Label Relation Graphs Enhanced Hierarchical Residual Network for Hierarchical Multi-Granularity Classification

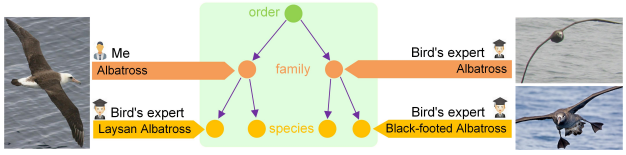
Jingzhou Chen[1], Peng Wang[1], Jian Liu[2], Yuntao Qian[1]*
[1]College of Computer Science, Zhejiang University, China
[2]Ant Financial Services Group
{11621038, pengwang18, ytqian}@zju.edu.cn, rex.lj@antgroup.com

Abstract

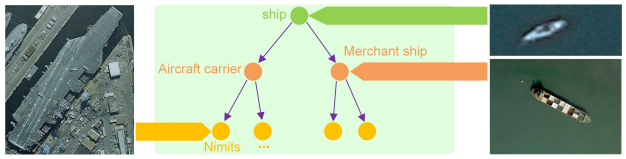
Hierarchical multi-granularity classification (HMC) assigns hierarchical multi-granularity labels to each object and focuses on encoding the label hierarchy, e.g., *["Albatross", "Laysan Albatross"] from coarse-to-fine levels. However, the definition of what is fine-grained is subjective, and the image 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. The code will be available at https://github.com/MonsterZhZh/HRN.*

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 [31], 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.



(a) Differences in domain knowledge and interference from the image occlusion.



(b) Large variations of image resolutions.

Figure 1. Different objects can be discerned at various levels in the label hierarchy due to differences in domain knowledge or image quality such as occlusion or resolution.

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 [28] aims to exploit hierarchical multi-granularity labels and embeds the label hierarchy

*Yuntao Qian is the corresponding author.

in loss function or network architecture. Whereas conventional 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 [12–14,29,34], *i.e*., features from parent levels acting as skip connections are added to features of children levels. In summary, we aim to tackle two challenges in HMC: (1) exploiting samples labeled at different levels; (2) designing the suitable experimental setting for this scenario. Accordingly, we propose a hierarchical loss on HRN and introduce class relabeling, image degeneration, and two evaluation metrics [30] in the experimental setting. Experiments on three commonly used FGVC datasets demonstrate the advantages of our approach compared to the state-of-the-art HMC approaches.

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, 30]. In text categorization, an increasing number of works [5, 15, 21, 23] 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, 11]. HMC with local multi-layer perceptrons (HMC-LMLP) [3] proposed to train a chain of 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) [11] 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, 27, 35, 37] 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*. [27] proposed a generalized large-

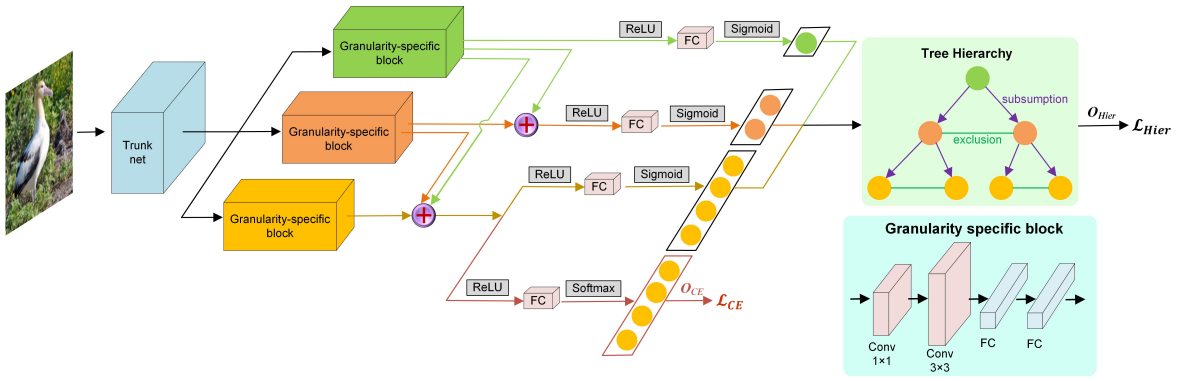


Figure 2. The network architecture consists of a trunk network, hierarchical feature interaction module, and two parallel output channels: O_{Hier} and O_{CE} forming the probabilistic classification loss ($\mathcal{L}_{Hier}$) and the cross-entropy loss ($\mathcal{L}_{CE}$), respectively. We illustrate the network architecture on CUB-200-2011 dataset that contains three hierarchical levels. Granularity-specific block for each hierarchical level process feature maps generated from the trunk network, then these hierarchical features interact via residual connections, *i.e.*, features from parent levels acting as skip connections are added to features of children levels. O_{Hier} organizes sigmoid outputs from three levels using the tree hierarchy, and O_{CE} generates softmax outputs corresponding to the fine-grained leaf categories.

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. 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. 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 exclusion 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.

2.3. Hierarchical Network Architecture

The proposed HRN is a feature learning model for class hierarchy, consisting of shared-specific feature representation and residual connection-based feature transfer. The shared-specific feature representation is commonly adopted in multi-task learning including tree-structured tasks as in our paper and [10, 18, 32, 36], where tree hierarchy is constructed by the semantics and the feature similarity, respectively. However, it is insufficient to integrate the knowledge from tree hierarchy, thus feature transfer between levels is an effective reinforcement. Wang *et al*. [32] use a linear combination of losses to fuse the features from different levels, while Li *et al*. [18] concatenate class predictions from lower levels to generate a new class prediction at the current level. Fan *et al*. [10] and Zhao *et al*. [36] define correlations among levels in the tree classifier rather than with a clear feature transfer module. Our method introduces a new solution to use the residual connection to transfer upper-level features to the current-level feature. The residual connection forces the network to learn residual features from identity mapping, which helps the HRN learn different features of upper levels compensated for lower 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 blocks 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 clas-

sification loss based on the tree hierarchy, in which each sigmoid node corresponds to a distinct label in the hierarchy. We perform the non-linear projection by sigmoid instead of softmax because sigmoid reflects the independent relations, whereas softmax implies mutual exclusion. The sigmoid nodes from each hierarchical level are then organized with the tree hierarchy to comply with the hierarchical constraints. 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_{Hier} and O_{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 the 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 $\mathbf{y} \in \{0, 1\}^n$ 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 $\mathbf{S} \in \mathcal{R}^{(n+1)\times n}$, where each row represents a legal binary label vector $\mathbf{y}$. 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 $\mathbf{y} \in \{0, 1\}^n$ is the binary label vector representing an assignment of all labels. Given an input image $\mathbf{x}$, the joint probability of all sigmoid nodes concerning the assignment $\mathbf{y}$ can be computed as:

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

where $\bar{x}_i$ is the sigmoid output of the i-th label node, $\tilde{P}(\mathbf{y}|\mathbf{x})$ 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 $\mathbf{y}$:

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

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

$$Z(\mathbf{x}) = \sum_{\bar{\mathbf{y}}\in\{0,1\}^n} \prod_{i=1}^{n} \phi_i(\bar{x}_i, \bar{y}_i) \prod_{i,j\in\{1,\ldots,n\}} \psi_{i,j}(\bar{y}_i, \bar{y}_j) \quad (3)$$

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

$$Pr(y_i = 1|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \sum_{\bar{\mathbf{y}}:\bar{y}_i=1} \prod_{i} \phi_i(\bar{x}_i, \bar{y}_i) \prod_{i,j} \psi_{i,j}(\bar{y}_i, \bar{y}_j) \quad (4)$$

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 $\mathbf{X} \in \mathcal{R}^{n \times k}$ from O_{Hier}, where n is the number of sigmoid nodes, and k stands for batch size. Each column in $\mathbf{X}$ is the output vector corresponding to a sample in the batch. The unnormalized joint probability can be computed as $\mathbf{J} = \exp(\mathbf{SX})$, and the partition function $\mathbf{z}$ can be calculated by summing each column of $\mathbf{J}$. 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 $\mathbf{S}$ that qualify $\mathbf{S}[:, i] > 0$, then we sum the corresponding elements in the j-th column of $\mathbf{J}$, finally, we normalize the summation by dividing the j-th element in $\mathbf{z}$.

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 $\mathcal{D} = \{\mathbf{x}^{(l)}, \mathbf{y}^{(l)}, g^{(l)}\}$, $l = 1, \ldots, m$, where $\mathbf{y}^{(l)}$ 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:

$$\mathcal{L}_{Hier}(\mathcal{D}) = -\frac{1}{m}\sum_{l}^{m} \ln(Pr(y^{(l)}_{g^{(l)}} = 1 | \mathbf{x}^{(l)})) \quad (5)$$

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:

$$\mathcal{L}_{com}(\mathbf{x}^{(l)}, y^{(l)}_{g^{(l)}}) = \begin{cases} \mathcal{L}_{CE} + \mathcal{L}_{Hier}, & \text{if } g^{(l)} \text{ is in} \\ & \text{leaf nodes} \\ \mathcal{L}_{Hier}, & \text{otherwise} \end{cases} \quad (6)$$

Depending on whether $\mathbf{x}^{(l)}$ 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:

$$\mathcal{L}_{total}(\mathcal{D}) = \sum_{l} \mathcal{L}_{com}(\mathbf{x}^{(l)}, y^{(l)}_{g^{(l)}}) \quad (7)$$

4. Experiments

4.1. Implementaion 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 pretrained 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 [19]. The learning rates of the trunk layers are maintained as 1/10 of the newly added layers.

4.2. Datasets and Experimental Designs

We evaluate our proposed method on three widely used FGVC datasets, *i.e.*, CUB-200-2011 [31], 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 covers 200 bird species grouped into a three-level label hierarchy with 13 orders, 38 families, and 200 species from the top layer to the bottom layer. Aircraft consists of a three-level label hierarchy with 30 makers, 70 families, and 100 plane models from the top layer to the bottom layer. Stanford Cars contains 196 car models that are re-organized into a two-level label hierarchy by adding 9 superordinate car types. We do not use any bounding box/part annotations in all our experiments and follow the official train/test splits for evaluation, *i.e.*, 5994/5794 images for CUB-200-2011, 6667/3333 images for Aircraft, and 8144/8041 images for Stanford Cars.

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 each hierarchical level as the predicted label and compute OA on the test set. The second criterion commonly used in HMC literature [11, 30, 33] measures the area under the average precision and recall curve $AU(\overline{PRC})$. Instead of calculating the precision and recall curve (PRC) for each class, $AU(\overline{PRC})$ computes an average PRC to evaluate the output probability vector of all classes in the hierarchy. Specifically, for a given threshold value, one point $(\overline{Prec}, \overline{Rec})$ in the average PRC is computed as:

$$\begin{aligned} \overline{Prec} &= \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \\ \overline{Rec} &= \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \end{aligned} \tag{8}$$

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 $AU(\overline{PRC})$ denotes the area under this curve. $AU(\overline{PRC})$ 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 conduct ablation studies to investigate two key designs of the proposed method on CUB-200-2011: 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 ($\mathcal{L}_{Hier}$) with the multi-class cross-entropy loss ($\mathcal{L}_{CE}$). Tab. 2 records OA on the species level with five relabeling proportions. In Tab. 2, it can be found that when more training samples are relabeled to coarse-grained classes, the fine-grained classification performance of $\mathcal{L}_{Hier}$ degenerates drastically. In contrast, the combinatorial loss consistently outperforms $\mathcal{L}_{Hier}$ by adding $\mathcal{L}_{CE}$ imposed on fine-grained leaf classes.

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**

Table 2. OA on the species level by analyzing the effectiveness of combining the probabilistic classification loss ($\mathcal{L}_{Hier}$) with the multi-class cross-entropy loss ($\mathcal{L}_{CE}$).

Relabeling	0%	30%	50%	70%	90%
$\mathcal{L}_{Hier}$	84.56	76.66	64.36	45.10	28.69
$\mathcal{L}_{Hier} + \mathcal{L}_{CE}$	**86.60**	**83.91**	**80.52**	**73.96**	**53.02**

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 [11], 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 $AU(\overline{PRC})$ results on test sets of three FGVC datasets: CUB-200-2011, Aircraft, and Stanford Cars, displayed in Tab. 3, Tab. 4, and Tab. 5, respectively.

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 $AU(\overline{PRC})$ 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 can transfer hierarchical knowledge during training. The proposed HRN disentangles hierarchical features by granularity-specific blocks, and these features

Table 3. OA(%)/$AU(\overline{PRC})$ results on **CUB-200-2011** by comparing to state-of-the-art methods.

Relabeling	Hierarchy	HMC-LMLP [3]		HMCN [33]		C-HMCNN [11]		Chang *et al*. [4]		Ours	
0%	Order	98.45		97.29		98.48		97.76		**98.67**	
	Family	94.24	0.945	93.15	0.934	94.63	0.960	94.17	0.968	**95.51**	**0.969**
	Specie	79.60		79.75		81.58		85.56		**86.60**	
30%	Order	98.17		96.82		97.98		97.81		**98.31**	
	Family	93.58	0.920	91.99	0.905	93.89	0.938	94.10	**0.962**	**94.79**	0.958
	Specie	71.30		71.68		74.91		82.53		**83.91**	
50%	Order	**98.36**		96.70		98.34		97.43		97.89	
	Family	93.84	0.895	90.85	0.874	94.10	0.909	93.47	**0.951**	**94.29**	0.944
	Specie	64.34		64.29		67.52		79.30		**80.52**	
70%	Order	98.27		97.22		98.02		96.65		**98.43**	
	Family	93.84	0.831	91.25	0.834	93.91	0.844	91.74	0.924	**93.94**	**0.936**
	Specie	47.98		52.90		50.05		70.03		**73.96**	
90%	Order	**98.38**		97.31		98.27		97.12		97.97	
	Family	**94.44**	0.716	86.85	0.725	94.37	0.772	91.91	**0.868**	93.32	0.865
	Specie	22.89		30.69		26.16		49.36		**53.02**	

Table 4. OA(%)/$AU(\overline{PRC})$ results on **Aircraft** by comparing to state-of-the-art methods.

Relabeling	Hierarchy	HMC-LMLP [3]		HMCN [33]		C-HMCNN [11]		Chang *et al*. [4]		Ours	
0%	Maker	97.09		96.07		**97.45**		96.88		**97.45**	
	Family	94.39	0.968	92.56	0.959	95.41	0.979	95.28	**0.981**	**95.79**	0.976
	Model	90.25		87.19		91.69		91.92		**92.58**	
30%	Maker	96.85		96.13		96.76		87.41		**97.27**	
	Family	93.34	0.950	92.74	0.952	94.27	**0.971**	94.44	0.957	**95.52**	0.970
	Model	85.42		85.42		88.39		89.33		**91.62**	
50%	Maker	97.24		95.71		96.49		73.56		**97.27**	
	Family	93.82	0.925	92.05	0.935	93.88	0.963	94.17	0.909	**95.67**	**0.965**
	Model	83.59		81.52		85.18		86.66		**89.66**	
70%	Maker	96.97		95.80		96.67		58.77		**96.75**	
	Family	93.70	0.898	90.49	0.900	94.00	**0.953**	93.78	0.816	**94.20**	**0.953**
	Model	81.61		78.37		80.11		82.96		**84.53**	
90%	Maker	**96.97**		93.40		96.76		49.88		95.43	
	Family	93.37	0.870	89.50	0.824	**94.36**	0.903	93.72	0.656	91.68	**0.904**
	Model	**74.41**		70.06		71.02		64.99		71.06	

interact via residual connections to fuse attributes following the hierarchy.

4.6. Generate Relabeled Images 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.

Moreover, for each method, we average its OA and $AU(\overline{PRC})$ results on all levels, relabeling proportions, and datasets in Tabs. 3 to 6 and summarize the averaged results in Tab. 7, which shows significant improvement over the compared state-of-the-art methods.

4.7. Evaluation on Traditional FGVC Setting

Considering hierarchical knowledge, FGVC approaches refine the feature representation related to hierarchical levels in the feature space [4, 6, 27, 35], *e.g*., measuring the distance between classes in the hierarchy [27, 35], learning finer-grained features with the prediction of higher level [6], or disentangling coarse-level features with fine-grained ones [4]. Nevertheless, they develop their loss functions based on the multi-class cross-entropy loss, which implies mutual exclusion at the same hierarchical level. On the other hand, encoding label relations like the parent-child correlation helps to utilize samples observed at different levels. In contrast, the proposed method specifies label relations with the tree hierarchy and computes the combinatorial loss to effectively exploit samples labeled at different levels.

We record the best results reported in their works in which each sample has complete hierarchical multi-granularity labels. However, some papers did not present results on all our datasets, resulting in the missing values. In Tab. 8, Chang *et al*. [4] achieve state-of-the-art

Table 5. OA(%)/$AU(\overline{PRC})$ results on **Stanford Cars** by comparing to state-of-the-art methods.

Relabeling	Hierarchy	HMC-LMLP [3]		HMCN [33]		C-HMCNN [11]		Chang *et al*. [4]		Ours	
0%	Type	96.98		95.21		96.75		96.40		**97.41**	
	Maker	87.65	0.953	88.71	0.938	90.64	0.971	93.65	0.977	**94.03**	**0.981**
30%	Type	**96.85**		94.38		96.23		96.23		96.13	
	Maker	79.16	0.909	81.59	0.887	81.92	0.927	**91.61**	**0.970**	90.55	0.969
50%	Type	**96.92**		93.46		95.95		95.60		95.88	
	Maker	66.45	0.842	73.03	0.832	70.22	0.850	88.10	0.960	**88.72**	**0.963**
70%	Type	**96.89**		93.02		95.67		92.90		96.06	
	Maker	41.52	0.705	52.66	0.713	43.17	0.708	76.13	0.905	**83.72**	**0.947**
90%	Type	96.38		93.42		**96.49**		92.25		94.32	
	Maker	13.51	0.572	19.89	0.560	13.54	0.577	45.79	0.761	**49.30**	**0.794**

Table 6. Compared OA(%)/$AU(\overline{PRC})$ results on **CUB-200-2011** by reducing the image resolution after relabeling.

Relabeling	Hierarchy	HMC-LMLP [3]		HMCN [33]		C-HMCNN [11]		Chang *et al*. [4]		Ours	
0%	Order	98.45		97.29		98.48		97.76		**98.67**	
	Family	94.24	0.945	93.15	0.934	94.63	0.960	94.17	0.968	**95.51**	**0.969**
	Specie	79.60		79.75		81.58		85.56		**86.60**	
30%	Order	97.86		96.32		97.81		97.62		**98.50**	
	Family	93.18	0.926	88.06	0.887	93.48	0.944	93.59	**0.961**	**94.75**	0.959
	Specie	74.32		70.78		76.04		82.33		**84.13**	
50%	Order	97.45		95.32		97.62		97.12		**98.20**	
	Family	92.25	0.907	85.93	0.853	92.51	0.925	91.79	0.947	**93.82**	**0.952**
	Specie	68.10		62.70		70.37		78.30		**81.18**	
70%	Order	**97.62**		94.43		97.20		96.32		97.58	
	Family	91.72	0.862	82.64	0.789	91.18	0.881	88.84	0.909	**92.42**	**0.926**
	Specie	53.11		45.75		55.14		68.06		**73.98**	
90%	Order	96.67		93.56		**96.79**		96.06		96.15	
	Family	**89.96**	0.695	78.60	0.694	89.44	0.801	87.52	**0.843**	88.29	0.837
	Specie	20.78		22.52		28.32		46.58		**50.10**	

Table 7. The average OA(%) and $AU(\overline{PRC})$ results on all levels, relabeling proportions, and datasets in Tabs. 3 to 6.

Metrics	HMC-LMLP	HMCN	C-HMCNN	Chang *et al*.	Ours
OA	83.66	82.34	84.44	86.29	**89.78**
$AU(\overline{PRC})$	0.861	0.846	0.887	0.910	**0.937**

Table 8. OA(%) results on the traditional FGVC setting with single leaf label.

Method	CUB-200-2011	Aircraft	Stanford Cars
Zhang *et al*. [35]	—	—	88.4
Shi *et al*. [27]	77.0	84.6	89.5
Chen *et al*. [6]	88.1	—	—
Chang *et al*. [4]	**89.9**	93.6	**95.1**
Ours	88.6	**94.1**	**95.1**

performances in the traditional single-label FGVC problem. Our approach reaches comparable results by simply replacing ResNet-50 with ResNeXt101-32×4d [34]. Other techniques that enrich the feature representation in the context of FGVC can be applied to boost the performance, which is beyond our scope.

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. 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. HRN manages to perform hierarchical feature interaction via residual connections, *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.

Acknowledgements This work was supported by National Key Research and Development Program of China under grant 2018AAA0100500 and National Natural Science Foundation of China under grant 62071421. We would like to thank the anonymous area chairs and reviewers for their useful feedback.

References

[1] Zafer Barutcuoglu, Robert E Schapire, and Olga G Troyanskaya. Hierarchical multi-label prediction of gene function. *BMC Bioinform.*, 22(7):830–836, 2006. 2

[2] Ricardo Cerri, Rodrigo C Barros, and André CPLF De Carvalho. Hierarchical multi-label classification using local neural networks. *J. Comput. Syst. Sci.*, 80(1):39–56, 2014. 2

[3] Ricardo Cerri, Rodrigo C Barros, André C PLF de Carvalho, and Yaochu Jin. Reduction strategies for hierarchical multi-label classification in protein function prediction. *BMC Bioinformat.*, 17(1):373, 2016. 2, 6, 7, 8

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11476–11485, 2021. 1, 2, 3, 5, 6, 7, 8

[5] Boli Chen, Xin Huang, Lin Xiao, Zixin Cai, and Liping Jing. Hyperbolic interaction model for hierarchical multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 7496–7503, 2020. 2

[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 *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 2023–2031, 2018. 2, 3, 7, 8

[7] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 48–64, 2014. 2

[8] Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Sašo Džeroski. Hierarchical annotation of medical images. *Pattern Recognit.*, 44(10-11):2436–2449, 2011. 2

[9] Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Sašo Džeroski. Hierarchical classification of diatom images using ensembles of predictive clustering trees. *Ecological Informat.*, 7(1):19–29, 2012. 2

[10] Jianping Fan, Tianyi Zhao, Zhenzhong Kuang, Yu Zheng, Ji Zhang, Jun Yu, and Jinye Peng. Hd-mtl: Hierarchical deep multi-task learning for large-scale visual recognition. *IEEE Trans. Image Process.*, 26(4):1923–1938, 2017. 3

[11] Eleonora Giunchiglia and Thomas Lukasiewicz. Coherent hierarchical multi-label classification networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9662–9673, 2020. 2, 6, 7, 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–645. Springer, 2016. 2

[13] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019. 2

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017. 2

[15] Wei Huang, Enhong Chen, Qi Liu, Yuying Chen, Zai Huang, Yang Liu, Zhou Zhao, Dan Zhang, and Shijin Wang. Hierarchical multi-label text classification: An attention-based recurrent network approach. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1051–1060, 2019. 2

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 554–561, 2013. 1, 5

[17] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5(Apr):361–397, 2004. 2

[18] Aoxue Li, Tiange Luo, Zhiwu Lu, Tao Xiang, and Liwei Wang. Large-scale few-shot learning: Knowledge transfer with class hierarchy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7212–7220, 2019. 3

[19] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[20] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1, 5

[21] Yuning Mao, Jingjing Tian, Jiawei Han, and Xiang Ren. Hierarchical text classification with reinforced label assignment. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 445–455, 2019. 2

[22] Andrew Mayne and Russell Perry. Hierarchically classifying documents with multiple labels. In *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 133–139. IEEE, 2009. 2

[23] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised hierarchical text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 6826–6833, 2019. 2

[24] Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In *Proceedings of the World Wide Web Conference (WWW)*, pages 1063–1072, 2018. 2

[25] Juho Rousu, Craig Saunders, Sandor Szedmak, and John Shawe-Taylor. Kernel-based learning of hierarchical multilabel classification models. *J. Mach. Learn. Res.*, 7:1601–1626, 2006. 2

[26] Leander Schietgat, Celine Vens, Jan Struyf, Hendrik Blockeel, Dragi Kocev, and Sašo Džeroski. Predicting gene func-

tion using hierarchical multi-label decision tree ensembles. *BMC Bioinformat.*, 11(1):1–14, 2010. 2

[27] Weiwei Shi, Yihong Gong, Xiaoyu Tao, De Cheng, and Nanning Zheng. Fine-grained image classification using modified dcnns trained by cascaded softmax and generalized large-margin losses. *IEEE Trans. Neural Netw. Learn. Syst.*, 30(3):683–694, 2018. 2, 7, 8

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

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, 2017. 2

[30] Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. Decision trees for hierarchical multi-label classification. *Mach. Learn.*, 73(2):185–214, 2008. 2, 6

[31] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *Technical Report CNS-TR-2010-001*, 2011. 1, 5

[32] Yu Wang, Ruonan Liu, Di Lin, Dongyue Chen, Ping Li, Qinghua Hu, and CL Philip Chen. Coarse-to-fine: Progressive knowledge transfer-based multitask convolutional neural network for intelligent large-scale fault diagnosis. *IEEE Trans. Neural Netw. Learn. Syst.*, 2021. 3

[33] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In *International Conference on Machine Learning (ICML)*, pages 5075–5084, 2018. 2, 6, 7, 8

[34] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1114–1123, 2016. 2, 7, 8

[36] Tianyi Zhao, Baopeng Zhang, Ming He, Wei Zhang, Ning Zhou, Jun Yu, and Jianping Fan. Embedding visual hierarchy with deep networks for large-scale visual recognition. *IEEE Trans. Image Process.*, 27(10):4740–4755, 2018. 3

[37] Feng Zhou and Yuanqing Lin. Fine-grained image classification by exploring bipartite-graph labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1124–1133, 2016. 2