Exploiting Category-Level Semantic Relationships for Fine-Grained Image Recognition

Xianjie Mo[1], Jiajie Zhu[1], Xiaoxuan Zhao[1], Min Liu[2], Tingting Wei[1], and Wei Luo[1(✉)]

[1] College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, GD, People's Republic of China
cedricmo.cs@gmail.com, {Zhujiajie,zhaobear}@stu.scau.edu.cn, weitingting@scau.edu.cn, cswluo@gmail.com

[2] School of Computer, National University of Defense Technology, Changsha 410003, HN, People's Republic of China
gfsliumin@gmail.com

Abstract. We present a label-based, semantic distance induced regularization learning method for Fine-grained image recognition (FGIR). In contrast to previous label-based methods that involve a nontrivial optimization in multi-task metric learning, our approach can be integrated into an end-to-end network without introducing any extra parameters, thus easy to be optimized. To this end, a category-level hierarchical distance matrix (HDM) that encodes semantic distance between subcategories through a tree-like label hierarchy is constructed. HDM is then incorporated into a DCNN to aggregate misclassified prediction probabilities for model learning, thus providing additional discriminative information for fine-grained feature learning. Experiments on three fine-grained benchmark datasets (Stanford Cars, FGVC-Aircraft, CUB-Birds) validate the effectiveness of our approach and demonstrate its improvements over previous methods.

Keywords: Fine-grained image recognition · Deep convolutional neural networks · Label structure

1 Introduction

Fine-grained image recognition (FGIR) aims at distinguishing images of subordinate categories that belong to the same base class, e.g., bird [22] species, car models [10], aircraft model variants [16], etc. Different from the base-class recognition, the differences between subordinate categories are subtle and usually exist in local regions, which make recognition models difficult to learn effective discriminative feature to distinguish them, e.g., for a car dataset with four-level

The first author is a student.

Z. Lin et al. (Eds.): PRCV 2019, LNCS 11857, pp. 50–62, 2019.
https://doi.org/10.1007/978-3-030-31654-9_5

hierarchical annotations—year, model, make and type, the visual differences in appearance between cars of the same type but from different makers and models are difficult to be ascertained. Intuitively, it is important to delve deeply into the rich semantic relationship inherited in fine-grained categories to learn better representations for FGIR.

Exploiting semantic relationships between categories for FGIR is usually proceeded from three aspects—parts [6,7,29,30], objects [24,34], and labels [23,25, 33]. Leveraging parts and bounding box annotations to propose geometry constraints for part detector learning [29,30] possesses the advantage of reducing the number of false positive parts compared to that of unsupervised part learning methods [20,26]. However, the application of this method is limited due to its requirements of annotations. Utilizing the relationship between objects is implicitly implemented in the configuration of weakly-supervised learning, which minimizes averaging prediction loss across training samples with only image-level labels available [24,34]. This methodology relaxes the requirements of data annotation but with a trade-off between data annotation and prediction accuracy. Constructing the hierarchy of fine-grained labels for FGIR was also studied in the community. A number of methods propose to learn different-granularity features from different label granularities [23], and several other approaches make efforts to utilize similarities among categories through multi-task metric learning [33]. Although the idea of learning fine-grained features by leveraging label relationships is straightforward, it is usually difficult to well define the relationships between labels and involves a nontrivial optimization procedure to learn the model, like the metric learning presented in [33].

In this paper, we propose a simple but effective regularization method that exploits the semantic relationship between categories by constructing a hierarchy distance matrix from fine-grained labels for FGIR. Our method leverages the hierarchical structure inherited in fine-grained labels to build a category-level hierarchical distance matrix (HDM), in which each entity represents the semantic distance between two fine-grained categories. To this end, a tree-like hierarchy based on semantics or domain knowledge is built with the coarsest and finest granularity labels located on the root and leaf nodes, respectively. For example, a car with year, model, maker, and type, like '2012 BMW M3 coupe', forms a four-level hierarchical structure—year, model, maker and type respectively represent the leaf, the penultimate layer, the second top, and the root nodes. Then a path from the root node to a leaf node naturally defines a kind of inclusion relationship in which a son node contains more fine-grained information than its parent node. Therefore, the semantic distance between any two fine-grained categories is defined as the smallest number of edges travelled through from one leaf node to another via the tree. In order to incorporate HDM into deep convolutional neural networks (DCNNs) for fine-grained feature learning, a regularization term based on the inner product between DCNNs' output class-probability and columns of HDM indexed by the true label is established. This regularizer can effectively aggregate large amounts of supervising information from misclassified categories in the training stage, especially from those with extreme similarities in appear-

ance but with large semantic distances. Experiments on three public available fine-grained datasets—Stanford Cars, CUB-Birds, and FGVC-aircraft, validate the effectiveness of our approach and demonstrate a clear improvement over existing methods. In summary, we make the following three concrete contributions:

- We propose a method to exploit the semantic relationship between categories by constructing a category-level hierarchical distance matrix (HDM) for fine-grained labels.
- We study a regularization method to aggregate misclassified information by using HDM to guide the fine-grained feature learning for FGIR.
- We construct a four-level tree-like label hierarchy—year, model, makers, and type—for images from Stanford Cars and make it publicly available for community research.

The remainder of this paper is organized as follows: Sect. 2 reviews related work. Section 3 details the construction of the hierarchical distance matrix (HDM) and our model learning with HDM. Experimental results and analysis are presented in Sect. 4 and we conclude our work in Sect. 5.

2 Related Work

2.1 Label Induced FGIR

Exploiting label relationships for performance improvement has been widely adopted in many applications [17,36]. The motivation behind this idea is that there is a latent semantic relationship between categories. For FGIR, the relationship between categories is apparent since all subcategories are derived from the same base category, namely, they generally share the same structure and attributes [10,16,22]. Therefore, [1] investigates attribute-based label embeddings for FGIR. [4,25] propose to augment training samples from external sources either to build super-type and factor-type super categories or combine with attributes for FGIR respectively. Developing similarities between categories from fine-grained labels was studied in [18,33,35], where [33] employs metric learning with triplet loss to facilitate feature learning, [35] groups fine-grained labels into several independent coarse label groups and learns features cooperatively, and [18] tries to maximize the entropy between visually very similar subcategories to prevent the classifier from being too confident in its outputs for feature learning. These methods normally involve a nontrivial optimization procedure to learn their models effectively. Besides, the hierarchy inherited in fine-grained labels was also studied, in which [2] incorporates predictions for high-level categories as prior knowledge to guide the feature learning of the low-level fine-grained categories while [23] combines features from different label granularities for prediction. Our work in this paper also focuses on exploiting label relationships and extends to man-made objects, which are limited in [2,23]. In addition, our work establishes a semantic distance between fine-grained categories while not a regularization relationship between different-granularity labels like that in [2,35].

Further, our model only involves a single network for model learning. This is different from previous work [2,23,33], which need to train an ensemble of networks.

2.2 Part Localization Based FGIR

Another line of research focuses on semantic parts localization [8,19,24,29]. The idea behind this viewpoint is that the discriminative structures are subtle and existed in local areas. Thus it is practical to first localize these areas and then extract feature from these local areas for FGIR. Early work in this line of research utilizes parts or bounding box annotations to guide part detectors learning [31,32] and then employ the learned detectors for part detection and feature extraction [29,30]. However, the requirements of part annotations limit its applications. With the increasing understanding of the functionality of neurons in DCNNs [3,15,28], developing detectors from DCNNs dominates the research. [20] finds constellations of neural activation patterns computed in DCNNs. [34] also studies this idea to select neural channels for detectors learning. Combining bottom-up and top-down information for part detectors discovering was also studied in [24]. Although improvements have been achieved by these methods, they involves a multi-stage optimization. Recent work unifies the detector learning and parts feature extraction in a single model, in which [13] explore the idea of visual attention for model learning by employing reinforcement learning while [14,27] learn fine-grained models by localizing semantic parts through exploiting the high-level feature maps. A weakness of these methods is that they either need to train an ensemble of networks or occasionally involve a separate initialization. Although our work in this paper does not involve localizing semantic parts, our model on the other hand can be trained end-to-end without introducing any extra parameters, and is thus easy to be extended to large-scale datasets. Moreover, our work is orthogonal to the part-based methods and can be easily integrated into these models.

3 Approach

We detail our approach in this section. Our approach includes two key components: (1) Constructing a category-level hierarchical distance matrix (HDM) based on a tree-like label hierarchy from fine to coarse (Sect. 3.1); and (2) Guiding fine-grained feature learning by aggregating misclassified output probabilities through HDM (Sect. 3.2). An overview of our approach is depicted in Fig. 1.

3.1 The Construction of HDM

Existing methods for FGIR by exploiting label relationships either utilize label granularities to augment the amount of training samples [23,25] or build similarity relationships between categories [33,35]. These methods usually need to train an ensemble of networks with each corresponding to one granularity or involve

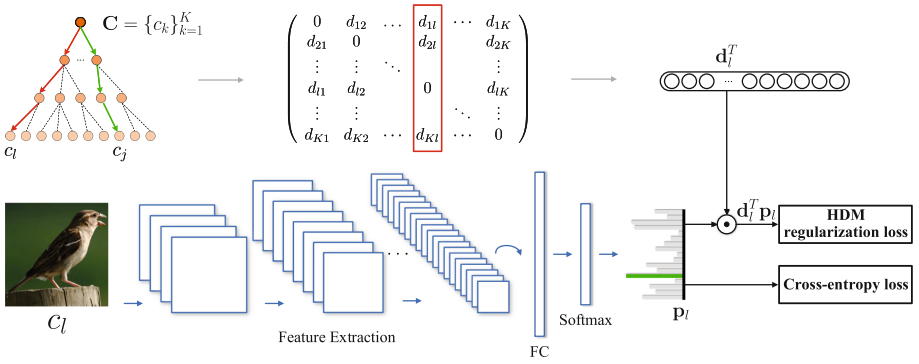

Fig. 1. An overview of our approach. The top row is an exemplification of the construction of HDM. The bottom row is a DCNN that extracts features and outputs prediction probabilities. The probabilities are regularized through a weighted sum, with weights from the column of HDM indexed by the ground-truth label, e.g., c_l and $\mathbf{d}_l^T$. The whole model is then trained end-to-end with gradients from the HDM regularization loss and the cross-entropy prediction loss.

a nontrivial optimization procedure. In this section, we propose a method to exploit label relationships for FGIR by constructing HDM from fine-grained labels, which can then be integrated into a network for end-to-end training.

Given a set of fine-grained categories with labels from set $\mathbf{C} = \{c_k\}_{k=1}^{K}$, we can construct a label hierarchy from $\mathbf{C}$ since the fine-grained label usually contain a full description of its derivative information, i.e., '2012 BMW M3 Coupe'. Supposing a M-layer relationship can be explored from $\mathbf{C}$, we can then construct a $M+1$-level hierarchy with the leaf and root nodes representing the finest and coarsest labels, respectively. Consequently, a path from the root node to a leaf node naturally defines a kind of inclusion relationship in which the son node contains more fine-grained information than its parent node. The distance between two fine-grained categories can then be defined as the smallest number of edges needed to be traveled through from one leaf node to another. Figure 2 illustrates the idea of our HDM.

Formally, for two different fine-grained categories c_i and c_j, represented by two leaf nodes i and j respectively, with a common parent node in layer L_m ($0 < m \leq M$), the category-level semantic distance between them can be simply defined as:

$$d_{ij}(m) = 2me \quad i, j \in L_0, \tag{1}$$

where e is a constant denoting the length of the edge. We set $e = 1$ in this work. Thus the distance between two categories is completely determined by m. Therefore, for a set of K fine-grained labels, we encode the semantic distances between every two fine-grained categories into a matrix $\mathbf{D}^{K\times K}$, in which the entries on the main diagonal are all zeros (that is, the matrix is a hollow matrix). Compared to previous designs of label relationships between fine-grained categories [33,35], our design is simple yet effective and easy to be optimized (see Sect. 4).

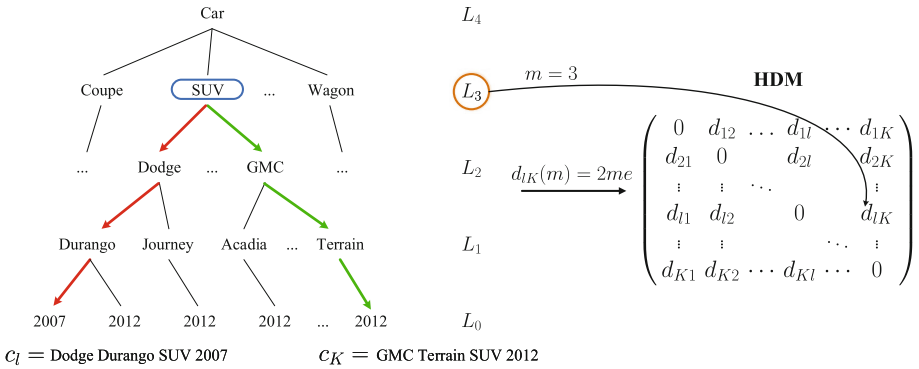

Fig. 2. An illustration of the construction of HDM on Stanford Cars. A 5-layer hierarchical tree is exemplified on the left by exploring derivative information inherited in the fine-grained labels. We build the 5-layer hierarchy according to the inclusion relationship of year-model-maker-type. Based on this hierarchy, the semantic distance between any pair of categories (e.g., c_l and c_K on the left) can be determined by the smallest number of edges needed to be traveled through between their corresponding nodes. The right is the constructed HDM.

3.2 Regularization Learning with HDM

HDM explicitly encapsulates the prior knowledge of category similarities into its design. Thus we can incorporate HDM into DCNNs for feature learning. Intuitively, this can be implemented in a regularization term that guides the fine-grained feature learning by providing more discriminative information through aggregating misclassified probabilities.

Given $\mathbf{D}^{K\times K} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_K]^T$, where each column $\mathbf{d}_t$ is a K-dimensional vector. Let p_{lj} be the output class-probability for c_j given an image with ground-truth label c_l, which could be predicted by the softmax layer. Then the regularization loss introduced by HDM can be defined as (see Fig. 1):

$$L_r = \frac{1}{N}\sum_{l=1}^{N}\sum_{k=1}^{K} b_l^{(k)} \mathbf{d}_k^T \mathbf{p}_l \tag{2}$$

where N is the total number of training samples, $b_l^{(k)} \in \{0, 1\}$ is a binary variable. $b_l^{(k)} = 1$ indicates that sample l belongs to c_k, otherwise $b_l^{(k)} = 0$. $\mathbf{d}_k$ denotes the k^{th} column from $\mathbf{D}^{K\times K}$. $\mathbf{p}_l$ represents the prediction probability from our model for sample l. Together with the HDM regularization loss, $\mathbf{p}_l$ is also employed in a cross-entropy loss for model training:

$$L_c = -\frac{1}{N}\sum_{l=1}^{N}\sum_{j=1}^{K} t_{lj} \log p_{lj} \tag{3}$$

Therefore, our model can be learned by minimizing:

$$L = L_c + \lambda L_r \tag{4}$$

where λ is a balance weight. Compared to canonical cross-entropy learning where misclassified prediction probabilities are unused, the HDM regularization term can effectively aggregate misclassified prediction probabilities to guide the model training. Therefore, it provides additional supervising discriminative information for fine-grained feature learning, especially for those categories with extreme similarities in appearance but with large semantic distance (see Sect. 4.6). We will develop our approach on DCNNs in this paper, e.g., ResNet [5] and SE-ResNet [9].

4 Experiments

4.1 Datasets

The empirical evaluation of our method is performed on Stanford Cars [10], FGVC-Aircraft [16], and CUB-200-2011 [22]. Statistics numbers of training and testing samples of all 3 datasets are shown in Table 1. FGVC-Aircraft organizes its labels in a Model-Family-Manufacturer hierarchy. We use the hierarchical labels for CUB-200-2011 from [2], which is organized based on a Species-Genera-Family-Order taxonomy. We construct a hierarchy based on a Year-Model-Maker-Type taxonomy for Stanford Cars since its fine-grained labels have already include such information. Finally, we obtain 16 years, 178 models, 49 makers, and 9 types on Stanford Cars. We will make the hierarchical labels publicly available.

Stanford Cars. Stanford Cars dataset contains 16,185 images of 196 classes of cars. The data is split into 8,144 training images and 8,041 testing images, where each class has been split roughly in a 50-50 split.

FGVC-Aircraft. FGVC-Aircraft is a dataset containing 10,000 images of aircraft, spanning 100 aircraft models, and organized in a three-level hierarchy. It includes 6667 and 3333 samples for training and testing respectively.

CUB-200-2011. CUB-200-2011 includes 11,788 bird images from 200 subspecies with 5,994 images for training and 5,794 images for testing.

4.2 Implementation

We experiment our model with ResNet-50 [5] and SE-ResNet-50 [9] on 4 NVIDIA GTX 1070 GPUs in PyTorch. We train our model 30 epochs with the batch size of 32 and momentum 0.9 by SGD [12]. The initial learning rate is 0.01 and decayed by 0.1 every 15 epochs. The images are first resized to 600×600 and then randomly cropped a region of size 448×448 as the input with a horizontal flipping probability of 0.5. A Center crop of 448×448 without horizontal flipping is used in testing. The balance weights, λ, are determined on the validation sets and set to 0.1, 0.2 and 0.1 for Stanford Cars, FGVC-Aircraft, and CUB-200-2011, respectively.

Table 1. Statistics of benchmark datasets

Datasets	#class	#Train	#Test
CUB-200-2011	200	5,994	5,794
Stanford Cars	196	8,144	8,041
FGVC Aircraft	100	6,667	3,333

Table 2. Performance evaluation on Stanford Cars. 1-stage means end-to-end learning.

Method	1-stage	Accuracy
TLAN [24]	×	–
Part-CNN w/o bbox [30]	×	–
MG-CNN w/o bbox [23]	×	–
PDFR [34]	×	–
HAR-CNN [25]	×	80.8%
ELS [33]	√	88.4%
ResNet-50 [5]	√	91.1%
SE-ResNet-50 [9]	√	91.2%
ResNet-50 + HDM (ours)	√	91.6%
SE-ResNet-50 + HDM (ours)	√	**92.2%**

Results of ResNet and SE-ResNet are from our reimplementation. The results of three label-induced methods—HAR-CNN [25], ELS [33], and MG-CNN [23], are from the authors reports. HAR-CNN acquires a large number of hyper-class-labeled images for mode training. ELS embeds label structures into a multi-task learning framework with a generalized triplet loss. MG-CNN learns an ensemble of networks for different label granularities. We develop our HDM regularization and report its performance based on ResNet-50 and SE-ResNet-50.

4.3 Results on Stanford Cars

The experimental results are presented in Table 2. HAR-CNN [25] and ELS [33] are correspondingly based on AlexNet [11] and GoogleNet [21]. From comparison, it shows clearly advantages of advanced network architectures, as we achieve 91.1% and 91.2% for ResNet-50 [5] and SE-ResNet-50 [9], respectively. The effectiveness of HDM is significant since it improves by 0.5% and 1.0% for ResNet-50 and SE-ResNet-50 respectively. Considering the simplicity of our HDM, the improvements effectively demonstrate the usefulness of misclassified probabilities in training for providing more supervised information to fine-grained feature learning.

Table 3. Performance evaluation on FGVC-Aircraft. The 1st group are methods with bbox annotations. The 2nd group are weakly-supervised methods. 1-stage means end-to-end learning.

Method	1-stage	Accuracy
Part-CNN w/bbox [30]	×	–
MG-CNN w/bbox [23]	×	86.6%
SPDA [29]	×	–
TLAN [24]	×	–
Part-CNN w/o bbox [30]	×	–
MG-CNN w/o bbox [23]	×	82.5%
PDFR [34]	×	–
HAR-CNN [25]	×	–
ELS [33]	√	–
ResNet-50 [5]	√	89.5%
SE-ResNet-50 [9]	√	90.8%
ResNet-50 + HDM (ours)	√	89.8%
SE-ResNet-50 + HDM (ours)	√	**91.2%**

4.4 Results on FGVC-Aircraft

Table 3 shows the performance of methods on this dataset. We achieve the best performance of 91.2%. Compared to MG-CNN [23], our approach only needs to train one network to make full use of information from different label granularities, while MG-CNN trains an ensemble of networks with each corresponding to one label granularity. This can save huge of training time for our model over that of MG-CNN. In addition, We can find the bounding box (bbox) annotations have a big influence on MG-CNN since it drops by 4.1% in performance when without bbox. Moreover, compared to the two supporting frameworks—ResNet-50 and SE-ResNet-50, training with HDM steadily improves their performance, which indicates the robustness of our approach.

4.5 Results on CUB-Birds

Table 4 demonstrates the prediction accuracy on CUB birds. Our approach achieves almost the best result on this dataset. Compared to the best result of methods with bbox (85.7%), our result is comparable (84.4%), considering only image-level labels are employed in our approach. Our approach surpasses almost all methods that employs multi-stage training while without annotations, like TLAN [24], Part-CNN [30], and MG-CNN [23]. The comparable performance with PDFR [34] advocates the effectiveness of our approach. In comparison with PDFR, our approach is much more simple and easy to extend to large-scale datasets since it does not introduce any extra parameters into its supporting

Table 4. Performance evaluation on CUB-200-2011. The 1st group are methods with bbox annotations. The 2nd group are weakly-supervised methods. 1-stage means end-to-end learning.

Method	1-stage	Accuracy
Part-CNN w/bbox [30]	×	76.4%
MG-CNN w/bbox [23]	×	83.0%
SPDA [29]	×	**85.7**%
TLAN [24]	×	69.7%
Part-CNN w/o bbox [30]	×	73.9%
MG-CNN w/o bbox [23]	×	81.7%
PDFR [34]	×	**84.5**%
HAR-CNN [25]	×	–
ELS [33]	√	–
ResNet-50 [5]	√	81.6%
SE-ResNet-50 [9]	√	83.6%
ResNet-50 + HDM (ours)	√	82.0%
SE-ResNet-50 + HDM (ours)	√	**84.4**%

(a) (b) (c)

Fig. 3. The left image in every group is misclassified into the category of the right image in the corresponding group by SE-ResNet since they have a very similar visual appearance. However, by exploiting the semantic relationships between categories, our SE-ResNet+HDM can correctly predict their true categories. The images in (a) Stanford Cars, (b) FGVC-Aircraft, and (c) CUB-Birds, from left to right, are respectively from categories 'Bentley Continental GT Coupe 2007', 'Aston Martin Virage Coupe 2012', 'A340-300', 'A340-200', 'Hooded Oriole', and 'Baltimore Oriole'.

networks and can be trained end-to-end. PDFR, however, involves a multi-stage training and needs to select filters in a DCNN to train part detectors for localizing parts before training a classification network.

4.6 Improvements Inspection

It is essential to inspect in which aspect HDM contributes to FGIR. To this end, we select the SE-ResNet-50 as our observation model since it surpasses ResNet-50 and achieves the best performance on all datasets. Figure 3 illustrates our

findings. We find that, by exploiting semantic distance between fine-grained categories, HDM can effectively distinguish objects that come from different categories but with an extremely similar visual appearance. The objects with this kind of properties are difficult to be correctly classified as demonstrated by SE-ResNet-50. Thus, we conclude that HDM contributes to correctly classify images that are almost indistinguishable in visual appearance.

5 Conclusion

In this paper, we proposed a method to exploit semantic relationships between subcategories by constructing a hierarchical distance matrix from fine-grained labels. In order to take advantage of this relationship to improve the performance of FGIR, we studied an HDM-induced regularization approach that aggregates misclassified prediction probabilities to guide the model learning. With more supervised discriminative information from the HDM regularizer, our approach improves the performance of FGIR significantly. Experiments on benchmark datasets validate the effectiveness of our approach and demonstrate its improvements over existing methods.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant 61702197, in part by the Natural Science Foundation of Guangdong Province under Grant 2017A030310261, in part by the program of China Scholarship Council.

References

1. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: CVPR (2015)
2. Chen, T., Wu, W., Gao, Y., Dong, L., Luo, X., Lin, L.: Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In: ACM MM (2018)
3. Fong, R., Vedaldi, A.: Net2vec: quantifying and explaining how concepts are encoded by filters in deep neural networks. In: CVPR (2018)
4. Gebru, T., Hoffman, J., Li, F.F.: Fine-grained recognition in the wild: a multi-task domain adaptation approach. In: ICCV (2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
6. He, X., Peng, Y.: Fine-grained image classification via combining vision and language. In: CVPR (2017)
7. He, X., Peng, Y.: Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In: AAAI (2017)
8. He, X., Peng, Y., Zhao, J.: Which and how many regions to gaze: focus discriminative regions for fine-grained visual categorization. Int. J. Comput. Vis., 1–21 (2019)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2017)
10. Krause, J., Stark, M., Deng, J., Li, F.F.: 3D object representations for fine-grained categorization. In: 4th IEEE Workshop on 3D Representation and Recognition at ICCV (2013)

11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS (2012)
12. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
13. Li, Z., Yang, Y., Liu, X., Wen, S., Xu, W.: Dynamic computational time for visual attention. In: ICCV (2017)
14. Liu, X., Xia, T., Wang, J., Lin, Y.: Fully convolutional attention localization networks: efficient attention localization for fine-grained recognition. In: arXiv preprint arXiv:1603.06765 (2016)
15. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: CVPR (2015)
16. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. In: arXiv preprint arXiv:1306.5151 (2013)
17. Mostajabi, M., Maire, M., Shakhnarovich, G.: Regularizing deep networks by modeling and predicting label structure. In: CVPR (2018)
18. Dubey, A., Gupta, O., Raskar, R., Naik, N.: Maximum entropy fine-grained classification. In: NIPS (2018)
19. Peng, Y., He, X., Zhao, J.: Object-part attention model for fine-grained image classification. IEEE Trans. Image Process. **27**(3), 1487–1500 (2017)
20. Simon, M., Rodner, E.: Neural activation constellations: unsupervised part model discovery with convolutional networks. In: ICCV (2015)
21. Szegedy, C., et al.: Going deeper with convolutions. In: CVPR (2015)
22. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD birds-200-2011 dataset. Technical report, California Institute of Technology (2011)
23. Wang, D., Shen, Z., Shao, J., Zhang, W., Xue, X., Zhang, Z.: Multiple granularity descriptors for fine-grained categorization. In: ICCV (2015)
24. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: CVPR (2015)
25. Xie, S., Yang, T., Wang, X., Lin, Y.: Hyper-class augmented and regularized deep learning for fine-grained image classification. In: CVPR (2015)
26. Yang, S., Bo, L., Wang, J., Shapiro, L.G.: Unsupervised template learning for fine-grained object recognition. In: NIPS (2012)
27. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to navigate for fine-grained classification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 438–454. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_26
28. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
29. Zhang, H., et al.: SPDA-CNN: unifying semantic part detection and abstraction for fine-grained recognition. In: CVPR (2016)
30. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 834–849. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_54
31. Zhang, N., Farrell, R., Darrell, T.: Pose pooling kernels for sub-category recognition. In: CVPR (2012)
32. Zhang, N., Farrell, R., Iandola, F., Darrell, T.: Deformable part descriptors for fine-grained recognition and attribute prediction. In: ICCV (2013)

33. Zhang, X., Zhou, F., Lin, Y., Zhang, S.: Embedding label structures for fine-grained feature representation. In: CVPR (2016)
34. Zhang, X., Xiong, H., Zhou, W., Lin, W., Tian, Q.: Picking deep filter responses for fine-grained image recognition. In: CVPR (2016)
35. Zhou, F., Lin, Y.: Fine-grained image classification by exploring bipartite-graph labels. In: CVPR (2016)
36. Zlateski, A., Jaroensri, R., Sharma, P., Durand, F.: On the importance of label quality for semantic segmentation. In: CVPR (2018)