FINE-GRAINED IMAGE CLASSIFICATION WITH COARSE AND FINE LABELS ON ONE-SHOT LEARNING

Qihan Jiao[†‡], Zhi Liu[†‡], Gongyang Li[†‡], Linwei Ye[§], Yang Wang[§]*

[†]Shanghai Institute for Advanced Communication and Data Science, Shanghai University, China
[‡]School of Communication and Information Engineering, Shanghai University, China
[§]Department of Computer Science, University of Manitoba, Canada
jiaoqihan@shu.edu.cn, liuzhisjtu@163.com, ligongyang@shu.edu.cn, yel3@cs.umanitoba.ca,
ywang@cs.umanitoba.ca

ABSTRACT

In this paper, we aim to solve the fine-grained image classification on one-shot learning, which only has one image provided from each class. Specifically, we introduce the hierarchical structure between coarse and fine labels to exploit the relationship among categories. First, we make coarse label prediction of the input image and utilize Attention Proposal Network (APN) to determine the attentive area for fine label prediction. Then, according to the result of coarse label prediction, we can automatically select the images belong to the same coarse category from all samples in the support set to form a subset, which will be sent to relation network. Finally, we fuse the results of relation network and those of fine label prediction to produce more robust and more accurate classification results. The superior fine-grained classification performance of our method is demonstrated on CUB-200-2011 dataset and miniImageNet dataset.

Index Terms— Fine-grained classification, one-shot learning, attention proposal network, relation network.

1. INTRODUCTION

Fine-grained image classification aims to recognize the fine classes belonging to the same coarse class. The difficulty of the task is that in some cases, the intra-class variance is large while the inter-class variance is small. Furthermore, the subtle differences between fine classes may only exist in local areas, which are difficult to detect and discriminate.

Existing methods [1-9] generally first locate the objects and discriminative parts, and then extract the features of these regions for recognizing the subcategories. Some works [1, 2] require extra bounding boxes and part annotations during the training phase. Recent deep learning based fine-grained image classification methods [3-9] gradually relieve the dependence of bounding boxes and part annotations. These methods can automatically locate the objects and discriminative parts with only image-level labels for training.

However, it is still difficult and expensive to obtain all accurate fine labels for all images in a huge training dataset which is critical for deep learning based methods. Oppositely, human is able to learn and recognize diverse objects effectively by only seeing a few examples. Thus, the few-shot learning [10-16] is introduced to solve fine-grained classification. Few-shot learning can be formulated as meta learning mechanism in supervised learning. Meta learning is also known as "learning to learn", which divides the dataset into different meta tasks to learn the generalization ability of the model. The meta task changes in each episode at the training phase, producing a robust trained model for testing phase. The classification is solved by few-shot learning with only a few samples are provided from per class. Some methods [10-13] focus on modeling the distance distribution between samples by finding a reasonable metric, so that the samples of the same classes are close while the samples of different classes are far away. Based on the label of image, a multi-attention network [14] is proposed to provide attention maps for obtaining representation of image, which can make full use of the information of category label. In [15], a multi-stage data augmentation method is used to improve the classification performance with only learning from one sample. In [16], a bilinear structure is exploited to obtain image representations, and a piecewise classifier mapping is used for fine-grained classification. However, these methods [10-16] lack consideration about the characteristics of fine-grained classification tasks, such as how to solve the problems of small differences among classes while large differences within classes in fine-grained classification task.

On the other hand, the hierarchical structure between coarse and fine labels, as shown in Fig. 1, can introduce more comprehensive information among categories and help to reduce the influence of intra-class and inter-class differences on fine-grained classification [17-22]. Specially, the

This work was supported by the National Natural Science Foundation of China under Grant No. 61771301.

*Corresponding author: Zhi Liu.

978-1-7281-1485-9/20/$31.00 ©2020 IEEE

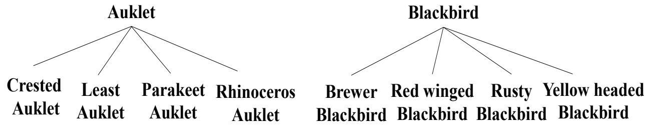

Fig. 1. The examples of hierarchical structure among coarse and fine labels.

taxonomy [17] or category hierarchy [18-20] are exploited to introduce the internal relationship among categories for improving the performance of fine-grained classification. And in [21, 22], the prediction of fine-grained categories is performed under a weakly supervised setting, the hierarchical structure among classes is introduced to provide extra information to improve the classification performance. The hierarchical structure has the ability of progressive classification, which can promote the accuracy of classification in stages.

Inspired by the few-shot learning and the hierarchical structure, in this paper, we take advantage of them and propose a one-shot learning based model with coarse and fine labels prediction for fine-grained classification. Our key idea is to introduce one-shot learning to solve the fine-grained classification with only learning from one sample of each class, which can reduce the dependence on a large amount of training data effectively. Furthermore, we explore the hierarchical structure as additional auxiliary information to effectively reduce the influence of large variance in intra-class and small variance in inter-class, which is always a difficulty in fine-grained classification.

In particular, we first predict the coarse label of input image. According the result of coarse label prediction, we choose the images belonging to the same coarse category as input image from the support set to form a subset. The selected subset and the input image are sent to the relation network. The relation network generates relation scores through comparing the input image with the images in the selected subset. Then an Attention Proposal Network (APN) [5] is used to locate the key area with effective features, and the attentive area is sent to the branch of fine label prediction. Finally, we fuse the results of fine label prediction branch and those of the relation network for final fined-grained classification. In this way, our method can achieve the fine-grained classification with only learning from one sample of each class and take advantage of the hierarchical structure among classes to reduce the influence of variance in intra-class and inter-class, resulting in excellent performance.

Our main contributions can be summarized as follows:

1) We introduce a hierarchical structure between coarse and fine categories for reducing the comparison range from all samples in the support set to a subset for the relation network making fine-grained classification prediction with samples of the same coarse label. It also relieves the influence of inter-class and intra-class differences on fine-grained classification.

2) We fuse the fine-grained classification results of the relation network with those of the fine label prediction branch to obtain more reliable classification results, which can further improve the final fine-grained classification performance.

3) The proposed method achieves the superior fine-grained classification performance on two datasets including CUB-200-2011 and miniImageNet.

2. PROPOSED METHOD

The architecture of our model is shown in Fig. 2. We first make coarse label prediction of the input image (marked with a red dashed box) and utilize the APN to determine the attentive area with more effective features for fine-grained classification. The attentive area is used as the input of fine label prediction (marked with a green dashed box). We select the images belong to the same coarse category as the input (marked by red solid wireframe in "Support set" of Fig. 2) according to the result of coarse label prediction, which is called as coarse label filter process in the following. This process can reduce the comparison range of all samples in support set, generating a subset of support set. The subset will be sent to the relation network for generating the relation score. Finally, we introduce mutual authentication method to fuse the results of the relation network and those of the fine label prediction branch to obtain more robust and more accurate classification results.

2.1. Problem definition

We first briefly introduce the classification task based on few-shot learning. There are three sets: a training set, a support set and a testing set. The support set and testing set have the same label space, while the training set has the label space which is disjoint with the label space of the testing set and the support set. Concretely, if $\boldsymbol{C}$ classes need to be recognized and $\boldsymbol{K}$ samples are provided for each class, the support set consists of $\boldsymbol{C}\times\boldsymbol{K}$ samples, and the problem is named $\boldsymbol{C}$-way-$\boldsymbol{K}$-shot. In this paper, we consider the fine-grained classification problem with only one sample provided for each class, *i.e.* $\boldsymbol{C}$-way-$\boldsymbol{1}$-shot. Theoretically, an image classification model can be trained to distinguish $\boldsymbol{C}$ categories according to $\boldsymbol{C}\times\boldsymbol{1}$ images. However, the labeled training data is too rare to obtain satisfactory performance. Thus, we utilize the few-shot learning method, which is the same as meta learning method in the training phase, which divides the training set into different meta tasks to learn the generalization ability of the classification model. The trained model has the ability of "learning to learn" to achieve better

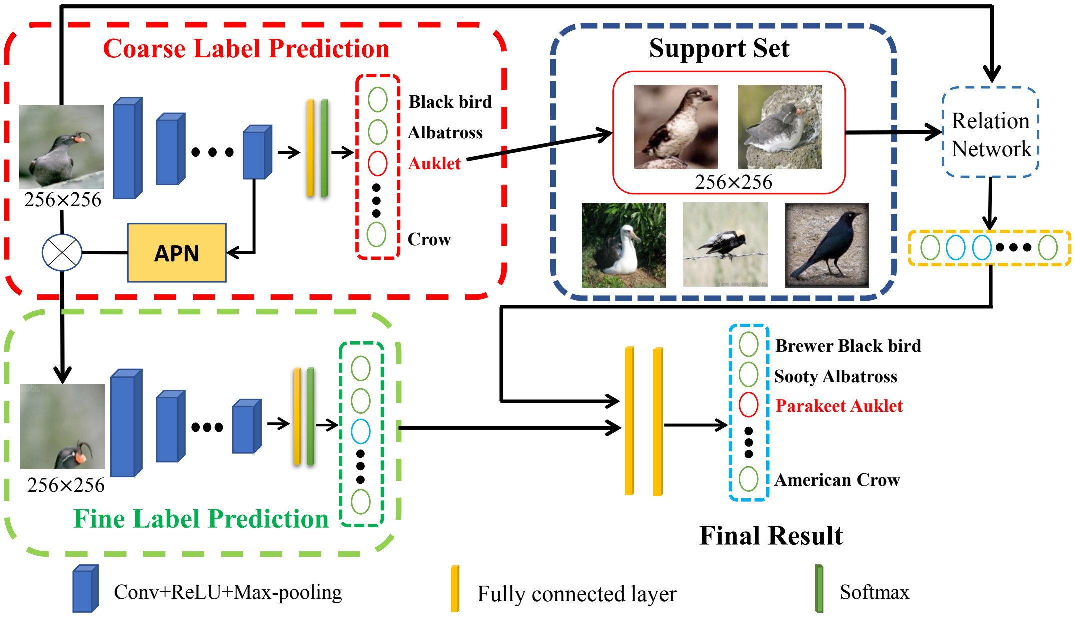

Fig. 2. The architecture of the proposed model. First, we predict the coarse label of the input image, and choose the images belong to the same coarse category with the input image from the support set and form a subset. Then the selected subset and the input image are sent to the relation network for classification. The Attention Proposal Network (APN) is used to locate the key area and the attentive area is sent to the branch of fine label prediction. Finally, the results of fine label prediction branch and those of the relation network are fused for final fined-grained classification.

performance on classification with only one sample.

2.2. Coarse and fine label prediction

The process of coarse and fine label prediction is shown in Fig. 2. The coarse label prediction consists of two parts, *i.e.* coarse label classification and APN. Concretely, the VGG network [24] is adopted to extract the features of the input image for coarse label classification, and it is also used to locate, crop and zoom the attentive area of the image for subsequent fine label prediction.

The attentive area is denoted as a square by three parameters: $\boldsymbol{t_x}$, $\boldsymbol{t_y}$ and $\boldsymbol{t_l}$, where $\boldsymbol{t_x}$, $\boldsymbol{t_y}$ are the center coordinates of the square separately, and $\boldsymbol{t_l}$ denotes the half of the side length of the square. APN can be implemented with two fully connected layers, and the last one outputs these three parameters of the attentive area. Then we search the region with the highest response value of the last convolutional layer of VGG as the initial attentive area.

When the attentive area is determined, we further crop out the attentive area as the input to the fine label prediction. The attentive area provides finer and more accurate representation of features for fine-grained label prediction. The process of cropping can be described as:

$$X^{att} = X \odot M(t_x, t_y, t_l), \qquad (1)$$

where X^{att} is the cropped area with the attention mask and $\odot$ means element-wise multiplication. $M(\cdot)$ is the attention mask, which is described as:

$$\begin{aligned} M(\cdot) &= [h(x-(t_x-t_l)) - h(x-(t_x+t_l))] \\ &\cdot [h(y-(t_y-t_l)) - h(y-(t_y+t_l))] \end{aligned}, \qquad (2)$$

where $h(\cdot)$ is:

$$h(x) = 1/\{1+\exp^{-10x}\}. \qquad (3)$$

Notably, the architecture of fine label prediction (*i.e.*, the lower branch in Fig. 2) is the same as the structure of coarse label prediction (the upper branch in Fig. 2), we will not repeat here.

2.3. Relation Network

After performing coarse label prediction on input image, we obtain the coarse category of input image. According to the coarse category, we can pick out the images from the support set to form a subset, whose coarse category is the same as input image's. The input image and the selected images are considered as the input to the relation network [13] for classification. Since the process takes advantage of the hierarchical structure between coarse and fine classes, it can reduce the comparison range (*i.e.* a completed support set to a subset of support set) and relieve the influence of intra-class

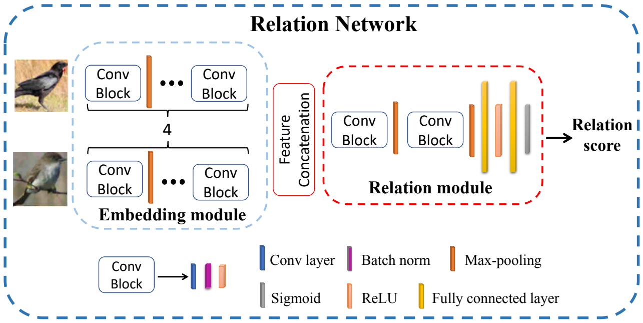

Fig. 3. The architecture of relation network.

and inter-class differences on fine-grained classification task.

Relation network [13] consists of the embedding module and the relation module, and it is proposed to measure the similarity of two images. The detailed structure of the relation network is presented in Fig. 3. Let x_t be the input image in testing set and x_s be the selected samples in support set. We first send x_t and x_s in parallel to the embedding module to obtain the corresponding feature representations $f_\varphi(x_t)$ and $f_\varphi(x_s)$.

As shown in Fig. 3, the embedding module consists of four Conv Blocks, and each Conv Block contains a convolutional layer with 3×3 kernel size and 64 channels, a batch normalization layer and a ReLu layer. The first two Conv Blocks additionally attach with a max-pooling layer while the latter two Conv Blocks have not. And the feature concatenation combine $f_\varphi(x_t)$ with $f_\varphi(x_s)$ in a cross-channel manner.

Then, the concatenated feature is sent to the relation module to produce a relation score $r_{t,s}$ between x_t and x_s, which measures the similarity between x_t and x_s. The larger the relation score is, the more likely x_t and x_s tend to be the same class. The relation module includes two Conv Blocks with a max-pooling layer attaching. Another two fully connected layers and the sigmoid function are designed at the end of the relation network for producing the relation score.

Finally, to take advantage of the relation score of the relation network and the initial results of the fine label prediction branch, we fuse them to obtain more reliable and more accurate classification results. Concretely, we connect the fine-grained classification results of the relation network and the results of the fine label prediction branch to two fully connected layers to obtain the final fused fine-grained classification results, which can be considered as the mutual authentication process.

2.4. Training strategy

During the training phase, we randomly select ***C*** classes with one sample from the training set as an episode. The selected ***C×1*** samples are defined as the sample set. A part of the remaining samples in the ***C*** classes of the training set are used as the query set. The sample set is used to imitate the support set and the query set is used to imitate the testing set. Each image in the query set is regarded as the testing image for making coarse and fine label prediction.

Specifically, the training phase contains three steps. Firstly, we jointly train the coarse and fine label prediction and APN. The total loss of this step can be described as follow:

$$L(X)=\alpha_1\cdot L_{cls}(Y^c,Y^{c*})+\alpha_2\cdot L_{cls}(Y^f,Y^{f*})+\alpha_3\cdot L_{rank}(p_t^c,p_t^f),\qquad (4)$$

where Y^c denotes the result of the coarse label prediction, Y^f denotes the result of the fine label prediction, and Y^{c*} and Y^{f*} denote the ground truth label of the coarse class and the fine class, respectively. L_{cls} is the cross-entropy loss. p_t^c and p_t^f represent the correct prediction probability of the coarse label and the fine label. L_{rank} is used to train APN, which is defined as:

$$L_{rank}(p_t^c,p_t^f)=\max\{0,p_t^c-p_t^f+margin\},\qquad (5)$$

where *margin* = 0.05. Notably, we adopt alternate training strategy in this step. The parameters of APN are firstly fixed (*i.e.* $\alpha_3=0$ in Eq. 4), and the coarse and fine label prediction model is trained until the total loss converges. Then we fix the parameters of the coarse and fine label prediction model (*i.e.* $\alpha_1=\alpha_2=0$ in Eq. 4), and train the APN until the total loss converges.

Secondly, we add the relation network to the model of the first step and fix their parameters. We adopt mean square error (MSE) loss to train the relation network and transfer the relation scores to the prediction vector.

Finally, we concatenate the results of the branch of fine label prediction and the relation network, and send them to the two newly added fully connected layers for final classification. Concretely, we fix the previous parameters, and adopt the softmax loss to train the parameters of the newly added fully connected layers.

3. EXPERIMENT RESULTS

3.1. Datasets and evaluation metrics

We conduct our experiments on two datasets: CUB-200-2011 [23] and miniImageNet [10]. CUB-200-2011 dataset includes 200 classes of birds and a total of 11,788 images. In order to be consistent with other methods for comparison, we divide the dataset into two parts: one is training set which contains 150 classes and another is testing set which contains 50 classes. For imitating the test process, in each training episode, we choose 50 classes in training set randomly, then

Table 1. The results on CUB-200-2011 dataset [23].

Method	Accuracy [%]
DHMM [25]	28.5
Pro Nets [12]	38.96
Siamese-Net [11]	37.38
Pcm [16]	42.1
Relation Network [13]	41.87
Ours	**43.83**

Table 2. The results on miniImageNet dataset [10].

Method	Accuracy [%]
Matching Nets [10]	43.56
Meta Nets [26]	49.21
Meta-learn LSTM [27]	43.44
Pro Nets [12]	49.42
Relation Network [13]	50.44
Ours	**52.13**

choose one image from each class as the sample set and 20 images from each class as the query set. The miniImageNet dataset contains 60,000 images with 100 classes, and each class consists of 600 images. The dataset is divided into three parts: 64 classes as the training set, 16 classes as the validation set and 20 classes as the testing set. Be different from CUB-200-2011, in each training episode, we choose 20 classes in the training set randomly, then choose 1 image from each class as the sample set and 15 images from each class as the query set. And the division methods of two datasets are different in order to be consistent with the respective comparison method on the corresponding dataset. All results of the experiments are evaluated by top-1 fine-grained classification accuracy.

3.2. Comparison with the state-of-the-arts

We compare our method with five methods, including DHMM [25], Pro Nets [12], Siamese-Net(FB) [11], Pcm [16] and Relation Network [13], on CUB-200-2011 dataset [23]. As shown in Table 1, we observe that our method obtains the best performance when compared with other methods. In more detail, our method is much better than DHMM [25] in terms of performance, which reaches 15.33%. Comparing to other four methods, our method still achieves significant improvement, which ranges from close to 2% to over 6%, *e.g.* our method improves the accuracy by 1.73% to Pcm [16].

For miniImageNet dataset [10], we compare our method with five methods: Matching Nets [10], Meta Nets [26], Meta-learn LSTM [27], Pro Nets [12] and Relation Network [13]. As shown in Table 2, our method also gets the best performance. Comparing to Meta-learn LSTM [27], our method has remarkable improvement in accuracy which reaches 8.69%. Our method improves the accuracy by nearly 2% to over 8% compared with other four methods, and it improves the performance by 1.69% compared to the Relation Network [13], which is the best method mentioned above.

Table 3. Ablation studies on CUB-200-2011 dataset [23] and miniImageNet dataset [10].

Variant	CUB-200-2011	miniImageNet
	Accuracy [%]	Accuracy [%]
Baseline [13]	41.87	50.44
B+Coarse	43.09	51.74
w/o APN	43.46	51.95
Ours	**43.83**	**52.13**

3.3. Ablation study

We conduct detailed examination to verify the contribution of each component in our method. The variants are provided as follows: (1) “Baseline” means that the baseline of our method is Relation Network [13]; (2) “B+Coarse” denotes to add the coarse label filter process to the baseline; (3) “w/o APN” means to remove the APN in the coarse label prediction.

As shown in Table 3, comparing “Baseline” and “B+Coarse”, we observe the contribution of the coarse label filter process is remarkable, which boosts the performance by 1.22% and 1.30% on CUB-200-2011 and miniImageNet, respectively. This means by choosing out the images belonging to the same coarse category with the input images from the support set can reduce the comparison range effectively and also can relieve the influence of intra-class and inter-class differences on fine-grained classification task by exploiting the extra information introduced by the hierarchical structure between coarse and fine classes. Our completed method adds the mutual authentication process to the “B+Coarse”. From Table 3, we observe that the mutual authentication process further improves the accuracy by 0.74% and 0.39% on CUB-200-2011 and miniImageNet, respectively. This means the more reliable classification results are obtained by fusing fine-grained classification prediction results of the relation network and the fine label prediction branch. Besides, comparing “w/o APN” and “Ours”, we observe that the APN promotes the accuracy by 0.37% and 0.18% on CUB-200-2011 and miniImageNet, respectively.

4. CONCLUSION

In this paper, we propose a fine-grained classification method under the setting of one-shot learning. First, the coarse label prediction is made to generate the coarse label of input image. According to the coarse label, we get a subset of support set which is the same coarse category as the input image. This process limits the comparison range for making fine-grained classification through relation network and reduces the influence of inter-class and intra-class differences on fine-grained classification. During the coarse label prediction, the attentive area can also be obtained by attention proposal network, and it provides the attentive part with effective features for fine-grained classification. Then, the attentive area will be sent to the branch of fine label prediction. Finally, the results of relation network and those of the fine label prediction branch are fused to make more reliable classification results. The experiments on CUB-200-2011 and miniImageNet demonstrate the superiority of our method.

5. REFERENCES

[1] R. Hong, Z. Hu, and R. Wang, "Part-based r-cnns for fine-grained category detection," in *ECCV*, 2014, pp. 834-849.

[2] S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," in *BMVC*, 2014, pp. 1-14.

[3] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *CVPR*, 2015, pp. 842-850.

[4] Y. Peng , X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1487-1500, 2018.

[5] J. Fu , H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *CVPR*, 2017, pp. 4476-4484.

[6] X. He, Y. Peng, and J. Zhao, "Fine-grained discriminative localization via saliency-guided faster r-cnn," in *ACM MM*, 2017, pp. 627-635.

[7] Y. Zhang, X. Wei, J. Wu, J. Cai, J. Lu, V. Nguyen, and M. N. Do, "Weakly supervised fine-grained categorization with part-based image representation," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1713-1725, 2016.

[8] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking deep filter responses for fine-grained image recognition," in *CVPR*, 2016, pp. 1134-1142.

[9] Z. Wang, S. Wang, P. Zhang, H. Li, and B. Liu, "Accurate and fast fine-grained image classification via discriminative learning," in *ICME*, 2019, pp. 634-639.

[10] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," in *NeurIPS*, 2016, pp. 3630–3638.

[11] K. Gregory, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML*, 2016, pp.1-8.

[12] S. Jake, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017, pp. 1-11.

[13] F. Sung, Y. Yang, L. Zhang, and T. Xiang, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018, pp. 1199-1208.

[14] P. Wang, L. Liu, C. Shen, Z. Huang, and A. Hengel, "Multi-attention network for one-shot learning," in *CVPR*, 2017, pp. 6212-6220.

[15] X. He and Y. Peng, "Only learn one sample: Fine-grained visual categorization with one sample training," in *ACM MM*, 2018, pp. 1372-1380.

[16] X. Wei, P. Wang, L. Liu, C. Shen, and J. Wu, "Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples," *arXiv preprint arXiv: 1805.04288*, 2018.

[17] W. Goo, J. Kim, G. Kim, and S. J. Hwang, "Taxonomy-regularized semantic deep convolutional neural networks," in *ECCV*, 2016, pp. 86-101.

[18] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu, "HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition," in *ICCV*, 2015, pp. 2740-2748.

[19] G. Zhang, L. Chen, and Y. Ding, "A multi-label classification model using convolutional neural networks," in *CCDC*, 2017, pp. 2151-2156.

[20] R. Hagawa, Y. Ishii, and S. Tsukizawa, "Multi-staged deep learning with created coarse and appended fine categories," in *ACPR*, 2015, pp. 36-40.

[21] J. Lei, Z. Guo, and Y. Wang, "Weakly supervised image classification with coarse and fine labels," in *CRV*, 2017, pp. 240-247.

[22] Q. Jiao, Z. Liu, L. Ye, and Y. Wang, "Weakly labeled fine-grained classification with hierarchy relationship of fine and coarse labels," *JVCIR*, vol. 63, pp. 102584, 2019.

[23] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015, pp. 1409-1556.

[25] F. Pahde, M. Nabi, T. Klein, and P. Jahnichen, "Discriminative hallucination for multi-modal few-shot learning," in *ICIP*, 2018, pp. 156-160.

[26] T. Munkhdalai and H. Yu, "Meta networks," in *ICML*, 2017, pp. 2554-2563.

[27] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *ICLR*, 2017, pp. 1-11.