Hierarchical multi-granularity classification (HMC) assigns hierarchical multi-granularity labels to each object and focuses on encoding the label hierarchy, e.g., [“Albatross”, “Laysan Albatross”] from coarse-to-fine levels. However, the definition of what is fine-grained is subjective, and the image quality may affect the identification. Thus, samples could be observed at any level of the hierarchy, e.g., [“Albatross’”] or [“Albatross”, “Laysan Albatross”], and examples discerned at coarse categories are often neglected in the conventional setting of HMC. In this paper, we study the HMC problem in which objects are labeled at any level of the hierarchy. The essential designs of the proposed method are derived from two motivations: (1) learning with objects labeled at various levels should transfer hierarchical knowledge between levels; (2) lower-level classes should inherit attributes related to upper-level superclasses. The proposed combinatorial loss maximizes the marginal probability of the observed ground truth label by aggregating information from related labels defined in the tree hierarchy. If the observed label is at the leaf level, the combinatorial loss further imposes the multi-class cross-entropy loss to increase the weight of fine-grained classification loss. Considering the hierarchical feature interaction, we propose a hierarchical residual network (HRN), in which granularity-specific features from parent levels acting as residual connections are added to features of children levels. Experiments on three commonly used datasets demonstrate the effectiveness of our approach compared to the state-of-the-art HMC approaches and fine-grained visual classification (FGVC) methods exploiting the label hierarchy.

1. Introduction

Traditional single-granularity classification usually assigns a single label to a given object from a set of mutually exclusive class labels. For instance, FGVC aims at distinguishing objects from different subordinate-level categories within a given object category, e.g., subcategories of birds [32], cars [16], aircraft [20]. However, the definition of what is fine-grained is subjective, and the image quality may affect the identification, as illustrated in Fig. 1. A bird can be discerned as Albatross or Laysan Albatross due to differences in domain knowledge. Moreover, a bird expert recognizes a bird as Albatross rather than Black-footed Albatross because of the occlusion of key parts. Airborne or satellite image resolutions often have large variations, causing objects to be recognized at different levels. These challenges increase the difficulty of constructing a dataset for single-granularity classification, while images annotated as coarse categories are also overlooked.

Compared to single-granularity classification, a more preferable solution is to employ hierarchical multi-granularity labels to describe an object, which provides more flexible options for annotators with different knowledge backgrounds [4]. HMC [15] aims to exploit hierarchical multi-granularity labels and embeds the label hierarchy in loss function or network architecture. Whereas conven-
tional HMC usually evaluates each sample with complete hierarchical labels from the coarsest to the finest granularity. A more robust HMC model should effectively utilize examples observed at various levels in the hierarchy, e.g., making use of bird images annotated as [“Albatross”] and [“Albatross”, “Laysan Albatross”].

In this paper, we study the HMC problem in which samples are labeled at any level of the hierarchy. We factorize this problem into two aspects: (1) how to effectively use instances labeled at different levels; (2) how to perform hierarchical feature interaction in the network architecture. For the first problem, we adopt a tree hierarchy that defines two kinds of semantic relationships between labels: parent-child correlations between levels and mutual exclusion at the same level. Inspired by the work of [7], if an instance is discerned at a label in the hierarchy, we maximize its marginal probability in the probability space constrained by the tree hierarchy. Such marginalization enjoys two benefits: learning with the coarse-level label could impact decisions of fine-grained subclasses while learning with the fine-level label aids the prediction of coarse-grained superclasses. Moreover, if the ground truth label is observed at the leaf level, we further impose the multi-class cross-entropy loss to enhance the discriminative power among fine-grained categories.

Another critical issue is to design appropriate hierarchical feature interaction that reflects the label hierarchy. A distinct characteristic of hierarchical categories is that from coarse-to-fine levels, fine-level classes not only have unique attributes but also inherit attributes related to coarse-level superclasses. Based on this property, we propose a hierarchical residual network (HRN) illustrated in Fig. 2. We first set up granularity-specific layers to disentangle hierarchical features from the trunk network. Then, these hierarchical features interact via residual connections [11–14, 19, 30, 34], i.e., features from parent levels acting as skip connections are added to features of children levels. Experiments on three commonly used FGVC datasets demonstrate the effectiveness of our approach compared to the state-of-the-art HMC approaches and FGVC methods exploiting hierarchical knowledge under two evaluation metrics [31].

2. Related Work

2.1. Hierarchical Multi-Granularity Classification

HMC problems naturally arise in many domains, such as text categorization [17, 22, 25] and functional genomics [1, 26, 31]. In text categorization, an increasing number of works [5, 21, 23, 29] leveraged the label hierarchy to improve accuracy. In image classification, HMC systems have been used to annotate medical images [8] and classify diatom images [9]. Based on deep neural networks (DNNs), the studies usually go along two paths: mapping the label hierarchy to network architectures [2, 3, 24, 33] or loss functions that impose the hierarchical constraints [7, 10]. HMC with local multi-layer perceptrons (HMC-LMLP) [3] proposed to train a chain of multi-layer perceptron (MLP) networks, each corresponding to a hierarchical level. The input of each MLP uses the output provided by the previously trained MLP to augment the feature vector of the instance. This supervised incremental greedy procedure continues until the last level of the hierarchy is reached. HMC network (HMCN) [33] comprised multiple local outputs, with one local output layer per hierarchical level of the class hierarchy plus a global output layer that captures the cumulative relationships forwarded across the entire network. All local outputs are then concatenated and pooled with the global output to generate a final consensual prediction. HMC-LMLP and HMCN embed label hierarchy in their network architecture. Their loss functions sum over binary cross-entropy losses from each hierarchical level, which assumes each label is independent of each other, causing the implicit hierarchical relations between two semantic labels to be ignored.

Another line of HMC works encodes the label hierarchy in loss functions by imposing the hierarchical constraints. Coherent HMC neural network (C-HMCNN) [10] revised the binary cross-entropy loss to satisfy the parent-child constraint. The revision ensures that no hierarchy violation happens, i.e., for any threshold, when C-HMCNN predicts a sample belonging to a class, this sample also belongs to its parent classes. Moreover, C-HMCNN can teach the network how to better make the prediction on the higher level classes using the prediction results on the lower level ones. While C-HMCNN only restricts the parent-child correlation, other kinds of semantic relations between hierarchical labels can be constructed using graphs. Deng et al. [7] formalized semantic connections between any two labels into a directed acyclic graph (DAG). They built a modified junction tree algorithm that contains multiple loops during message passing on the junction tree to compute the probabilistic classification loss defined on the DAG.

2.2. Fine-Grained Visual Classification

Since FGVC inherently forms a hierarchy with different levels of concept abstraction, many approaches [4, 6, 28, 35, 36] proposed to exploit the hierarchical label structure of FGVC. Zhang et al. [35] generalized the triplet loss by describing inequalities of the distance between images belonging to the same fine-grained class, different fine-grained classes but the same coarse class, and different coarse classes. Shi et al. [28] proposed a generalized large-margin loss that not only reduces between-class similarity and within-class variance of the learned features but also makes the subclasses belonging to the same coarse class be more similar than those belonging to different coarse
classes in the feature space. Chen et al. [6] developed a novel hierarchical semantic embedding framework that incorporates the predicted score vector of the higher level as prior knowledge to learn finer-grained feature representation at each hierarchical level. During training, the predicted score vector of the higher level is also employed to regularize sub-categories prediction by using it as soft targets. Chang et al. [4] leveraged level-specific classification heads to disentangle coarse-level features with fine-grained ones and allowed fine-grained features to participate in coarser-grained label predictions but constraining the gradient flow to only update the parameters within each classification head. Their method reaches the state-of-the-art results on the traditional single-label FGVC problem. These approaches refine the feature representation related to hierarchical levels in the feature space. They developed the loss function based on the multi-class cross-entropy loss that implies the mutual exclusions among classes at the same hierarchical level. However, they neglect to encode other label relations like the parent-child correlation to transfer hierarchical knowledge between levels using samples observed at different levels.

3. Proposed Methods

3.1. Network Architecture

Our network architecture includes a trunk network, a hierarchical feature interaction module, and two parallel output channels, see Fig. 2. The trunk network is used to extract features from the input images and any common network is applicable. Here, we adopt the ResNet-50 as the trunk network since it is widely used for feature extraction. The hierarchical feature interaction module contains granularity-specific block and residual connections. These blocks share the same structure that comprises two convolutional layers and two fully connected (FC) layers. Each block is designed to extract the specialized feature for one hierarchical level. The residual connections first linearly combine features of fine-level subclasses with features of coarse-level superclasses. Accordingly, subclasses not only have unique attributes but also inherit the attributes from their superclasses. Then, non-linear transformation (ReLU) is applied to combined features.

We set up two output channels in our model. The first output channel is utilized to compute the probabilistic classification loss based on the tree hierarchy, in which each sigmoid node corresponds to a distinct label in the hierarchy. The second output channel computes the multi-class cross-entropy loss imposed on the leaf level so that the mutually exclusive fine-grained classes gain more attention during training. For simplicity, we denote the first and the second output channel as $O_{\text{Hier}}$ and $O_{\text{CE}}$, respectively.

3.2. Loss Function

The proposed combinatorial loss integrates two forms of losses: the probabilistic classification loss and the multi-class cross-entropy loss. We first formalize the tree hierarchy to encode semantic relations between hierarchical labels. The probabilistic classification loss defined on the tree hierarchy aims to transfer hierarchical knowledge during training. We empirically find that if the training samples labeled at the leaf level are few, the probabilistic classification
loss fails to well separate the fine-grained leaf classes. One simple but feasible solution is to increase the weight of fine-grained classification loss. Therefore, we further impose the multi-class cross-entropy loss on the leaf categories, which obeys the mutually exclusive constraint among fine-grained classes defined in the tree hierarchy.

3.2.1 The Formalism of Tree Hierarchy

The tree hierarchy $G = (V, E_h, E_e)$ consists of a set of nodes $V = \{v_1, \ldots, v_n\}$, directed edges $E_h \subseteq V \times V$, and undirected edges $E_e \subseteq V \times V$. Each node $v \in V$ corresponds to a distinct class label. The number of nodes $n$ equals the number of all labels in the hierarchy. A directed edge $(v_i, v_j) \in E_h$ is a subsumption edge, indicating that class $i$ subsumes label $j$, e.g., Albatross is a parent or superclass of Black-footed Albatross. An undirected edge $(v_i, v_j) \in E_e$ is an exclusion edge, denoting that classes $v_i$ and $v_j$ are mutually exclusive, e.g., a bird cannot be both Black-footed Albatross and Laysan Albatross simultaneously. Any two nodes share a subsumption edge or an exclusion edge.

Each class label takes binary values, i.e., $v_i \in \{0, 1\}$, representing whether an object belongs to this class or not. Each edge then defines a constraint on the binary values that two labels of its incident nodes can take. An assignment of $(v_i, v_j) = (0, 1)$ (e.g. a Black-footed Albatross but not a Albatross) for a subsumption edge $(v_i, v_j) \in E_h$ is illegal, while $(v_i, v_j) = (1, 1)$ (it is both Black-footed Albatross and Laysan Albatross) is also an illegal assignment for an exclusion edge $(v_i, v_j) \in E_e$. Defined by these local constraints of individual edges, a legal global assignment of all labels in the hierarchy is a binary label vector $y \in \{0, 1\}^n$ for an object. The set of all legal global assignments forms the state space $S_G \subseteq \{0, 1\}^n$ of tree $G$. We can infer $S_G$ to be a matrix $S \in \mathbb{R}^{(n+1)\times n}$, where each row represents a legal binary label vector $y$. We traverse all legal assignments by assigning each label a value of 1, along with an assignment that is all zeros.

3.2.2 Probabilistic Classification Loss

We calculate the probabilistic classification loss from $O_{Hier}$, and each sigmoid node in $O_{Hier}$ corresponds to a class label in the tree hierarchy. Suppose the number of sigmoid nodes is $n$, and $y \in \{0, 1\}^n$ is the binary label vector representing an assignment of all labels. Given an input image $x$, the joint probability of all sigmoid nodes concerning the assignment $y$ can be computed as:

$$P(y|x) = \prod_{i=1}^n \phi_i(\bar{x}_i, y_i) \prod_{i,j \in \{1, \ldots, n\}} \psi_{i,j}(y_i, y_j)$$  \hspace{1cm} (1)

where $\bar{x}_i$ is the sigmoid output of the $i$-th label node, $\tilde{P}(y|x)$ is the unnormalized probability, and $\phi_i(\bar{x}_i, y_i) = e^{\bar{x}_i(y_i=1)}$, $\psi_{i,j}(y_i, y_j)$ is the constraint defined in the tree hierarchy between any two labels in $y$:

$$\psi_{i,j}(y_i, y_j) = \begin{cases} 0, & \text{if violates constraints} \\ 1, & \text{otherwise} \end{cases}$$ \hspace{1cm} (2)

The joint probability is then normalized by $Pr(y|x) = \frac{\tilde{P}(y|x)}{Z(x)}$, where $Z(x)$ is the partition function that sums over all legal assignments $y \in S_G$ in the state space of tree $G$:

$$Z(x) = \sum_{y \in \{0,1\}^n} \prod_{i=1}^n \phi_i(\bar{x}_i, y_i) \prod_{i,j \in \{1, \ldots, n\}} \psi_{i,j}(y_i, y_j)$$ \hspace{1cm} (3)

If input image $x$ is observed at the $i$-th label in the tree hierarchy, i.e., $y_i = 1$, we can obtain the marginal probability $Pr(y_i = 1|x)$ of label $i$ by summing over all legal assignments $y \in S_G$ that include $y_i = 1$:

$$Pr(y_i = 1|x) = \frac{1}{Z(x)} \sum_{y \in \{0,1\}^n} \phi_i(\bar{x}_i, y_i) \prod_{i,j \in \{1, \ldots, n\}} \psi_{i,j}(y_i, y_j)$$ \hspace{1cm} (4)

The marginal probability of a leaf label in tree $G$ relies on the sum of its ancestors’ scores because all its ancestors must be 1 if the label of this leaf node takes value 1, which enables the parents’ scores to impact the descendants’ decisions. On the other hand, the marginal probability of a parent label is marginalized over all possible states of its descendants, i.e., aggregating the information from all its subclasses.

We propose to compute marginalization via matrix multiplication. Suppose the network outputs $X \in \mathbb{R}^{n \times k}$ from $O_{Hier}$, where $n$ is the number of sigmoid nodes, and $k$ stands for batch size. Each column in $X$ is the output vector corresponding to a sample in the batch. The unnormalized joint probability can be computed as $J = \exp(SX)$, and the partition function $z$ can be calculated by summing each column of $J$. To obtain the marginal probability of the $j$-th sample labeled at $i$, we first search for eligible rows in the $i$-th column of $S$ that qualify $S[j, i] > 0$, then we sum the corresponding elements in the $j$-th column of $J$, finally, we normalize the summation by dividing the $j$-th element in $z$.

In the training process, the observed label can be at any level of the hierarchy, and we maximize the marginal likelihood of the observed ground truth label. Given $m$ training samples $D = \{x(l), y(l), g(l)\}$, $l = 1, \ldots, m$, where $y(l)$ is the complete ground truth label vector and $g(l) \in \{1, \ldots, n\}$ is the index of the observed label, the probabilistic classification loss is defined as:

$$L_{Hier}(D) = -\frac{1}{m} \sum_{l=1}^m \ln(Pr(y(l)|g(l)) = 1|x(l)))$$ \hspace{1cm} (5)
3.2.3 Combinatorial Loss

The multi-class cross-entropy loss is commonly used in FGVC to separate fine-grained categories. We add $\mathcal{L}_{CE}$ to our model to further increase the discriminative power for fine-grained leaf classes. $\mathcal{L}_{CE}$ employs softmax outputs from $O_{CE}$, in which each node corresponds to a fine-grained leaf label in the tree hierarchy. Softmax outputs imply mutually exclusive relations among fine-grained classes, which is consistent with the hierarchy constraint defined in the tree hierarchy. We combine $\mathcal{L}_{CE}$ with $\mathcal{L}_{Hier}$ as follows:

$$
\mathcal{L}_{com}(x^{(l)}, y^{(l)}_{g(l)}) = \begin{cases} 
\mathcal{L}_{CE} + \mathcal{L}_{Hier}, & \text{if } g^{(l)} \text{ is in leaf nodes} \\
\mathcal{L}_{Hier}, & \text{otherwise}
\end{cases}
$$

Depending on whether $x^{(l)}$ is labeled at fine-grained leaf categories, the combined loss decides whether it needs to incorporate $\mathcal{L}_{CE}$ or not. Finally, the total loss on $\mathcal{D}$ is:

$$
\mathcal{L}_{total}(\mathcal{D}) = \sum_{l} \mathcal{L}_{com}(x^{(l)}, y^{(l)}_{g(l)})
$$

4. Experiments

4.1. Implementation Details

In all our experiments, we resize input images to 448 × 448 and train every single experiment for 200 epochs. Random horizontal flipping and random cropping (random cropping for training and center cropping for testing) are applied for data augmentation. We adopt ResNet-50 pre-trained on ImageNet as our trunk network and 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. Meanwhile, the learning rates of the convolution layers and the FC layers newly added for hierarchical interaction are initialized as 0.002 and adjusted by the cosine annealing strategy [18]. The learning rates of the trunk layers are maintained as 1/10 of the newly added layers. The code will be made publicly accessible.

4.2. Datasets and Experimental Designs

We evaluate our proposed method on three widely used FGVC datasets, i.e., CUB-200-2011 [32], Aircraft [20], and Stanford Cars [16]. However, CUB-200-2011 and Stanford Cars only provide one fine-grained label for each image. To construct a taxonomy of label hierarchy for these two datasets, we learn from the work of Chang et al. [4], in which they trace parent nodes in Wikipedia pages. CUB-200-2011 is the most widely used benchmark for FGVC. It covers 11,788 bird images that are re-organized into a three-level label hierarchy with 13 orders, 38 families, and 200 species. Aircraft contains 10,000 images consisting of a three-level label hierarchy with 30 makers, 70 families, and 100 models. Stanford Cars contains 16,185 images of 196 classes of cars. It is re-organized into a two-level label hierarchy with 9 car types and 196 specific models. We do not use any bounding box/part annotations in all our experiments and adopt the official train and test splits for evaluation.

Besides assigning hierarchical multi-granularity labels for each image, experimental designs simulate the aforementioned situation where samples are observed at different levels of the hierarchy. To imitate the lack of domain knowledge, we select 0%, 30%, 50%, 70%, and 90% samples from each fine-grained class and relabel their last-level fine-grained classes to immediate parent classes in the training set, respectively. Considering the impact of image quality, we conduct another experiment by reducing the image resolution of selected samples using the nearest-neighbor interpolation with a factor of 4 after relabeling. The extreme case 0% represents the conventional setting of HMC or fine-grained classification that exploits the label hierarchy. Other cases indicate that part of samples are observed at internal levels of the tree hierarchy, and the rest owns the complete label hierarchy from the highest level to the lowest fine-grained level. All images in the test set are tested with the complete label hierarchy.

4.3. Evaluation Metrics

To reasonably evaluate the performance of HMC on FGVC datasets, we employ two evaluation metrics. The first metric follows the convention of FGVC and uses the overall accuracy (OA). The output of HMC models is a probability vector for each class. Concerning the hierarchical label structure, we take the maximum value of the output probability vector corresponding to the actual hierarchical level as the predicted label and compute OA on the test set. The second criterion commonly used in HMC literature [10, 31, 33] measures the area under the average precision and recall curve $AU(\text{AP}_R)$. Instead of calculating the precision and recall curve (PRC) for each class, $AU(\text{AP}_R)$ computes an average PRC to evaluate the output probability vector of all classes in the hierarchy. Specifically, for a given threshold value, one point $(\text{Prec}, \text{Rec})$ in the average PRC is computed as:

$$
\text{Prec} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i}, \quad \text{Rec} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i}
$$

where $i$ ranges over all classes, and $TP_i$, $FP_i$, and $FN_i$ are the numbers of true positives, false positives, and false negatives for class label $i$, respectively. By varying the threshold, an average PRC is obtained and $AU(\text{AP}_R)$ denotes the...
Table 1. OA on the species level with the relabeling proportion of 0% by gradually adding each component in HRN: granularity-specific block (GSB), the linear combination of hierarchical features (LC), and non-linear transformation of combined features (ReLU).

| Component                        | OA     |
|----------------------------------|--------|
| Combinatorial Loss               | 84.32  |
| Combinatorial Loss + GSB         | 85.77  |
| Combinatorial Loss + GSB + LC    | 86.17  |
| Combinatorial Loss + GSB + LC + ReLU | 86.60  |

area under this curve. AU(PRC) also has the advantage of being independent of the threshold used to predict when a sample belongs to a particular class (which is often heavily application-dependent).

4.4. Ablation Study

In this section, we first conduct ablation studies to investigate two key designs of the proposed method on CUB-200-2011: HRN and combinatorial loss. Then, we conduct visualization experiments by illustrating the attention regions for each hierarchical level using Grad-Cam [27] to demonstrate the role of HRN and combinatorial loss.

4.4.1 Significance of HRN

As displayed in Fig. 2, we analyze three components of HRN: granularity-specific block (GSB), the linear combination of hierarchical features (LC), and non-linear transformation of combined features (ReLU). We report OA on the species level with the relabeling proportion of 0% in Tab. 1. The model that only contains ResNet-50 and combinatorial loss obtained a result of 84.32. As more components of HRN are integrated into the model, we gradually achieve better results.

4.4.2 Contribution of Combinatorial Loss

In this subsection, we validate the effectiveness of combining the probabilistic classification loss ($L_{Hier}$) with the multi-class cross-entropy loss ($L_{CE}$). Tab. 2 records OA on the species level with five relabeling proportions. It can be found that when more training samples are relabeled to coarse-grained classes, the fine-grained classification performance of $L_{Hier}$ degenerates drastically. In contrast, the combinatorial loss consistently outperforms $L_{Hier}$ by adding $L_{CE}$ imposed on fine-grained leaf classes.

4.4.3 Visualization Experiments

We conduct visualization experiments to demonstrate that granularity-specific blocks can capture different regions of interest while hierarchical knowledge can be transferred across levels via hierarchical residual interaction and supervision of combinatorial loss. To this end, we adopt Grad-Cam to visualize different attention regions of each hierarchical level by propagating their respective gradients back to feature maps generated from the trunk network. Fig. 3 illustrates two species of birds: House Wren and Marsh Wren from the same family (Troglodytidae) and order (Passeriformes). Domain knowledge reveals that Marsh Wrens have a more distinctive eyebrow than House Wrens. In Fig. 3, the order level focuses on the head and legs, and the family level pays close attention to the head and feathers on the wing and tail. In addition to feathers, the species level concentrates on the head. Three hierarchical levels give similar attention to the head with a distinctive eyebrow, while they have unique regions of interest.

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

To fairly evaluate the proposed method, we compare it to state-of-the-art HMC methods: HMC-LMLP [3], HMCN [33], and C-HMCNN [10], and the state-of-the-art FGVC approach that exploits the label hierarchy: Chang et al. [4]. In our hierarchical settings, we train all methods with different relabeling proportions. Chang et al. [4] sum the multi-class cross-entropy loss from each hierarchical level. When adapting their approach to hierarchical settings, we neglect the last-level loss if a sample has been relabeled to its parent class. We report OA of each hierarchical level and AU(PRC) results on test sets of three FGVC.
From Tab. 3, Tab. 4, and Tab. 5, we can observe that the proposed method achieves the best OA results of each hierarchical level and the best $AU(\text{PRC})$ results in most cases. In other cases, our results are also comparable to the best results. Chang et al. [4] use level-specific classification heads to disentangle coarse-level features with fine-grained ones, but they only consider mutually exclusion in each hierarchical level without examining subsumption relations between hierarchical levels in their loss function. C-HMCNN only constrains subsumption relations. HMC-LMLP and HMCN embed label hierarchy in their network architecture and train with the binary cross-entropy loss that implies all classes are independent. By contrast, in our framework, the tree hierarchy specifies the relation between any two labels with mutually exclusion or subsumption, and the corresponding probabilistic loss combined with the multi-class cross-entropy loss that implies all classes are independent. Considering the practical limitation of image quality, we reduce the image resolution of selected samples corresponding to different relabeling proportions. Tab. 6 displays the experimental results, and our method consistently outperforms compared methods in most cases under two evaluation metrics.

### 4.6. Analyze State-of-the-art Methods by Reducing Image Resolution

Except for domain knowledge, samples captured at low-resolution can hardly be identified with the last-level fine-grained categories, and thus they are more likely to be inferred as upper-level coarse classes. Considering the practical limitation of image quality, we reduce the image resolution of selected samples corresponding to different relabeling proportions. Tab. 6 displays the experimental results, and our method consistently outperforms compared methods in most cases under two evaluation metrics.

### 4.7. Comparison with FGVC Methods Exploiting Hierarchical Knowledge

Considering hierarchical knowledge, FGVC approaches refine the feature representation related to hierarchical levels in the feature space [4, 6, 28, 35], e.g., measuring the distance between classes in the hierarchy [28, 35], learn-
We study the HMC problem in which different objects can be discerned at various levels in the label hierarchy due to the differences in domain knowledge or image quality. To address this problem, we propose combinatorial loss and HRN. The combinatorial loss combines the probabilistic classification loss defined on the tree hierarchy that encodes semantic relations between any two hierarchical labels with the multi-class entropy loss imposed on the fine-grained leaf categories. The probabilistic classification loss can transfer hierarchical knowledge across levels, and the multi-class entropy loss increases the discriminative power on the leaf classes. In addition, HRN manages to perform hierarchical feature interaction via residual connections, \textit{i.e.,} features from parent levels acting as skip connections are added to features of children levels. Comprehensive experiments on three commonly used datasets demonstrated the effectiveness of the proposed method compared to state-of-the-art HMC and FGVC methods.

5. Conclusion

We study the HMC problem in which different objects can be discerned at various levels in the label hierarchy due to the differences in domain knowledge or image quality.
References

[1] Z. Barutcuoglu, R. E. Schapire, and O. G. Troyanskaya. Hierarchical multi-label prediction of gene function. *BMC Bioinform.*, 22(7):830–836, 2006.

[2] R. Cerri, R.C. Barros, and A. C. P. L. F. de Carvalho. Hierarchical multi-label classification using local neural networks. *J. Comput. Syst. Sci.*, 80(1):39–56, 2014.

[3] R. Cerri, R. C. Barros, A. C. P. L. F. de Carvalho, and Y. Jin. Reduction strategies for hierarchical multi-label classification in protein function prediction. *BMC Bioinform.*, 17(1):373, 2016.

[4] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your “flamingo” is my “bird”: Fine-grained, or not. In CVPR, pages 11476–11485, 2021.

[5] B. Chen, X. Huang, L. Xiao, Z. Cai, and L. Jing. Hyperbolic interaction model for hierarchical multi-label classification. In AAAI, volume 34, pages 7496–7503, 2020.

[6] Tianshui Chen, Wenxi Wu, Yuefang Gao, Le Dong, Xiaonan Luo, and Liang Lin. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In *ACM MM*, pages 2023–2031, 2018.

[7] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *ECCV*, pages 48–64, 2014.

[8] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Dzeroski. Hierarchical annotation of medical images. *Pattern Recognit.*, 44(10-11):2436–2449, 2011.

[9] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Dzeroski. Hierarchical classification of diatom images using ensembles of predictive clustering trees. *Ecological Informatics*, 7(1):19–29, 2012.

[10] E. Giunchiglia and T. Lukasiewicz. Coherent hierarchical multi-label classification networks. In *NeurIPS*, volume 33, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–635, Springer, 2016.

[12] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *ICCV*, pages 1314–1324, 2019.

[13] Andrew R Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kevin Q Weinberger. Densely connected convolutional networks. In CVPR, pages 4700–4708, 2017.

[15] C. N. Silla Jr and A. A. Freitas. A survey of hierarchical classification across different application domains. *Data Min. Knowl. Discov.*, 22(1-2):31–72, 2011.

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In CVPR workshops, pages 554–561, 2013.

[17] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5(Apr):361–397, 2004.

[18] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[19] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, pages 116–131, 2018.

[20] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[21] Y. Mao, J. Tian, J. Han, and X. Ren. Hierarchical text classification with reinforced label assignment. In *EMNLP-IJCNLP*, pages 445–455, 2019.

[22] A. Mayne and R. Perry. Hierarchically classifying documents with multiple labels. In *Proc. IEEE Symp. Comput. Intell. Data Mining*, pages 133–139, IEEE, 2009.

[23] Y. Meng, J. Shen, C. Zhang, and J. Han. Weakly-supervised hierarchical text classification. In AAAI, volume 33, pages 6826–6833, 2019.

[24] H. Peng, J. Li, Y. He, Y. Liu, M. Bao, L. Wang, Y. Song, and Q. Yang. Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In *WWW*, pages 1063–1072, 2018.

[25] J. Rouss, C. Saunders, S. Szedmak, and J. Shawe-Taylor. Kernel-based learning of hierarchical multilabel classification models. *J. Mach. Learn. Res.*, 7:1601–1626, 2006.

[26] L. Schieltgat, C. Vens, J. Struyf, H. Blockeel, D. Kocev, and S. Dzeroski. Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinform.*, 11(1):2, 2010.

[27] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[28] Weiwei Shi, Yihong Gong, Xiaoyu Tao, De Cheng, and Nanqing Zheng. Fine-grained image classification using modified dcnns trained by cascaded softmax and generalized large-margin losses. *NeurIPS*, 30(3):683–694, 2018.

[29] W. Huang et al. Hierarchical multi-label text classification: An attention-based recurrent network approach. In *Proc. 28th ACM Int. Conf. Inf. Knowl. Manag.*, pages 1051–1060, 2019.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[31] C. Vens, J. Struyf, L. Schieltgat, S. Dzeroski, and H. Blockeel. Decision trees for hierarchical multi-label classification. *Mach. Learn.*, 73(2):185–214, 2008.

[32] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
[33] J. Wehrmann, R. Cerri, and R.C. Barros. Hierarchical multi-label classification networks. In ICML, pages 5075–5084, 2018. 2, 5, 6, 7, 8

[34] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In CVPR, pages 1492–1500, 2017. 2, 8

[35] Xiaofan Zhang, Feng Zhou, Yuanqing Lin, and Shaoting Zhang. Embedding label structures for fine-grained feature representation. In CVPR, pages 1114–1123, 2016. 2, 7, 8

[36] Feng Zhou and Yuanqing Lin. Fine-grained image classification by exploring bipartite-graph labels. In CVPR, pages 1124–1133, 2016. 2