Research Progress of Fine-grained Visual Classification: Basic Framework, Challenges, and Future Development

Congying Ma[1, †]
[1]Department of Computer Science
University of Bath
Bath, Somerset, UK
cm2654@bath.ac.uk

Yuhan Pu[1, †]
[1]Department of Statistics
University of California, Davis
Davis, CA, USA
ypu@ucdavis.edu

Abstract— **With the progress of multimedia technology, fine-grained visual classification (FGVC) has gradually attracted extensive attention in the field of computer vision. Thanks to the success of deep neural networks, both the accuracy and speed have achieved an unprecedented breakthrough. Hundreds of fine-grained classification techniques have been developed from the early, fully supervised methods to the current weakly ones. To this end, in this article, we provide a comprehensive survey of the recent achievements of FGVC. Specifically, we describe the formulation and setting of the FGVC, including the background, basic framework, and challenges. Then, we introduce the widely used datasets and compare the performance of different models. Lastly, we discuss the potential future directions of FGVC.**

Keywords: Fine-grained visual classification; deep learning; weakly supervised learning; zero-shot learning;

I. Introduction

Image classification is a basic task with active research and has achieved remarkable success in computer vision. Thanks to the development of convolutional neural networks (CNNs), classification accuracy has made a great breakthrough. Furthermore, with the customization and refinement of user requirements in search engines and social networks, fine-grained visual classification (FGVC) has gradually received widespread attention from academia and industry in recent years.

In contrast to traditional image classification, which simply distinguishes the general categories (e.g., dogs vs. cats), FGVC aims to classify subcategories within the same general category, such as identifying breeds of dogs [1], species of plants [2], or models of cars [3]. An effective deep learning model is inseparable from sufficient learning data and high-quality feature representation. The existing FGVC methods have benefited from the construction of benchmark datasets [1, 2, 4], which provide a good foundation for the training and evaluation of models. New datasets [5] and challenges [6] have also reflected the growing demand of FGVC in realistic settings and brought new possibilities and complexity. Similar to the basic image classification, the advance of FGVC also benefits a great deal from the vigorous development of deep neural networks in recent years [7-10], which provide technical support for high-quality object feature expression.

However, FGVC is extremely challenging compared with basic level image classification due to the following aspects: (1) the inherently large intra-class variations and subtle inter-class variations in fine-grained datasets, which makes FGVC focus on more subtle distinctions and requires more discriminative methods; (2) the labels of fine-grained datasets often depend on high-level professional knowledge and require annotations manually by experts rather than ordinary people for general categorization, which is always expensive and non-scalable; (3) the large number of sub-categories can lead to a long tail distribution of datasets, with limited data for training in most categories; (4) the backgrounds from different categories are more similar compared with general categorization, which makes them far less useful and even become noises. To cope with these specific challenges, several types of approaches have been developed, with enormous success been made in the past decade.

A fundamental step of FGVC is to identify discriminative regions and extract informative features from images, which is the main idea of most early research. Early methods relied on accurate annotations heavily, while a trend of annotation-free has emerged to break the restrictions. With the vigorous developments of deep learning, network architecture, design, and ensemble aim at better feature representation are also important considerations. And extra information has been integrated to support the task, such as human-in-the-loop in early research and web data which is popular in recent years. More advanced alternative solutions for data challenges include few-shot learning or zero-shot learning. The main tasks of FGVC focus on three issues: (1) how to localize discriminative regions; (2) how to learn sophisticated representations containing informative and discriminative features; (3) how to tackle the problem of limited data and dependency on annotations.

In the following chapters, we will review the milestones of progress in deep FGVC, analyze and compare the performance of representative models based on popular benchmark datasets, and discuss current challenges and development directions in the future.

II. Part-based Approaches

An essential challenge of FGVC is to extract discriminative regions and weaken the influence of background noise. Many early approaches aimed at training a part detector and gradually adopted the pipeline as shown in Fig.1 [11], which are known as part-based approaches. As seen in the pipeline, semantic parts

978-1-6654-0267-5/21/$31.00 ©2021 IEEE

(e.g., head or torso of birds) are localized, aligned, based on which features are extracted for classification. In this way, the subtle differences can be extracted for training, against the influence by variations such as poses, viewpoints, scales. Early part-based algorithms rely on dense annotations heavily, which is expensive and time-consuming to generate. Gradually, a trend of annotation-free approaches has drawn more attention. This development is known as a transfer from strongly supervision to weakly supervision. This section will review the development of the representative part-based deep learning approaches and discuss the most recent achievements.

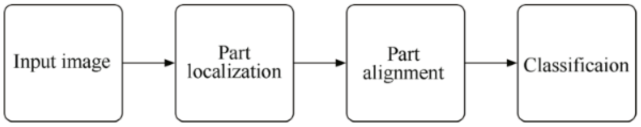

Fig. 1 Pipeline of part-based approaches [11].

A. *Fully Supervised Part-based Approaches*

Early research for FGVC belongs to strongly supervision methods, depending on extra annotations such as bounding boxes or part annotations to localize significant regions. The bounding box has also been a requirement to guarantee accuracy. Early part-based methods, segmentation-based methods, and their combination, all need ground truth bounding boxes at the test stage to achieve adequate accuracy. Hence, the development of them has been based on benchmark datasets with dense manual annotations.

Traditional approaches for FGVC include template matching models [12, 13] and sparse coding models [14].With the introduction of large-scale image databases [15] and the power of GPU, CNN networks have been the most accurate methods for basic-level categorization with its strong ability to extract discriminative features. The breakthrough success of R-CNN to make use of deep CNN features in a region proposal framework for object detection [16] has inspired an early milestone [17] for deep FGVC. The part-based R-CNN approach proposed by Zhang et al. leverages deep CNN features computed on bottom-up region proposals, to free the dependency on bounding box annotations at test time [17]. The approach achieves an accuracy on the CUB bird dataset as 73.9% and 76.4%, with and without annotations at the test stage.

Based on the comparison of various approaches for FGVC from perspectives as pose normalization schemes, feature representations, learning algorithms, and annotation types, Branson et al. [18] proposes the pose-normalized CNN approach, which combines and extended the most useful methods and techniques for FGVC at that time. The approach is inspired by the work of Donahue et al. [19], which extracts CNN features from part regions detected by the DPM algorithm. The approach also combines further optimizations based on related FGVC and CNN research. They make use of a similarity-based pose warping function [20] to achieve the best performance and improves it by using more parts for warping estimation. In addition, they investigates different CNN fine-tuning methods on the CUB dataset, based on the work of Girshick et al. [16]. They have achieved 85.4% and 75.7% classification accuracy, with and without annotations at the test stage.

Including part-based R-CNN and pose-normalized CNN, early research depends on annotations at the test stage to achieve higher accuracy and did not evaluate deep convolutional descriptors, most of which are not useful for FGVC. CNN networks with fully connected layers usually have high dimensionality and a great deal of parameters to learn. Wei et al. proposes an end-to-end Mask-CNN model discarding the fully connected layers for bird species categorization with a novel part detection and descriptor selection scheme [21]. By discarding fully connected layers, the method is computationally efficient with a smaller feature representation. With part annotations at the training stage, images are parted into the head mask, torso mask, and the background, with which the part localization can be treated as a three-class segmentation task. Fully convolutional networks (FCN) are adopted for both part localization and descriptor selection through mask weighting. At last, a three-stream Mask-CNN is formed for joint training and aggregating both object-level and part-level cues for fine-grained categorization. Without annotations provided at the test stage, the Mack-CNN with Residual Nets achieves 85.7% classification accuracy on the CUB dataset.

Later, more research has tried to free the dependency of annotations at the test stage. At the same time, the dependency on ground truth part annotations at the training stage has also limited the applications of strongly supervised methods, which lost attention from researchers gradually.

B. *Weakly Supervised Part-based Approaches*

The weakly supervision is to use only category labels or weak annotations such as auxiliary information. Remarkable success has been made in this direction, and advanced approaches include attention-based approaches and transformer-based approaches.

1) Attention-based Approaches

The two-level attention approach proposed by Xiao *et al.* [22] is an early competitive achievement of weakly supervision without relying on bounding boxes or part annotations to train or test. Three types of visual attention are adopted in the approach. Firstly, a bottom-up attention based on selective search [23] is used to propose candidate image patches that have high objectness. To filter out noisy patches, two top-down attention models (object-level and part-level) are designed for finding foreground object and object parts. Two separate pipelines are processed: in the object-level attention, a pre-trained CNN is turned into a FilterNet to select object-relevant patches, based on which another CNN DomainNet will be trained as a domain classifier; in the part-level attention, clustering patterns inside the DonmainNet are observed and corresponding filters are chosen as part-detector. At last, the two pipelines are merged for final classification to take advantage of both level attentions. The approach achieves 69.7% and 77.9% accuracy on the CUB dataset with AlexNet and the more advanced VGGNet respectively. The attention derived from the CNN trained with categorization task allows the approach to be conducted under a weakly supervision setting.

Without explicit part annotations, the performance of part localization and feature extraction are heavily restricted by the discrimination ability of the category-level CNN. Instead of learning independent channels as in the two-level attention approach, a multi-attention convolutional neural network (MA-CNN) [24] proposes to take use of a group of spatial-correlated convolutional channels for stronger discriminative power. Furthermore, it is found that the part localization and fine-grained feature learning are mutually correlated. The MA-CNN conducts part localization and feature leaning in a mutual reinforced way, which optimizes feature maps consistently, improving the accuracy of the detection for multiple representative parts. The approach achieves 86.5% classification accuracy on the CUB dataset.

2) Transformer-based Approaches

Most of the current work for FGVC reuses the backbone network to extract features of selected regions. Despite its powerfulness, CNN networks cannot learn invariance in a computational and parameter efficient manner. A Spatial Transformer module is proposed [25] to explicitly allow the spatial manipulation of data with the network. The module can actively spatially transform feature maps, conditional on the feature map itself, without extra supervision or modification. It can be inserted into existing convolutional architectures, with the recent progress in vision transformer, which with a powerful performance in the traditional classification task. He Chen et al. propose a novel transformer-based framework TansFG to gain a state-of-the-art accuracy on the CUB dataset [26]. They integrate all raw attention weights of the transformer into an attention map to guide the network to select discriminative image patches and compute their relations effectively and accurately. A contrastive loss is added to further enlarge the distance between feature representations of subcategories. They yield tremendous detailed and small discriminative regions. The TransFG method is the current winner on the CUB dataset.

III. Feature Representation

The choice of feature representation is a major factor in both generic and fine-grained classification performance. Branson et al. [18] have conducted a controlled comparison between different feature implementations. It is shown that early methods with bag-of-words features [4] only achieved accuracy in 10-30% range, and the employments of modern features like POOF [20], Fisher-encoded SIFT [27] and Kernel Descriptors (KDES) [28] brought a breakthrough with accuracy achieving 50-62%, and the CNN features [8] has brought another big jump to yield performance of 65-76%, which acts as a foundation for a new era for FGVC. Network architecture is also an important branch in the development of deep learning algorithm for computer vision. Many FGVC methods have aimed at enhancing feature learning on CNN-based approaches by investigating the hierarchy relationship of classes and ensemble of sub neural networks [29-31].

Localization and feature representation are closely related and complementary [32]. Semantic relations between hierarchy labels (e.g., family-level, genus-level, and species-level) can be used to extract discriminative regions and learn feature representations. Wang et al. first integrate taxonomic hierarchy into feature learning for FGVC [32]. They leverage the free labels from the ontology tree of subordinate-level objects to train a series of CNN-based classifiers for each grain level. The internal representation from this network can generate different regions of interest (ROIs) to construct per-granularity descriptors encoding discriminative information across different grain levels. The network can free the dependency on bounding boxes or part annotations with surpassing accuracy compared with models with strong labels. The method achieves 81.7% classification accuracy on the CUB dataset.

IV. Challenge of Data

Though the quality of standard datasets is better than the data in actual application, instead of using expensive, strongly annotated data, web data has demonstrated to be a convenient and free resource for fine-grained classification. How to leverage the rich web images with tags to learn an FGVC model has attracted increasing attention. Another direction is few-shot learning (FSL) or more extremely zero-shot learning (ZSL), which aim at relieving the dependency on dense annotations and the problem of the long-tailed distribution of datasets. In this section, we will introduce and discuss the FGVC research dealing with data challenges.

A. With Web Data

The quality of web data labels is not as good as annotations by experts, and web data are always noisy for training. While it is proved, the noisy web data are effective for fine-grained classification [33]. Adversarial learning [34] and attention mechanisms [35] have been commonly adopted to tackle the challenge of web data. Web data can also be used to generate samples for unseen classes, which is the case in ZSL [36].

A further benefit of the web is the booming of multi-media data. Diverse types of data, such as images, texts, and knowledge bases, can be accessed from the web and be integrated for analysis. Multimodal analysis has drawn attention in recent years, which can boost the accuracy of fine-grained recognition. Commonly used multimodal data for fine-grained recognition are text descriptions and knowledge bases. Text descriptions can be collected from ordinary people while with decent quality, and knowledge bases are well-structured data with rich information of attributes. They can be combined with image data to enrich embedding space [37, 38] or achieve joint training [39].

B. Zero-shot Learning

ZSL has drawn great attention in recent years. In FGVC, ZSL aims to circumvent the requirement of labeled data for all classes and achieve image classification by transferring knowledge from seen classes to unseen ones, always based on auxiliary semantic information such as text descriptions, attributes [40], and online information [35]. Hence, it can relieve the dependency on dense annotations and the problem of long-tailed distribution. The key points for the success of ZSL include improving discriminative power and finding a suitable deviation for unseen classes from seen classes [41]. Typical extant ZSL strategies can be categorized into the following two directions: (1) visual-semantic embedding-based methods; (2) generative ZSL.

1) Visual-semantic Embedding-based Methods

The rich developments of embedding-based ZSL models originate from early pioneering trials of ZSL in computer vision [40, 42]. Lampert et al. proposes a Direct Attribute Prediction (DAP) model that combined attribute probabilities to estimate the posterior of test classes based on the assumption of attribute independence [40]. However, this assumption is not so realistic, and Akata et al. proposes an Attribute Label Embedding (ALE) approach, which cast DAP into a linear joint visual-semantic embedding fashion [42].

The core of visual-semantic embedding-based approaches is to generate the intermediate semantic representations based on side information, which enable a sharing space and knowledge transfer between classes. Specifically, this strategy will learn a mapping function between the visual feature space and semantic space, mapping from one to another [43, 44], or jointly learn an embedding function of both and map them into a common space [42, 45]. Embedding-based ZSL approaches typically include two steps. With an intermediary space, a test sample will be projected. The classification will be carried out based on the similarities between the projected sample and target class feature vectors in the space, during which information loss is inevitable. Guo et al. proposes a one-step framework to avoid this loss by transferring, and pseudo labeling selected samples and performing recognition in the original space, which demonstrates improvement over the two-step framework and transferred ZSL to conventional supervised learning [46].

2) Generative ZSL Methods

Most visual-semantic models cannot guarantee a good generalization from seen to unseen classes and tends to misclassify unseen test instances into seen classes. Another strategy is the generative ZSL approach, which creates artificial examples [46, 47] or features of unseen classes [48], which can alleviate the above problem to some extent, and further enable to make use of advances of many conventional supervised learning models. Early approaches assumed Gaussian distribution before every class, learned and extrapolated the probability distribution of each seen class to unseen class [46, 49]. Long et al. proposes a one-to-one strategy to synthesized visual data by mapping attributes to the visual space, during which synthesized examples are rigidly restricted by the size of the dataset [50]. Later, many research combine the Generative Adversarial Networks (GANs) with ZSL to relax the data distribution assumption. Zhu et al. proposes a generative adversarial ZSL approach that can create arbitrary pseudo data [51]. To enhance the adaptability of transfer from seen to unseen classes, Yu et al. proposes an episode-based prototype generating network (PGN) for zero-shot learning, with each episode as a simulation of a fake ZSL task, and the generalization from seen to unseen classes is enhanced by accumulative experiences through episodes [52].

With additional information as an important part of ZSL, there has been a development of semantic representations in ZSL. Early ZSL methods are attribute-based and have achieved great progress [40, 42, 44]. However, semantic attributes rely on manually collecting, which restricts the generalization of these methods. Hence, a trend of using online text descriptions has become increasingly popular, especially Wikipedia articles. Online data are free, easy to access, and be assigned to different classes automatically. However, online data are also quite noisy and challenging to use. Research has also aimed to tackle this problem [53, 54].

V. Performance Comparison and Analysis

In this section, the performances of representative methods are compared and analyzed, based on the widely-used Caltech-UCSD Birds dataset (CUB200-2011) [4].

A. Benchmark Dataset

The CUB200-2011 dataset includes 11788 images in 200 bird categories, divided into the training set with 5794 images and the test set with 5994 images [4]. The dataset has image level label for each image, and detailed annotations including object bounding boxes, part annotations, and auxiliary information including attribute labels and text descriptions.

B. Performance Analysis

As shown in Table 1, the classification performances of typical methods on the dataset are listed. The methods are divided into two groups: strongly supervised learning (SSL) and weakly supervised learning (WSL). The backbone networks and major features are shown in the table. For the SSL part, the training annotation of each method is mostly the joint of bounding box (BBox) with part landmarks (Parts), and some of them only use single Parts. While for the WSL part, there is no use of annotations in both training and test stages.

Theoretically speaking, SSL algorithms should have more accurate performance than WSL because of the use of more annotations. But with the development of WSL algorithms, the overall performance of WSL has exceeded SSL. The recent WSL methods have beaten most of the early SSL methods. Especially for TransFG, which has a 91.7% accuracy. While the end-to-end Mask CNN with 87.3% accuracy from SSL algorithms maintains an advantage over most WSLs. In a further comparison that all methods performed on VggNet are grouped, it can be found that the method performs better if Softmax is put into use. Moreover, if the variables are controlled, the Two-level Attention and multiple granularity CNN performed on VggNet have different performance accuracy. This is because the internal representation from the network of multiple granularity CNN can generate different regions of interest to construct per-granularity descriptors encoding discriminative information across different grain levels and can still earn surpassing accuracy without depending on bounding boxes or part annotations. Besides, it can be found that the choice of descriptors affects the performance significantly. The use of VggNet has better accuracy than AlexNet. Although end-to-end Mask CNN is an SSL method, it still greatly exceeds most WSLs because of its descriptors. And if both methods and descriptors are the same, the use of annotation or extra data can boost the final accuracy, which can be seen from the comparison of the pairs of the same method from the table. The result below shows that the descriptors are the key factor for performance accuracy.

TABLE 1. THE CLASSIFICATION ACCURACY OF REPRESENTATIVE METHODS ON DIFFERENT DATASETS

Method	**Supervision**	**Description**	**Train annotation**	**Test annotation**	**Accuracy**
Part-based R-CNN[17]	SSL	AlexNet + Fine-Tune + SVM	BBox + Parts	BBox	76.4
Part-based R-CNN[17]	SSL	AlexNet + Fine-Tune + SVM	BBox + Parts	/	73.9
Pose-normalized CNN[18]	SSL	AlexNet + Fine-Tune + SVM	BBox + Parts	BBox + Parts	85.4
Pose-normalized CNN[18]	SSL	AlexNet + Fine-Tune + SVM	BBox + Parts	/	75.7
End-to-end Mask CNN[21]	SSL	VGGNet + Softmax	Parts	/	85.7
Two-level Attention[22]	WSL	AlexNet	/	/	69.7
Two-level Attention[22]	WSL	VGGNet	/	/	77.9
MA-CNN[24]	WSL	VGGNet + Softmax	/	/	86.5
TransFG[26]	WSL	ViT-B 16	/	/	91.7
Multiple granularity CNN[32]	SSL	VGGNet	BBox	/	83.0
Multiple granularity CNN[32]	WSL	VGGNet	/	/	81.7
Xu et al. [54]	SSL	AlexNet	BBox + Parts	/	78.6
Xu et al. [54]	SSL	AlexNet	BBox + Parts + Web	/	84.6

VI. FUTRUE WORK

FGVC is in the basement of most computer vision fields. However, the problems it may encounter during the process are significantly different from other visual tasks. In this case, people are facing more unique and severe challenges in the process of FGVC improvement. With the development of deep learning and progress made by data analysis, FGVC did achieve a breakthrough. But beyond that, various potential problems and unknown improvement ideas still exist. Here we will conclude the current problems that need to be solved and prospect the future direction.

(1) The greatest challenge FGVC faces would be distinguishing different subcategories from subtle differences through accurate feature expression. Researchers are now using the fine-grained aggregation model to improve the two-dimensional approximation model. Fine-grained image recognition assisted by external information makes full use of text, network data, multimodal data, and human help to effectively carry out fine-grained tasks. Due to the high cost of human labeled data, and with the current success with external or alternative data, researchers would continue to make more use of data with weak labels from the internet, as well as zero-shot learning.

(2) The goal of fine-grained recognition is to correctly identify targets in various subcategories of large categories. However, the movement, posture, and gait of objects in the same subcategories may be the same, and objects from different subcategories may have the same posture, which is a major difficulty in recognition. To deal with this kind of problem, a selective search algorithm can be adopted to generate candidate boxes that may appear in organisms or object parts on fine-grained images. Object detection is conducted according to the R-CNN process. With the help of object bounding box and part annotation in fine-grained images, three detection models can be trained: one corresponds to fine-grained object level detection, one corresponds to the detection of the head of the object, and the other corresponds to the detection of the body trunk of the object. Next, the obtained image patches can be taken as input to train a CNN respectively, which learns the features of the object or part.

(3) To carry out more accurate fine-grained image analysis, more appropriate datasets are in demand. Now, there are many benchmark datasets provided for the classification of varied tasks. With fine-grained image recognition requiring to accurately identify differences among the same category, gigantic datasets are still a fundamental support.

(4) How to obtain high-performance neural networks. The quality of these networks directly affects the quality of the fine-grained recognition model. Now the choice of network is full of randomness. Improving the overall quality of the network is particularly important. Moreover, neural networks have different performances when applied to different tasks. Hence, its unity for various situations needs to be considered.

VII. CONCLUSION

This article focuses on the core and technical problems of fine-grained visual classification. From the advantages and disadvantages of mainstream FGVC algorithms to optimizing and improving their accuracy, from strongly supervision to weakly supervision, we summarize the different core problems and methods. And we also compare and analyze the experimental data and results of typical algorithms on benchmark datasets. Moreover, we analyze and prospect the four different research challenges and development direction in the future.

REFERENCES

[1] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC), 2011, vol. 2, no. 1: Citeseer.

[2] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, 2008: IEEE, pp. 722-729.

[3] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in Proceedings of the IEEE international conference on computer vision workshops, 2013, pp. 554-561.

[4] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[5] J. Kay and M. Merrifield, "The Fishnet Open Images Database: A Dataset for Fish Detection and Fine-Grained Categorization in Fisheries," arXiv preprint arXiv:2106.09178, 2021.

[6] S. Beery, A. Agarwal, E. Cole, and V. Birodkar, "The iWildCam 2021 Competition Dataset," arXiv preprint arXiv:2105.03494, 2021.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, pp. 1097-1105, 2012.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[10] C. Qiu and W. Zhou, "A Survey of Recent Advances in CNN-Based Fine-Grained Visual Categorization," in 2020 IEEE 20th International Conference on Communication Technology (ICCT), 2020: IEEE, pp. 1377-1384.

[11] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," International Journal of Automation and Computing, vol. 14, no. 2, pp. 119-135, 2017.

[12] S. Yang, L. Bo, J. Wang, and L. Shapiro, "Unsupervised template learning for fine-grained object recognition," Advances in neural information processing systems, vol. 25, pp. 3122-3130, 2012.

[13] B. Yao, A. Khosla, and L. Fei-Fei, "Combining randomization and discrimination for fine-grained image categorization," in CVPR 2011, 2011: IEEE, pp. 1577-1584.

[14] S. Gao, I. W.-H. Tsang, and Y. Ma, "Learning category-specific dictionary and shared dictionary for fine-grained image categorization," IEEE Transactions on Image Processing, vol. 23, no. 2, pp. 623-634, 2013.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition, 2009: Ieee, pp. 248-255.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580-587.

[17] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in European conference on computer vision, 2014: Springer, pp. 834-849.

[18] S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," arXiv preprint arXiv:1406.2952, 2014.

[19] J. Donahue et al., "Decaf: A deep convolutional activation feature for generic visual recognition," in International conference on machine learning, 2014: PMLR, pp. 647-655.

[20] T. Berg and P. N. Belhumeur, "Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 955-962.

[21] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization," Pattern Recognition, vol. 76, pp. 704-714, 2018.

[22] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 842-850.

[23] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," International journal of computer vision, vol. 104, no. 2, pp. 154-171, 2013.

[24] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5209-5217.

[25] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," Advances in neural information processing systems, vol. 28, pp. 2017-2025, 2015.

[26] J. He et al., "TransFG: A Transformer Architecture for Fine-grained Recognition," arXiv preprint arXiv:2103.07976, 2021.

[27] Y. Chai, V. Lempitsky, and A. Zisserman, "Symbiotic segmentation and part localization for fine-grained categorization," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 321-328.

[28] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 729-736.

[29] Z. Ge, C. McCool, C. Sanderson, and P. Corke, "Subset feature learning for fine-grained category classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 46-52.

[30] Z. Ge, A. Bewley, C. McCool, P. Corke, B. Upcroft, and C. Sanderson, "Fine-grained classification via mixture of deep convolutional neural networks," in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 2016: IEEE, pp. 1-6.

[31] Z. Wang, X. Wang, and G. Wang, "Learning fine-grained features via a CNN tree for large-scale classification," Neurocomputing, vol. 275, pp. 1231-1240, 2018.

[32] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang, "Multiple granularity descriptors for fine-grained categorization," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2399-2406.

[33] J. Krause et al., "The unreasonable effectiveness of noisy data for fine-grained recognition," in European Conference on Computer Vision, 2016: Springer, pp. 301-320.

[34] X. Sun, L. Chen, and J. Yang, "Learning from web data using adversarial discriminative neural networks for fine-grained classification," in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, vol. 33, no. 01, pp. 273-280.

[35] B. Zhuang, L. Liu, Y. Li, C. Shen, and I. Reid, "Attend in groups: a weakly-supervised deep learning framework for learning from web data," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1878-1887.

[36] L. Niu, A. Veeraraghavan, and A. Sabharwal, "Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7171-7180.

[37] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 49-58.

[38] H. Xu, G. Qi, J. Li, M. Wang, K. Xu, and H. Gao, "Fine-grained Image Classification by Visual-Semantic Embedding," in IJCAI, 2018, pp. 1043-1049.

[39] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5994-6002.

[40] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 3, pp. 453-465, 2013.

[41] M. Elhoseiny and M. Elfeki, "Creativity inspired zero-shot learning," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5784-5793.

[42] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 7, pp. 1425-1438, 2015.

[43] R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. D. Manning, and A. Y. Ng, "Zero-shot learning through cross-modal transfer," arXiv preprint arXiv:1301.3666, 2013.

[44] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2021-2030.

[45] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in International conference on machine learning, 2015: PMLR, pp. 2152-2161.

[46] Y. Guo, G. Ding, J. Han, and Y. Gao, "Synthesizing samples fro zero-shot learning," 2017: IJCAI.

[47] S. Kousha and M. A. Brubaker, "Zero-shot Learning with Class Description Regularization," arXiv preprint arXiv:2106.16108, 2021.

[48] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5542-5551.

[49] W. Wang et al., "Zero-shot learning via class-conditioned deep generative models," in Thirty-second AAAI conference on artificial intelligence, 2018.

[50] Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, and J. Han, "From zero-shot learning to conventional supervised classification: Unseen visual data synthesis," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1627-1636.

[51] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal, "A generative adversarial approach for zero-shot learning from noisy texts," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1004-1013.

[52] Y. Yu, Z. Ji, J. Han, and Z. Zhang, "Episode-based prototype generating network for zero-shot learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14035-14044.

[53] R. Qiao, L. Liu, C. Shen, and A. Van Den Hengel, "Less is more: zero-shot learning from online textual documents with noise suppression," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2249-2257.

[54] Z. Xu, S. Huang, Y. Zhang, and D. Tao, "Webly-supervised fine-grained visual categorization via deep domain adaptation," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 5, pp. 1100-1113, 2016.