Multimedia Tools and Applications (2021) 80:215 – 225
https://doi.org/10.1007/s11042-020-09679-8

Bilinear pyramid network for flower species categorization

Cheng Pang[1] · Wenhao Wang[1] · Rushi Lan[2] · Zhuo Shi[1] · Xiaonan Luo[1]

Received: 26 September 2019 / Revised: 14 June 2020 / Accepted: 20 August 2020 /
Published online: 1 September 2020

Abstract

It is a challenging task to distinguish between numerous species of flowers due to their visually similarities and variations of the pose and structure. Thanks to properly modeling of the local feature interactions, bilinear CNN has succeeded in classifying of many non-rigid fine-grained species including flowers. However, bilinear CNN only computes the feature in a straightforward way without exploring the interactions between features from multiple layers in the network. In this paper,we present a novel Bilinear Pyramid Network (BPN)for flower categorization. Instead of passing through the network and directly feeding the final classifier, features from a convolutional layer are resized and multiplied with that from the former layer, which alternates multiple times to generates prediction vectors using the features from distinct layers. These features encoded from the feature pyramid spontaneously carry multi-level semantic cues, which yields stronger discriminative powers than single-layer features. Experiments show that the proposed network obtains superior classification results on the challenging dataset of flowers.

Keywords Fine-grained image classification · Fine-grained visual categorization (FGVC) · Image classification · Convolutional neural network (CNN) · Deep learning

1 Introduction

Fine-grained visual categorization refers to differentiating species from the same basic-level categories (e.g., species of dogs and models of cars) [4–7, 10, 12, 19, 30, 33]. Different from basic-level categories, fine-grained species present extremely high visual similarities and large pose variations, and often belong to considerably numerous species. As a result, classifying fine-grained species with large intra-class variations and small between-class differences is ambitious.

✉ Rushi Lan
rslan2016@163.com

1 School of Computer Science and Information Security,Guilin University of Electronic Technology, Guilin, 541004, China

2 School of Computer Science and Engineering, South China University of Technology, Guangzhou, 510641, China

Researches on cognitive science indicate that basic-level categories are distinguished by their differences of the body parts while fine-grained species are differentiated via the different properties of the same parts [24]. Follow this hypothesis,many part-based methods have gained their success in classifying rigid objects like birds, dogs and cars. In [15], a HSnet is proposed in which the classification problem is formulated as a sequential search for informative parts over a deep feature map produced by a CNN. In [29],part matching is introduced in traditional Bag-of-Words pipeline to inference discriminative foregrounds of the objects and eliminate background noises. In [32],multiscale part proposals are generated from object proposals, and then filtered to form the image representations. However, these methods fail in flower categorization due to the ambiguous structures of flowers.

Existing methods for flower species categorization are mostly devoted to precise detection of the foregrounds using attention mechanisms and designing discriminative features for the for flowers [2, 13, 21]. Chai et al. segment flowers out by performing within-image and across-image appearance propagation consecutively rather than jointly [3]. Pang et al. introduce the eye shifting of human beings into the feature encoding procedure when classifying flowers [22]. Britto et al. show that a combination of carefully designed features, including shape, color and texture descriptors, is essential for flower categorization [18].

Although these methods use attention mechanisms to explore the structures of flowers and fuse multiple feature to enhance the discrimination, they fail to get satisfactory results for flower categorization. Problems of these methods are mainly two-fold:1)hand-craft features are incapable of distinguishing subtle differences between visually similar species, 2) the feature extraction and final classification are performed independently, which prevents the global optimization of the classification model, and 3) only high-level visual cues are used in the prediction, ignoring the importance of low-level visual cues.

In this paper, we present a novel Bilinear Pyramid Network (BPN) for flower species categorization. The proposed method is built on the prevalent CNN architecture which yields superior performance via discriminative deep descriptors and end-to-end learning. Moreover, instead of directly passing through the network and feeding the final classifier, features from a convolutional layer are fused with that from the former layer. These features encoded from the feature pyramid spontaneously carry multi-level semantic cues, which introduces additional robustness to the variation of the pose and scale and outperforms single-layer features in discrimination for categorization. We conduct extensive validation on the bench-marked dataset to show the effectiveness of the proposed method.

2 Related work

2.1 Deep convolutional neural networks

Deep Convolutional Neural Networks (DCNN) has significantly boosted the performance of supervised and semi-supervised classification [9, 14, 25]. Fu et al. propose a effective graph convolutional network which preserves the local geometry of samples via hypergraph p-Laplacian [8]. Lin et al. develop the feature pyramid network (FPN) building high-level semantic feature maps at all scales, which shows significant improvement in object detection [16]. Inspired by the feature pyramid, Yu et al. present the spatial pyramid-enhanced NetVLAD for place recognition, which is shown to be an optimal feature encoding method built on DCNN [31]. Unlike the FPN based on conventional architectures, the proposed BPN is built on a bilinear CNN. Moreover, we give prediction fused by all scales of the feature pyramid instead of predicting at each scales.

2.2 Attention based method

Some recent deep methods for FGVC also adopt attention mechanisms, for example, the methods based on attention proposal sub-network (APN) and bilinear CNN [17]. Instead of explicitly detecting or discovering body parts of the fine-grained species, bilinear CNN investigates local feature interactions by multiplying feature maps generated by two CNNs. These two CNNs share the same architecture while capture distinct patterns of an input image. Thus, the outer product of their features explicitly depicts the feature co-occurrence in each location without any precondition of body parts. However,bilinear CNN only takes advantages of features from the last layer,ignoring cues from the other layers. As a comparison, the proposed BPN encodes information from the feature pyramid and spontaneously incorporates cues from multiple layers, generating more discriminative features for flower categorization.

3 Proposed model

When passing an image through a CNN, feature maps keep decreasing in their sizes, forming a feature pyramid [16]. Feature maps computed by the deeper convolutional layers have smaller size but carry more semantic cues. Most of the existing methods only adapt feature maps from the last layer for classification, while the proposed BPN utilizes the feature pyramid: first up-samples the feature maps from the deeper layer, and then multiplies them with the feature map from the adjacent layer. This procedure is repeated multiple times from deeper layers to upper layers, collecting semantic cues specifying different scales and visual patterns,which introduces additional robustness to the variation of the pose and scale of flowers and generates discriminative local features .

It is shown that the two CNNs of the bilinear model actually play different roles: one detects object parts (salient patterns) while the other extracts local features of the detected parts. Thus, the bilinear model incorporates feature detection and feature extraction in an unify framework,with each detector couples with an extractor. CNNs in the BPN are similar to that of the bilinear network, but differs in the scale: the BPN assembles features from multiple layers and thus utilizes multi-scale detectors and feature extractors. The proposed network is illustrated in Fig. 1.

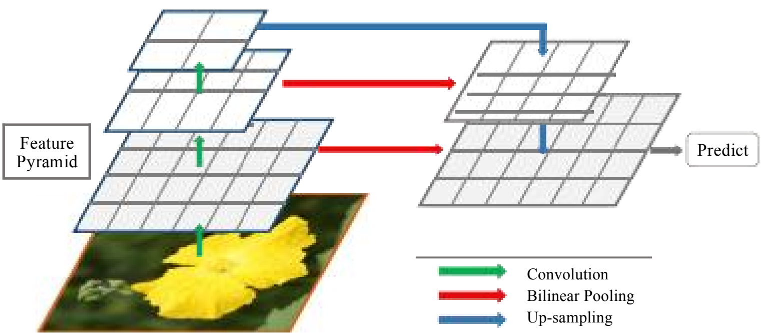

Fig. 1 The bilinear pooling and up-sampling architecture of the proposed Bilinear Pyramid Network. Multi- level semantic cues from distinct layers are incorporated from the feature pyramid, highlighting multi-scale visual patterns and thus benefiting flower categorization

Specifically, we detail the last three convolutional layers which are used as the feature pyramid for bilinear pooling in Fig. 2. As is seen, the compared convolutional layers in a typical VGG-16 network [26] packaged in a cascading way only produce single-scale convolutional features,which is sensitive to scale and pose variations of flowers (the left network in Fig. 2). We also give examples of the related layers with different configurations in a 2-level BPN and 3-level BPN which pools features from 2 layers and 3 layers respectively (the central and right networks in Fig. 2). In contrast to the VGG network, these two BPNs pool features from multi-layers of the feature pyramid. Besides, the layers of these three models before conv5_1 share the same configuration,serving as the basic feature extractors. Notice that the feature maps are resized by the fully-connected layer bilinear_proj before element-wise production via the eltwise layer.

4 Formulation

Let $\mathbf{x} = [x_1, x_2, ..., x_c]^T$ denotes a feature with c channels at a spatial location of the feature map from a convolutional layer. Then the bilinear model is given by:

$$b_i = \mathbf{x}^t M_i \mathbf{x} \tag{1}$$

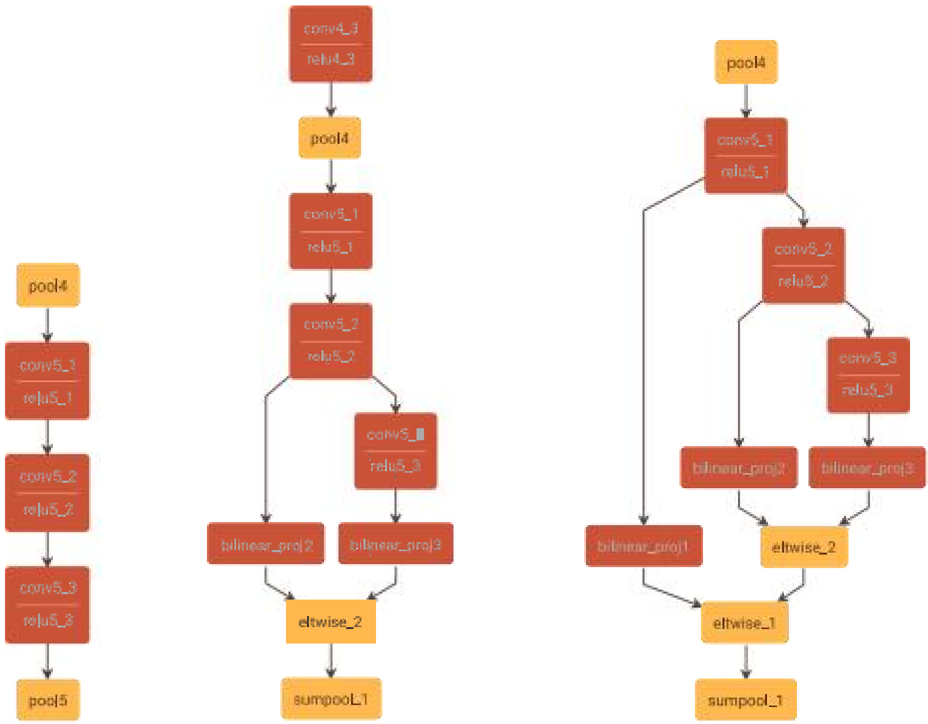

Fig. 2 Comparison of the last three convolutional layers in the original VGG network, 2-level BPN and 3- level BPN. The bilinear pooling of multi-layer features improves the robustness to pose and scale variations of flowers

where b_i is the output of the bilinear model specifying location i . $\mathbf{M} = [M_1, M_2, ..., M_j] \in R_{c \times c \times j}$, a set of learnable matrices, are used for projection. As shown in [23], a projection matrix M_i can be further factorized into two low-rank vectors D_i and Q_i^t. Thus, formulation (1) can be updated using the vectors, and rewritten as the outputs of the bilinear model specifying all the locations:

$$\mathbf{b} = \mathbf{C}^t(\mathbf{D}^t\mathbf{x} * \mathbf{Q}^t\mathbf{x}) \tag{2}$$

where **C** denotes the classification matrix, **D** and **Q** are two projection matrices consisted by low-rank vectors,$*$ is the Hadamard product. Different to the traditional bilinear model, we pool the features from two adjacent layers:

$$\mathbf{b} = \mathbf{C}^t(\mathbf{D}^t S(\mathbf{x}) * \mathbf{Q}^t\mathbf{y}) \tag{3}$$

where $S(\mathbf{x})$ represents the up-sampled feature of **x**, and **y** is the feature from the adjacent convolutional layer.

Let $B(\cdot)$ denotes the above bilinear pooling operation be-tween features from two adjacent layers in the network, $\mathbf{x}_b$ denotes the feature produced by the basic feature extractor. Then $\mathbf{x}_i$, the bilinear descriptor of the i-th convolutional layer after basic feature extraction module can be obtained by:

$$\begin{aligned} \mathbf{x}_i &= B(H_i(\mathbf{x}_b)) \\ H_1(\cdot) &= F(\cdot) \\ &\vdots \\ H_i(\cdot) &= F(H_{i-1}(\cdot)) \end{aligned} \tag{4}$$

where $F(\cdot)$ is the convolutional operation in the network, and $H_i(\cdot)$ is the combination of i convolutions.

In the experiments, x_2 encoded from a 3-layer feature pyramid is used for the final classification. Encoding features from more layers is not encouraged, as it will introduce heavy computational costs and hurt the discriminative power of the descriptor.

5 Experiments

The proposed BPN is built on the VGG-16 architecture [26], in which we fix all the convolutional layers and fine-tune the fully-connected layers using the Oxford 102 category and 17 category flower datasets [20] (Fig. 3). We first introduce up-sampling layers and bilinear pooling layers to the last three convolutional layers conv5_1, conv5_2and conv5_3, obtaining the bilinear product of their features. Then the product is fed through the fully connected layers fc_7and fc_8, where fc_8 is the predefined VGG-16 layer and fc_7 is newly added for enhancing the non-linear property of the network. All the images are resized to 224*224 and randomly cropped to 192*192 before being fed into the network. Mirror flipping is adopted to double the training samples. The training process follows the standard stochastic gradient descent (SGD) with a learning rate of 0.01, and will be terminated when reach the maximum iteration of 100000. The basic feature extractor (the layer conv5_1 and its former layers) generates a feature with 512 channels, which are further processed by two convolutional layers to obtain multi-scale extensions. Then these features specifying 3 scales are projected to have 8192 channels before element-wise multiplication. Finally, the fused feature is projected to vector twice, reducing its dimension from 1024 to the number of flower species.

We train the BPN using a desktop with a Intel i7 CUP (3.5GHZ)and a 1080-ti GPU,and the architecture is implemented by Caffe toolkit. The network converged in 630 minutes

Fig. 3 Some samples in the Oxford 102 category and 17 category flower datasets which present significant variations of appearance and pose. Samples belong to the same species are listed in one column, which highlights large intra-class dissimilarities of the species

after 10 epochs of training. Finally, the trained network could classify 23 frames of images per second. Table 1 lists the training time and computational performance of the proposed BPN.

To explicitly show the benefits of the proposed bilinear concatenating of the feature pyramid, we investigate the performance of distinct concatenations using the features from different layers (see Table 2). When using conv5_3 exclusively, the BPN degenerates to the VGG-16 networks which achieves top-1 precision and top-5 precision of 92.1 % and 98.5 % respectively. The top-1 and top-5 accuracy boost to 93.5 % and 99.02 % after applying bilinear pooling on conv5_3 and conv5_2, showing the effectiveness of the proposed architecture. While the top-5 precision does not hit the top, the top-1 precision further goes up when three convolutional layers in the feature pyramid are used.

Intuitively,the top-5 prediction using the 3-layer features should present the highest precision. However, it goes to 98.8 %, versus 99.0 % using the 2-layer features. We argue this is caused by the differences of the discrimination of these two types of features. The 3-layer features are more discriminative, which can be seen from the top-1 precision. The 2-layer features, as a comparison, are less discriminative. Top-5 results of the 2-layer features may incorporate some possibly correct samples which is not so similar to the query species. On the other hand, flowers present large intra-class variations, introducing some outliers which are dissimilar to their ground-truth categories (see Fig. 3). Consequently, when classifying

Table 1 Training time and computational performance of the BPN

Training time	processing speed
630 min	23 f/s

Table 2 Comparison of the BPN features trained on 102 category dataset

Feature	conv5_3	conv5_3+ conv5_2	conv5_3+ conv5_2+ conv5_1
Top-1 precision (%)	92.16	93.53	94.02
Top-5 precision (%)	98.53	99.02	98.82

an outlier, the 2-layer features may generate a correct prediction whereas the 3-layer features may give an incorrect prediction. Notice that we only pool the last three layers of the feature pyramid. Pooling more layers will introduce numerous parameters brought by the fully-connected layers, which is a disaster for network training .

We compare the proposed BPN with other mainstream methods on 102 category dataset in Table 3. The first six methods are traditional methods using hand-craft features, most of which follow the pipeline of Bag-of-Words (BOW). A BOW-based method first samples features across the training set,and obtains a set of visual words by off-line clustering using the sampled features. Then,all the image instances are encoded by the visual words to form the final descriptors of the instances. Distinct methods only differ in the way they design and encode the features or the application of different classifiers. Nilsback and Zisserman [21] combine features specifying shape, texture and color, developing a kernel learning method for flower categorization. Pang et al. [22] introduce a saliency-based method which encodes the attention shifts of human beings into the features. However, due to the absence of jointly optimization of the feature learning and classification, these traditional methods fundamentally underperform those based on deep learning.

VGG-16 network [26], our baseline model, achieves 92 % accuracy in top-1 precision. This model becomes an ideal feature extractor after being trained on the large-scale ImageNet dataset. The original bilinear network yields a slightly worse accuracy of 91 %,which may be caused by its training strategy. Empirically, architectures fine-tuned on large scale dataset perform better on feature extraction due to the abundance of training data. However, the bilinear network can't adopt the fine-tuned layers as its feature extractors. The network encourages its twin feature extractors to be trained with different weights, and the training can only be done on the target dataset. As a result, the trained feature extractors might be

Table 3 Comparison of 102 category flower categorization

Method	Top-1 precision (%)
Nilsback and Zisserman [21]	72.8
Kanan and Cottrell [13]	75.2
Angelova et al. [2]	76.9
Chai et al. [3]	80.0
Britto et al. [18]	80.8
Pang et al. [22]	82.6
Bilinear Network [17]	91.3
Simonyan and Zisserman [26]	92.2
Gogul et al. [11]	93.4
Xia et al. [28]	94.0
Ours	94.2

Table 4 Comparison of 17 category flower categorization

Method	Top-1 precision (%)
Varma and Ray [27]	82.6
Nilsback and Zisserman [21]	88.3
Angelova and Zhu [1]	85.0
Zou and Nagy [34]	89.0
Chai et al. [3]	90.4
Xia et al. [28]	95.0
Britto et al. [18]	95.0
Bilinear Network [17]	98.8
Ours	99.1

suboptimal if the target dataset is small-scale. Similar to our method,Gogul et al. also build their recognition system on a deep neural network, using a transfer learning strategy [11]. Xia et al. introduce the Inception-v3 module in flower classification and obtain satisfied accuracy on 102 category dataset [28]. However, they fail to further promote the performance on the smaller 17 category dataset due to their intricate architecture which relies on numerous training data. The proposed BPN powered by a simple VGG-16 feature extraction network achieves superior results when compared with more complicated networks, showing the effectiveness of bilinear pooling of the multi-layer feature pyramid and the significance of exploring semantic cues with distinct scales. Generally,methods based on deep learning architectures obtain better results on flower categorization.

Table 4 compares the categorization results of the 17 category dataset. The number of category in this dataset is considerably less than that of the 102 category dataset, which

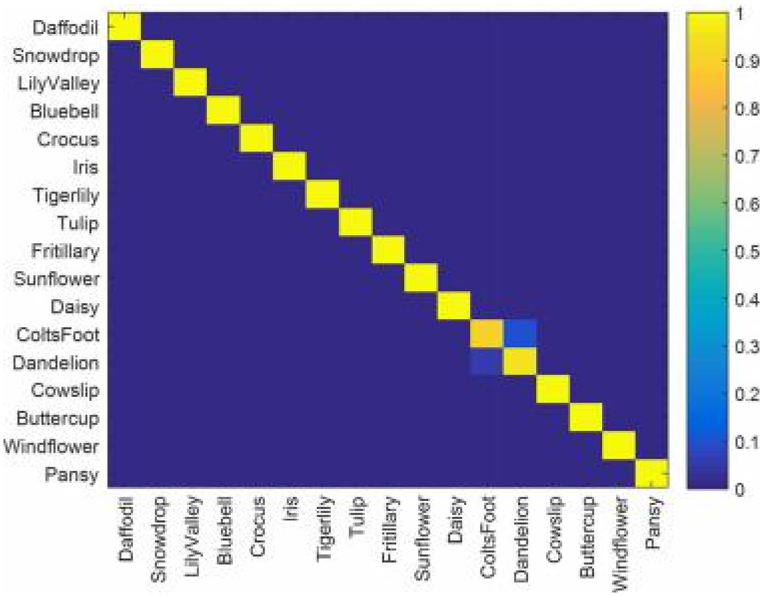

Fig. 4 Confusion matrix of 17-category categorization obtained by the proposed BPN. Most of the samples of flowers are correctly classified

ColtsFoot Dandelion

Fig. 5 Confusing species of flowers in 17 category dataset, which share similar color and structures. The proposed BPN fails in distinguishing few samples of these species

lead to higher accuracy for all the methods. BPN doesn't win the bilinear network with large margin because the small-scale dataset is easily saturated. Similar to the categorization on 102 category dataset, these two methods based on deep network achieve the best results. Although it is hard to completely explore the discriminative power of distinct methods using this dataset quantitatively, some qualitative analysis can be made. In Fig. 4, we illustrate the confusion matrix specifying the classification results of all 17 categories in the dataset. Categories with brighter colors in the energy map are with higher accuracy. It is clear that the proposed method achieves satisfactory results on most of the categories (energies of the categories concentrate in the diagonal of the matrix),except for two species of the flowers: ColtsFoot and Dandelion. The confusion matrix indicates that some samples of ColtsFoot are classified to Dandelion while some of Dandelion are recognized as ColtsFoot (see Fig. 5). These two species of flowers share similar delicate long-strip petals. Their difference only lies in the stamen. As a result,it can be confused to differentiate some over-exposure and low-resolution samples with ambiguous details.

6 Conclusion

In this paper, we present a novel Bilinear Pyramid Network for flower categorization. The network pools multiple features from distinct layers of the feature pyramid generated by the convolutional network, which spontaneously incorporates distinct semantic cues of different layers. The bilinear pooling computes the outer product of the multi-layer features in each spatial location, explicitly depicting the co-occurrence of different visual cues and thus benefiting the categorization. Experiments conducted on Oxford 102 category and 17 category flower datasets show the proposed BPN is able to differentiate numerous species of flower with superior accuracy.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China under Grant 61962014, Grant 61702129, Grant 61772149, and Grant U1701267, in part by the China Postdoctoral Science Foundation under Grant 2018M633047,and in part by the Guangxi Science and Technology Project under Grant AB17195057 and Grant AA18118039.

References

1. Angelova A, Zhu S (2013) Efficient object detection and segmentation for fine-grained recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 811–818
2. Angelova A, Zhu S, Lin Y, Wong J, Shpecht C (2012) Development and deployment of a large-scale flower recognition mobile app. NEC Labs America Technical Report
3. Chai Y, Lempitsky V, Zisserman A (2011) Bicos: a bi-level co-segmentation method for image classification. In: 2011 International conference on computer vision. IEEE, pp 2579–2586
4. Cui Y, Song Y, Sun C, Howard A, Belongie S (2018) Large scale fine-grained categorization and domain-specific transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4109–4118
5. Dat H, Ehsan E (2020) Fine-grained generalized zero-shot learning via dense attribute-based attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition
6. Dubey A, Gupta O, Guo P, Raskar R, Farrell R, Naik N (2018) Pairwise confusion for fine-grained visual classification. In: Proceedings of the European conference on computer vision (ECCV), pp 70–86
7. Fu J, Zheng H, Mei T (2017) Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4438–4446
8. Fu S, Liu W, Zhou Y, Nie L (2019) Hplapgcn: Hypergraph p-laplacian graph convolutional networks. Neurocomputing 362:166–174
9. Fu S, Liu W, Tao D, Zhou Y, Nie L (2020) Hesgcn: Hessian graph convolutional networks for semi-supervised classification. Inf Sci 514:484–498
10. Ge W, Lin X, Yu Y (2019) Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In:Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3034–3043
11. Gogul I, Kumar VS (2017) Flower species recognition system using convolution neural networks and transfer learning. In: 2017 fourth international conference on signal processing, communication and networking (ICSCN)
12. Huang Z, Li Y (2020) Interpretable and accurate fine-grained recognition via region grouping. In: Proceedings of the IEEE conference on computer vision and pattern recognition
13. Kanan C, Cottrell G (2010) Robust classification of objects, faces, and flowers using natural image statistics. In:2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, pp 2472–2479
14. Khan N, Chaudhuri U, Banerjee B, Chaudhuri S (2019) Graph convolutional network for multi-label vhr remote sensing scene recognition. Neurocomputing 357:36–46
15. Lam M, Mahasseni B, Todorovic S (2017) Fine-grained recognition as hsnet search for informative image parts. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2520–2529
16. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125
17. Lin TY, RoyChowdhury A, Maji S (2017) Bilinear convolutional neural networks for fine-grained visual recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1309–1322
18. Mottos AB, Feris RS (2014) Fusing well-crafted feature descriptors for efficient fine-grained classification. In: 2014 IEEE international conference on image processing (ICIP). IEEE, pp 5197–5201
19. Munro J, Damen D (2020) Multi-modal domain adaptation for fine-grained action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition
20. Nilsback ME, Zisserman A (2007) Delving into the whorl of flower segmentation. In: British machine vision conference, vol 2007, pp 1–10
21. Nilsback ME, Zisserman A (2008) Automated flower classification over a large number of classes. In: 2008 sixth Indian conference on computer vision, graphics & image processing. IEEE, pp 722–729
22. Pang C, Yao H, Sun X, Zhao S, Yu W (2018) Rediscover flowers structurally. Multimed Tools Appl 77(7):7851–7863
23. Rendle S (2010) Factorization machines. In: 2010 IEEE international conference on data mining. IEEE, pp 995–1000
24. Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P (1976) Basic objects in natural categories. Cogn Psychol 8(3):382–439
25. Sichao F, Weifeng L, Shuying L, Yicong Z (2019) Two-order graph convolutional networks for semi-supervised classification. IET Image Process 13(14):2763–2771

26. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
27. Varma M, Ray D (2007) Learning the discriminative power-invariance trade-off. In: 2007 IEEE 11th international conference on computer vision. IEEE, pp 1–8
28. Xia X,Cui X,Bing N (2017)Inception-v3 for flower classification. In:International conference on image
29. Xie L, Tian Q, Hong R, Yan S, Zhang B (2013) Hierarchical part matching for fine-grained visual categorization. In: Proceedings of the IEEE international conference on computer vision, pp 1641–1648
30. Yu J, Kuang Z, Zhang B, Zhang W, Lin D, Fan J (2018) Leveraging content sensitiveness and user trustworthiness to recommend fine-grained privacy settings for social image sharing. IEEE Transactions on Information Forensics and Security 13(5):1317–1332
31. Yu J, Zhu C, Zhang J, Huang Q, Tao D (2019) Spatial pyramid-enhanced netvlad with weighted triplet loss for place recognition. IEEE Transactions on Neural Networks and Learning Systems 31:661–674
32. Zhang Y, Wei XS, Wu J, Cai J, Lu J, Nguyen VA, Do MN (2016) Weakly supervised fine-grained categorization with part-based image representation. IEEE Trans Image Process 25(4):1713–1725
33. Zheng H, Fu J, Mei T, Luo J (2017) Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE international conference on computer vision, pp 5209–5217
34. Zou J, Nagy G (2004) Evaluation of model-based interactive flower recognition. In: Proceedings of the international conference on pattern recognition, vol 2. IEEE, pp 311–314