between them. To tackle this problem, some works [4, 5, 20] use hierarchical multi-granularity labels to enhance the fine-grained object features. Chang et al. [4] integrate fine-grained features with coarse-grained label prediction and restrict the gradient flow to update parameters within each classification head. Labels of different granularity will become distinct learning directions for the network. A robust multi-granularity image classification (MIC) model should facilitate interaction between information at different granularities through hierarchical classification.

Recent studies in HMC have bifurcated into two primary domains: Based on convolutional neural networks (CNNs) study the label hierarchy mapping into the network structure [8, 13] and design of hierarchically constrained loss functions [29]. Wang [29] uses a linear combination of losses to fuse features from different layers. HRN [5] uses residual connections to convert coarse-hierarchy features to current features. However, these approaches lack constraints on the prediction results between hierarchies when investigating the relationships between hierarchies.

In this paper, for the automatic labeling of hierarchical labels, we instruct the network to construct hierarchical pseudo-labels from fine-grained to coarse-grained within the learned distribution of fine-grained features. We propose a hierarchical label auto-labeling clustering algorithm (HLA), specifically designed for the automatic labeling of hierarchical labels. Image features can be automatically divided into hierarchical structures by emulating experts. **Fig. 1** shows the difference between HLA and the manual labeling of the multi-granularity datasets. In **Fig. 1**, (a) Illustration of ornithologists grouping fine-grained labels. (b) Illustration of network grouping fine-grained labels., based on domain knowledge of birds, ornithologists' group ["Black footed Albatross", "Laysan Albatross", "Sooty Albatross"] in the family "Diomedecidae" and "Frigatebirds" in the family "Fregatidae". Conversely, in **Fig. 1**(b), the neural network is

devoid of explicit prior knowledge about families of birds. Based on the inherent representation relationship of the bird image, the network groups ["Frigatebird"] in the same class as ["Black footed Albatross", "Laysan Albatross", "Sooty Albatross"].

We investigate the constraint relationships between hierarchies. We introduce hierarchical constrained loss (HCL) to constrain the relationship between predicted values of classes in the upper and lower hierarchies. Additionally, we propose a hierarchical prediction constrained loss (HPCL), which utilizes a class prediction value at the upper hierarchy to constrain the output at the lower hierarchy. HPCL aims to prevent serious prediction bias in network learning. Specifically, the lower hierarchy considers the upper hierarchy's prediction more when generating the prediction results.

In summary, the primary contributions of this paper are as follows:

- We propose a hierarchical label auto-labeling clustering algorithm (HLA) to automatically generate hierarchical multi-granularity labels. Extensive experiments on three widely used datasets for the MIC task demonstrate our auto-labeling can significantly reduce labor and time costs.
- We introduce hierarchical constrained loss (HCL) and propose hierarchical prediction constrained loss (HPCL) to constrain and enhance the relationship between different hierarchical features.
- Extensive experiments on three widely used FGVC datasets illustrate that HCL and HPCL achieve promising performance. Compared to manual labeling method, our HLA exhibit better convenience and reliability. Furthermore, by applying our method to object re-identification (ReID) baseline models, we demonstrate our method significantly improves the baseline.

## 2 Related work

### 2.1 Automatic labeling

The most widely used auto-labeling approach is the clustering algorithm. K-means algorithm [14] obtains the clustering results by an iterative method. It results in different results for each clustering. The number of clusters needs to be specified in advance. Hierarchical clustering [1] performs clustering based on distance rules, which can reveal the hierarchical relationships of classes. However, hierarchical clustering has a high computational complexity. The number of clusters in each layer still needs to be known in advance when constructing the label hierarchy. Therefore, we propose a hierarchical label auto-labeling clustering algorithm (HLA) to obtain the predicted label order for each class. After quickly locking the range of the clusters number, the optimal number of clusters is determined by subsequent tests, which is used to auto-labeling the hierarchical labels.

### 2.2 Fine-grained visual classification

Recent works [4, 6, 25, 34] applied hierarchical label structures to FGVC. Zhang [34] proposed a new triplet loss between different coarse-grained classes, the same coarse-

grained class but different fine-grained classes, and the same fine-grained class. Shi [25] proposed a generalized large-margin loss that makes subclasses of the same coarse-grained class more similar than subclasses of different coarse-grained classes in the feature space. Chen et al. [6] use coarse-level prediction score vectors as a prior knowledge to learn feature representations on finer levels. Chang et al. [4] fuse fine-grained features with coarse-grained label prediction and restrict the gradient flow to update the parameters within each classification head. All these methods enhance feature representation at the same hierarchy, but they ignore the relationship between hierarchical labels. For instance, parent class information should contain the child class information, and the child class information should inherit the relevant properties of the parent class.

### 2.3 Hierarchical multi-granularity classification

HMC is a classification task for hierarchical structures. In image classification, HMC has been applied to diatomic image classification [9] and fine-grained image classification [4, 5]. In the recent studies, the research direction of HMC is divided into two main parts: Based on deep neural networks (DNNs) study the label hierarchy mapping into the network structure [8, 13] and design of hierarchically constrained loss functions. In HMC with local multi-layer perceptron (HMC-LMLP) [2], each MLP network corresponds to a hierarchical level. The input of each MLP network is the output of the previously trained MLP. This process is carried out from the first level to the last level. HMC Network (HMCN) [31] proposes a combination of local and global information to solve the HMC problem. Each level corresponds to a local output layer. The global output layer captures information across the entire network. Then, the entire local output is converged with the global output to generate a final consistent prediction.

In HMC-LMLP and HMCN network structures, the labels between the levels are independent of each other, and there is no semantic information interaction between the levels. Coherent HMC neural network (C-HMCNN) [13] modifies the cross-entropy loss between the levels to constrain the relationship between them. According to C-HMCNN, if a sample is predicted to belong to a certain class, it also belongs to the parent class of that class.

Fan [18] and Zhao et al. [35] defined the correlation between levels in a tree classifier. We propose a simple and effective inter-level feature fusion model. Concretely, the finer-level subclass features are fused with the coarse-level superclass features to obtain more comprehensive information. Meanwhile, the training process can facilitate coarse-level superclass feature learning. In addition, our HCL and HPCL constrain the relationship between the predicted probability values of each hierarchy output.

### 3 Method

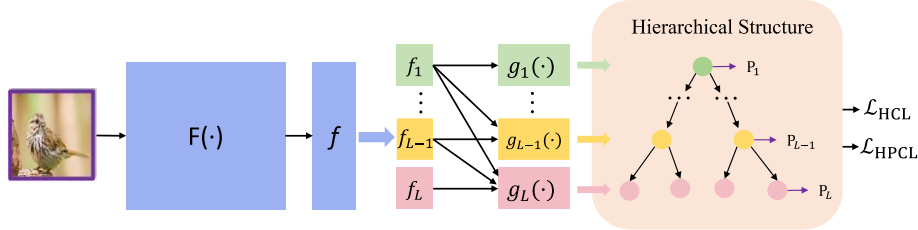
#### 3.1 Hierarchical Dataset Auto-Labeling

To save the cost of hierarchical dataset labeling, we propose a hierarchical label auto-labeling clustering algorithm (HLA) to automatically generate hierarchical multi-granularity labels. Once the network obtains the prediction results for all fine-grained classes, HLA constructs hierarchical pseudo-labels from fine to coarse based on this prediction. The specific methods are as follows.

Firstly, we use the CNN-based network which is pre-trained on ImageNet-1K to train the original fine-grained image training set. Next, we load the model training weights into the parameters of the network and retest the training dataset.

Then, we save the prediction results of each class to the prediction labels order table. In each row of the prediction labels order table, the larger the predicted value of the class probability output, the higher the ranking of the prediction labels. Inspired by  $k$ -reciprocal nearest neighbors, we intercept the results after Top 2 for  $k$ -nearest neighbor clustering. If the features of two images are  $k$ -nearest neighbors to each other, then they should be clustered into one class.

Finally, with HLA clustering algorithm (the detail is described in Section 3.2), we can cluster this  $k$ -reciprocal nearest neighbor labels to obtain the hierarchical multi-granularity labels. For example, if we set the number of Hierarchies to be 3, then we must use HLA twice to construct the hierarchical labels, i.e., Hierarchy1, Hierarchy2 and Species, where "Species" is usually denoted as the original single-label of the dataset. Therefore, the two  $k$  values of input in HLA clustering algorithm are different. The second input  $k$  value must be smaller than the first input  $k$  value.



**Fig. 2.** A schematic illustration of our model with multi-granularity class probability outputs.

#### 3.2 HLA Clustering Algorithm

The input is prediction label order list  $List\_Preds$ , and class number  $N$ . The output is hierarchical label set  $S$ . We obtain  $k\_sets$ , a list of reciprocal nearest neighbors, from  $List\_Preds$  according to  $k$ . We define a list of length  $N$  called  $Tag$  ( $Tag = [1] * N$ ). It records whether each class is accessed, where 0 is accessed and 1 is not accessed. When the 1 is still in  $Tag$ , we create a new queue  $q$ . Then, we define  $empty\_n$  records the class when the  $q$  is empty and  $class\_idx$  records the hierarchical label number (The

initial value of  $class\_idx$  is 1). Next, we find the  $k$ -reciprocal nearest neighbor prediction labels for class  $i$  by traversing the class  $N$  and store them to  $q$ . Finally, since the queue has a first-in-first-out (FIFO) principle, we iterate over  $q$ . The class  $i$  is then stored in  $empty\_n$  and  $k\_sets$ . When  $q$  is empty,  $Tag[empty\_n] = 0$  and  $S[N + class\_idx] = k\_sets$ . When  $S$  stores  $k\_sets$  for each time, the value of  $class\_idx$  is incremented by 1 and  $k\_sets = []$ . Repeat until there are no number 1 in  $Tag$ . Ultimately, the hierarchical label set  $S$  is obtained.

### 3.3 Loss Function

As shown in **Fig. 2**, given any CNN-based network backbone  $F(\cdot)$ , we feed image  $x$  as input to extract its feature embedding  $f = F(x)$ . Our goal is then to correctly predict classes across  $L$  independent classifiers,  $g_1(\cdot)$ ,  $g_2(\cdot)$ , ...,  $g_l(\cdot)$ , ...,  $g_L(\cdot)$  based on  $f$ , i.e.,  $\hat{y}^l = y^l$ , where  $L$  indicates the number of hierarchical granularity,  $\hat{y}^l = g_l(f)$  and  $y^l$  denotes the correct class of the  $l$ th hierarchy. Our optimization objective is  $L$  independent cross-entropy loss  $\sum_{l=1}^L L_{CE}(\hat{y}^l, y^l)$ .

Inspired by the hierarchical node relationships of decision trees [26], we introduce the HCL to enhance the robustness of the hierarchical feature. Specifically, as shown in **Fig. 2**, in the hierarchical structure, the prediction probability value  $P_l$  of the finer-hierarchy subclasses should be lower than the coarse-hierarchy superclasses prediction probability values  $P_1, P_2, \dots, P_{l-1}$ . HCL is defined as:

$$\mathcal{L}_{HCL} = \sum_{i=2}^L \sum_{j=1}^i \frac{1}{2} \max\{0, P_i - P_j\}^2 \quad (1)$$

In addition, there is a certain constraint on the predicted values between the hierarchies. Specifically, in the same hierarchy  $l$ , when the predicted value  $P_l^S$  of class  $S$  is maximal and the predicted label is correct, the set of subclasses of the next hierarchy  $l+1$  corresponding to class  $S$  is  $s = \{s_1, s_2, \dots, s_m\}$  and the maximum value of  $s$  is  $\max\{s\}$ , where  $m$  denotes the number of subclasses that belong to  $S$ . In hierarchy  $l+1$ , except for  $s$ , the value of  $\max\{s\}$  must be higher than the predicted values of the other classes  $q = \{q_1, q_2, \dots, q_n\}$ , where  $n$  denotes the number of classes in the hierarchy  $l+1$  that do not belong to  $s$ . Thus, the HPCL is defined as:

$$\mathcal{L}_{HPCL} = \sum_{i=1}^n \max\{0, P_{q_i} - \max\{s\}\} \quad (2)$$

The total loss function Loss can be defined as:

$$Loss = \sum_{l=1}^L L_{CE}(\hat{y}^l, y^l) + \lambda \mathcal{L}_{HCL} + \mathcal{L}_{HPCL} \quad (3)$$

## 4 Experiments

In our experiments, we resized the input image to  $448 \times 448$ . We train each experiment for 200 epochs. In training phrase, data augmentation is performed via Random Crop and Random Horizontal Flip. In testing phrase, data augmentation is center cropping. We use Resnet-50 as the backbone network, and we concatenate upper hierarchical

features to lower hierarchies to enhance the lower hierarchical feature representations as baseline, as shown in **Fig. 2**. We use stochastic gradient descent (SGD) with a momentum of 0.9, weight decay of 0.0005 to optimize our model. The batch size is set to 8. The learning rate of the network is initialized as 0.0002. The learning rate is adjusted by cosine annealing strategy [21]. All codes are implemented using the PyTorch library and run on a single NVIDIA v40 GPU.

#### 4.1 Datasets and metrics.

We evaluate our proposed method on three widely used FGVC datasets. A taxonomy for manually constructing label hierarchies by tracing their parent nodes (superclasses) in Wikipedia pages. CUB-200-2011 (CUB) [27] is a dataset that contains 11877 images belonging to 200 bird species, including a three-hierarchy label hierarchy with 13 orders, 38 families, and 200 species. FGVC-Aircraft (Air) [23] is an aircraft dataset with 10000 images covering 100 model variants, including a three-hierarchy label hierarchy with 30 makers, 70 families, and 100 models. Stanford Cars (Car) [17] contains 8144 car images categorized by 196 car makers, including a two-hierarchy label hierarchy with 9 car types and 196 specific models.

We follow the standard train/test splits in existing works. We do not use any bounding box annotations in all our experiments.

We use two evaluation metrics. The first criterion follows the FGVC convention and uses fine-grained accuracy to evaluate the designed model. The second evaluation metric is the Top-1 precision of all hierarchical classes. Then, the hierarchical classification performance can be evaluated by the weighted average precision (wAP) of all hierarchical classes:

$$\text{wAP} = \sum_{l=1}^L \frac{\text{class\_num}_l}{\sum_{k=1}^L \text{class\_num}_k} P_l \quad (4)$$

where  $\text{class\_num}_l$  and  $P_l$  denote the number of class and Top-1 classification accuracy at hierarchy  $l$ , respectively. The finer the classification, the greater its weight in performance evaluation. The best results for each indicator are shown in bold.

#### 4.2 Effects of HCL and HPCL

To verify the effectiveness of HCL and HPCL, we conduct experiments on three manually labeled multi-granularity datasets. As shown in **Table 1** and **Table 2**, compared to existing work, Ours(HCL+HPCL) achieves the best performance in terms of both fine-grained accuracy and wAP across all three datasets. In addition, in comparison to Baseline, HCL and HPCL each obtain significant performance improvements. Experiments demonstrate that HCL and HPCL are effective in constraining the relationships between hierarchies and enhancing the performance of the network.

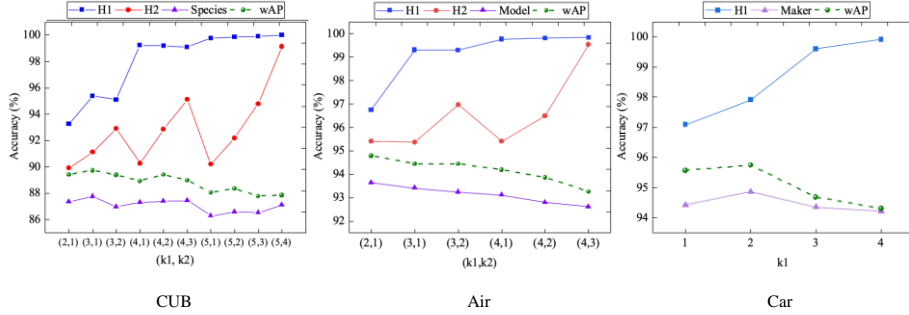
**Table 1.** Performance of our methods in the MIC task on manually labeled hierarchical CUB dataset.

Method	CUB			
	Order	Family	Specie	wAP
HMC-LMLP [2]	98.45	94.24	79.60	82.79
HMCN [31]	97.29	93.15	79.15	82.69
C-HMCNN [13]	98.48	94.63	81.58	84.43
FGN [4]	97.76	94.17	85.56	87.50
HRN [5]	98.67	95.51	86.60	88.57
Baseline	98.48	95.65	85.50	87.71
Ours(HCL)	98.86	95.74	87.53	89.36
Ours(HPCL)	98.81	95.70	87.32	89.18
Ours(HCL+HPCL)	<b>98.93</b>	<b>95.89</b>	<b>88.02</b>	<b>89.78</b>

**Table 2.** Performance of our methods in the MIC task on manually labeled hierarchical Air and Car datasets.

Method	Air				Car		
	Maker	Family	Model	wAP	Type	Maker	wAP
HMC-LMLP	97.09	94.39	90.25	92.72	96.98	87.65	88.06
HMCN	96.07	92.56	87.19	90.40	95.21	88.71	88.99
C-HMCNN	<b>97.45</b>	95.41	91.69	93.86	96.75	90.64	90.91
FGN	96.88	95.28	91.92	93.84	96.40	93.65	93.77
HRN	<b>97.45</b>	<b>95.79</b>	92.58	94.43	<b>97.41</b>	94.03	94.18
Baseline	95.90	93.82	91.54	92.99	96.24	93.49	93.61
Ours(HCL)	96.48	94.41	92.73	93.88	96.54	94.38	94.47
Ours(HPCL)	96.41	94.43	92.54	93.78	96.55	93.98	94.09
Ours(HCL+HPCL)	97.05	95.38	<b>93.33</b>	<b>94.61</b>	97.39	<b>94.96</b>	<b>95.07</b>





**Fig. 3.** The effect of different  $k$  values on the CUB, Air and Car datasets. The H1 of CUB, Air and Car denotes Order, Maker and Type respectively. The H2 of CUB and Air denote their respective Family.

**Table 3.** Performance of different proportions for manually labeled training set of CUB dataset.

Proportion	CUB			
	Order	Family	Specie	wAP
0%	-	-	87.75	-
25%	98.53	94.80	87.15	88.90
50%	98.79	95.71	87.53	89.35
75%	98.70	95.85	87.96	89.71
100%	<b>98.84</b>	<b>95.88</b>	<b>88.04</b>	<b>89.79</b>

**Table 4.** Performance of different proportions for manually labeled training set of Air and Car datasets.

Proportion	Air				Car		
	Maker	Family	Model	wAP	Type	Maker	wAP
0%	-	-	93.64	-	-	94.87	-
25%	95.97	95.45	93.53	94.57	94.31	94.69	94.67
50%	96.52	95.30	93.67	94.67	96.42	94.99	95.05
75%	<b>97.10</b>	<b>95.65</b>	93.76	94.92	96.68	95.03	95.10
100%	97.02	95.59	<b>93.85</b>	<b>94.94</b>	<b>96.75</b>	<b>95.12</b>	<b>95.19</b>

### 4.3 Validity of HLA

We use HLA on the three datasets to evaluate the performance. The value of  $k$  varies for each hierarchy within each dataset. The results are shown in **Fig. 3**, for CUB dataset, both fine-grained accuracy and wAP are the best when  $k_1=3$  and  $k_2=1$ , with 87.75% and 89.73%, respectively. Similarly, for Air dataset, both fine-grained accuracy and wAP are best when  $k_1=2$  and  $k_2=1$ , with 93.64% and 94.78%, respectively. For Car dataset, when  $k_1=2$ , both fine-grained accuracy and wAP are the best, with 94.87% and 95.74%, respectively.

We further study the ability of our HLA in reducing labor and time costs. We choose the best wAP accuracy model in **Fig. 3** as our pre-training model, and then fine-tune it on the manually labeled training set at various proportions (e.g., 25%, 50%, 75%, 100%). Then test the fine-tuned model on the manually labeled test set. As shown in **Table 3** and **Table 4**, on the Air and Car datasets, the performance of the manually labeled training set using 50% and 75%, respectively, is better than the performance of Ours(HCL+HPCL) of **Table 1** and **Table 2**. Besides, On the CUB dataset, the performance of the manually labeled training set using 75% is close to the performance of Ours(HCL+HPCL) of **Table 1** and **Table 2**. Experiments demonstrate that HLA can diminish the necessity for manually labeled hierarchical labels in hierarchical multi-granularity classification tasks.

**Table 5.** Performance comparisons on traditional FGVC setting with single fine-grained label output.

Method	Precision (%)		
	CUB	Air	Car
NTS-Net [33]	87.5	91.4	93.9
PC [12]	86.9	89.2	92.9
DCL [7]	87.8	93.0	94.5
S3N [10]	88.5	92.8	94.7
ACNet [16]	88.1	92.4	94.6
SPS [15]	88.7	92.7	94.9
CHRF [20]	89.4	93.6	95.2
AAM [28]	88.6	93.5	94.0
PMG [11]	89.6	93.4	95.1
Ours	88.0	93.9	95.2
Ours_PMG	<b>89.9</b>	<b>94.1</b>	<b>95.6</b>

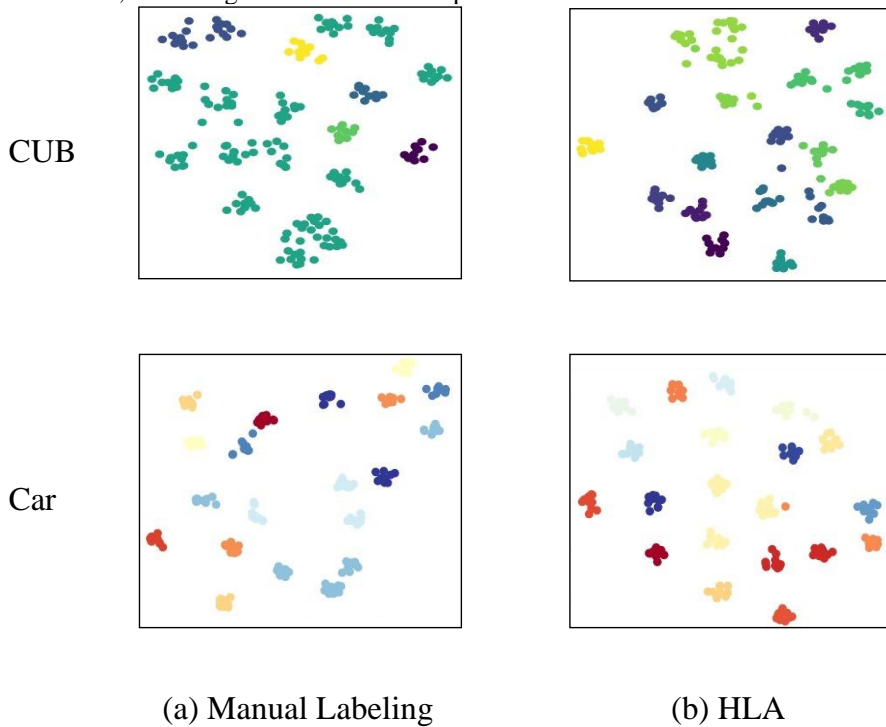
#### 4.4 Comparison with FGVC works

We compare our methods with recent FGVC works. The results are shown in **Table 5**. HCL and HPCL effectively constrain the relationships between hierarchies to enhance the fine-grained feature representation and improve the fine-grained performance. Furthermore, PMG [11] introduces a progressive training strategy that effectively fuses features from different granularities. Consequently, we apply the proposed HCL and HPCL to PMG to obtain Ours\_PMG. Experiments show that Ours\_PMG further improves performance on three datasets.

#### 4.5 Visualization

We further show visualizations to demonstrate the validity of our methods.

**Fig. 4** visualizes image feature distribution of 200 images (20 species for CUB dataset, 20 makers for Car dataset and 10 images per class) randomly sampled from the test dataset with t-SNE [22]. Features extracted by manual labeling approach and our HLA are compared. Our HLA can make the features of the same granularity category closer to each other, indicating the better feature representation.



**Fig. 4.** Visualization of image feature distribution. The first row represents the distribution of features where the hierarchy is Order in the CUB dataset. The second row represents the distribution of features where the hierarchy is Type in the Car dataset. (a) The method of manual labeling. (b) The approach of our HLA.

**Table 6.** Performance comparisons on object re-identification tasks.

Method	Market-1501		DukeMTMC-reID		VeRi-776	
	mAP	CMC@1	mAP	CMC@1	mAP	CMC@1
Baseline	87.5	95.0	78.7	89.0	79.3	96.5
Ours	<b>88.6</b>	<b>95.5</b>	<b>79.9</b>	<b>89.2</b>	<b>80.2</b>	<b>96.8</b>

## 5 Object Re-identification

To validate the generalizability of our proposed method, we conduct experiments on object re-identification (ReID) tasks (person re-identification and vehicle re-identification), where the datasets include Market-1501[36], DukeMTMC-reID [24] and VeRi-776 [19]. We employ the widely adopted metrics including mean Average Precision (mAP) and Rank1 (CMC@1) to quantify the ReID performance.

As shown in **Table 6**, HLA automatically generates a single hierarchy of coarse-grained pseudo-labels for object ReID datasets. Compared to the baseline, our method improves the mAP by about 1.1% on average. The experiments demonstrate that our proposed method can also improve the performance of object ReID. Especially in the context of datasets characterized by high similarity and potential confounding features, HLA effectively facilitates the clustering and differentiation of these features.

## 6 Conclusion

This paper addresses the challenge of automatically generating hierarchical multi-granularity labels for fine-grained visual classification at different granularities. We propose the hierarchical label auto-labeling (HLA) method, specifically designed to generate hierarchical multi-granularity labels. In the MIC tasks, HLA achieves the equivalent performance of full manual labeling while requiring a manual labeling effort of only 25% or less. This demonstrates HLA's effectiveness in mitigating the cost associated with manual labeling. Furthermore, we introduce the HCL and propose HPCL to constrain and enhance relationships between different hierarchical features. Extensive experimentation validates the effectiveness of our approach across FGVC, MIC, and Object ReID tasks. Furthermore, the categorization mentioned in this article is typically carried out in outdoor environments, generally using mobile devices for portability. How to ensure accuracy while simultaneously the model is worthy of further research.

## 7 Acknowledgements

This work was supported in part by Beijing Natural Science Foundation (4232025), the National Natural Science Foundation of China (62202061;62171043), and R&D Program of Beijing Municipal Education Commission (KM202311232002).

## 8 Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Bridges Jr, C.C.: Hierarchical cluster analysis. *Psychological reports* 18(3), 851–854 (1966)
2. Cerri, R., Barros, R.C., PLF de Carvalho, A.C., Jin, Y.: Reduction strategies for hierarchical multi-label classification in protein function prediction. *BMC bioinformatics* 17(1), 1–24 (2016)
3. Chang, D., Ding, Y., Xie, J., Bhunia, A.K., Li, X., Ma, Z., Wu, M., Guo, J., Song, Y.Z.: The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing* 29, 4683–4695 (2020)
4. Chang, D., Pang, K., Zheng, Y., Ma, Z., Song, Y.Z., Guo, J.: Your "flamingo" is my "bird": Fine-grained, or not. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11476–11485 (2021)
5. Chen, J., Wang, P., Liu, J., Qian, Y.: Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4858–4867 (2022)
6. Chen, T., Wu, W., Gao, Y., Dong, L., Luo, X., Lin, L.: Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In: *Proceedings of the 26th ACM international conference on Multimedia*. pp. 2023–2031 (2018)
7. Chen, Y., Bai, Y., Zhang, W., Mei, T.: Destruction and construction learning for fine-grained image recognition. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 5157–5166 (2019)
8. Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., Adam, H.: Large-scale object classification using label relation graphs. In: *European conference on computer vision*. pp. 48–64. Springer (2014)
9. Dimitrovski, I., Kocev, D., Loskovska, S., Džeroski, S.: Hierarchical classification of diatom images using ensembles of predictive clustering trees. *Ecological Informatics* 7(1), 19–29 (2012)
10. Ding, Y., Zhou, Y., Zhu, Y., Ye, Q., Jiao, J.: Selective sparse sampling for fine-grained image recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 6599–6608 (2019)
11. Du, R., Chang, D., Bhunia, A.K., Xie, J., Ma, Z., Song, Y.Z., Guo, J.: Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*. pp. 153–168. Springer (2020)
12. Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., Naik, N.: Pairwise confusion for fine-grained visual classification. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 70–86 (2018)
13. Giunchiglia, E., Lukasiewicz, T.: Coherent hierarchical multi-label classification networks. *Advances in Neural Information Processing Systems* 33, 9662–9673 (2020)
14. Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28(1), 100–108 (1979)

15. Huang, S., Wang, X., Tao, D.: Stochastic partial swap: Enhanced model generalization and interpretability for fine-grained recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 620–629 (2021)
16. Ji, R., Wen, L., Zhang, L., Du, D., Wu, Y., Zhao, C., Liu, X., Huang, F.: Attention convolutional binary neural tree for fine-grained visual categorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10468–10477 (2020)
17. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
18. Kuang, Z., Li, Z., Zhao, T., Fan, J.: Deep multi-task learning for large-scale image classification. In: 2017 IEEE Third International Conference on Multimedia Big Data (BigMM). pp. 310–317. IEEE (2017)
19. Liu, H., Tian, Y., Yang, Y., Pang, L., Huang, T.: Deep relative distance learning: Tell the difference between similar vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2167–2175 (2016)
20. Liu, Y., Zhou, L., Zhang, P., Bai, X., Gu, L., Yu, X., Zhou, J., Hancock, E.R.: Where to focus: Investigating hierarchical attention relationship for fine-grained visual classification. In: European Conference on Computer Vision. pp. 57–73. Springer (2022)
21. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
22. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of machine learning research* 9(11) (2008)
23. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
24. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. pp. 17–35. Springer (2016)
25. Shi, W., Gong, Y., Tao, X., Cheng, D., Zheng, N.: Fine-grained image classification using modified dcnn trained by cascaded softmax and generalized large-margin losses. *IEEE transactions on neural networks and learning systems* 30(3), 683–694 (2018)
26. Vens, C., Struyf, J., Schietgat, L., Džeroski, S., Blockeel, H.: Decision trees for hierarchical multi-label classification. *Machine learning* 73(2), 185–214 (2008)
27. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
28. Wang, X., Shi, J., Fujita, H., Zhao, Y.: Aggregate attention module for fine-grained image classification. *Journal of Ambient Intelligence and Humanized Computing* 14(7), 8335–8345 (2023)
29. Wang, Y., Liu, R., Lin, D., Chen, D., Li, P., Hu, Q., Chen, C.P.: Coarse-to-fine: progressive knowledge transfer-based multitask convolutional neural network for intelligent large-scale fault diagnosis. *IEEE Transactions on Neural Networks and Learning Systems* (2021)
30. Wang, Z., Wang, S., Yang, S., Li, H., Li, J., Li, Z.: Weakly supervised fine-grained image classification via gaussian mixture model oriented discriminative learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9749–9758 (2020)
31. Wehrmann, J., Cerri, R., Barros, R.: Hierarchical multi-label classification networks. In: International conference on machine learning. pp. 5075–5084. PMLR (2018)

32. Xu, S., Chang, D., Xie, J., Ma, Z.: Grad-cam guided channel-spatial attention module for fine-grained visual classification. In: 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2021)
33. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to navigate for fine-grained classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 420–435 (2018)
34. Zhang, X., Zhou, F., Lin, Y., Zhang, S.: Embedding label structures for fine-grained feature representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1114–1123 (2016)
35. Zhao, T., Zhang, B., He, M., Zhang, W., Zhou, N., Yu, J., Fan, J.: Embedding visual hierarchy with deep networks for large-scale visual recognition. *IEEE Transactions on Image Processing* 27(10), 4740–4755 (2018)
36. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. pp. 1116–1124 (2015)