Your “Flamingo” is My “Bird”: Fine-Grained, or Not

Dongliang Chang[1], Kaiyue Pang[2], Yixiao Zheng[1], Zhanyu Ma[1*], Yi-Zhe Song[2], and Jun Guo[1]
[1] The Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence,
Beijing University of Posts and Telecommunications, Beijing, China
[2]SketchX, CVSSP, University of Surrey, United Kingdom
[1]{changdongliang,zhengyixiao,mazhanyu,guojun}@bupt.edu.cn
[2]{kaiyue.pang,y.song}@surrey.ac.uk

Abstract

Whether what you see in Figure 1 is a “flamingo” or a “bird”, is the question we ask in this paper. While fine-grained visual classification (FGVC) strives to arrive at the former, for the majority of us non-experts just “bird” would probably suffice. The real question is therefore – how can we tailor for different fine-grained definitions under divergent levels of expertise. For that, we re-envisage the traditional setting of FGVC, from single-label classification, to that of top-down traversal of a pre-defined coarse-to-fine label hierarchy – so that our answer becomes “bird” ⇒ “Phoenicopteriformes” ⇒ “Phoenicopteridae” ⇒ “flamingo”.

To approach this new problem, we first conduct a comprehensive human study where we confirm that most participants prefer multi-granularity labels, regardless whether they consider themselves experts. We then discover the key intuition that: coarse-level label prediction exacerbates fine-grained feature learning, yet fine-level feature betters the learning of coarse-level classifier. This discovery enables us to design a very simple albeit surprisingly effective solution to our new problem, where we (i) leverage level-specific classification heads to disentangle coarse-level features with fine-grained ones, and (ii) allow finer-grained features to participate in coarser-grained label predictions, which in turn helps with better disentanglement. Experiments show that our method achieves superior performance in the new FGVC setting, and performs better than state-of-the-art on the traditional single-label FGVC problem as well. Thanks to its simplicity, our method can be easily implemented on top of any existing FGVC frameworks and is parameter-free.

1. Introduction

Fine-grained visual classification (FGVC) was first introduced to the vision community almost two decades ago with the landmark paper of [2]. It brought out a critical question that was largely overlooked back then – that can machines match up to humans on recognising objects at fine-grained level (e.g., a “flamingo” other than a “bird”). Great strides have been made over the years, starting with the conventional part-based models [51, 14, 1, 3], to the recent surge of deep models that either explicitly or implicitly tackle part learning with or without strong supervision [26, 34, 52, 55, 57, 48]. Without exception, the focus has been on mining fine-grained discriminative features to better classification performances.

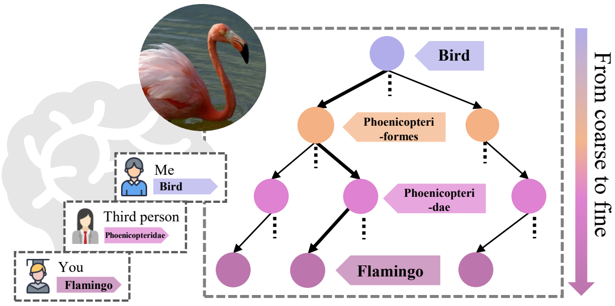


Figure 1: Definition of what is fine-grained is subjective. Your “flamingo” is my “bird”.

In this paper, we too are interested in the fine-grained rationale at large – yet we do not set out to pursue performance gains, we instead question the very definition of fine-grained classification itself. In particular, we ask whether the fine-grained expert labels commonplace to current FGVC datasets indeed convey what end users are accustomed to – i.e., are the “Florida scrub jay”, “Fisker Karma Sedan 2012”, “Boeing 737-200” are indeed the desired, or would “bird”, “car”, “aircraft” suffice for many – my “flamingo” can be just your “bird”. The answer is of course subjective [31], and largely correlates with expert knowledge – the more you are a bird lover, the more fine-grained labels you desire, some might even look for “American flamingo” other than just “flamingo”. The follow-up question is therefore, how can we tailor for the various sub-

[*] indicates corresponding author.

jective definitions of what is fine-grained, and design a system that best accommodates practical usage scenarios of FGVC.

To answer this, we first conduct a human study on the popular CUB-200-2011 bird dataset [42] with two questions in mind (i) how useful are the pre-defined fine-grained labels to a general user, and (ii) whether a single label output is in fact a preferred solution. We first build a hierarchical taxonomy of bird, by tracing existing fine-grained labels in CUB-200-2011 to its parent sub-category, all the way to the super node of "bird" using Wikipedia. We then recruited 50 participants with various background of bird knowledge, each of whom rated 100 bird photos by (i) picking a label amongst fine- and coarse-grained ones relating to the bird, and (ii) indicating whether more label choices are desirable other than just the single label previously selected. We find that (i) participants do not necessarily choose the pre-defined fine-grained (bottom) labels as their preferred choice, (ii) only 36.4% of all returned choices prefer just a single label, and (iii) although domain experts tend to choose finer-grained labels while amateurs prefer coarser ones, close to 80% of choices from experts also turn to the option of multi-granularity labels.

Following results from the human study, we propose to re-instantiate the FGVC problem by extending it from a *single-label classification* problem, to that of *multiple label predictions* on a pre-defined label hierarchy. The central idea is while people tend to feel baffled facing a single expert label, a chain of coarse-to-fine labels that describe an object can potentially be more practical – we leave it to the users to decide which fine-grained level along the hierarchy best suits their needs. Compared with a single label telling you it is a "flamingo" (as per conventional FGVC), our model offers a coarse-to-fine series of labels such as "bird" ⇒ "Phoenicopteriformes" ⇒ "Phoenicopteridae" ⇒ "flamingo" (See Figure 1).

On the outset, classifying an image into multiple cross-granularity labels seems an easy enough extension to the well-studied problem of FGVC with single-label output. One can simply train a single model for classifying all nodes in the hierarchy, or better yet use separate classifiers for each hierarchy level. Although these do work as baselines, they do not benefit from the inherent coarse-fine hierarchical relationship amongst labels – we show exploring these relationships not only helps to solve for the new FGVC setting, but also in turn benefits the learning of fine-grained features which then helps the conventional task.

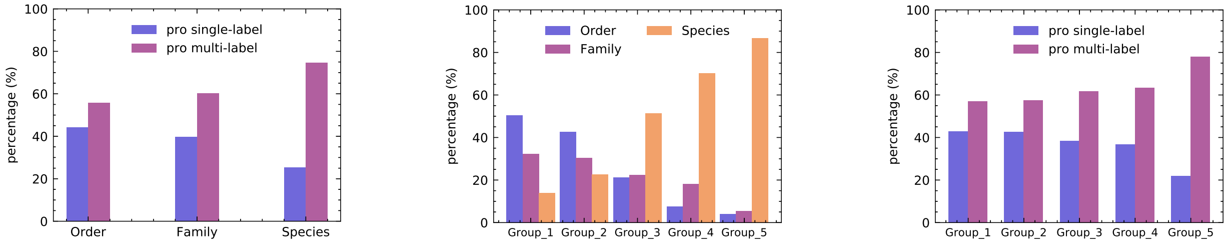
Our design is based on the discovery of two key observations on the label hierarchy: (i) coarse-level features in fact exacerbates the learning of fine-grained features, and (ii) finer-grained label learning can be exploited to enhance the discriminability of coarser-grained label classifier. Our first technical contribution is therefore a multi-task learning framework to perform level-wise feature disentanglement, with the aim to separate the adverse effect of coarse feature from fine-grained ones. To further encourage the disentanglement, we then resort to the clever use of gradients to reflect our second observation. Specifically, during the forward pass only, we ask finer-grained features to participate in the classification of coarser-grained labels via feature concatenation. We, however, constrain the gradient flow to only update the parameters within each multi-task head. Our method is generic to any existing FGVC works and experiments show that it yields stronger classifiers across all granularities. Interestingly, our model also delivers state-of-the-art result when evaluated on the traditional FGVC setting, while not introducing any additional parameters.

Our contributions are as follows: (i) we re-envisage the problem setting of FGVC, to accommodate the various subjective definitions of "fine-grained", where we advocate for top-bottom traversal of a coarse-to-fine label hierarchy, other than the traditional single-label classification; (ii) we discover important insights on the inherent coarse-fine hierarchical relationship to drive our model design, and (iii) we show by disentangling coarse-level feature learning with that of fine-grained, state-of-the-art performances can be achieved both on our new problem, and on the traditional problem of FGVC.

2. Related Work

Fine-grained image classification Deep learning has emerged as powerful tool that led to remarkable breakthroughs in FGVC [53, 50, 46, 27, 8, 59, 22]. Compared with generic image recognition task [10, 43], FGVC requires a model to pay special attention on the very subtle and local image regions [50, 5], which are usually hard to notice in human eyes. A major stream of FGVC works thus undergoes two stages by first adopting a localisation subnetwork to localise key visual cues and then a classification subnetwork to perform label prediction. Earlier works on localisation module rely heavily on additional dense part/bounding box annotations to perform detection [1, 4], and gradually move towards weakly supervised setting that only requires image labels [50, 5]. Relevant techniques including unsupervised detection/segmentation, utilisation of deep filters and attention mechanism have been proposed to guide the extraction of the most discriminative image regions [49, 45, 21]. Another line of FGVC research focuses on end-to-end feature encoding [13, 39, 38]. This saves the effort of explicit image localisation but asks for extra effort to encourage feature discriminability, e.g., high-order feature interactions [26, 56]. In this paper, we study a different setting for FGVC that generates multiple output labels at different granularities for an image.

Multi-task learning Multi-task learning (MTL) aims to leverage the common information among tasks to improve



(a) (b) (c)

Figure 2: Human study on CUB-200-2011 bird dataset. Order, family, species are three coarse-to-fine label hierarchy for a bird image. A higher group_id represents a group of people with better domain knowledge of birds, with group_5 interpreted as domain experts. (a) Human preference between single and multiple labels. (b) Impact of human familiarity with birds on single-label choice. (c) Impact of human familiarity with birds on multi-label choice.

Choice	**Order**	**Family**	**Species**	**None**
Percentage	29.2%	30.4%	36.4%	4%

Table 1: Human preference between labels at different granularity on CUB-200-2011 bird dataset.

the generalisability of the model [6, 7, 28, 58]. Under the context of deep learning, MTL translates to designing and optimising networks that encourage shared representations under multi-task supervisory signals. There are two types of parameter sharing. The hard way is to divide the parameter set into shared and task-specific operators [24, 23, 9]. In soft parameter sharing, however, each task is assigned its own set of parameters and further regularisation technique are introduced to encourage cross-task talk [30, 35, 15]. Joint learning of multiple tasks is prone to negative transfer if the task dictionary contains unrelated tasks [24, 19]. This problem triggers another line of MTL research with numerous solutions proposed, including reweighing the individual task loss [23, 37], tailoring task-specific gradient magnitudes [9] and disentangling features between irrelevant tasks [17, 54]. We approach the multi-task learning in FGVC following a similar underlying motivation - by identifying impacts of transfer between label predictions at different granularities. More specifically, we propose a novel solution to simultaneously reinforce positive and mitigate negative task transfer.

3. Human Study

To inform the practical necessity of our multiple cross-granularity label setting, we conduct a human study [16] on the CUB-200-2011 bird dataset. This is in order to show (i) single fine-grained label generated by existing FGVC models does not meet the varying subjective requirements for label granularity in practice; (ii) multiple label outputs covering a range of granularity are able to bridge the perceived gaps amongst different populations.

Data & Participant Setup CUB-200-2011 is a bird dataset commonly used by the FGVC community. It contains $11,877$ images each labelled as a fine-grained bird species by the domain expert. We extend it by adding two new hierarchy levels on top of the species with reference to Wikipedia pages, i.e., identifying the *family* and *order* name for a bird image. This makes each image annotated with three labels at different granularity, in an increasing fineness level from order to species. We performed an initial test amongst 200 participants across different ages, genders and education levels, to find out their familiarity with birds. We discover that there exists a considerable "long tail" problem in their distribution of scores – there are naturally less bird experts. This motivates us to manually filter for a population that serves as a better basis for statistical analysis. We therefore sample 50 participants from the original 200 and encourage the distribution of their expertise (scores) to follow a Gaussian-like shape. We then divide them into 5 groups ([group_1, group_2, ..., group_5]) based on their scores, where a higher group id corresponds to a population of better domain knowledge. These 50 participants are included for the task below.

Experiment setting Designing experiments to validate people's preference on one single label across all granularities is straightforward. But it requires extra consideration for making comparative choices between single and multiple labels. For example, it would not be ideal if we show participants an image with two options of single and multiple labels, since people are naturally biased towards multiple labels as they contain more information [40]. We therefore design a two-stage experiment, with both stages showing a participant the same image but with different questions.

Stage 1: *This is a bird. Which one of the labels further defines this bird? You can only choose one option. [A] order_name [B] family_name [C] species_name [D] none of above*

Stage 2: *At stage 1, do you have the impulse to choose more than one label? [A] yes [B] no*

Note that participants selecting option D in stage 1 will be

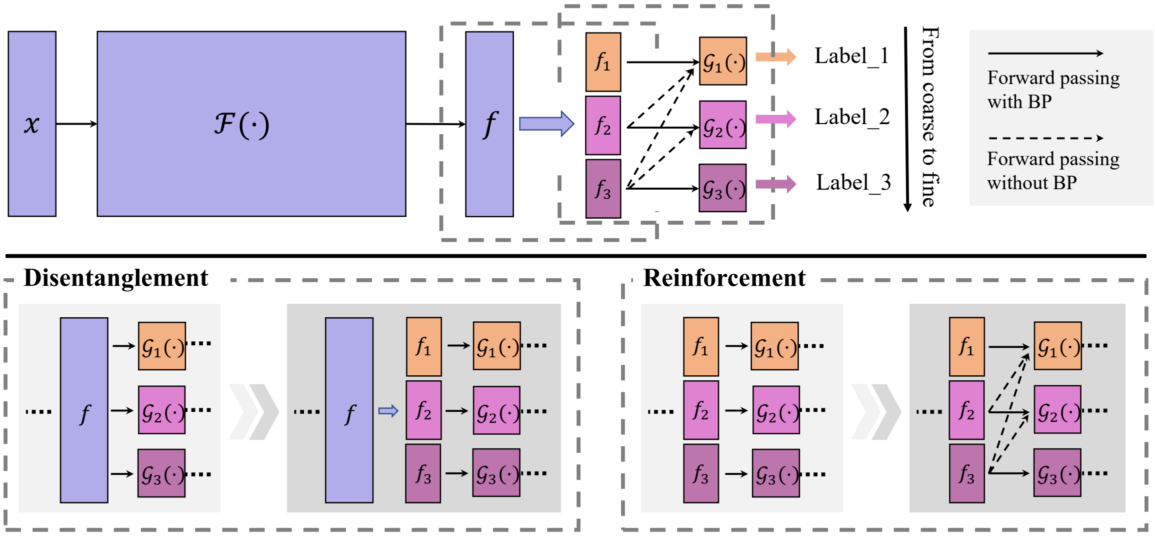


Figure 3: A schematic illustration of our FGVC model with multi-granularity label output. BP: backpropagation.

directly guided to the next image, skipping stage 2 all together.

Results We select 1000 images from CUB-200-2011 and from which, a set of random 100 images is assigned to each participant. Images received less then three responses are excluded for statistical significance. We analyse the results as follows:

Your label is not mine Table 1 shows the percentage of each option being selected in Stage 1. We can see that (i) participants have varying demands for label granularity; and (ii) The single fine-grained labels (Species option) optimised by existing FGVC models only constitute 36.4% of participant choices in our experiment, while leaving the rest 59.6% (order + family) potentially catered for under a multi-label setting.

Multiple labels work In Figure 2(a), we show the distribution of preference between single and multiple labels in the second stage. It can be seen that no matter what label (excluding "None") is chosen in the first stage, the majority of participants turn to embrace multiple labels. This is especially true for participants once selecting species as their single choice, who are the target audience under traditional FGVC setting, and yet still consider multiple cross-granularity labels a better way to interpret an image.

Further analysis Figure 2(b) and (c) further show how populations with different familiarity levels with birds lead to different choices in stage 1 and stage 2 respectively. We can see that (i) participants with more domain knowledge (e.g., group_4) tend to choose finer-grained single labels while amateurs (e.g., group_1) prefer more interpretable coarser-grained counterparts; (ii) choices under multiple labels have greatly converged regardless of the gaps of domain knowledge. In summary, it is hard to have one level of label granularity that caters to every participant. Multiple cross-granularity labels, however, are found to be meaningful to the many.

4. Methodology

Conclusions from our human study motivate us to go beyond the single label output as found in most existing FGVC works, and move towards generating multi-granularity labels. This makes our new setting fall naturally under the multi-task learning framework. Our first goal is to investigate the impact of transfer between label prediction tasks at different granularities. We next build on the insight gained and propose a simple but effective solution that improves the accuracy of label prediction at all granularities. A schematic illustration of our model is shown in Figure 3.

Definition Suppose for each image x, we have one fine-grained label y^K from the existing FGVC dataset. To tailor it for our new FGVC setting, we build upon y^K to form $(K-1)$ label hierarchies by finding its superclasses in the Wikipedia pages. This gives us a re-purposed dataset where each image x is annotated with a chain of K labels defined across different granularities, $y^1, y^2, ..., y^k, ..., y^K$. We denote the number of categories within each label granularity as $C_1, C_2, ..., C_k, ..., C_K$, so that y^k is a one-hot vector of length C_k. Given any CNN-based network backbone $\mathcal{F}(\cdot)$, We feed x as input to extract its feature embedding $f = \mathcal{F}(x)$. Our goal is then to correctly predict labels across K independent classifiers, $\mathcal{G}_1(\cdot), \mathcal{G}_2(\cdot), ..., \mathcal{G}_k(\cdot), ..., \mathcal{G}_K(\cdot)$ based on f, i.e., $\hat{y}^k = y^k$, where $\hat{y}^k = \mathcal{G}_k(f)$. Our optimisation objective is K independent cross-entropy loss $\sum_{k=1}^{K} L_{CE}(\hat{y}^k, y^k)$, and during inference, we take the maximum output probability from each classifier as its label, $l^k = \underset{C_k}{\operatorname{argmax}}\, \hat{y}^k$.

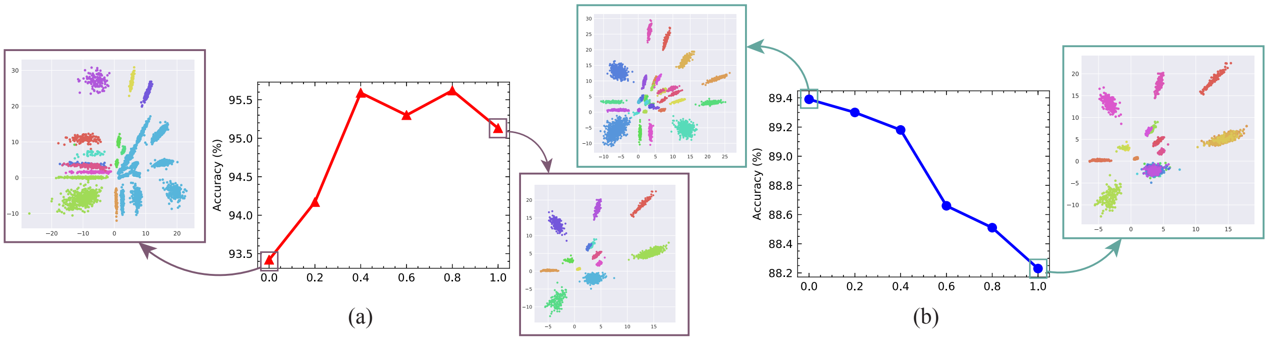


Figure 4: Joint learning of two-granularity labels under different weighting strategy on CUB-200-2011 bird dataset. (a) x-axis: β value that controls the relative importance of a fine-grained classifier; y axis: performance of the coarse-grained classifier. (b) x-axis: α value that controls the relative importance of a coarse-grained classifier; y axis: performance of the fine-grained classifier.

4.1. Cooperation or Confrontation?

To explore the transfer effect in the joint learning of multi-granularity labels, we design an image classification task for predicting two labels at different granularities, i.e., $K = 2$. We form our train/test set from CUB-200-2011 bird dataset and assign each image with two labels at order and family level. During training, we introduce two weights as hyperparameters to control the relative importance of each task. This is formulated as:

$$\alpha L_{CE}(\hat{y}^1, y^1) + \beta L_{CE}(\hat{y}^2, y^2) \tag{1}$$

where a larger value of α and β then prioritise feature learning towards predicting coarse-grained and fine-trained labels respectively.

Figure 4(a) shows that by keeping $\alpha = 1.0$ and gradually increasing the value of β from 0.0 to 1.0, coarse-grained classifier is constantly reinforced when the features is optimised towards fineness. This is in a stark contrast with Figure 4(b) where the performance of fine-grained classifier becomes consistently worse with the increasing proportions of coarse-level features. This provides compelling evidence to the discovery we mentioned earlier: coarse-level label prediction in fact hurts fine-grained feature learning, yet fine-level feature betters the learning of coarse-level classifiers. Such finding is also intuitively understandable because models optimised towards finer-grained recognition are forced to interpret and analyse more local and subtle discriminative regions. They thus comprise additional useful information for coarse-grained classifiers as well. In comparison, features optimised for predicting coarse-grained labels are less likely to generalise.

To provide further proof, we visualise the feature embeddings learned under four weighting strategies using t-SNE, i.e., $\{\alpha = 1, \beta = 0\}$, $\{\alpha = 1, \beta = 1\}$, $\{\alpha = 0, \beta = 1\}$, $\{\alpha = 1, \beta = 1\}$. Same conclusions still hold. The decision boundaries for coarse-grained label classifiers become more separated with the help of finer-grained features, while fine-grained classifiers are getting worse in this sense given the increasing involvement of coarser-grained features.

4.2. Disentanglement and Reinforcement

Observations in Section 4.1 suggests that there involves both positive and negative task transfer in multi-granularity label predictions. This leads to our two technical considerations: (i) To restrain from the negative transfer between label predictions at different granularity, we first explicitly disentangle the decision space by constructing granularity-specific classification heads. (ii) We then implement the potential of positive transfer by allowing fine-grained features to participate in the coarse-grained label predictions and make smart use of gradients to enable better disentanglement.

Specifically, We first split f into K equal parts, with each representing a feature f_k independently responsible for one classifier $\mathcal{G}_k(\cdot)$. To allow finer-grained features in jointly predicting a coarse-grained label y^k, we concatenate feature f_k and all the other finer features $f_{k+1}, f_{k+2},...,f_K$ as input to the classifier $\mathcal{G}_k(\cdot)$. One issue remains unsolved. While we have adopted finer-grained features to improve coarse-grained label predictions, this risks the fact that features belonging to fine-grained classifiers will be biased towards coarse-grained recognition during model optimisation and undermines our efforts on disentanglement. We therefore introduce a gradient controller $\Gamma(\cdot)$. That is during the model backward passing stage, we only propagate the gradients flow of one classifier along its own feature dimensions and stop other gradients via $\Gamma(\cdot)$. This gives us final representation of predicting a label:

$$\hat{y}^k = \mathcal{G}_k(CONCAT(f_k, \Gamma(f_k + 1), ..., \Gamma(f_K))) \tag{2}$$

5. Experimental Settings

Datasets We evaluate our proposed method on three widely used FGVC datasets. While some dataset only of-

Method	**CUB-200-2011**				**FGVC-Aircraft**				**Stanford Cars**		
	order_acc	**family_acc**	**specie_acc**	**avg_acc**	**maker_acc**	**family_acc**	**model_acc**	**avg_acc**	**maker_acc**	**model_acc**	**avg_acc**
Vanilla_single	95.38 ± 0.47	87.70 ± 0.79	74.24 ± 0.86	85.77	90.82 ± 1.02	88.73 ± 1.17	86.26 ± 1.37	88.60	95.30 ± 0.11	88.66 ± 0.45	91.98
Vanilla_multi	95.13 ± 0.53	89.70 ± 0.13	78.31 ± 0.35	87.71	90.69 ± 0.48	89.23 ± 0.53	88.10 ± 0.10	89.34	95.24 ± 0.20	89.14 ± 0.16	92.19
Ours_single	95.63 ± 0.27	88.50 ± 0.15	77.46 ± 0.10	87.50	90.73 ± 0.23	89.39 ± 0.11	87.96 ± 0.27	89.36	95.23 ± 0.09	89.12 ± 0.29	92.18
Ours	96.37 ± 0.16	90.39 ± 0.15	77.95 ± 0.04	88.24	**93.04±0.25**	90.73 ± 0.19	**88.35±0.18**	**90.71**	95.58 ± 0.06	89.66 ± 0.16	92.62
Ours_MC	96.58 ± 0.15	90.36 ± 0.07	77.85 ± 0.38	88.26	92.86 ± 0.12	90.74 ± 0.11	88.19 ± 0.11	90.59	95.56 ± 0.17	89.62 ± 0.21	92.59
Ours_NTS	**96.57±0.07**	**91.58±0.57**	**80.45±0.68**	**89.53**	92.48 ± 0.16	**90.75±0.07**	88.31 ± 0.23	90.51	**95.96±0.39**	**90.64±0.37**	**93.30**
Ours_PMG	**97.98±0.12**	**93.50±0.10**	**82.26±0.13**	**91.25**	**94.57±0.10**	**92.44±0.07**	**89.62±0.15**	**92.21**	**96.42±0.05**	**91.05±0.15**	**93.74**

Table 2: Comparisons with different baselines for FGVC task under multi-granularity label setting.

fers one fine-grained label for each of its images, we manually construct a taxonomy of label hierarchy by tracing their parent nodes (superclasses) in Wikipedia pages. Details are as follows. (i) **CUB-200-2011** [42] is a dataset that contains 11, 877 images belonging to 200 bird species. We re-organise this dataset into three-level label hierarchy with 13 orders (e.g., "Passeriformes" and "Anseriformes"), 38 families (e.g., "Icteridae" and "Cardinalidae") and 200 species (e.g., "Brewer Blackbird" and "Red winged Blackbird"). (ii) **FGVC-Aircraft** [29] is an aircraft dataset with 10, 000 images covering 100 model variants. It comes with three-level label hierarchy with 30 makers (e.g., "Boeing" and "Douglas Aircraft Company"), 70 families (e.g.," Boeing 767"," Boeing 777"), and 100 models (e.g., "767-200", "767-300"), which we directly adopt for our setting. (iii) **Stanford Cars** [25] contains 8, 144 car images categorised by 196 car makers. We re-organise this dataset into two-level label hierarchy with 9 car types (e.g., "Cab" and "SUV") and 196 specific models (e.g., "Cadillac Escalade EXT Crew Cab 2007" and "Chevrolet Avalanche Crew Cab 2012"). We follow the standard train/test splits as laid out in the original datasets. We do not use any bounding box/part annotations in all our experiments.

Implementation details For fair comparisons, we adopted ResNet50 pre-trained on ImageNet as our network backbone and resize each input image to 224 × 224 throughout the experiments unless otherwise specified. We set the number of hidden units in f as 512 when a single model is asked to predict one label only, and 600 when that is adapted for multiple labels. To deal with the imbalance between ImageNet pre-trained convolutional layers and newly added fully-connected layers in the classification heads, we adopt different learning rates starting from 0.01 and 0.1 respectively. Common training augmentation approaches including horizontal flipping and random cropping, as well as colour jittering are applied. We train every single experiment for 100 epochs with weight decay value as 5×10^{-4}. `MomentumOptimizer` is used with momentum value 0.9 throughout. The code will be made publicly accessible.

Evaluation metrics Following community convention, FGVC performance is quantified by acc, the percentage of images whose labels are correctly classified. We use avg_acc to calculate the mean of the performance across label granularities. Each experiment is run three times. The mean and standard deviation of the results obtained over three trials are then reported.

Baselines As our focus is on how to adapt an image classification model with single label output into multiple ones, our baselines comprise alternative multi-label classification models. To show our proposed solution is generic to any existing FGVC frameworks, we also include three other baselines by replacing the backbone of our model with different advanced FGVC-specific components. **Vanilla_single**: this corresponds to one single shared network backbone with multiple classification heads appended to the end. **Vanilla_multi** adopts one independent network backbone for each label prediction. **Ours_single** improves upon Vanilla_single aiming to disentangle the decision space in multi-granularity label predictions. This is achieved by splitting f into equal number of segments as that of classifiers, with each independently responsible for one classifier at one granularity. **Ours** advances Ours_single in better feature disentanglement by reinforcing coarse-grained classifiers with fine-grained features. Finally, **Ours_MC** [5], **Ours_NTS** [50], **Ours_PMG** [12], represent three means of training our proposed method on top of state-of-the-art FGVC frameworks.

6. Results and Analysis

6.1. Comparison with Baselines

Our experimental discovery coincides well with our intuition that compared with classifying one fine-grained label, there exists additional issue that needs to be taken care of in multi-granularity label predictions. Our proposed method can not only effectively solve this problem, but also generic in terms of the network backbone used. Belows is more detailed analysis of the results with reference to Table 2.

Is our model effective in solving FGVC problem with multi-granularity label output? Yes. It is evident that the proposed model (Ours) outperforms all other baselines under the metric of avg_acc on all three datasets. Furthermore, the consistent performance gain from Our_MC to Ours_NTS, and to Ours_PMG tells one important message: our solution not only supports easy drop-in to existing

FGVC models, but also does not undermine their original functionality when adapted.

Are the proposed technical contributions appropriate? Yes. The significant gap between Vanilla_single and Ours_single confirms the severity of feature entanglement between label granularities - that can be alleviated by simply splitting a feature into several parts with each corresponding to an independent classifier. The proposed Reinforce module (Ours_single vs. Ours) is effective in boosting the classification performance at coarse granularity (e.g., order_acc and family_acc in CUB-200-2011). The fact that it can also achieve higher accuracy on the finest labels (e.g., species_acc), a task which has not been explicitly designed to improve on, provides direct evidence of how better feature disentanglement is further taking place.

What does Vanilla_multi tell us? The performance of Vanilla_multi draws our attention. On one hand, its accuracy on the finest label prediction crushes all opponents by significant margins across the datasets. On the other, it performs the worst on classifying coarsest labels. Such contrast, however, echoes our observation that underlies the technical considerations of this paper: finer-grained classifier performs the best when it is portrayed as a single independent task itself, while coarser-level label predictions can benefit significantly from a multi-granularity task setting. Note that since Vanilla_multi requires equal number of unshared network backbones as that for classification tasks, it is not a strictly fair comparison in terms of its model capacity. The purpose here is to show solving disentanglement between label prediction at different granularities remains challenging, albeit we have greatly advanced the problem.

What does it look like? We further carry out model visualisation to demonstrate that classifiers $[\mathcal{G}_1, ..., \mathcal{G}_K]$ under Vanilla_single and Ours indeed capture different regions of interests that are useful for FGVC, and offer insight on how better disentanglement is taking place. To this end, we adopt Grad-Cam [36] to visualise the different image supports for each $\mathcal{G}_k$ by propagating their gradients back to x. It can be seen from the bottom half of Figure 5 that our classifiers at different hierarchy levels attends to different scales of visual regions – a clear sign of the model's awareness on coarse-fine disentanglement. In contrast, the top half of Figure 5 shows that Vanilla_single appears to focus on similar un-regularised image parts across label granularity.

6.2. Evaluation on traditional FGVC setting

Our model can be evaluated for FGVC without any changes – we just need to report classification accuracy for fine-grained labels at the bottom of the hierarchy. However, for fair comparison with other state-of-the-art FGVC works, we also resize image input to a size of 448 × 448. We leave all other implementation settings unchanged, and do not perform grid search for performance gain. The results are reported in Table 3. We can see that by building our method upon the backbone of PMG, new state-of-the-art results (Ours_PMG) for traditional FGVC setting are gained on CUB-200-2011 and FGVC-Aircraft datasets. Improvements over state-of-the-art on Stanford Cars dataset is less significant. We attribute this to the relatively shallow hierarchy (two levels) on Stanford Cars. Note that we do not introduce any extra parameters when implemented on top of traditional FGVC methods.

Method	**CUB-200-2011**	**FGVC-Aircraft**	**Stanford Cars**
FT ResNet [45]	84.1	88.5	91.7
DFL [45]	87.4	91.7	93.1
NTS [50]	87.5	91.4	93.9
TASN [57]	87.9	-	93.8
API-Net [59]	87.7	93.0	**94.8**
MC-Loss [5]	87.3	92.6	93.7
PMG [12]	**89.6**	**93.4**	**95.1**
Ours	86.8	92.8	94.3
Ours_PMG	**89.9**	**93.6**	**95.1**
Ours_HC	85.0	90.6	92.8
Ours_DFT	85.5	91.7	93.2

Table 3: Performance comparisons on traditional FGVC setting with single fine-grained label output.

The role of label hierarchy To investigate the impact of label hierarchy on the traditional FGVC performance, we compare our manual method of constructing label hierarchy based on Wikipedia pages with two variants, Hierarchical Clustering (Ours_HC) and Deep Fuzzy Tree (Ours_DFT) [44]. These are two clustering methods that automatically mine hierarchical structures from data, which mainly differ in how to measure the distance between clusters and whether there are tree structures explicitly modelled. For both methods, we stop the discovery process when three-level label hierarchy has been formed. From the last two rows in Table 3, the following observations can be made: (i) Manual hierarchies achieves the best performance across all three datasets, suggesting semantically defined parent-child relationships tend to encourage cross granularity information change. (Ours vs. Ours_HC vs. Ours_DFT); (ii) Traditional FGVC problem (FT ResNet) benefits from multi-granularity label setting, regardless of what label hierarchy is used.

7. Discussions

Here, we offer discussions on some potentially viable future research directions, with the hope to encourage follow up research.

Beyond multi-task learning While our MTL framework has shown promise as a first stab, other means of encouraging information exchange/fusion across hierarchy levels can be explored. One possible alternative is via meta learning [20]. In this sense, rather than learning multi-granularity label prediction task in one shot, we can treat them as a

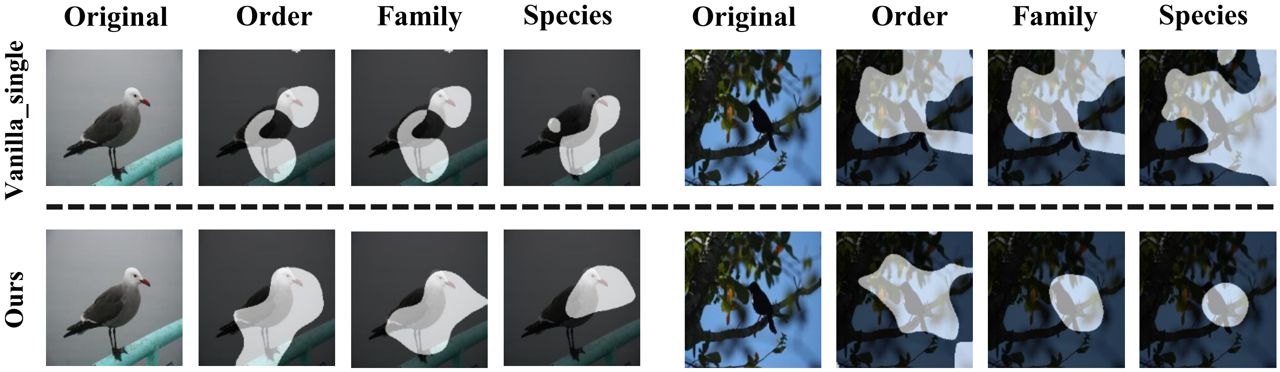


Figure 5: We highlight the supporting visual regions for classifiers at different granularity of two compared models. Order, Family, Species represent three coarse-to-fine classifiers trained on CUB-200-2011 bird dataset.

sequence of related tasks optimised over multiple learning episodes. An idea could be that in the inner loop, we find a meta-learner that serves as good initialisation with few gradients away to each task (as per disentanglement). We then ask the outer task-specific learners to quickly adapt from it (as per reinforcement).

From classification to retrieval. Formulating the problem of fine-grained visual analysis as a classification task itself underlies certain limitations: the fixed number of labels makes it rigid to be applied in some open-world scenarios [47]. By projecting images into a common embedding space (as per retrieval) however, we will not only grant the flexibility but also potentially relax the ways of granularity interpretation into model design. Pretending that we were to address the goal of this paper from a retrieval perspective, we can associate label granularity with the model's receptive field – the finer the label, the more local the regions of interest. We can also potentially directly use label granularity as an external knowledge to dynamically parameterise the embedding space (as per hypernetworks [18]). More importantly, a successfully-trained model now has a chance to learn a smooth interpolation between label granularities, which is of great practical value but infeasible under the formulation of classifiers.

Rethinking ImageNet pre-training FGVC datasets remain significantly smaller than modern counterparts on generic classification [11, 33]. This is a direct result of the bottleneck on acquiring expert labels. Consequently, almost all contemporary competitive FGVC models rely heavily on pre-training: the model must be fine-tuned upon the pre-trained weights of an ImageNet classifier. While useful in ameliorating the otherwise fatal lack of data, such practice comes with a cost of potential mismatch to the FGVC task – model capacity for distinguishing between "dog"' and "cat" is of little relevance with that for differentiating "Giant Ibis" and "flamingo". In fact, our paper argues otherwise – that coarse-level feature learning is best disentangled from that of fine-grained. Recent advances on self-supervised representation learning provide a promising label-efficient way to tailor pre-training approaches for downstream tasks [32, 41]. However, its efficacy remains unknown for FGVC.

8. Conclusion

Following a human study, we re-envisaged the problem of fine-grained visual classification, from the conventional single label output setting, to that of coarse-fine multi-granularity label prediction. We discovered important insights on how positive information exchange across granularities can be explored. We then designed a rather simple yet very effective solution following these insights. Extensive experiments on three challenging FGVC datasets validate the efficacy of our approach. When evaluated on the traditional FGVC setting, we also report state-of-the-art results while not introducing any extra parameters. We will release all human study data, and make our code publicly accessible. Last but not least, we hope to have caused a stir, and trigger potential discussions on the very title of this paper – that whether my "Flamingo" should or should not be your "Bird".

Acknowledgements This work was supported in part by the National Key R&D Program of China under Grant 2019YFF0303300 and under Subject II No. 2019YFF0303302, in part by National Natural Science Foundation of China (NSFC) No. 61922015, 61773071, U19B2036, in part by Beijing Natural Science Foundation Project No. Z200002, in part by the Beijing Academy of Artificial Intelligence (BAAI) under Grant BAAI2020ZJ0204, in part by the Beijing Nova Programme Interdisciplinary Cooperation Project under Grant Z191100001119140, in part by a scholarship from China Scholarship Council (CSC) under Grant CSC No. 202006470036, and in part by BUPT Excellent Ph.D. Students Foundation No. CX2020105.

References

[1] Thomas Berg and Peter Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013. 1, 2

[2] Irving Biederman, Suresh Subramaniam, Moshe Bar, Peter Kalocsai, and Jozsef Fiser. Subordinate-level object classification reexamined. *Psychological research*, 1999. 1

[3] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014. 1

[4] Yuning Chai, Victor Lempitsky, and Andrew Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *ICCV*, 2013. 2

[5] Dongliang Chang, Yifeng Ding, Jiyang Xie, Ayan Kumar Bhunia, Xiaoxu Li, Zhanyu Ma, Ming Wu, Jun Guo, and Yi-Zhe Song. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, 2020. 2, 6, 7

[6] Liangfu Chen, Zeng Yang, Jianjun Ma, and Zheng Luo. Driving scene perception network: Real-time joint detection, depth estimation and semantic segmentation. In *WACV*, 2018. 3

[7] Xin Chen, Lingxi Xie, Jun Wu, and Qi Tian. Progressive differentiable architecture search: Bridging the depth gap between search and evaluation. In *CVPR*, 2019. 3

[8] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *CVPR*, 2019. 2

[9] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018. 3

[10] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 8

[12] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Yi-Zhe Song, Zhanyu Ma, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *ECCV*, 2020. 6, 7

[13] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. Pairwise confusion for fine-grained visual classification. In *ECCV*, 2018. 2

[14] Shenghua Gao, Ivor Wai-Hung Tsang, and Yi Ma. Learning category-specific dictionary and shared dictionary for fine-grained image categorization. *IEEE Transactions on Image Processing*, 2013. 1

[15] Yuan Gao, Jiayi Ma, Mingbo Zhao, Wei Liu, and Alan L Yuille. Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *CVPR*, 2019. 3

[16] Nina Granqvist, Stine Grodal, and Jennifer L Woolley. Hedging your bets: Explaining executives' market labeling strategies in nanotechnology. *Organization science*, 2013. 3

[17] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *ECCV*, 2018. 3

[18] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 8

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 3

[20] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020. 7

[21] Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *CVPR*, 2020. 2

[22] Ruyi Ji, Longyin Wen, Libo Zhang, Dawei Du, Yanjun Wu, Chen Zhao, Xianglong Liu, and Feiyue Huang. Attention convolutional binary neural tree for fine-grained visual categorization. In *CVPR*, 2020. 2

[23] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 3

[24] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, 2017. 3

[25] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshops*, 2013. 6

[26] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, 2015. 1, 2

[27] Chaohao Lu and Yuexian Zou. Using coarse label constraint for fine-grained visual classification. In *MMM*, 2019. 2

[28] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, 2020. 3

[29] Subhransu Maji, Juho Kannala, Esa Rahtu, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. arXiv:1306.5151. 6

[30] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, 2016. 3

[31] Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C Berg, and Tamara L Berg. Predicting entry-level categories. *International Journal of Computer Vision*, 2015. 1

[32] Kaiyue Pang, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In *CVPR*, 2020. 8

[33] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 8

[34] Yuxin Peng, Xiangteng He, and Junjie Zhao. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing*, 2017. 1

[35] Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Latent multi-task architecture learning. In *AAAI*, 2019. 3

[36] 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*, 2017. 7

[37] Ayan Sinha, Zhao Chen, Vijay Badrinarayanan, and Andrew Rabinovich. Gradient adversarial training of neural networks. *arXiv preprint arXiv:1806.08028*, 2018. 3

[38] Guolei Sun, Hisham Cholakkal, Salman Khan, Fahad Shahbaz Khan, and Ling Shao. Fine-grained recognition: Accounting for subtle differences between similar classes. In *AAAI*, 2020. 2

[39] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *ECCV*, 2018. 2

[40] Lehana Thabane, Jinhui Ma, Rong Chu, Ji Cheng, Afisi Ismaila, Lorena P Rios, Reid Robson, Marroon Thabane, Lora Giangregorio, and Charles H Goldsmith. A tutorial on pilot studies: the what, why and how. *BMC medical research methodology*, 2010. 3

[41] Michael Tschannen, Josip Djolonga, Marvin Ritter, Aravindh Mahendran, Neil Houlsby, Sylvain Gelly, and Mario Lucic. Self-supervised learning of video-induced visual invariances. In *CVPR*, 2020. 8

[42] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *Caltech, Technical Report*, 2011. 2, 6

[43] Li Wang, Dong Li, Yousong Zhu, Lu Tian, and Yi Shan. Dual super-resolution learning for semantic segmentation. In *CVPR*, 2020. 2

[44] Yu Wang, Qinghua Hu, Pengfei Zhu, Linhao Li, Bingxu Lu, Jonathan M Garibaldi, and Xianling Li. Deep fuzzy tree for large-scale hierarchical visual classification. *IEEE Transactions on Fuzzy Systems*, 2019. 7

[45] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *CVPR*, 2018. 2, 7

[46] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *ICCV*, 2019. 2

[47] Xiu-Shen Wei, Jianxin Wu, and Quan Cui. Deep learning for fine-grained image analysis: A survey. *arXiv preprint arXiv:1907.03069*, 2019. 8

[48] Junfeng Wu, Li Yao, Bin Liu, and Zheyuan Ding. Leveraging fine-grained labels to regularize fine-grained visual classification. In *ICCMS*, 2019. 1

[49] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, 2015. 2

[50] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *ECCV*, 2018. 2, 6, 7

[51] Bangpeng Yao, Gary Bradski, and Li Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *CVPR*, 2012. 1

[52] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *CVPR*, 2016. 1

[53] Yu Zhang, Xiu-Shen Wei, Jianxin Wu, Jianfei Cai, Jiangbo Lu, Viet-Anh Nguyen, and Minh N Do. Weakly supervised fine-grained categorization with part-based image representation. *IEEE Transactions on Image Processing*, 2016. 2

[54] Xiangyun Zhao, Haoxiang Li, Xiaohui Shen, Xiaodan Liang, and Ying Wu. A modulation module for multi-task learning with applications in image retrieval. In *ECCV*, 2018. 3

[55] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, 2017. 1

[56] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning deep bilinear transformation for fine-grained image representation. In *NeurIPS*, 2019. 2

[57] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *CVPR*, 2019. 1, 7

[58] Ling Zhou, Zhen Cui, Chunyan Xu, Zhenyu Zhang, Chaoqun Wang, Tong Zhang, and Jian Yang. Pattern-structure diffusion for multi-task learning. In *CVPR*, 2020. 3

[59] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. In *AAAI*, 2020. 2, 7