Edinburgh Research Explorer

Fine-Grained Image Analysis with Deep Learning: A Survey

Fine-Grained Image Analysis with Deep Learning: A Survey

Xiu-Shen Wei, *Member, IEEE*, Yi-Zhe Song, *Senior Member, IEEE*, Oisin Mac Aodha, Jianxin Wu, *Member, IEEE*, Yuxin Peng, *Senior Member, IEEE*, Jinhui Tang, *Senior Member, IEEE*, Jian Yang, *Member, IEEE*, Serge Belongie

Abstract—Fine-grained image analysis (FGIA) is a longstanding and fundamental problem in computer vision and pattern recognition, and underpins a diverse set of real-world applications. The task of FGIA targets analyzing visual objects from subordinate categories, *e.g.*, species of birds or models of cars. The small inter-class and large intra-class variation inherent to fine-grained image analysis makes it a challenging problem. Capitalizing on advances in deep learning, in recent years we have witnessed remarkable progress in deep learning powered FGIA. In this paper we present a systematic survey of these advances, where we attempt to re-define and broaden the field of FGIA by consolidating two fundamental fine-grained research areas – fine-grained image recognition and fine-grained image retrieval. In addition, we also review other key issues of FGIA, such as publicly available benchmark datasets and related domain-specific applications. We conclude by highlighting several research directions and open problems which need further exploration from the community.

Index Terms—Fine-Grained Images Analysis; Deep Learning; Fine-Grained Image Recognition; Fine-Grained Image Retrieval.

✦

1 INTRODUCTION

THE human visual system is inherently capable of fine-grained image reasoning – we are not only able to tell a dog from a bird, but also know the difference between a Siberian Husky and an Alaskan Malamute (see Figure 1). Fine-grained image analysis (FGIA) was introduced to the academic community for the very same purpose, *i.e.*, to teach machine to "see" in a fine-grained manner. FGIA approaches are present in a wide-range of applications in both industry and research, with examples including automatic biodiversity monitoring [1], [2], [3], intelligent retail [4], [5], [6], and intelligent transportation [7], [8], and have resulted in a positive impact in areas such as conservation [9] and commerce [10].

The goal of FGIA in computer vision is to retrieve and recognize images belonging to multiple subordinate categories of a super-category (*aka* a meta-category or a basic-level category), *e.g.*, different species of animals/plants, different models of cars, different kinds of retail products, etc. The key challenge therefore lies with understanding fine-grained visual differences that sufficiently discriminate between objects that are highly similar in overall appearance, but differ in *fine-grained* features. Great strides has been made since its inception almost two decades ago [11], [12], [13]. Deep learning [14] in particular has emerged as a powerful method for discriminative feature learning, and has led to
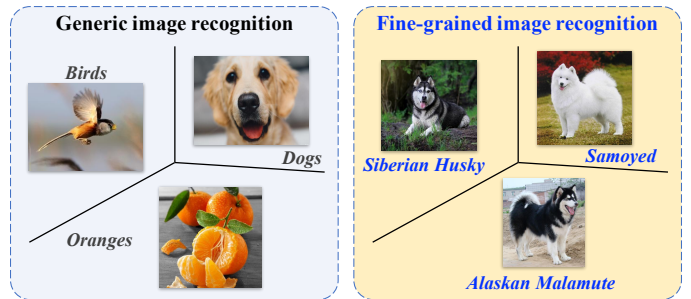

Figure 1. Fine-grained image analysis *vs*. generic image analysis (using visual classification as an example).

remarkable breakthroughs in the field of FGIA. Deep learning enabled FGIA has greatly advanced the practical deployment of these methods in a diverse set of application scenarios [5], [7], [8], [9].

There has been significant interest in FGIA from both the computer vision and machine learning research communities in recent years. Rough statistics indicate an average of > 10 conference papers on deep learning based FGIA are published every year in each of the premium vision and machine learning conferences. There have also been a number of special issues addressing FGIA [15], [16], [17], [18], [19]. Additionally, a number of influential competitions on FGIA are frequently held on online platforms. Representatives include the series of iNaturalist Competitions (for large numbers of natural species) [20], the Nature Conservancy Fisheries Monitoring (for fish species categorization) [21], Humpback Whale Identification (for whale identity categorization) [22], among others. Each competition attracted hundreds of participants from around the world, and some even exceeding 2,000 teams. Specific tutorials and workshops aimed at FGIA topics have also been held at top-tier international conferences, *e.g.*, [23], [24].

- *X.-S. Wei and J. Yang are with PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, China. Y.-Z. Song is with University of Surrey, UK. O. Mac Aodha is with the University of Edinburgh, UK. J. Wu is the State Key Laboratory for Novel Software Technology, Nanjing University, China. Y. Peng is with Peking University, China. J. Tang is with Nanjing University of Science and Technology, China. S. Belongie is with the University of Copenhagen and the Pioneer Centre for AI, Denmark.*
- *X.-S. Wei and J. Yang are the corresponding authors.*

```
                        ┌─────────────────────────────────────────┐
                        │  Fine-Grained Image Analysis with Deep   │
                        │            Learning                       │
                        └─────────────────────────────────────────┘
```

Fine-Grained Image Recognition

Recognition by Localization-Classification Subnetworks (*Sect. 5.1*)

Employing Detection or Segmentation Methods [31, 64-75]

Utilizing Deep Filters [76-82]

Leveraging Attention Mechanisms [27, 39, 83-89]

Other Methods [90-94]

Recognition by End-to-End Feature Encoding (*Sect. 5.2*)

Performing High-Order Feature Interactions [95-106]

Designing Specific Loss Functions [27, 107-112]

Other Methods [50, 113-115]

Recognition with External Information (*Sect. 5.3*)

Recognition with Web Data [62, 149-156]

Recognition by using Multi-Modality Data [59, 157-160]

Recognition with Humans-in-the Loop [166-167]

Fine-Grained Image Retrieval

Content-based Fine-Grained Retrieval (*Sect. 6.1*) [28, 32, 170-171]

Sketch-based Fine-Grained Retrieval (*Sect. 6.2*) [52, 54-55, 175-180]

Future Directions of FGIA

Precise Definition of "Fine-Grained"

Next-Generation Fine-Grained Datasets

Application to 3D Fine-Grained Tasks

Robust Fine-Grained Representations

Interpretations of Fine-Grained Models

Fine-Grained Few-Shot Learning

Fine-Grained Hashing

Automatic Fine-Grained Models

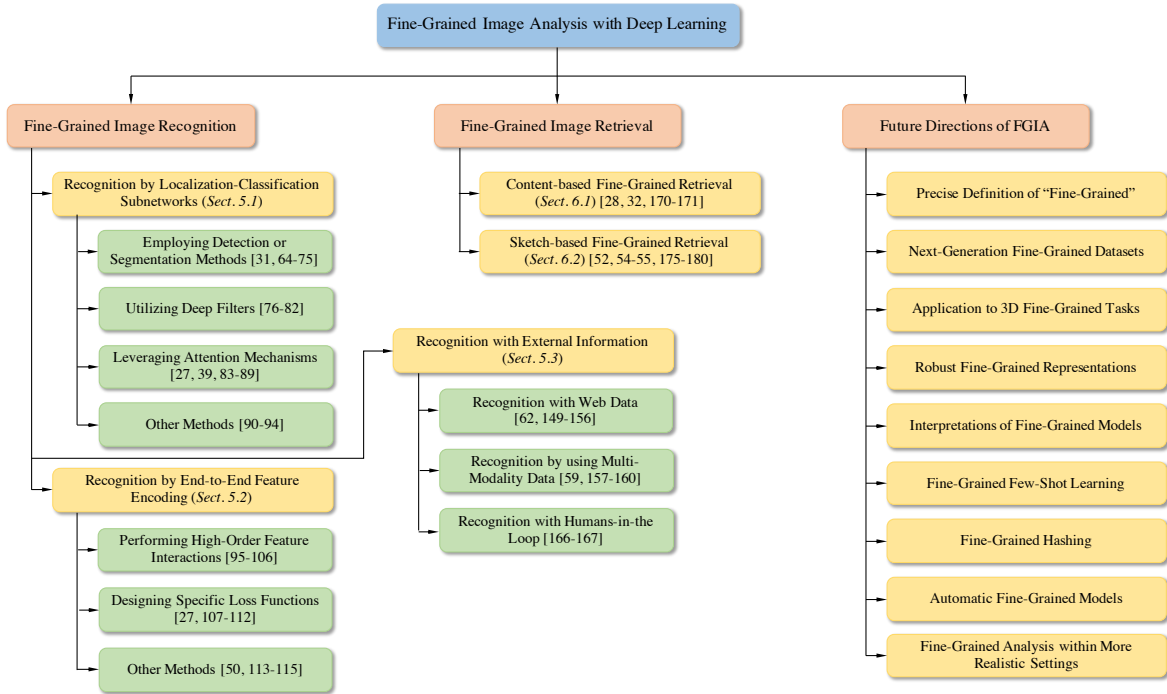Fine-Grained Analysis within More Realistic Settings

Figure 2. Overview of the landscape of deep learning based fine-grained image analysis (FGIA), as well as future directions.

Despite such salient interest, the study of FGIA with deep learning remains fragmented. It is therefore the purpose of this survey to (i) provide a comprehensive survey of recent achievements in FGIA, especially those brought by deep learning techniques, and more importantly (ii) to propose a unified research front by consolidating research from different aspects of FGIA. Our approach is significantly different to existing surveys [25], [26] that focus solely on the problem of fine-grained *recognition/classification*, which as we argue only constitutes part of the larger study of FGIA. In particular, we attempt to re-define and broaden the field of fine-grained image analysis, by highlighting the synergy between fine-grained *recognition*, and the parallel but complementary task of fine-grained image *retrieval*, which is also an integral part of FGIA.

Our survey takes a unique deep learning based perspective to review recent advances in FGIA in a broad, systematic, and comprehensive manner. Our main contributions are summarized as follows:

- We broaden the field of FGIA, by offering a consolidated landscape that promotes synergies between related problems in fine-grained image analysis.
- We provide a comprehensive review of FGIA techniques based on deep learning, including commonly accepted problem definitions, benchmark datasets, different families of FGIA methods, along with covering domain-specific FGIA applications. Particularly, we organize these approaches taxonomically (see Figure 2) to provide readers with a quick snapshot of the state-of-the-art in this area.
- We consolidate the performance of existing methods on several publicly available datasets and provide discussion and insights to inform future research.
- We finish by discussing existing challenges and open issues, and identify new trends and future directions

to provide a plausible road map for the community to address these problems.

- Finally, in an attempt to continuously track recent developments in this fast advancing field, we provide an accompanying webpage which catalogs papers addressing FGIA problems, according to our problem-based taxonomy: http://www.weixiushen.com/project/Awesome_FGIA/Awesome_FGIA.html.

2 RECOGNITION VS. RETRIEVAL

Previous surveys of FGIA, *e.g.*, [25], [26], predominately focused on fine-grained recognition, and as a result do not expose all facets of the FGIA problem. In this survey, we cover two fundamental areas of fine-grained image analysis for the first time (*i.e.*, recognition and retrieval) in order to give a comprehensive review of recent advances in deep learning based FGIA techniques. In Figure 2, we provide a new taxonomy that reflects the current FGIA landscape.

Fine-Grained Recognition: We organize the different families of fine-grained *recognition* approaches into three paradigms, *i.e.*, 1) recognition by localization-classification subnetworks, 2) recognition by end-to-end feature encoding, and 3) recognition with external information. Fine-grained recognition is the most studied area in FGIA, since recognition is a fundamental ability of most visual systems and is thus worthy of long-term continuous research.

Fine-Grained Retrieval: Based on the type of query image, we separate fine-grained *retrieval* methods into two groups, *i.e.*, 1) content-based fine-grained image retrieval and 2) sketch-based fine-grained image retrieval. Compared with fine-grained recognition, fine-grained retrieval is an emerging area of FGIA in recent years, one that is attracting more and more attention from both academia and industry.

Recognition and Retrieval Differences: Both fine-grained recognition and retrieval aim to identify the discrim-

inative, but subtle, differences between different fine-grained objects. However, fine-grained recognition is a closed-world task with a *fixed* number of subordinate categories. In contrast, fine-grained retrieval extends the problem to an open-world setting with unlimited sub-categories. Furthermore, fine-grained retrieval also aims to rank all the instances so that images depicting the concept of interest (*e.g.*, the same sub-category label) are ranked highest based on the fine-grained details in the query.

Recognition and Retrieval Synergies: Advances in fine-grained recognition and retrieval have commonalities and can benefit each other. Many common techniques are shared by both fine-grained recognition and retrieval, *e.g.*, deep metric learning methods [27], [28], multi-modal matching methods [29], [30], and the basic ideas of selecting useful deep descriptors [31], [32], etc. Detailed discussions are elaborated in Section 7. Furthermore, in real-world applications, fine-grained recognition and retrieval also compliment each other, *e.g.*, retrieval techniques are able to support novel sub-category recognition by utilizing learned representations from a fine-grained recognition model [5], [33].

3 BACKGROUND: PROBLEM AND CHALLENGES

Fine-grained image analysis (FGIA) focuses on dealing with objects belonging to multiple *subordinate categories* of the same meta-category (*e.g.*, different species of birds or different models of cars), and generally involves two central tasks: fine-grained image recognition and fine-grained image retrieval. As illustrated in Figure 3, fine-grained analysis lies in the continuum between basic-level category analysis (*i.e.*, generic image analysis) and instance-level analysis (*e.g.*, the identification of individuals).

Specifically, what distinguishes FGIA from generic image analysis is that in generic image analysis, target objects belong to coarse-grained meta-categories (*i.e.*, basic-level categories) and are thus visually quite different (*e.g.*, determining if an image contains a bird, a fruit, or a dog). However, in FGIA, since objects typically come from sub-categories of the same meta-category, the fine-grained nature of the problem causes them to be visually similar. As an example of fine-grained recognition, in Figure 1, the task is to classify different breeds of dogs. For accurate image recognition, it is necessary to capture the subtle visual differences (*e.g.*, discriminative features such as ears, noses, or tails). Characterizing such features is also desirable for other FGIA tasks (*e.g.*, retrieval). Furthermore, as noted earlier, the fine-grained nature of the problem is challenging because of the *small inter-class variations* caused by highly similar sub-categories, and the *large intra-class variations* in poses, scales and rotations (see Figure 4). It is as such the opposite of generic image analysis (*i.e.*, the small intra-class variations and the large inter-class variations), and what makes FGIA a unique and challenging problem.

While instance-level analysis typically targets a *specific instance* of an object not just object categories or even object sub-categories, if we move down the spectrum of granularity, in the extreme, individual identification (*e.g.*, face identification) can be viewed as a special instance of fine-grained recognition, where the granularity is at the
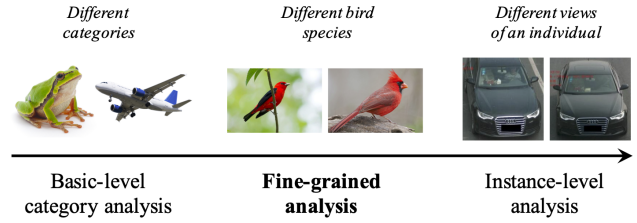

Figure 3. An illustration of fine-grained image analysis which lies in the continuum between the basic-level category analysis (*i.e.*, generic image analysis) and the instance-level analysis (*e.g.*, car identification).

individual identity level. For instance, person/vehicle re-identification [7], [34] can be considered a fine-grained task, which aims to determine whether two images are taken of the same specific person/vehicle. In practice, these works solve the corresponding domain-specific problems using related methods to FGIA, *e.g.*, by capturing the discriminative parts of objects (faces, people, and vehicles) [8], [35], [36], discovering coarse-to-fine structural information [37], developing attribute-based models [38], [39], and so on. Research in these instance-level problems is also very active. However, since such problems are not within the scope of classical FGIA (see Figure 3), for more information, we refer readers to survey papers of these specific topics, *e.g.*, [7], [34], [40]. In the following, we start by formulating our definition of classical FGIA.

Formulation: In generic image recognition, we are given a training dataset $\mathcal{D} = \left\{ \left(\boldsymbol{x}^{(n)}, y^{(n)} \right) | i = 1, ..., N \right\}$, containing multiple images and associated class labels (*i.e.*, $\boldsymbol{x}$ and y), where $y \in [1, ..., C]$. Each instance $(\boldsymbol{x}, y)$ belongs to the joint space of both the image and label spaces (*i.e.*, $\mathcal{X}$ and $\mathcal{Y}$, respectively), according to the distribution of $p_r(\boldsymbol{x}, y)$

$$(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}. \tag{1}$$

In particular, the label space $\mathcal{Y}$ is the union space of all the C subspaces corresponding to the C categories, *i.e.*, $\mathcal{Y} = \mathcal{Y}_1 \cup \mathcal{Y}_2 \cup \cdots \cup \mathcal{Y}_c \cup \cdots \cup \mathcal{Y}_C$. Then, we can train a
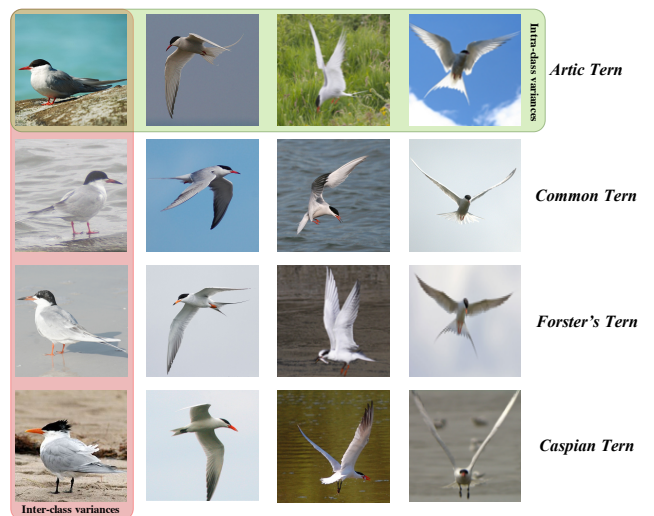

Figure 4. Key challenges of fine-grained image analysis, *i.e.*, small inter-class variations and large intra-class variations. Here we present four different `Tern` species from [13], one species per row, with different instances in the columns.

predictive/recognition deep network $f(\boldsymbol{x};\theta)$ parameterized by θ for generic image recognition by minimizing the expected risk

$$\min_{\theta} \mathbb{E}_{(\boldsymbol{x},y)\sim p_r(\boldsymbol{x},y)}\left[\mathcal{L}(y, f(\boldsymbol{x};\theta))\right], \qquad (2)$$

where $\mathcal{L}(\cdot, \cdot)$ is a loss function that measures the match between the true labels and those predicted by $f(\cdot;\theta)$. While, as aforementioned, fine-grained recognition aims to accurately classify instances of different subordinate categories from a certain meta-category, *i.e.*,

$$(\boldsymbol{x}, y') \in \mathcal{X} \times \mathcal{Y}_c, \qquad (3)$$

where y' denotes the fine-grained label and $\mathcal{Y}_c$ represents the label space of class c as the meta-category. Therefore, the optimization objective of fine-grained recognition is as

$$\min_{\theta} \mathbb{E}_{(\boldsymbol{x},y')\sim p'_r(\boldsymbol{x},y')}\left[\mathcal{L}(y', f(\boldsymbol{x};\theta))\right]. \qquad (4)$$

Compared with fine-grained recognition, in addition to getting the sub-category correct, fine-grained retrieval also must rank all the instances so that images belonging to the same sub-category are ranked highest based on the fine-grained details in the query of retrieval tasks. Given an input query $\boldsymbol{x}^q$, the goal of a fine-grained retrieval system is to rank all instances in a retrieval set $\Omega = \{\boldsymbol{x}^{(i)}\}_{i=1}^{M}$ (whose label $y' \in \mathcal{Y}_c$) based on their fine-grained relevance to the query. Let $\mathcal{S}_{\Omega} = \{s^{(i)}\}_{i=1}^{M}$ represent the similarity between $\boldsymbol{x}^q$ and each $\boldsymbol{x}^{(i)}$ measured via a pre-defined metric applied to the corresponding fine-grained representations, *i.e.*, $h(\boldsymbol{x}^q;\delta)$ and $h(\boldsymbol{x}^{(i)};\delta)$. Here, δ denotes the parameters of a retrieval model h. For the instances whose labels are consistent with the fine-grained category of $\boldsymbol{x}^q$, we form them into a positive set $\mathcal{P}_q$ and obtain the corresponding $\mathcal{S}_P$. Then, the retrieval model $h(\cdot;\delta)$ can be trained by maximizing the ranking based score

$$\max_{\delta} \frac{\mathcal{R}(i, \mathcal{S}_P)}{\mathcal{R}(i, \mathcal{S}_{\Omega})}, \qquad (5)$$

w.r.t. all the query images, where $\mathcal{R}(i, \mathcal{S}_P)$ and $\mathcal{R}(i, \mathcal{S}_{\Omega})$ refer to the rankings of the instance i in $\mathcal{P}_q$ and Ω, respectively.

4 BENCHMARK DATASETS

In recent years, the vision community has released many fine-grained benchmark datasets covering diverse domains, *e.g.*, birds [1], [13], [41], dogs [27], [42], cars [43], airplanes [44], flowers [45], vegetables [46], fruits [46], foods [47], fashion [6], [33], [38], retail products [5], [48], etc. Additionally, it is worth noting that even the most popular large-scale image classification dataset, *i.e.*, ImageNet [49], also contains fine-grained classes covering a lot of dog and bird sub-categories.

Representative images from some of these fine-grained benchmark datasets can be found in Figure 5. In Table 1, we summarize the most commonly used image datasets, and indicate their meta-category, the amount of images, the number of categories, their main task, and additional available supervision, *e.g.*, bounding boxes, part annotations, hierarchical labels, attribute labels, and text descriptions (cf. Figure 6). These datasets have been one of the most important factors for the considerable progress in the field, not only as a common ground for measuring and comparing performance

of competing approaches, but also pushing this field towards increasingly complex, practical, and challenging problems.

The fine-grained bird classification dataset *CUB200-2011* [13] is one of the most popular fine-grained datasets. The majority of FGIA approaches choose it for comparisons with the state-of-the-art. Moreover, continuous contributions are made upon *CUB200-2011* for advanced tasks, *e.g.*, collecting text descriptions of the fine-grained images for multi-modal analysis, cf. [58], [59] and Section 5.3.2.

In recent years, more challenging and practical fine-grained datasets have been proposed, *e.g.*, *iNat2017* containing different species of plants and animals [2], and *RPC* for retail products [5]. Novel properties of these datasets include the fact that they are large-scale, have a hierarchical structure, exhibit a domain gap, and form a long-tailed distribution. These challenges illustrate the practical requirements of FGIA in the real-world and motivate new interesting research challenges (cf. Section 8).

Beyond that, a series of fine-grained sketch-based image retrieval datasets, *e.g.*, *QMUL-Shoe* [52], *QMUL-Chair* [52], *QMUL-handbag* [54], *SBIR2014* [51], *SBIR2017* [55], *Sketchy* [53], *QMUL-Shoe-V2* [56], were constructed to further advance the development of fine-grained retrieval, cf. Section 6.2. Furthermore, some novel datasets and benchmarks, such as *FG-Xmedia* [57], were constructed to expand fine-grained image retrieval to fine-grained cross-media retrieval.

5 FINE-GRAINED IMAGE RECOGNITION

Fine-grained image recognition has been by far the most active research area of FGIA in the past decade. Fine-grained recognition aims to discriminate numerous visually similar subordinate categories that belong to the same basic category, such as the fine distinction of animal species [2], cars [43], fruits [46], aircraft models [44], and so on. It has been frequently applied in real-world tasks, *e.g.*, ecosystem conservation (recognizing biological species) [9], intelligent retail systems [5], [10], etc. Recognizing fine-grained categories is difficult due to the challenges of discriminative region localization and fine-grained feature learning. Researchers have attempted to deal with these challenges from diverse perspectives. In this section, we review the main fine-grained recognition approaches since the advent of deep learning.

Broadly, existing fine-grained recognition approaches can be organized into the following three main paradigms:

- Recognition by localization-classification subnetworks;
- Recognition by end-to-end feature encoding;
- Recognition with external information.

Among them, the first two paradigms restrict themselves by only utilizing the supervisions associated with fine-grained images, such as image labels, bounding boxes, part annotations, etc. To further resolve ambiguous fine-grained problems, there is a body of work that uses additional information such as where and when the image was taken [60], [61], web images [62], [63], or text description [58], [59]. In order to present these representative deep learning based fine-grained recognition methods intuitively, we show a chronological overview in Figure 7 by organizing them into the three aforementioned paradigms.

For performance evaluation, when the test set is balanced (*i.e.*, there is a similar number test examples from each class),

Figure 5. Examples of fine-grained images belonging to different species of flowers/vegetables [46], different models of cars [43] and aircraft [44] and different kinds of retail products [5]. Accurate identification of these fine-grained objects requires the extraction of discriminative, but subtle, object parts or image regions. (Best viewed in color and zoomed in.)

Table 1
Summary of popular fine-grained image datasets organized by their major applicable topics and sorted by their release time. Note that, "♯ images" means the total number of images of these datasets. "BBox" indicates whether this dataset provides object bounding box supervisions. "Part anno." means that key parts annotations are provided. "HRCHY" corresponds to hierarchical labels. "ATR" represents attribute labels (*e.g.*, wing color, male, female, etc). "Texts" indicates whether fine-grained text descriptions of images are supplied. Several datasets are listed here twice since they are commonly used in both recognition and retrieval tasks.

Topic	Dataset name	Year	Meta-class	♯ images	♯ categories	BBox	Part anno.	HRCHY	ATR	Texts
Recog.	Oxford Flowers [45]	2008	Flowers	8,189	102					✓
	CUB200-2011 [13]	2011	Birds	11,788	200	✓	✓		✓	✓
	Stanford Dogs [42]	2011	Dogs	20,580	120	✓				
	Stanford Cars [43]	2013	Cars	16,185	196	✓				
	FGVC Aircraft [44]	2013	Aircrafts	10,000	100	✓		✓		
	Birdsnap [41]	2014	Birds	49,829	500	✓	✓		✓	
	Food101 [47]	2014	Food dishes	101,000	101					
	NABirds [1]	2015	Birds	48,562	555	✓	✓			
	Food-975 [50]	2016	Foods	37,885	975				✓	
	DeepFashion [38]	2016	Clothes	800,000	1,050	✓	✓		✓	
	Fru92 [46]	2017	Fruits	69,614	92			✓		
	Veg200 [46]	2017	Vegetable	91,117	200			✓		
	iNat2017 [2]	2017	Plants & Animals	857,877	5,089	✓		✓		
	Dogs-in-the-Wild [27]	2018	Dogs	299,458	362					
	RPC [5]	2019	Retail products	83,739	200	✓		✓		
	Products-10K [48]	2020	Retail products	150,000	10,000	✓		✓		
	iNat2021 [3]	2021	Plants & Animals	3,286,843	10,000			✓		
Retriev.	Oxford Flowers [45]	2008	Flowers	8,189	102					✓
	CUB200-2011 [13]	2011	Birds	11,788	200	✓	✓		✓	✓
	Stanford Cars [43]	2013	Cars	16,185	196	✓				
	SBIR2014* [51]	2014	Multiple	1,120/7,267	14	✓	✓		✓	
	DeepFashion [38]	2016	Clothes	800,000	1,050	✓	✓		✓	
	QMUL-Shoe* [52]	2016	Shoes	419/419	1				✓	
	QMUL-Chair* [52]	2016	Chairs	297/297	1				✓	
	Sketchy* [53]	2016	Multiple	75,471/12,500	125					
	QMUL-Handbag* [54]	2017	Handbags	568/568	1					
	SBIR2017* [55]	2017	Shoes	912/304	1		✓		✓	
	QMUL-Shoe-V2* [56]	2019	Shoes	6,730/2,000	1					
	FG-Xmedia† [57]	2019	Birds	11,788	200					✓

* For these fine-grained sketch-based image retrieval datasets, normally they have sketch-and-image pairs (*i.e.*, not only images). Thus, we present the numbers of sketches and images separately (the numbers of sketches first). Regarding "♯ categories", we report the number of meta-categories in these datasets.
† Except for text descriptions, *FG-Xmedia* also contains multiple other modalities, *e.g.*, videos and audios.
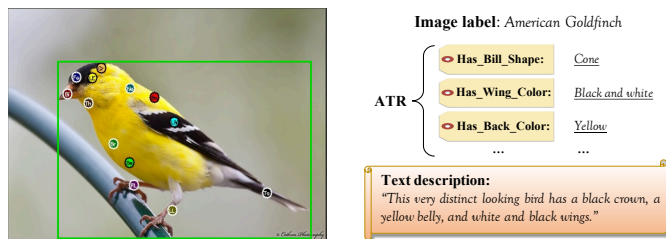

Figure 6. An example image from *CUB200-2011* [13] with multiple different types of annotations *e.g.*, category label, part annotations (*aka* key point locations), object bounding box shown in green, attribute labels (*i.e.*, "ATR"), and a text description.

the most commonly used metric in fine-grained recognition is classification *accuracy* across all subordinate categories of the datasets. It is defined as

$$\text{Accuracy} = \frac{|I_{\text{correct}}|}{|I_{\text{total}}|}, \tag{6}$$

where $|I_{\text{total}}|$ represents the number of images across all sub-categories in the test set and $|I_{\text{correct}}|$ represents the number of images which are correctly categorized by the model.

5.1 Recognition by Localization-Classification Subnetworks

Researchers have attempted to create models that capture the discriminative semantic parts of fine-grained objects and then construct a mid-level representation corresponding to these parts for the final classification, cf. Figure 8. More specifically, a localization subnetwork is designed for locating key parts, and then the corresponding part-level (local) feature vectors are obtained. This is usually combined with object-level
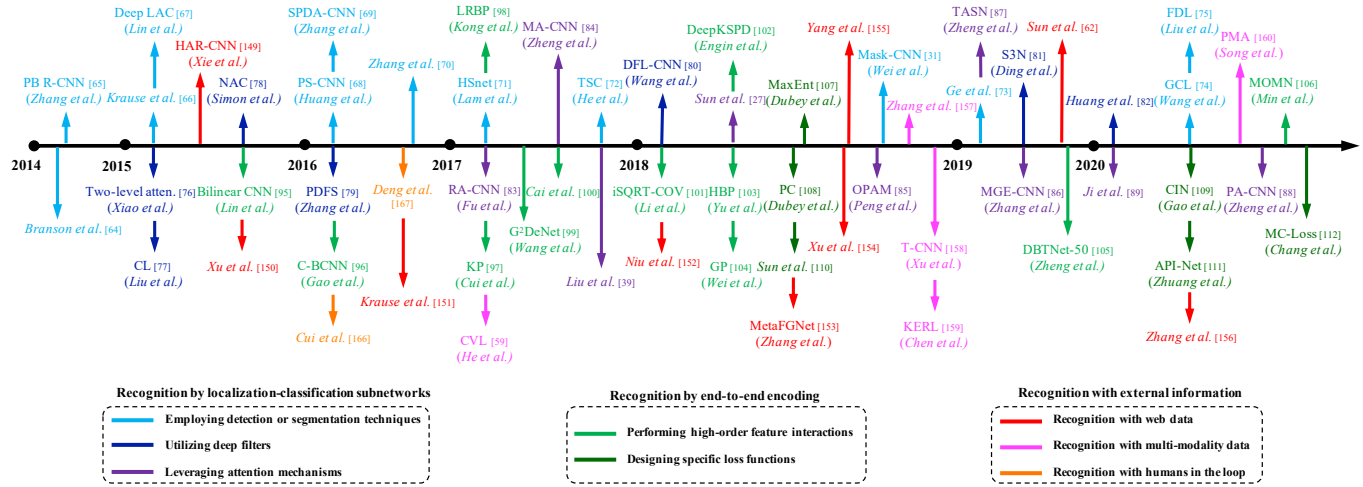
Figure 7. Chronological overview of representative deep learning based fine-grained recognition methods which are categorized by different learning approaches. (Best viewed in color.)
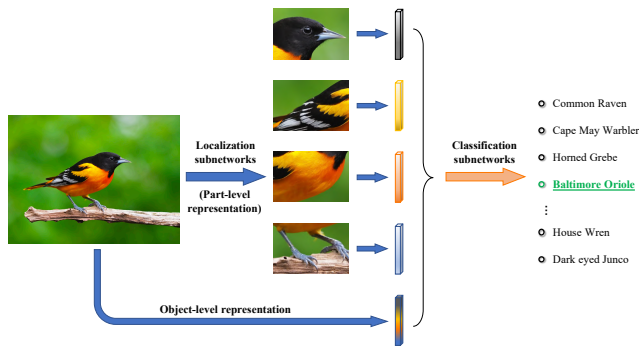

Figure 8. Illustration of the high-level pipeline of the fine-grained recognition by localization-classification subnetworks paradigm.

(global) image features for representing fine-grained objects. This is followed by a classification subnetwork which performs recognition. The framework of such two collaborative subnetworks forms the first paradigm, *i.e.*, fine-grained recognition with *localization-classification subnetworks*. The motivation for these models is to first find the corresponding parts and then compare their appearance. Concretely, it is desirable to capture semantic parts (*e.g.*, heads and torsos) that are shared across fine-grained categories and for discovering the subtle differences between these part representations.

Existing methods in this paradigm can be divided into four broad types: 1) employing detection or segmentation techniques, 2) utilizing deep filters, 3) leveraging attention mechanisms, and 4) other methods.

5.1.1 Employing Detection or Segmentation Techniques

It is straightforward to employ detection or segmentation techniques [116], [117], [118] to locate key image regions corresponding to fine-grained object parts, *e.g.*, bird heads, bird tails, car lights, dog ears, dog torsos, etc. Thanks to localization information, *i.e.*, part-level bounding boxes or segmentation masks, the model can obtain more discriminative mid-level (part-level) representations w.r.t. these parts. Thus, it could further enhance the learning capability of the

classification subnetwork, thus significantly boost the final recognition accuracy.

Earlier works in this paradigm made use of additional dense part annotations (*aka* key point localization, cf. Figure 6 on the left) to locate semantic key parts of objects. For example, Branson *et al.* [64] proposed to use groups of detected part keypoints to compute multiple warped image regions and further obtained the corresponding part-level features by pose normalization. In the same period, Zhang *et al.* [65] first generated part-level bounding boxes based on ground truth part annotations, and then trained a R-CNN [118] model to perform part detection. Di *et al.* [67] further proposed a Valve Linkage Function, which not only connected all subnetworks, but also refined localization according to the *part alignment* results. In order to integrate both semantic part detection and abstraction, SPDA-CNN [69] designed a top-down method to generate *part-level* proposals by inheriting prior geometric constraints and then used a Faster R-CNN [116] to return part localization predictions. Other approaches made use of segmentation information. PS-CNN [68] and Mask-CNN [31] employed segmentation models to get part/object masks to aid part/object localization. Compared with detection techniques, segmentation can result in more accurate part localization [31] as segmentation focuses on the finer pixel-level targets, instead of just coarse bounding boxes.

However, employing traditional detectors or segmentation models requires dense part annotations for training, which is labor-intensive and would limit both scalability and practicality of real-world fine-grained applications. Therefore, it is desirable to accurately locate fine-grained parts by only using image level labels [70], [72], [73], [74], [75]. These set of approaches are referred to as "weakly-supervised" as they only use image level labels. It is interesting to note that since 2016 there is an apparent trend in developing fine-grained methods in this weakly-supervised setting, rather than the strong-supervised setting (*i.e.*, using part annotations and bounding boxes), cf. Table 2.

Recognition methods in the weakly-supervised localization based classification setting always rely on unsupervised approaches to obtain semantic groups which correspond to object parts. Specifically, Zhang *et al.* [70] adopted the

Table 2
Comparative fine-grained recognition results of two learning paradigms (cf. Section 5.1 and Section 5.2) on the fine-grained benchmark datasets, *i.e.*, Birds (*CUB200-2011* [13]), Dogs (*Stanford Dogs* [42]), Cars (*Stanford Cars* [43]), and Aircrafts (*FGVC Aircraft* [44]). Note that, "Train anno." and "Test anno." mean which supervised signals used in the training and test phases, respectively. The symbol "–" means the results are unavailable.

Methods	Published in	Train anno.	Test anno.	Backbones	Img. resolution	Accuracy			
						Birds	*Dogs*	*Cars*	*Aircrafts*
Branson *et al.* [64]	BMVC 2014	BBox+Parts		CaffeNet	224 × 224	75.7%	–	–	–
PB R-CNN [65]	ECCV 2014	BBox+Parts	BBox	Alex-Net	224 × 224	76.4%	–	–	–
Krause *et al.* [66]	CVPR 2015	BBox		CaffeNet	224 × 224	82.0%	–	92.6%	–
Deep LAC [67]	CVPR 2015	BBox+Parts	BBox	Alex-Net	227 × 227	80.3%	–	–	–
PS-CNN [68]	CVPR 2016	BBox+Parts	BBox	CaffeNet	227 × 227	76.6%	–	–	–
SPDA-CNN [69]	CVPR 2016	BBox+Parts	BBox	CaffeNet	Longer side to 800px	81.0%	–	–	–
SPDA-CNN [69]	CVPR 2016	BBox+Parts	BBox	VGG-16	Longer side to 800px	85.1%	–	–	–
Zhang *et al.* [70]	IEEE TIP 2016			Alex-Net	224 × 224	78.9%	80.4%	–	–
HSnet [71]	CVPR 2017	Parts		GoogLeNet	224 × 224	87.5%	–	93.9%	–
TSC [72]	AAAI 2017			VGG-16	Not given	84.7%	–	–	–
Mask-CNN [31]	PR 2018	Parts		VGG-16	448 × 448	85.7%	–	–	–
Ge *et al.* [73]	CVPR 2019			GoogLeNet+BN	Shorter side to 800px	90.3%	93.9%	–	–
GCL [74]	AAAI 2020			ResNet-50+BN	448 × 448	88.3%	–	94.0%	93.2%
FDL [75]	AAAI 2020			ResNet-50	448 × 448	88.6%	85.0%	94.3%	93.4%
Two-level atten. [76]	CVPR 2015			VGG-16	Not given	77.9%	–	–	–
CL [77]	CVPR 2015			Alex-Net	448 × 448	73.5%	–	–	–
NAC [78]	ICCV 2015			VGG-19	Not given	81.0%	–	–	–
PDFS [79]	CVPR 2016			VGG-16	Not given	84.5%	72.0%	–	–
DFL-CNN [80]	CVPR 2018			VGG-16	448 × 448	86.7%	–	93.8%	92.0%
S3N [81]	ICCV 2019			ResNet-50	448 × 448	88.5%	–	94.7%	92.8%
Huang *et al.* [82]	CVPR 2020			ResNet-101	Shorter side to 448px	87.3%	–	–	–
RA-CNN [83]	CVPR 2017			VGG-19	448 × 448	85.3%	87.3%	92.5%	–
MA-CNN [84]	ICCV 2017			VGG-19	448 × 448	86.5%	–	92.8%	89.9%
Liu *et al.* [39]	AAAI 2017	Parts+Attr.		ResNet-50	448 × 448	85.4%	–	–	–
Sun *et al.* [27]	ECCV 2018			ResNet-50	448 × 448	86.5%	84.8%	93.0%	–
OPAM [85]	IEEE TIP 2018			VGG-16	Not given	85.8%	–	92.2%	–
MGE-CNN [86]	ICCV 2019			ResNet-50	448 × 448	88.5%	–	93.9%	–
TASN [87]	CVPR 2019			ResNet-50	224 × 224	87.9%	–	93.8%	–
PA-CNN [88]	IEEE TIP 2020			VGG-19	448 × 448	87.8%	–	93.3%	91.0%
Ji *et al.* [89]	CVPR 2020			ResNet-50	448 × 448	88.1%	–	94.6%	92.4%
STN [90]	NeurIPS 2015			GoogLeNet+BN	448 × 448	84.1%	–	–	–
BoT [91]	CVPR 2016			Alex-Net	Not given	–	–	92.5%	88.4%
NTS-Net [92]	ECCV 2018			ResNet-50	448 × 448	87.5%	–	93.9%	91.4%
M2DRL [93]	IJCV 2019			VGG-16	448 × 448	87.2%	–	93.3%	–
DF-GMM [94]	CVPR 2020			ResNet-50	448 × 448	88.8%	–	94.8%	93.8%
Bilinear CNN [95]	ICCV 2015			VGG-16+VGG-M	448 × 448	84.1%	–	91.3%	84.1%
C-BCNN [96]	CVPR 2016			VGG-16	448 × 448	84.3%	–	91.2%	84.1%
KP [97]	CVPR 2017			VGG-16	224 × 224	86.2%	–	92.4%	86.9%
LRBP [98]	CVPR 2017			VGG-16	224 × 224	84.2%	–	90.0%	87.3%
G^{2}DeNet [99]	CVPR 2017			VGG-16	Longer side to 200px	87.1%	–	92.5%	89.0%
Cai *et al.* [100]	ICCV 2017			VGG-16	448 × 448	85.3%	–	91.7%	88.3%
iSQRT-COV [101]	CVPR 2018			ResNet-101	224 × 224	88.7%	–	93.3%	91.4%
DeepKSPD [102]	ECCV 2018			VGG-16	448 × 448	86.5%	–	93.2%	91.0%
HBP [103]	ECCV 2018			VGG-16	448 × 448	87.1%	–	93.7%	90.3%
GP [104]	ECCV 2018			VGG-16	448 × 448	85.8%	–	92.8%	89.8%
DBTNet-50 [105]	NeurIPS 2019			VGG-16	448 × 448	87.5%	–	94.1%	91.2%
MOMN [106]	IEEE TIP 2020			VGG-16	448 × 448	87.3%	–	92.8%	90.4%
MaxEnt [107]	NeurIPS 2018			VGG-16	224 × 224	77.0%	65.4%	83.9%	78.1%
MaxEnt [107]	NeurIPS 2018			Bilinear CNN	224 × 224	85.3%	83.2%	92.8%	86.1%
PC [108]	ECCV 2018			Bilinear CNN	224 × 224	85.6%	83.0%	92.4%	85.7%
Sun *et al.* [27]	ECCV 2018			ResNet-50	448 × 448	86.5%	84.8%	93.0%	–
CIN [109]	AAAI 2020			ResNet-101	448 × 448	88.1%	–	94.5%	92.8%
Sun *et al.* [110]	AAAI 2020			ResNet-50	448 × 448	88.6%	87.7%	94.9%	93.5%
API-Net [111]	AAAI 2020			ResNet-50	448 × 448	87.7%	88.3%	94.8%	93.0%
API-Net [111]	AAAI 2020			DenseNet-161	448 × 448	90.0%	89.4%	95.3%	93.9%
MC-Loss [112]	IEEE TIP 2020			Bilinear CNN	448 × 448	86.4%	–	94.4%	92.9%
BGL [50]	CVPR 2016			VGG-16	224 × 224	75.9%	–	86.0%	–
BGL [50]	CVPR 2016		BBox	VGG-16	224 × 224	80.4%	–	90.5%	–
DCL [113]	CVPR 2019			ResNet-50	448 × 448	87.8%	–	94.5%	93.0%
Cross-X [114]	ICCV 2019			ResNet-50	448 × 448	87.7%	88.9%	94.6%	92.6%
PMG [115]	ECCV 2020			ResNet-50	448 × 448	89.6%	–	95.1%	93.4%

(Left vertical groupings: "Fine-grained recognition by localization-classification subnetworks" — subgroups: *Employing detection or segmentation techniques*, *Utilizing deep filters*, *Attention mechanisms*, *Others*; "Fine-grained recognition by end-to-end feature encoding" — subgroups: *High-order feature interactions*, *Specific loss functions*, *Others*)

spatial pyramid strategy [119] to generate part proposals from object proposals. Then, by using a clustering approach, they generated part proposal prototype clusters and further selected useful clusters to get discriminative part-level features. Co-segmentation [120] based methods are also commonly used in this weakly supervised case. One approach is to use co-segmentation to obtain object masks without supervision, and then perform heuristic strategies, *e.g.*, part constraints [72] or part alignment [66], to locate fine-grained parts.

It is worth noting that the majority of previous work over-looks the internal semantic correlation among discriminative part-level features. Concretely, the aforementioned methods pick out the discriminative regions independently and utilize their features directly, while neglecting the fact that an object's features are mutually semantic correlated and region groups can be more discriminative. Therefore, very recently, some methods attempt to jointly learn the interdependencies among part-level features to obtain more universal and powerful fine-grained image representations. By performing different feature fusion strategies (*e.g.*, LSTMs [71], [73], graphs [74], or knowledge distilling [75]) these joint part fea-

ture learning methods yield significantly higher recognition accuracy over previous independent part feature learning methods.

5.1.2 Utilizing Deep Filters

In deep convolutional neural networks (CNNs), deep filters (*i.e.*, CNN filters) refer to the learned weights of the convolution layers [14]. The responses/activations from these deep filters can be viewed as localized descriptors. The deep descriptors have the following properties [121]: 1) Locality: they describe and correspond to local image regions w.r.t. the whole input image and 2) Spatiality: they are also able to encode spatial information. As works started exploring the use of CNNs in computer vision, researchers gradually discovered that intermediate CNN outputs (*e.g.*, local deep descriptors) could be linked to semantic parts of common objects [122]. Therefore, the fine-grained community attempted to employ these filter outputs as part detectors [76], [77], [78], [79], [80], [81], [82], and thus rely on them to conduct localization-classification fine-grained recognition. One of the main advantages here is that this does not require any part-level annotations.

Xiao *et al.* [76] performed spectral clustering [123] on these deep filters to form groups and then used the filter groups to serve as part detectors. Similarly, NAC [78] exploits the channels of a CNN as part detectors. Liu *et al.* [77] developed a cross-layer pooling method by aggregating the part-based deep descriptors using activations from two successive convolutional layers as guidance. PDFS [79] was proposed to select deep filters corresponding to parts and then iteratively update the learned "detectors" which resulted in the discovery of discriminative and consistent part-based image regions. For these aforementioned methods, after obtaining detected parts using deep filters from pre-trained classification CNNs, they typically trained off-line classifiers, *e.g.*, SVMs or decision trees [123], using the part-based feature vectors to conduct the final recognition task.

To facilitate the learning of both part detection and part-based classification, unified end-to-end trained fine-grained models [80], [81], [82] were developed. As a result, significant recognition improvements were observed, cf. Table 2. Wang *et al.* [80] utilized an additional learnable 1×1 convolutional filter as a small patch (*i.e.*, part) detector. This was followed by a global max-pooling to keep the highest activations w.r.t. that filter for the final classification. Later, based on class response maps [124], S3N [81] leveraged the class peak responses, *i.e.*, local maximums, as the basis of part localization. Similar to [71], [73], [74], S3N also considers part correlations in a mutual part learning way.

5.1.3 Leveraging Attention Mechanisms

Even though the previous localization-classification fine-grained methods have shown strong classification performance, one of their major drawbacks is that they require meaningful definitions of the object parts. In many applications however, it may be hard to represent or even define common parts of some object classes, *e.g.*, non-structured objects like food dishes [47] or flowers with repeating parts [45]. Compared to these localization-classification methods, a more natural solution of finding parts is to leverage attention mechanisms [125] as sub-modules. This enables CNNs to attend to loosely defined regions for fine-grained objects and as a result have emerged as a promising direction.

It is common knowledge that attention plays an important role in human perception [125], [126]. Humans exploit a sequence of partial glimpses and selectively focus on salient parts of an object or a scene in order to better capture visual structure [127]. Inspired by this, Fu *et al.* and Zheng *et al.* [83], [84] were the first to incorporate attention processing to improve the fine-grained recognition accuracy of CNNs. Specifically, RA-CNN [83] uses a recurrent visual attention model to select a sequence of attention regions (corresponding to object "parts"[1]). RA-CNN iteratively generates region attention maps in a coarse to fine fashion by taking previous predictions as a reference. MA-CNN [84] is equipped with a multi-attention CNN, and can return multiple region attentions in parallel. Subsequently, Peng *et al.* [85] and Zheng *et al.* [88] proposed multi-level attention models to obtain hierarchical attention information (*i.e.*, both object- and part-level). He *et al.* [128] applied multi-level attention to localize multiple discriminative regions simultaneously for each image via an n-pathway end-to-end discriminative localization network that simultaneously localizes discriminative regions and encodes their features. This multi-level attention can result in diverse and complementary information compared to the aforementioned single-level attention methods. Sun *et al.* [27] incorporated channel attentions [129] and metric learning [130] to enforce the correlations among different attended regions. Zheng *et al.* [87] developed a trilinear attention sampling network to learn fine-grained details from hundreds of part proposals and efficiently distill the learned features into a single CNN. Recently, Ji *et al.* [89] presented an attention based convolutional binary neural tree, which incorporates attention mechanisms with a tree structure to facilitate coarse-to-fine hierarchical fine-grained feature learning. Although the attention mechanism achieves strong accuracy in fine-grained recognition, it tends to overfit in the case of small-scale data.

5.1.4 Other Methods

Many other approaches in the localization-classification paradigm have also been proposed for fine-grained recognition. Spatial Transformer Networks (STN) [90] were originally introduced to explicitly perform spatial transformations in an end-to-end learnable way. They can also be equipped with multiple transformers in parallel to conduct fine-grained recognition. Each transformer in an STN can correspond to a part detector with spatial transformation capabilities. Later, Wang *et al.* [91] developed a triplet of patches with geometric constraints as a template to automatically mine discriminative triplets and then generated mid-level representations for classification with the mined triplets. In addition, other methods have achieved better accuracy by introducing feedback mechanisms. Specifically, NTS-Net [92] employs a multi-agent cooperative learning scheme to address the core problem of fine-grained recognition, *i.e.*, accurately identifying informative regions in an image. M2DRL [93], [131] was the first to utilize deep reinforcement learning [132]

1. Note that here "parts" refers to the loosely defined attention regions for fine-grained objects, which is different from the clearly defined object parts from manual annotations, cf. Section 5.1.1.

at both the object- and part-level to capture multi-granularity discriminative localization and multi-scale representations using their tailored reward functions. Inspired by low-rank mechanisms in natural language processing [133], Wang *et al.* [94] proposed the DF-GMM framework to alleviate the region diffusion problem in high-level feature maps for fine-grained part localization. DF-GMM first selects discriminative regions from the high-level feature maps by constructing low-rank bases, and then applies spatial information of the low-rank bases to reconstruct low-rank feature maps. Part correlations can also be modeled by reorganization processing, which brings accuracy improvements.

Figure 9. Illustration of feature maps and deep descriptors in CNNs.

5.2 Recognition by End-to-End Feature Encoding

The second learning paradigm of fine-grained recognition is *end-to-end feature encoding*. As with other vision tasks, feature learning also plays a fundamental role in fine-grained recognition. Since the differences between sub-categories are typically very subtle and local, capturing global semantic information using only fully connected layers limits the representation capacity of a fine-grained model, and hence restricts further improvements in final recognition performance. Therefore, methods have been developed that aim to learn a unified, yet discriminative, image representation for modeling subtle differences between fine-grained categories in the following ways: 1) by performing high-order feature interactions, 2) by designing novel loss functions, and 3) through other means.

5.2.1 Performing High-Order Feature Interactions

Feature learning plays a crucial role in almost all vision tasks such as retrieval, detection, tracking, etc. The success of deep convolutional networks is mainly due to the learned discriminative deep features. In the initial era of deep learning, the features (*i.e.*, activations) of fully connected layers were commonly used as image representations. Later, the feature maps of deeper convolutional layers were discovered to contain mid- and high-level information, *e.g.*, object parts or complete objects [134], which led to the widespread use of convolutional features/descriptors [77], [135], cf. Figure 9. Additionally, applying encoding techniques for these local convolutional descriptors has resulted in significant improvements compared with fully-connected outputs [135], [136], [137]. To some extent, these improvements in encoding techniques come from the *higher-order* statistics encoded in the final features. Particularly for fine-grained recognition, where the need for end-to-end modeling of higher-order statistics became evident when the Fisher Vector encodings of SIFT features outperformed a fine-tuned AlexNet in several fine-grained tasks [138].

The covariance matrix based representation [139], [140] is a representative higher-order (*i.e.*, second-order) feature interaction technique, which has been used in computer vision and machine learning. Let $\boldsymbol{V}_{d \times n} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$ denote a data matrix, in which each column contains a local descriptor $\boldsymbol{v}_i \in \mathcal{R}^d$, extracted from an image. The corresponding $d \times d$ sample covariance matrix over $\boldsymbol{V}$ is denoted as $\boldsymbol{\Sigma} = \bar{\boldsymbol{V}}\bar{\boldsymbol{V}}^\top$ (or simply $\boldsymbol{V}\boldsymbol{V}^\top$), where $\bar{\boldsymbol{V}}$ denotes the centered $\boldsymbol{V}$. Originally, this covariance matrix is proposed as a regio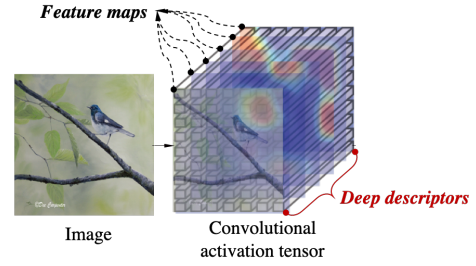n descriptor, *e.g.*, characterizing the covariance of the color intensities of pixels in an image patch. In recent years, it has been used as a promising second-order pooled image representation for visual recognition.

By integrating the covariance matrix based representation with deep descriptors, a series of methods showed promising accuracy in fine-grained recognition in the past few years. The most representative method among them is Bilinear CNNs [95], [141], which represents an image as a pooled outer product of features derived from two deep CNNs, and thus encodes second-order statistics of convolutional activations, resulting in clear improvements in fine-grained recognition. This outer product essentially leads to a covariance matrix (in the form of $\boldsymbol{V}\boldsymbol{V}^\top$) when the two CNNs are set as the same. However, the outer product operation results in extremely high dimensional features, *i.e.*, the bilinear pooled feature is reshaped into a vector $\boldsymbol{z} = \text{vec}(\boldsymbol{V}\boldsymbol{V}^\top) \in \mathcal{R}^{d^2}$. This results in a large increase in the number of parameters in the classification module of the deep network, which can cause overfitting and make it impractical for realistic applications, especially for large-scale ones. To address this problem, Gao *et al.* [96] applied Tensor Sketch [142] to both approximate the second-order statistics of the original bilinear pooling operation and reduce feature dimensions. Kong *et al.* [98] adopted a low-rank approximation to the covariance matrix and further learned a low-rank bilinear classifier. The resulting classifier can be evaluated without explicitly computing the bilinear feature matrix which results in a large reduction on the parameter size. Li *et al.* [143] also modeled pairwise feature interaction by performing a quadratic transformation with a low-rank constraint. Yu *et al.* [103] used a dimension reduction projection before bilinear pooling to alleviate dimension explosion. Zheng *et al.* [105] applied bilinear pooling to feature channel groups where the bilinear transformation is represented by calculating pairwise interactions within each group. This also results in large saving in computation cost.

Beyond these approaches, some methods attempt to capture much higher-order (more than second-order) interactions of features to generate stronger and more discriminative feature representations. Cui *et al.* [97] introduced a kernel pooling method that captures arbitrarily ordered and non-linear features via compact feature mapping. Cai *et al.* [100] proposed a polynomial kernel based predictor to model higher-order statistics of convolutional activations across multiple layers for modeling part interactions. Subsequently, DeepKSPD [102] was developed to jointly learn the deep local descriptors and the kernel-matrix based covariance representation in an end-to-end trainable manner.

As ℓ_2 feature normalization can suppress the common

patterns of high response and thereby enhance those discriminative features (*i.e.*, the visual burstiness problem [104], [144]), the aforementioned bilinear pooling based methods typically perform element-wise square root normalization followed by ℓ_2-normalization on covariance matrix to improve performance. However, merely employing ℓ_2-normalization can cause unstable high-order information and also lead to slow convergence. To this end, many methods have explored non-linearly scaling based on the singular value decomposition (SVD) or eigendecomposition (EIG) to obtain more stability for second-order representations. Specifically, Li *et al.* [145] proposed to apply the power exponent to the eigenvalues of bilinear features to achieve better recognition accuracy. G^2DeNet [99] further combined complementary first-order and second-order information via a Gaussian embedding and matrix square root normalization. iSQRT-COV [101] and the improved B-CNN [146] used the Newton-Schulz iteration to approximate matrix square-root normalization with only matrix multiplication to decrease training time. Recently, MOMN [106] was proposed to simultaneously normalize a bilinear representation in terms of square-root, low-rank, and sparsity all within a multi-objective optimization framework.

5.2.2 Designing Specific Loss Functions

Loss functions play an important role in the construction of deep networks. They can directly affect both the learned classifiers and features. Thus, designing fine-grained tailored loss functions is an important direction for fine-grained image recognition.

Distinct from generic image recognition, in fine-grained classification, where samples that belong to different classes can be visually very similar, it is reasonable to prevent the classifier from being too confident in its outputs (*i.e.*, discourage low entropy). Following this intuition, [107] also maximized the entropy of the output probability distribution when training networks on fine-grained tasks. Similarly, Dubey *et al.* [108] used a pairwise confusion optimization procedure to solve both overfitting and sample-specific artifacts in fine-grained recognition by bringing the different class-conditional probability distributions closer together and confusing the deep network. This allows a reduction in prediction over-confidence, therefore resulting in improved generalization performance.

Humans can effectively identify contrastive clues by comparing image pairs, and this type of metric / contrastive learning is also common in fine-grained recognition. Specifically, Sun *et al.* [27] first learned multiple part-corresponding attention regions and then leveraged metric learning for pulling same-attention same-class features closer, while pushing different-attention or different-class features away. Furthermore, their approach can coherently enforce the correlations among different object parts during training. CIN [109] pulls positive pairs closer while pushing negative pairs away via a contrastive channel interaction module which also exploits channel correlations between samples. API-Net [111] was also built upon a metric learning framework, and can adaptively discover contrastive cues from a pair of images and distinguish them via pairwise attention based interactions.

Designing a single loss function for localizing part-level patterns and further aggregating image-level representations

has also been explored in the literature. Specifically, Sun *et al.* [110] developed a gradient-boosting based loss function along with a diversification block to force the network to move swiftly to discriminate the hard classes. Concretely, the diversification block suppresses the discriminative regions of the class activation maps, and hence the network is forced to find alternative informative features. The gradient-booting loss focuses on difficult (*i.e.*, confusing) classes for each image and boosts their gradient. MC-Loss [112] encourages the feature channels to be more discriminative by focusing on various part-level regions. They propose a single loss that does not require any specific network modifications for partial localization of fine-grained objects.

These aforementioned loss function based fine-grained recognition methods are backbone-agnostic and their performance can typically be improved by using more powerful backbone network architectures.

5.2.3 Other Methods

Beyond modelling the interactions between higher-order features and designing novel loss functions, another set of approaches involve constructing fine-grained tailored auxiliary tasks for obtaining unified and discriminative image representations.

BGL [50] was proposed to incorporate rich bipartite-graph labels into CNN training to model the important relationships among fine-grained classes. DCL [113] performed a "destruction and construction" process to enhance the difficulty of recognition to guide the network to focus on discriminative parts for fine-grained recognition (*i.e.*, by destruction learning) and then model the semantic correlation among parts of the object (*i.e.*, by construction learning). Similar to DCL, Du *et al.* [115] tackled fine-grained representation learning using a jigsaw puzzle generator proxy task to encourage the network to learn at different levels of granularity and simultaneously fuse features at these levels together. Recently, a more direct fine-grained feature learning method [147] was formulated with the goal of generating *identity-preserved* fine-grained images in an adversarial learning manner to directly obtain a unified fine-grained image representation. The authors showed that this direct feature learning approach not only preserved the identity of the generated images, but also significantly boosted the visual recognition performance in other challenging tasks like fine-grained few-shot learning [148].

5.3 Recognition with External Information

Beyond the conventional recognition paradigms, which are restricted to using supervision associated with the images themselves, another paradigm is to leverage external information, *e.g.*, web data, multi-modal data, or human-computer interactions, to further assist fine-grained recognition.

5.3.1 Noisy Web Data

Large and well-labeled training datasets are necessary in order to identify subtle differences between various fine-grained categories. However, acquiring accurate human labels for fine-grained categories is difficult due to the need for domain expertise and the myriads of fine-grained categories (*e.g.*, potentially more than tens of thousands of subordinate

Table 3

Comparison of fine-grained "recognition with external information" (cf. Section 5.3) on multiple fine-grained benchmark datasets, including Birds (*CUB200-2011* [13]), Dogs (*Stanford Dogs* [42]), Cars (*Stanford Cars* [43]), and Aircrafts (*FGVC Aircraft* [44]). "External info." denotes which kind of external information is used by the respective approach. "Train anno." and "Test anno." indicate the supervision used during training and testing, and "–" means the results are unavailable.

		Methods	Published in	Train anno.	Test anno.	External info.	Backbones	Img. resolution	Accuracy			
									Birds	*Dogs*	*Cars*	*Aircrafts*
Fine-grained recognition with external information	With web / auxiliary data	HAR-CNN [149]	CVPR 2015	BBox	BBox	Web data	Alex-Net	224×224	–	49.4%	80.8%	–
		Xu *et al.* [150]	ICCV 2015	BBox+Parts		Web data	CaffeNet	224×224	84.6%	–	–	–
		Krause *et al.* [151]	ECCV 2016			Web data	Inception-v3	224×224	92.3%	80.8%	–	93.4%
		Niu *et al.* [152]	CVPR 2018			Web data	VGG-16	224×224	76.5%	85.2%	–	–
		MetaFGNet [153]	ECCV 2018			Auxiliary data	ResNet-34	224×224	87.6%	96.7%	–	–
		Xu *et al.* [154]	IEEE TPAMI 2018	BBox+Parts		Web data	Alex-Net	224×224	84.6%	–	–	–
		Yang *et al.* [155]	IEEE TIP 2018			Web data	ResNet-50	224×224	–	87.4%	–	–
		Sun *et al.* [62]	AAAI 2019			Web data	ResNet-50	224×224	–	87.1%	–	–
		Zhang *et al.* [156]	AAAI 2020			Web data	VGG-16	224×224	77.2%	–	78.7%	72.9%
	With multi-modal data	CVL [59]	CVPR 2017			Language texts	VGG-16	Not given	85.6%	–	–	–
		Zhang *et al.* [157]	AAAI 2018	BBox		Audio	VGG-16	227×227	85.6%	–	–	–
		Zhang *et al.* [157]	AAAI 2018	BBox	BBox	Audio	VGG-16	227×227	86.6%	–	–	–
		T-CNN [158]	IJCAI 2018	BBox		Knowledge base + Texts	ResNet-50	224×224	86.5%	–	–	–
		KERL [159]	IJCAI 2018			Knowledge base	VGG-16	224×224	87.0%	–	–	–
		PMA [160]	IEEE TIP 2020			Language texts	VGG-16	448×448	88.2%	–	–	–
		PMA [160]	IEEE TIP 2020			Language texts	ResNet-50	448×448	88.7%	–	–	–

categories in a meta-category). As a result, some fine-grained recognition methods seek to utilize freely available, but noisy, web data to boost recognition performance. The majority of existing work in this line can be roughly grouped into two directions.

The first direction involves scraping noisy labeled web data for the categories of interest as training data, which is regarded as *webly supervised learning* [62], [154], [161]. These approaches typically concentrate on: 1) overcoming the domain gap between easily acquired web images and the well-labeled data from standard datasets; and 2) reducing the negative effects caused by the noisy data. For instance, HAR-CNN [149] utilized easily annotated meta-classes inherent in the fine-grained data and also acquired a large number of meta-class-labeled images from the web to regularize the models for improving recognition accuracy in a multi-task manner (*i.e.*, for both the fine-grained and the meta-class data recognition task). Xu *et al.* [150] investigated if fine-grained web images could provide weakly-labeled information to augment deep features and thus contribute to robust object classifiers by building a multi-instance (MI) learner, *i.e.*, treating the image as the MI bag and the proposal part bounding boxes as the instances of MI. Krause *et al.* [151] introduced an alternative approach to combine a generic classification model with web data by excluding images that appear in search results for more than one category to combat cross-domain noise. Inspired by adversarial learning [162], [62] proposed an adversarial discriminative loss to encourage representation coherence between standard and web data.

The second direction is to transfer the knowledge from auxiliary categories with well-labeled training data to the test categories, which usually employs zero-shot learning [63] or meta learning [163]. Niu *et al.* [63] exploited zero-shot learning to transfer knowledge from annotated fine-grained categories to other fine-grained categories. Subsequently, Zhang *et al.* [153], Yang *et al.* [155], and Zhang *et al.* [156] investigated different approaches for selecting high-quality web training images to expand the training set. Zhang *et al.* [153] proposed a novel regularized meta-learning objective to guide the learning of network parameters so they are optimal for adapting to the target fine-grained categories. Yang *et al.* [155] designed an iterative method that progressively
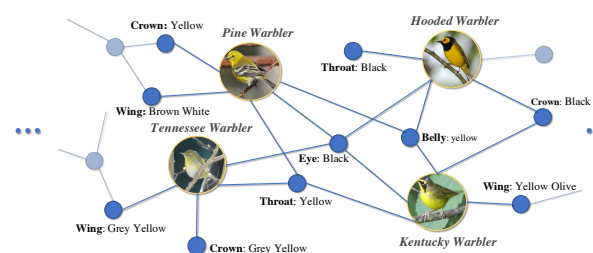

Figure 10. An example knowledge graph for modeling category-attribute correlations in *CUB200-2011* [13].

selects useful images by modifying the label assignment using multiple labels to lessen the impact of the labels from the noisy web data. Zhang *et al.* [156] leveraged the prediction scores in different training epochs to supervise the separation of useful and irrelevant noisy web images.

5.3.2 Multi-Modal Data

Multi-modal analysis has attracted a lot of attention with the rapid growth of multi-media data, *e.g.*, image, text, knowledge bases, etc. In fine-grained recognition, multi-modal data can be used to establish joint-representations/embeddings by incorporating multi-modal data in order to boost fine-grained recognition accuracy. Compared with strong semantic supervision from fine-grained images (*e.g.*, part annotations), text descriptions are a weak form of supervision (*i.e.*, they only provide image-level supervision). One advantage however, is that text descriptions can be relatively accurately generated by non-experts. Thus, they are both easy and cheap to be collected. In addition, high-level knowledge graphs, when available, can contain rich knowledge (*e.g.*, *DBpedia* [164]). In practice, both text descriptions and knowledge bases are useful extra guidance for advancing fine-grained image representation learning.

Reed *et al.* [58] collected text descriptions, and introduced a structured joint embedding for zero-shot fine-grained image recognition by combining text and images. Later, He and Peng [59] combined vision and language bi-streams in a joint end-to-end fashion to preserve the intra-modality and inter-modality information for generating complementary fine-grained representations. Later, PMA [160] proposed a

mask-based self-attention mechanism to capture the most discriminative parts in the visual modality. In addition, they explored using out-of-visual-domain knowledge using language with query-relational attention. Multiple PMA blocks for the vision and language modalities were aggregated and stacked using the proposed progressive mask strategy. For fine-grained recognition with knowledge bases, some existing works [158], [159] have introduced knowledge base information (using attribute label associations, cf. Figure 10) to implicitly enrich the embedding space, while also reasoning about the discriminative attributes of fine-grained objects. Concretely, T-CNN [158] explored using semantic embeddings from knowledge bases and text, and then trained a CNN to linearly map image features to the semantic embedding space to aggregate multi-modal information. To incorporate the knowledge representation into image features, KERL [159] employed a gated graph network to propagate node messages through the graph to generate the knowledge representation. Finally, [157] incorporated audio information related to the fine-grained visual categories of interest to boost recognition accuracy.

5.3.3 Humans-in-the-Loop

Human-in-the-loop methods [165] combine the complementary strengths of human knowledge with computer vision algorithms. Fine-grained recognition with humans-in-the-loop is typically posed in an iterative fashion and requires the vision system to be intelligent about when it queries the human for assistance. Generally, for these kinds of recognition methods, in each round the system is seeking to understand how humans perform recognition, *e.g.*, by asking expert humans to label the image class [166], or by identifying key part localization and selecting discriminative features [167] for fine-grained recognition.

5.4 Summary and Discussion

The *CUB200-2011* [13], *Stanford Dogs* [42], *Stanford Cars* [43], and *FGVC Aircraft* [44] benchmarks are among the most influential datasets in fine-grained recognition. Tables 2 and 3 summarize results achieved by the fine-grained methods belonging to three recognition learning paradigms outlined above, *i.e.*, "recognition by localization-classification subnetworks", "recognition by end-to-end feature encoding", and "recognition with external information". A chronological overview can be seen in Figure 7. The main observations can be summarized as follows:

- There is an explicit correspondence between the reviewed methods and the aforementioned challenges of fine-grained recognition. Specifically, the challenge of capturing subtle visual differences can be overcome by localization-classification methods (cf. Section 5.1) or via specific construction-based tasks [113], [115], [147], as well as human-in-the-loop methods. The challenge of characterizing fine-grained tailored features is alleviated by performing high-order feature interactions or by leveraging multi-modality data. Finally, the challenging nature of FGIA can be somewhat addressed by designing specific loss functions [107], [108], [110] for achieving better accuracy.

Table 4
Comparative fine-grained recognition results on *CUB200-2011* using different input image resolutions. The results in this table are conducted based on a vanilla ResNet-50 trained at the respective resolution.

Resolution	224×224	280×280	336×336	392×392
Accuracy	81.6%	83.3%	85.0%	85.6%

- Among the different learning paradigms, the "recognition by localization-classification subnetworks" and "recognition by end-to-end feature encoding" paradigms are the two most frequently investigated ones.
- Part-level reasoning of fine-grained object categories boosts fine-grained recognition accuracy especially for non-rigid objects, *e.g.*, birds. Modeling the internal semantic interactions/correlations among discriminative parts has attracted increased attention in recent years, cf. [27], [71], [73], [74], [75], [81], [94], [109].
- Non-rigid fine-grained object recognition (*e.g.*, birds or dogs) is more challenging than rigid fine-grained objects (*e.g.*, cars or aircrafts), which is partly due to the variation on object appearance.
- Fine-grained image recognition performance improves as image resolution increases [168]. Comparison results on *CUB200-2011* of different image resolutions are reported in Table 4.
- There is a trade-off between recognition and localization ability for the "recognition by localization-classification subnetworks" paradigm, which might impact a single integrated network's recognition accuracy. Such a trade-off is also reflected in practice when trying to achieve better recognition results, in that training usually involves alternating optimization of the two networks or separately training the two followed by joint tuning. Alternating or multistage strategies complicate the tuning of the integrated network.
- While effective, most end-to-end encoding networks are less human-interpretable and less consistent in their accuracy across non-rigid and rigid visual domains compared to localization-classification subnetworks. Recently, it has been observed that several higher-order pooling methods attempt to understand such kind of methods by presenting visual interpretation [141] or from an optimization perspective [169].
- "Recognition by localization-classification subnetworks" based methods are challenging to apply when the fine-grained parts are not consistent across the meta-categories (*e.g.*, *iNaturalist* [2]). Here, unified end-to-end feature encoding methods are more appropriate.

6 FINE-GRAINED IMAGE RETRIEVAL

Fine-grained retrieval is another fundamental aspect of FGIA that has gained more traction in recent years. What distinguishes fine-grained retrieval from fine-grained recognition is that in addition to estimating the sub-category correctly, it is also necessary to rank all the instances so that images belonging to the same sub-category are ranked highest based on the fine-grained details in the query. Specifically, in fine-grained retrieval we are given a database of images of the same meta-category (*e.g.*, birds or cars) and a query, and the goal is to return images related to the query based on relevant

fine-grained features. Compared to generic image retrieval, which focuses on retrieving near-duplicate images based on similarities in their content (*e.g.*, texture, color, and shapes), fine-grained retrieval focuses on retrieving the images of the same category type (*e.g.*, the same subordinate species of animal or the same model of vehicle). What makes it more challenging is that objects of fine-grained categories have exhibit subtle differences, and can vary in pose, scale, and orientation or can contain large cross-modal differences (*e.g.*, in the case of sketch-based retrieval).

Fine-grained retrieval techniques have been widely used in commercial applications, *e.g.*, e-commerce (searching fine-grained products [172]), touch-screen devices (searching fine-grained objects by sketches [53]), crime prevention (searching face photos [173]), among others. Depending on the type of query image, the most studied areas of fine-grained image retrieval can be separated into two groups: fine-grained content-based image retrieval (FG-CBIR, cf. Figure 11) and fine-grained sketch-based image retrieval (FG-SBIR, cf. Figure 12). Fine-grained image retrieval can also be expanded into fine-grained cross-media retrieval [57], which can utilize one media type to retrieve any media types, for example using an image to retrieve relevant text, video, or audio.

For performance evaluation, following the standard convention, FG-CBIR performance is typically measured using *Recall@K* [174] which is the average recall score over all M query images in the test set. For each query, the top K relevant images are returned. The recall score will be 1 if there are at least one positive image in the top K returned images, and 0 otherwise. By formulation, the definition of *Recall@K* is as follows

$$Recall@K = \frac{1}{M} \sum_{i=1}^{M} \text{score}_i \,. \tag{7}$$

For measuring FG-SBIR performance, *Accuracy@K* is commonly used, which is the percentage of sketches whose true-match photos are ranked in the top K:

$$Accuracy@K = \frac{|I_{\text{correct}}^K|}{K} \,, \tag{8}$$

where $|I_{\text{correct}}^K|$ is the number of true-match photos in top K.

6.1 Content-based Fine-Grained Image Retrieval

SCDA [32] is one of the earliest examples of fine-grained image retrieval that used deep learning. It employs a pre-trained CNN to select meaningful deep descriptors by localizing the main object in an image without using explicit localization supervision. Unsurprisingly, they show that selecting only useful deep descriptors, by removing background features, can significantly benefit retrieval performance in such an unsupervised retrieval setting (*i.e.*, requiring no image labels). Recently, supervised metric learning based approaches (*e.g.*, [28], [170]) have been proposed to overcome the retrieval accuracy limitations of unsupervised retrieval. These methods still include additional sub-modules specifically tailored for fine-grained objects, *e.g.*, the weakly-supervised localization module proposed in [170], which is in turn inspired by [32]. CRL-WSL [170] employed a centralized ranking loss with a weakly-supervised localization approach

to train their feature extractor. DGCRL [28] eliminated the gap between inner-product and the Euclidean distance in the training and test stages by adding a Normalize-Scale layer to enhance the intra-class separability and inter-class compactness with their Decorrelated Global-aware Centralized Ranking Loss. Recently, the Piecewise Cross Entropy loss [171] was proposed by modifying the traditional cross entropy function by reducing the confidence of the fine-grained model, which is similar to the basic idea of following the natural prediction confidence scores for fine-grained categories [107], [108].

The performance of recent fine-grained content-based image retrieval approaches are reported in Table 5. Although supervised metric learning based retrieval methods outperform their unsupervised counterparts, the absolute recall scores (*i.e.*, *Recall@K*) of the retrieval task still has room for further improvement. One promising direction is to integrate advanced techniques, *e.g.*, attention mechanisms or higher-order feature interactions, which are successful for fine-grained recognition into FG-CBIR to achieve better retrieval accuracy. However, new large-scale FG-CBIR datasets are required to drive future progress, which are also desirable to be associated with other characteristics or challenges, *e.g.*, open-world sub-category retrieval (cf. Section 2).

6.2 Sketch-based Fine-Grained Image Retrieval

The goal of fine-grained sketch-based image retrieval (FG-SBIR) is to match specific photo instances using a free-hand sketch as the query modality. Distinct from the previously discussed content-based approach, the additional sketch-photo domain gap lies at the centre of FG-SBIR. Thus, the key is introducing a cross-modal representation that not only captures fine-grained characteristics present in the sketches, but also possesses the ability to traverse the sketch and image domain gap.

Existing FG-SBIR solutions generally aim to train a joint embedding space where sketches and photos can be compared in a nearest neighbor fashion. Specifically, Li *et al.* [51] proposed the first method for FG-SBIR, where they learned a deformable part-based models (DPM) [181] as a mid-level representation to discover and encode the various poses in the sketch and image domains independently. This was followed by a graph matching operation on the DPM to establish pose correspondences across the two domains. Yu *et al.* [52] first employed deep learning for this task, leveraging a triplet ranking model with a staged pre-training strategy to learn a joint embedding space for the two domains. This triplet ranking model was further augmented with auxiliary semantic attribute prediction and attribute ranking layers [175] to improve generalization [52]. The implicit challenge encountered due to the large sketch-photo domain gap in FG-SBIR was tackled via a discriminative-generative hybrid model [182] using cross-domain translation. Inspired on these approaches, Song *et al.* [54] introduced an attention module and encoded the spatial position of visual features and combined both coarse and fine-grained semantic knowledge. Li *et al.* [55] introduced a part-aware learning approach to reduce the instance-level domain-gap by performing subspace alignment with fine-grained attributes. While research on fine-grained SBIR has flourished
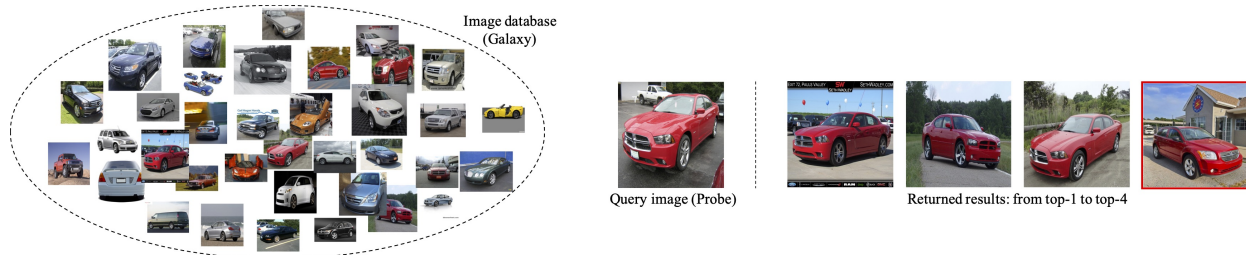
Figure 11. An illustration of fine-grained content-based image retrieval (FG-CBIR). Given a query image (*aka* probe) depicting a "`Dodge Charger Sedan 2012`", fine-grained retrieval is required to return images of the same car model from a car database (*aka* galaxy). In this figure, the fourth returned image, marked with a red outline, is incorrect as it is a different car model, it is a "`Dodge Caliber Wagon 2012`".

Table 5
Comparison of recent fine-grained content-based image retrieval methods on *CUB200-2011* [13] and *Stanford Cars* [43]. *Recall@K* is the average recall over all query images in the test set.

Methods	Published in	Supervised	Backbones	Img. Resolution	Recall@K							
					Birds				*Cars*			
					1	2	4	8	1	2	4	8
SCDA [32]	TIP 2017		VGG-16	224×224	62.2%	74.2%	83.2%	90.1%	58.5%	69.8%	79.1%	86.2%
CRL-WSL [170]	IJCAI 2018	✓	VGG-16	224×224	65.9%	76.5%	85.3%	90.3%	63.9%	73.7%	82.1%	89.2%
DGCRL [28]	AAAI 2019	✓	ResNet-50	Not given	67.9%	79.1%	86.2%	91.8%	75.9%	83.9%	89.7%	94.0%
Zeng *et al.* [171]	Image and Vis. Comp. 2020	✓	ResNet-50	224×224	70.1%	79.8%	86.9%	92.0%	86.7%	91.7%	95.2%	97.0%

Table 6
Comparison of fine-grained sketch-based image retrieval methods on *QMUL-Shoe* [52], *QMUL-Chair* [52], *QMUL-Handbag* [54], *Sketchy* [53], and *QMUL-Shoe-V2* [56]. *Accuracy@K* is the percentage of sketches whose true-match photos are ranked in the top K.

| Methods	Published in	Accuracy@K									
		QMUL-Shoe		*QMUL-Chair*		*QMUL-Handbag*		*Sketchy*		*QMUL-Shoe-V2*	
		1	10	1	10	1	10	1	10	1	10
Yu *et al.* [52]	CVPR 2016	39.1%	87.8%	69.1%	97.9%	–	–	–	–	–	–
Song *et al.* [175]	BMVC 2016	50.5%	91.3%	78.3%	98.9%	–	–	–	–	–	–
Song *et al.* [54]	ICCV 2017	61.7%	94.8%	81.4%	95.9%	49.4%	82.7%	–	–	–	–
Li *et al.* [55]	IEEE TIP 2017	51.3%	90.4%	76.3%	97.9%	–	–	45.3%	98.2%	–	–
Radenovic *et al.* [176]	ECCV 2018	54.8%	92.2%	85.7%	97.9%	51.2%	85.7%	–	–	–	–
Zhang *et al.* [177]	ECCV 2018	35.7%	84.3%	67.1%	99.0%	–	–	–	–	–	–
Pang *et al.* [178]	CVPR 2020	56.5%	–	85.9%	–	62.9%	–	–	–	36.5%	–
Bhunia *et al.* [179]	CVPR 2020	–	–	–	–	–	–	–	–	–	79.6%
Sain *et al.* [180]	BMVC 2020	–	–	–	–	–	–	–	–	36.3%	80.6%
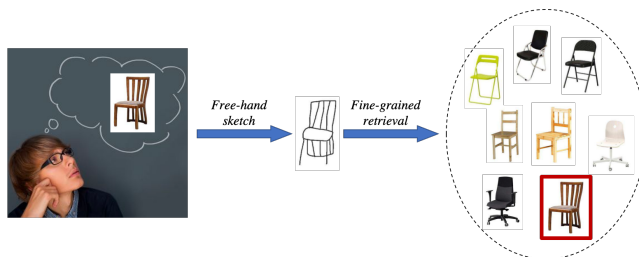
Figure 12. An illustration of fine-grained sketch-based image retrieval (FG-SBIR), where a free-hand human sketch serves as the query for instance-level retrieval of images. FG-SBIR is challenging due to 1) the fine-grained and cross-domain nature of the task and 2) free-hand sketches are highly abstract, making fine-grained matching even more difficult.

over the years, a dilemma still remains – whether sketching is a better input modality for fine-grained image retrieval compared to other alternatives, *e.g.*, texts or attributes. To answer this question, Song *et al.* [183] showed the superiority of sketch over text for fine-grained retrieval, illustrating that each modality can complement the other when they are jointly modeled.

In contrast to these earlier works that were mostly based on Siamese-triplet networks, there have been several recent attempts to advance FG-SBIR performance. For instance, Pang *et al.* [56] identified that a baseline triplet based

model fails to generalize when exposed to unseen categories. Thus, cross-category generalization was introduced through a domain generalization setup, where a universal manifold of prototypical visual sketches was modeled to dynamically represent the sketch/photo. On the other hand, Bhunia *et al.* [179] developed an on-the-fly retrieval setup via reinforcement learning that begins retrieving photos as soon as the user starts drawing. ImageNet pre-trained weights have been considered as the de-facto choice for initializing sketch/photo embedding networks for FG-SBIR. However, following the recent progress in self-supervised learning [184], [185], [186], Pang *et al.* [178] performed a jigsaw solving strategy over mixed-modal patches between a particular photo and its edge-map. They showed this to be an effective pre-text task for self-supervised pre-training strategy and it improves FG-SBIR performance. Instead of performing independent sketch and photo embeddings, as in almost all previous works, Sain *et al.* [180] used paired-embeddings by employing cross-modal co-attention and hierarchical stroke/region-wise feature fusion in order to deal with varying levels of sketch detail. In spite of a significant performance boost, it is noteworthy that paired-embeddings incur a significant computational overhead, by a multiple of at least the number of gallery photos, compared to other state-of-the-art methods. Recently, a new scene-level fine-grained SBIR task [187] was explored. Liu *et al.* [187]

proposed to utilize local features such as object instances and their visual detail, as well as global context (*e.g.*, the scene layout), to deal with such a task.

In addition to the previous FG-SBIR methods, Zhang *et al.* [177] and Radenovic *et al.* [176] also evaluated on popular FG-SBIR datasets, in addition to category-level SBIR ones. We compare the results of recent FG-SBIR methods in Table 6. Note that we only present methods whose evaluation strategies are uniform and consistent, and thus exclude those that involve problem specific evaluation protocols, such as zero-shot FG-SBIR [56].

7 COMMON TECHNIQUES SHARED BY BOTH FINE-GRAINED RECOGNITION AND RETRIEVAL

The tasks of fine-grained image recognition and retrieval are complementary. As a result, there exists common techniques shared by both problems in the FGIA literature. In this section, we discuss these common techniques, in terms of methods and basic ideas, with the aim of motivating further work in these areas.

Common Methods: In the context of deep learning, both fine-grained recognition and retrieval tasks require discriminative deep feature embeddings to distinguish subtle differences between fine-grained objects. While recognition aims to distinguish category labels, retrieval aims to return an accurate ranking. To achieve these goals, deep metric learning and multi-modal matching can be viewed as two common sets of techniques that are applicable for both fine-grained recognition and retrieval.

Specifically, deep metric learning [188] attempts to map image data to an embedding space, where similar fine-grained images are close together and dissimilar images are far apart. In general, fine-grained recognition realizes metric learning by classification losses, where it includes a weight matrix to transform the embedding space into a vector of fine-grained class logits, *e.g.*, [27], [109], [111]. While metric learning on fine-grained retrieval tasks (most cases without explicit image labels) is always achieved by means of embedding losses which operate on the relationships between fine-grained samples in a batch, *e.g.*, [28], [170].

Recently, multi-modal matching methods [29], [30] have emerged as powerful representation learning approaches to simultaneously boost fine-grained recognition and retrieval tasks. In particularly, Mafla *et al.* [29] leveraged textual information along with visual cues to comprehend the existing intrinsic relation between the two modalities. [30] employed a graph convolutional network to perform multi-modal reasoning and obtain relationship-enhanced multi-modal features. Such kind of methods reveal a new development trend of multi-modal learning in FGIA.

Common Basic Ideas: Beyond these common methods, many basic ideas are shared by both fine-grained recognition and retrieval, *e.g.*, selecting useful deep descriptors [31], [32], reducing the uncertainty of fine-grained predictions [107], [108], [171], and deconstructing/constructing fine-grained images for learning fine-grained patterns [113], [178]. These observations further illustrate the benefit of consolidating work in fine-grained recognition and fine-grained retrieval in this survey paper.

8 FUTURE DIRECTIONS

Advances in deep learning have enabled significant progress in fine-grained image analysis (FGIA). Despite the success however, there are still many unsolved problems. Thus, in this section, we aim to explicitly point out these problems, and highlight some open questions to motivate future progression of the field.

Precise Definition of "Fine-Grained": Although FGIA has existed for many years as a thriving sub-field of computer vision and pattern recognition, one fundamental problem in FGIA persists, *i.e.*, the designation lacks a precise definition [189], [190]. Specifically, the community always *qualitatively* describes a so-called fine-grained dataset/problem as being "fine-grained" by roughly stating that its target objects belong to one meta-category. However, a precise definition of "fine-grained" could enable us to *quantitatively* evaluate the granularity of a dataset. In addition, it would not only provide a better understanding of current algorithmic performance on tasks of different granularities, but also bring additional insights to the fine-grained community.

Next-Generation Fine-Grained Datasets: Classic fine-grained datasets such as *CUB200-2011*, *Stanford Dogs*, *Stanford Cars*, and *Oxford Flowers*, are overwhelmingly used for benchmarking performance in FGIA. However, by modern standards these datasets are relatively small-scale and are largely saturated in terms of performance. In the future, it would be valuable to see additional large-scale fine-grained datasets being promoted to replace these existing benchmarks, *e.g.*, *iNat2017* [2], *Dogs-in-the-wild* [27], *RPC* [5]. State-of-the-art results on these datasets are 72.6% [191], 78.5% [27], 80.5% [192], respectively, which reveals the substantial room for further improvement. Moreover, these new benchmarks should embrace and embody all the challenges associated with fine-grained learning in addition to being large-scale (in terms of both the number of sub-categories and images), contain diverse images captured in realistic settings, and include rich annotations. However, high-quality fine-grained datasets usually require domain-experts to annotate. This limits the development of fine-grained datasets to a certain extent. As a potential solution, constructing an unlabeled large-scale fine-grained database and employing unsupervised feature learning techniques (*e.g.*, self-supervised learning [193]) could benefit discriminative features learning and further promote unsupervised downstream tasks [3]. Also, synthetic data [194] is an increasingly popular tool for training deep models, and offers promising opportunities to be explored further for FGIA.

Application to 3D Fine-Grained Tasks: Most existing FGIA approaches target 2D fine-grained images and the value of 3D information (*e.g.*, 3D pose labels or 3D shape information) is under-explored. How 3D object representations boost the performance of 2D FGIA approaches is an interesting and important problem. On the other hand, how 2D FGIA approaches generalize to 3D fine-grained applications [195], *e.g.*, robotic bin picking, robot perception or augmented reality, is also worthy of future attention. To make progress in this area there are open questions associated with difficulties in obtaining accurate 3D annotations for many object categories, *e.g.*, animals or other non-rigid objects.

Robust Fine-Grained Representations: One important

factor which makes FGIA uniquely challenging is the overwhelming variability in real-world fine-grained images, including changes in viewpoint, object scale, pose, and factors such as part deformations, background clutter, and so on. Meanwhile, FGIA often dictates fine-grained patterns (derived from discriminative but subtle parts) to be identified to drive predictions, which makes it more sensitive to image resolutions, corruptions and perturbations, and adversarial examples. Despite advances in deep learning, current models are still exhibit a lack of robustness to these variations or disturbances, and this significantly constrains their usability in many real-world applications where high accuracy is essential. Therefore, how to effectively obtain robust fine-grained representations (*i.e.*, not only containing discriminative fine-grained visual clues, but also resisting the interference of irrelevant information) requires further in-depth exploration. As discussed above, when coupled with next generation fine-grained datasets, there are questions related to the utility of self-supervised learning in these domains [196]. For example, will self-supervised learning help improving FGIA, and do we perform self-supervision on fine-grained data to generate robust fine-grained representations?

Interpretable Fine-Grained Learning: Unilaterally achieving higher accuracy compared to non-expert humans may no longer be the primary goal of FGIA. It is very important for fine-grained models to not only result in high accuracy but also to be interpretable [197]. More interpretable FGIA could help the community to address several challenges when dealing with fine-grained tasks using deep learning, *e.g.*, semantically debugging network representations, learning via human-computer communications at the semantic level, designing more effective, task-specific deep models, etc.

Fine-Grained Few-Shot Learning: Humans are capable of learning a new fine-grained concept with very little supervision (*e.g.*, with a few image of a new species of bird), yet our best deep learning fine-grained systems need hundreds or thousands of labeled examples. Even worse, the supervision of fine-grained images are both time-consuming and expensive to obtain, as fine-grained objects need to be accurately labeled by domain experts. Thus, it is desirable to develop fine-grained few-shot learning (FGFS) methods [148], [198], [199], [200]. The task of FGFS requires the learning system to build classifiers for novel fine-grained categories from few examples (*e.g.*, less than 10). Robust FGFS methods could significantly strengthen the usability and scalability of fine-grained recognition.

Fine-Grained Hashing: Recently, larger-scale fine-grained datasets have been released, *e.g.*, [2], [3], [5], [38], [41], [46]. In real applications like fine-grained image retrieval, the computational cost of finding the exact nearest neighbor can be prohibitively high in cases where the reference database is very large. Image hashing [201], [202] is a popular and effective technique for approximate nearest neighbor search, and has the potential to help with large-scale fine-grained data too. Therefore, targeting the big data challenge, fine-grained hashing [203], [204] is a promising direction worthy of further explorations.

Automatic Fine-Grained Models: Automated machine learning (AutoML) [205] and neural architecture search (NAS) [206] have attracted growing attention of late. AutoML targets automating the process of applying machine learning to real-world tasks and NAS is the process of automating neural network architecture design. Recent methods for AutoML and NAS could be comparable or even outperform hand-designed architectures in various computer vision applications. Thus, it is also logical that AutoML and NAS techniques could find better, and more tailor-made deep models for FGIA.

Fine-Grained Analysis in More Realistic Settings: In the past decade, FGIA related techniques have been developed and have achieved good performance in standard computer vision benchmarks, *e.g.*, [13], [42], [43]. The vast majority of these existing FGIA benchmarks are however defined in a static and closed environment. One other big limitation of current FGIA datasets is that they typically contain large instances of the object (*i.e.*, the objects of interest occupy most of the image frame). However, these settings are not representative of many real-world applications, *e.g.*, recognizing retail products in storage racks by models trained with images collected in controlled environments [5] or recognizing/detecting tends of thousands of natural species in the wild [2]. More research is needed in areas such as domain adaptation [207], [208], [209], long-tailed distributions [210], [211], open world settings [212], scale variations [2], fine-grained video understanding [213], [214], knowledge transfer, and resource constrained embedded deployment, to name a few.

9 CONCLUSION

We have presented a comprehensive survey on recent advances in deep learning based fine-grained image analysis (FGIA). Specifically, we advocated a broadened definition of FGIA, by consolidating work in fine-grained recognition and fine-grained retrieval. We enumerated gaps in existing research, pointed out a series of emerging topics, highlighted important future research directions, and illustrated that the problem of FGIA is still far from solved. However, given the significant improvements in performance over the past decade, we remain optimistic about future progress as we move towards more realistic and impactful applications.

REFERENCES

[1] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *CVPR*, 2015, pp. 595–604.

[2] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The iNaturalist species classification and detection dataset," in *CVPR*, 2017, pp. 8769–8778.

[3] G. Van Horn, E. Cole, S. Beery, K. Wilber, S. Belongie, and O. Mac Aodha, "Benchmarking representation learning for natural world image collections," in *CVPR*, 2021.

[4] L. Karlinsky, J. Shtok, Y. Tzur, and A. Tzadok, "Fine-grained recognition of thousands of object categories with single-example training," in *CVPR*, 2017, pp. 4113–4122.

[5] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, "RPC: A large-scale retail product checkout dataset," *arXiv preprint arXiv:1901.07249*, 2019.

[6] M. Jia, M. Shi, M. Sirotenko, Y. Cui, C. Cardie, B. Hariharan, H. Adam, and S. Belongie, "Fashionpedia: Ontology, segmentation, and an attribute localization dataset," in *ECCV*, 2020, pp. 316–332.

[7] S. D. Khan and H. Ullah, "A survey of advances in vision-based vehicle re-identification," *CVIU*, vol. 182, pp. 50–63, 2019.

[8] J. Yin, A. Wu, and W.-S. Zheng, "Fine-grained person re-identification," *IJCV*, vol. 128, pp. 1654–1672, 2020.

[9] ICCV 2019 Workshop on Computer Vision for Wildlife Conservation. [Online]. Available: https://openaccess.thecvf.com/ICCV2019_workshops/ICCV2019_CVWC

[10] Y. Wei, S. Tran, S. Xu, B. Kang, and M. Springer, "Deep learning for retail product recognition: Challenges and techniques," *Computational Intelligence and Neuroscience*, vol. 128, pp. 1–23, 2020.

[11] K. E. Johnson and A. T. Eilers, "Effects of knowledge and development on subordinate level categorization," *Cognitive Dev.*, pp. 515–545, 1998.

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

[13] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," *Tech. Report CNS-TR-2011-001*, 2011.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[15] IEEE TPAMI Special Issue on Fine-Grained Visual Categorization. [Online]. Available: https://www.computer.org/digital-library/journals/tp/fine-grained-visual-categorization

[16] Pattern Recognition Special Issue on Fine-Grained Object Retrieval, Matching and Ranking. [Online]. Available: https://www.journals.elsevier.com/pattern-recognition/call-for-papers/fine-grained-object-retrieval-matching-and-ranking

[17] ACM TOMM Special Issue on Fine-Grained Visual Recognition and Re-Identification. [Online]. Available: https://dl.acm.org/pb-assets/static_journal_pages/tomm/pdf/CFP_FGVRreID-1592406610240.pdf

[18] Pattern Recognition Letters Special Issue on Fine-Grained Categorization in Ecological Multimedia. [Online]. Available: https://www.sciencedirect.com/journal/pattern-recognition-letters/special-issue/10Z519ZW5DJ

[19] Neurocomputing Special Issue on Fine-grained Visual Understanding and Reasoning. [Online]. Available: https://www.journals.elsevier.com/neurocomputing/call-for-papers/fine-grained-visual-understanding-and-reasoning

[20] The competition homepage of "iNaturalist". [Online]. Available: https://www.kaggle.com/c/inaturalist-2019-fgvc6/overview

[21] The competition homepage of "Nature Conservancy Fisheries Monitoring". [Online]. Available: https://www.kaggle.com/c/the-nature-conservancy-fisheries-monitoring

[22] The competition homepage of "Humpback Whale Identification". [Online]. Available: https://www.kaggle.com/c/humpback-whale-identification

[23] CVPR 2021 Tutorial on Fine-Grained Visual Analysis with Deep Learning. [Online]. Available: https://fgva-cvpr21.github.io/

[24] The Eight Workshop on Fine-Grained Visual Categorization. [Online]. Available: https://sites.google.com/view/fgvc8/home

[25] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," *Int. J. Autom. and Comput.*, vol. 14, no. 2, pp. 119–135, 2017.

[26] M. Sun, Q. Li, Y. Geng, H. Yu, J. Wang, J. Gan, and W. Xue, "A survey of fine-grained image categorization," in *Proc. IEEE Conf. Signal Process.*, 2018, pp. 533–538.

[27] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *ECCV*, 2018, pp. 834–850.

[28] X. Zheng, R. Ji, X. Sun, B. Zhang, Y. Wu, and F. Huang, "Towards optimal fine grained retrieval via decorrelated centralized loss with normalize-scale layer," in *AAAI*, 2019, pp. 9291–9298.

[29] A. Mafla, S. Dey, A. F. Biten, L. Gomez, and D. Karatzas, "Fine-grained image classification and retrieval by combining visual and locally pooled textual features," in *WACV*, 2020, pp. 2950–2959.

[30] ——, "Multi-modal reasoning graph for scene-text based fine-grained image classification and retrieval," in *WACV*, 2020, pp. 4023–4033.

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

[32] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE TIP*, vol. 26, no. 6, pp. 2868–2881, 2017.

[33] Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, and P. Luo, "A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *CVPR*, 2019, pp. 5337–5345.

[34] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *arXiv preprint arXiv:2001.04193*, 2020.

[35] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[36] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *ECCV*, 2018, pp. 402–419.

[37] X.-S. Wei, C.-L. Zhang, L. Liu, C. Shen, and J. Wu., "Coarse-to-fine: A RNN-based hierarchical attention model for vehicle re-identification," in *ACCV*, 2018, pp. 575–591.

[38] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *CVPR*, 2016, pp. 1096–1104.

[39] X. Liu, J. Wang, S. Wen, E. Ding, and Y. Lin, "Localizing by describing: Attribute-guided attention localization for fine-grained recognition," in *AAAI*, 2017, pp. 4190–4196.

[40] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[41] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *CVPR*, 2014, pp. 2019–2026.

[42] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization," in *CVPR Workshop on Fine-Grained Visual Categorization*, 2011, pp. 806–813.

[43] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *ICCV Workshop on 3D Representation and Recognition*, 2013.

[44] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[45] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Indian Conf. on Comput. Vision, Graph. and Image Process.*, 2008, pp. 722–729.

[46] S. Hou, Y. Feng, and Z. Wang, "VegFru: A domain-specific dataset for fine-grained visual categorization," in *ICCV*, 2017, pp. 541–549.

[47] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101 – mining discriminative components with random forests," in *ECCV*, 2014, pp. 446–461.

[48] Y. Bai, Y. Chen, W. Yu, L. Wang, and W. Zhang, "Products-10K: A large-scale product recognition dataset," *arXiv preprint arXiv:2008.10545*, 2020.

[49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *IJCV*, vol. 115, pp. 211–252, 2015.

[50] F. Zhou and Y. Lin, "Fine-grained image classification by exploring bipartite-graph labels," in *CVPR*, 2016, pp. 1124–1133.

[51] Y. Li, T. M. Hospedales, Y.-Z. Song, and S. Gong, "Fine-grained sketch-based image retrieval by matching deformable part models," in *BMVC*, 2014, pp. 1–12.

[52] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C. C. Loy, "Sketch me that shoe," in *CVPR*, 2016, pp. 799–807.

[53] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: learning to retrieve badly drawn bunnies," *ACM Trans. on Graphics*, vol. 35, no. 4, pp. 1–12, 2016.

[54] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *ICCV*, 2017, pp. 5551–5560.

[55] K. Li, K. Pang, Y.-Z. Song, T. M. Hospedales, T. Xiang, and H. Zhang, "Synergistic instance-level subspace alignment for fine-grained sketch-based image retrieval," *IEEE TIP*, vol. 26, no. 12, pp. 5908–5921, 2017.

[56] K. Pang, K. Li, Y. Yang, H. Zhang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Generalising fine-grained sketch-based image retrieval," in *CVPR*, 2019, pp. 677–686.

[57] X. He, Y. Peng, and X. Liu, "A new benchmark and approach for fine-grained cross-media retrieval," in *ACM MM*, 2019, pp. 1740–1748.

[58] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *CVPR*, 2016, pp. 49–58.

[59] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," in *CVPR*, 2017, pp. 5994–6002.

[60] O. Mac Aodha, E. Cole, and P. Perona, "Presence-only geographical priors for fine-grained image classification," in *ICCV*, 2019, pp. 9596–9606.

[61] G. Chu, B. Potetz, W. Wang, A. Howard, Y. Song, F. Brucher, T. Leung, and H. Adam, "Geo-aware networks for fine-grained recognition," in *ICCV Workshops*, 2019.

[62] X. Sun, L. Chen, and J. Yang, "Learning from web data using adversarial discriminative neural networks for fine-grained classification," in *AAAI*, 2019, pp. 273–280.

[63] L. Niu, A. Veeraraghavan, and A. Sabharwal, "Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification," in *CVPR*, 2018, pp. 7171–7180.

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

[65] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *ECCV*, 2014, pp. 834–849.

[66] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *CVPR*, 2015, pp. 5546–5555.

[67] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep LAC: Deep localization, alignment and classification for fine-grained recognition," in *CVPR*, 2015, pp. 1666–1674.

[68] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," in *CVPR*, 2016, pp. 1173–1182.

[69] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas, "SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition," in *CVPR*, 2016, pp. 1143–1152.

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

[71] M. Lam, B. Mahasseni, and S. Todorovic, "Fine-grained recognition as HSnet search for informative image parts," in *CVPR*, 2017, pp. 2520–2529.

[72] X. He and Y. Peng, "Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification," in *AAAI*, 2017, pp. 4075–4081.

[73] W. Ge, X. Lin, and Y. Yu, "Weakly supervised complementary parts models for fine-grained image classification from the bottom up," in *CVPR*, 2019, pp. 3034–3043.

[74] Z. Wang, S. Wang, H. Li, Z. Dou, and J. Li, "Graph-propagation based correlation learning for weakly supervised fine-grained image classification," in *AAAI*, 2020, pp. 12 289–12 296.

[75] C. Liu, H. Xie, Z.-J. Zha, L. Ma, L. Yu, and Y. Zhang, "Filtration and distillation: Enhancing region attention for fine-grained visual categorization," in *AAAI*, 2020, pp. 11 555–11 562.

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

[77] L. Liu, C. Shen, and A. van den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," in *CVPR*, 2015, pp. 4749–4757.

[78] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *ICCV*, 2015, pp. 1143–1151.

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

[80] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a CNN for fine-grained recognition," in *CVPR*, 2018, pp. 4148–4157.

[81] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao, "Selective sparse sampling for fine-grained image recognition," in *ICCV*, 2019, pp. 6599–6608.

[82] Z. Huang and Y. Li, "Interpretable and accurate fine-grained recognition via region grouping," in *CVPR*, 2020, pp. 8662–8672.

[83] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *CVPR*, 2017, pp. 4438–4446.

[84] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *ICCV*, 2017, pp. 5209–5217.

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

[86] L. Zhang, S. Huang, W. Liu, and D. Tao, "Learning a mixture of granularity-specific experts for fine-grained categorization," in *ICCV*, 2019, pp. 8331–8340.

[87] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *CVPR*, 2019, pp. 5012–5021.

[88] H. Zheng, J. Fu, Z.-J. Zha, J. Luo, and T. Mei, "Learning rich part hierarchies with progressive attention networks for fine-grained image recognition," *IEEE TIP*, vol. 29, pp. 476–488, 2020.

[89] R. Ji, L. Wen, L. Zhang, D. Du, Y. Wu, C. Zhao, X. Liu, and F. Huang, "Attention convolutional binary neural tree for fine-grained visual categorization," in *CVPR*, 2020, pp. 10 468–10 477.

[90] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *NIPS*, 2015, pp. 2017–2025.

[91] Y. Wang, J. Choi, V. I. Morariu, and L. S. Davis, "Mining discriminative triplets of patches for fine-grained classification," in *CVPR*, 2016, pp. 1163–1172.

[92] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *ECCV*, 2018, pp. 438–454.

[93] X. He, Y. Peng, and J. Zhao, "Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization," *IJCV*, vol. 127, pp. 1235–1255, 2019.

[94] Z. Wang, S. Wang, S. Yang, H. Li, J. Li, and Z. Li, "Weakly supervised fine-grained image classification via guassian mixture model oriented discriminative learning," in *CVPR*, 2020, pp. 9749–9758.

[95] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *ICCV*, 2015, pp. 1449–1457.

[96] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *CVPR*, 2016, pp. 317–326.

[97] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, and S. Belongie, "Kernel pooling for convolutional neural networks," in *CVPR*, 2017, pp. 2921–2930.

[98] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *CVPR*, 2017, pp. 365–374.

[99] Q. Wang, P. Li, and L. Zhang, "G^2DeNet: Global Gaussian distribution embedding network and its application to visual recognition," in *CVPR*, 2017, pp. 2730–2739.

[100] S. Cai, W. Zuo, and L. Zhang, "Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization," in *ICCV*, 2017, pp. 511–520.

[101] P. Li, J. Xie, Q. Wang, and Z. Gao, "Towards faster training of global covariance pooling networks by iterative matrix square root normalization," in *CVPR*, 2018, pp. 947–955.

[102] M. Engin, L. Wang, L. Zhou, and X. Liu, "DeepKSPD: Learning kernel-matrix-based SPD representation for fine-grained image recognition," in *ECCV*, 2018, pp. 629–645.

[103] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *ECCV*, 2018, pp. 595–610.

[104] X. Wei, Y. Zhang, Y. Gong, J. Zhang, and N. Zheng, "Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification," in *ECCV*, 2018, pp. 365–380.

[105] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Learning deep bilinear transformation for fine-grained image representation," in *NIPS*, 2019, pp. 4277–4286.

[106] S. Min, H. Yao, H. Xie, Z.-J. Zha, and Y. Zhang, "Multi-objective matrix normalization for fine-grained visual recognition," *IEEE TIP*, vol. 29, pp. 4996–5009, 2020.

[107] A. Dubey, O. Gupta, R. Raskar, and N. Naik, "Maximum entropy fine-grained classification," in *NIPS*, 2018, pp. 637–647.

[108] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik, "Pairwise confusion for fine-grained visual classification," in *ECCV*, 2018, pp. 71–88.

[109] Y. Gao, X. Han, X. Wang, W. Huang, and M. R. Scott, "Channel interaction networks for fine-grained image categorization," in *AAAI*, 2020, pp. 10 818–10 825.

[110] G. Sun, H. Cholakkal, S. Khan, F. S. Khan, and L. Shao, "Fine-grained recognition: Accounting for subtle differences between similar classes," in *AAAI*, 2020, pp. 12 047–12 054.

[111] P. Zhuang, Y. Wang, and Y. Qiao, "Learning attentive pairwise interaction for fine-grained classification," in *AAAI*, 2020, pp. 2457–2463.

[112] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, M. Wu, J. Guo, and Y.-Z. Song, "The devil is in the channels: Mutual-channel loss for fine-grained image classification," *IEEE TIP*, vol. 29, pp. 4683–4695, 2020.

[113] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *CVPR*, 2019, pp. 5157–5166.

[114] W. Luo, X. Yang, X. Mo, Y. Lu, L. S. Davis, J. Li, J. Yang, and S.-N. Lim, "Cross-X learning for fine-grained visual categorization," in *ICCV*, 2019, pp. 8242–8251.

[115] R. Du, D. Chang, A. K. Bhunia, J. Xie, Y.-Z. Song, Z. Ma, and J. Guo, "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *ECCV*, 2020, pp. 153–168.

[116] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[117] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[118] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[119] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006, pp. 1–8.

[120] M. Guillaumin, D. Küttel, and V. Ferrari, "ImageNet auto-annotation with segmentation propagation," *IJCV*, vol. 110, pp. 328–348, 2014.

[121] X.-S. Wei, C.-L. Zhang, J. Wu, C. Shen, and Z.-H. Zhou, "Unsupervised object discovery and co-localization by deep descriptor transformation," *PR*, vol. 88, pp. 113–126, 2019.

[122] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014, pp. 818–833.

[123] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006, vol. 1.

[124] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.

[125] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE TPAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[126] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Reviews Neuroscience*, vol. 3, pp. 201–215, 2002.

[127] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order boltzmann machine," in *NIPS*, 2010, pp. 1243–1251.

[128] X. He, Y. Peng, and J. Zhao, "Fast fine-grained image classification via weakly supervised discriminative localization," *IEEE TCSVT*, vol. 29, no. 5, pp. 1394–1407, 2019.

[129] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[130] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," in *NIPS*, 1993, pp. 737–744.

[131] X. He, Y. Peng, and J. Zhao, "StackDRL: Stacked deep reinforcement learning for fine-grained visual categorization," in *IJCAI*, 2018, pp. 741–747.

[132] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *JAIR*, vol. 4, pp. 237–285, 1996.

[133] J. Mu, S. Bhat, and P. Viswanath, "Representing sentences as low-rank subspaces," in *ACL*, 2017, pp. 629–634.

[134] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *ICCV*, 2011, pp. 2018–2025.

[135] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in *CVPR*, 2015, pp. 1798–1807.

[136] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *CVPR*, 2015, pp. 3828–3836.

[137] B.-B. Gao, X.-S. Wei, J. Wu, and W. Lin, "Deep spatial pyramid: The devil is once again in the details," *arXiv preprint arXiv:1504.05277*, 2015.

[138] P. H. Gosselin, N. Murray, H. Jégou, and F. Perronnin, "Revisiting the fisher vector for fine-grained classification," in *Pattern Recogn. Letters*, no. 49, 2015, pp. 92–98.

[139] L. Wang, J. Zhang, L. Zhou, C. Tang, and W. Li, "Beyond covariance: Feature representation with nonlinear kernel matrices," in *ICCV*, 2015, pp. 4570–4578.

[140] Q. Wang, J. Xie, W. Zuo, L. Zhang, and P. Li, "Deep CNNs meet global covariance pooling: Better representation and generalization," *IEEE TPAMI*, 2020.

[141] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear convolutional neural networks for fine-grained visual recognition," *IEEE TPAMI*, vol. 40, no. 6, pp. 1309–1322, 2018.

[142] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *KDD*, 2013, pp. 239–247.

[143] Y. Li, N. Wang, J. Liu, and X. Hou, "Factorized bilinear models for image recognition," in *ICCV*, 2017, pp. 2079–2087.

[144] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *IJCV*, vol. 105, no. 3, pp. 222–245, 2013.

[145] P. Li, J. Xie, Q. Wang, and W. Zuo, "Is second-order information helpful for large-scale visual recognition?" in *ICCV*, 2017, pp. 2070–2078.

[146] T.-Y. Lin and S. Maji, "Improved bilinear pooling with CNNs," in *BMVC*, 2017, pp. 1–12.

[147] W. Xiong, Y. He, Y. Zhang, W. Luo, L. Ma, and J. Luo, "Fine-grained image-to-image transformation towards visual recognition," in *CVPR*, 2020, pp. 5840–5849.

[148] X.-S. Wei, P. Wang, L. Liu, C. Shen, and J. Wu, "Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples," *IEEE TIP*, vol. 28, no. 12, pp. 6116–6125, 2019.

[149] S. Xie, T. Yang, X. Wang, and Y. Lin, "Hyper-class augmented and regularized deep learning for fine-grained image classification," in *CVPR*, 2015, pp. 2645–2654.

[150] Z. Xu, S. Huang, Y. Zhang, and D. Tao, "Augmenting strong supervision using web data for fine-grained categorization," in *CVPR*, 2015, pp. 2524–2532.

[151] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei, "The unreasonable effectiveness of noisy data for fine-grained recognition," in *ECCV*, 2016, pp. 301–320.

[152] L. Niu, A. Veeraraghavan, and A. Sabharwal, "Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification," in *CVPR*, 2018, pp. 7171–7180.

[153] Y. Zhang, H. Tang, and K. Jia, "Fine-grained visual categorization using meta-learning optimization with sample selection of auxiliary data," in *ECCV*, 2018, pp. 241–256.

[154] Z. Xu, S. Huang, Y. Zhang, and D. Tao, "Webly-supervised fine-grained visual categorization via deep domain adaptation," *IEEE TPAMI*, vol. 40, no. 4, pp. 769–790, 2018.

[155] J. Yang, X. Sun, Y.-K. Lai, L. Zheng, and M.-M. Cheng, "Recognition from web data: A progressive filtering approach," *IEEE TIP*, vol. 27, no. 11, pp. 5303–5315, 2018.

[156] C. Zhang, Y. Yao, H. Liu, G.-S. Xie, X. Shu, T. Zhou, Z. Zhang, F. Shen, and Z. Tang, "Web-supervised network with softly update-drop training for fine-grained visual classification," in *AAAI*, 2020, pp. 12781–12788.

[157] H. Zhang, X. Cao, and R. Wang, "Audio visual attribute discovery for fine-grained object recognition," in *AAAI*, 2018, pp. 7542–7549.

[158] 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.

[159] T. Chen, L. Lin, R. Chen, Y. Wu, and X. Luo, "Knowledge-embedded representation learning for fine-grained image recognition," in *IJCAI*, 2018, pp. 627–634.

[160] K. Song, X.-S. Wei, X. Shu, R.-J. Song, and J. Lu, "Bi-modal progressive mask attention for fine-grained recognition," *IEEE TIP*, vol. 29, pp. 7006–7018, 2020.

[161] 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 *CVPR*, 2017, pp. 1878–1887.

[162] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[163] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *arXiv preprint arXiv:2004.05439*, 2020.

[164] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer, "DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web Journal*, pp. 167–195, 2015.

[165] F. M. Zanzotto, "Viewpoint: Human-in-the-loop artificial intelligence," *JAIR*, vol. 64, pp. 243–252, 2019.

[166] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop," in *CVPR*, 2016, pp. 1153–1162.

[167] J. Deng, J. Krause, M. Stark, and L. Fei-Fei, "Leveraging the wisdom of the crowd for fine-grained recognition," *IEEE TPAMI*, vol. 38, no. 4, pp. 666–676, 2016.

[168] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *CVPR*, 2018, pp. 4109–4118.

[169] Q. Wang, L. Zhang, B. Wu, D. Ren, P. Li, W. Zuo, and Q. Hu, "What deep CNNs benefit from global covariance pooling: An optimization perspective," in *CVPR*, 2020, pp. 10771–10780.

[170] X. Zheng, R. Ji, X. Sun, Y. Wu, F. Huang, and Y. Yang, "Centralized ranking loss with weakly supervised localization for fine-grained object retrieval," in *IJCAI*, 2018, pp. 1226–1233.

[171] X. Zeng, Y. Zhang, X. Wang, K. Chen, D. Li, and W. Yang, "Fine-grained image retrieval via piecewise cross entropy loss," *Image and Vision Computing*, vol. 93, pp. 2868–2881, 2020.

[172] CVPR 2020 Workshop on RetailVision. [Online]. Available: https://retailvisionworkshop.github.io/

[173] H. Idrees, M. Shah, and R. Surette, "Enhancing camera surveillance using computer vision: a research note," *arXiv preprint arXiv:1808.03998*, 2018.

[174] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *CVPR*, 2016, pp. 4004–4012.

[175] J. Song, Y.-Z. Song, T. Xiang, T. M. Hospedales, and X. Ruan, "Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval," in *BMVC*, 2016, pp. 1–11.

[176] F. Radenovic, G. Tolias, and O. Chum, "Deep shape matching," in *ECCV*, 2018, pp. 751–767.

[177] J. Zhang, F. Shen, L. Liu, F. Zhu, M. Yu, L. Shao, H. Tao Shen, and L. Van Gool, "Generative domain-migration hashing for sketch-to-image retrieval," in *ECCV*, 2018, pp. 297–314.

[178] K. Pang, Y. Yang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval," in *CVPR*, 2020, pp. 10347–10355.

[179] A. K. Bhunia, Y. Yang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Sketch less for more: On-the-fly fine-grained sketch-based image retrieval," in *CVPR*, 2020, pp. 9779–9788.

[180] A. Sain, A. K. Bhunia, Y. Yang, T. Xiang, and Y.-Z. Song, "Cross-modal hierarchical modelling for fine-grained sketch based image retrieval," in *BMVC*, 2020, pp. 1–14.

[181] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *CVPR*, 2015, pp. 437–446.

[182] K. Pang, Y.-Z. Song, T. Xiang, and T. M. Hospdales, "Cross-domain generative learning for fine-grained sketch-based image retrieval," in *BMVC*, 2017, pp. 1–12.

[183] J. Song, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Fine-grained image retrieval: the text/sketch input dilemma," in *BMVC*, 2017, pp. 1–12.

[184] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *ICLR*, 2018, pp. 1–16.

[185] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9729–9738.

[186] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.

[187] F. Liu, C. Zou, X. Deng, R. Zuo, Y.-K. Lai, C. Ma, Y.-J. Liu, and H. Wang, "SceneSketcher: Fine-grained image retrieval with scene sketches," in *ECCV*, 2020, pp. 718–734.

[188] K. Musgrave, S. Belongie, and S.-N. Lim, "A metric learning reality check," in *ECCV*, 2020, pp. 681–699.

[189] Y. Cui, Z. Gu, D. Mahajan, L. van der Maaten, S. Belongie, and S.-N. Lim, "Measuring dataset granularity," *arXiv preprint arXiv:1912.10154*, 2019.

[190] D. Chang, K. Pang, Y. Zheng, Z. Ma, Y.-Z. Song, and J. Guo, "Your "flamingo" is my "bird": Fine-grained, or not," *arXiv preprint arXiv:2011.09040*, 2020.

[191] X. Wang, L. Lian, Z. Miao, Z. Liu, and S. X. Yu, "Long-tailed recognition by routing diverse distribution-aware experts," in *ICLR*, 2021, pp. 1–15.

[192] C. Li, D. Du, L. Zhang, T. Luo, Y. Wu, Q. Tian, L. Wen, and S. Lyu, "Data priming network for automatic check-out," in *ACM MM*, 2019, pp. 2152–2160.

[193] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE TPAMI*, 2020.

[194] S. I. Nikolenko, "Synthetic data for deep learning," *arXiv preprint arXiv:1909.11512*, 2019.

[195] X. Liu, Z. Han, Y.-S. Liu, and M. Zwicker, "Fine-grained 3D shape classification with hierarchical part-view attention," *IEEE TIP*, vol. 30, pp. 1744–1758, 2021.

[196] E. Cole, X. Yang, K. Wilber, O. Mac Aodha, and S. Belongie, "When does contrastive visual representation learning work?" *arXiv preprint arXiv:2105.05837*, 2021.

[197] Q. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: a survey," *arXiv preprint arXiv:1802.00614*, 2018.

[198] S. Tsutsui, Y. Fu, and D. Crandall, "Meta-reinforced synthetic data for one-shot fine-grained visual recognition," in *NIPS*, 2019, pp. 3063–3072.

[199] L. Tang, D. Wertheimer, and B. Hariharan, "Revisiting pose-normalization for fine-grained few-shot recognition," in *CVPR*, 2020.

[200] J.-C. Su, S. Maji, and B. Hariharan, "When does self-supervision improve few-shot learning?" in *ECCV*, 2020, pp. 645–666.

[201] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *IEEE TPAMI*, vol. 40, no. 4, pp. 769–790, 2018.

[202] W.-J. Li, S. Wang, and W.-C. Kang, "Feature learning based deep supervised hashing with pairwise labels," in *IJCAI*, 2016, pp. 1711–1717.

[203] Q. Cui, Q.-Y. Jiang, X.-S. Wei, W.-J. Li, and O. Yoshie, "ExchNet: A unified hashing network for large-scale fine-grained image retrieval," in *ECCV*, 2020, pp. 189–205.

[204] S. Jin, H. Yao, X. Sun, S. Zhou, L. Zhang, and X. Hua, "Deep saliency hashing for fine-grained retrieval," *IEEE TIP*, vol. 29, pp. 5336–5351, 2020.

[205] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *NIPS*, 2015, pp. 2962–2970.

[206] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *arXiv preprint arXiv:1808.05377*, 2018.

[207] T. Gebru, J. Hoffman, and L. Fei-Fei, "Fine-grained recognition in the wild: A multi-task domain adaptation approach," in *ICCV*, 2017, pp. 1349–1358.

[208] S. Wang, X. Chen, Y. Wang, M. Long, and J. Wang, "Progressive adversarial networks for fine-grained domain adaptation," in *CVPR*, 2020, pp. 9213–9222.

[209] Y. Wang, R.-J. Song, X.-S. Wei, and L. Zhang, "An adversarial domain adaptation network for cross-domain fine-grained recognition," in *WACV*, 2020, pp. 1228–1236.

[210] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *CVPR*, 2019, pp. 9268–9277.

[211] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *CVPR*, 2020, pp. 9719–9728.

[212] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *CVPR*, 2019, pp. 2537–2546.

[213] C. Zhu, X. Tan, F. Zhou, X. Liu, K. Yue, E. Ding, and Y. Ma, "Fine-grained video categorization with redundancy reduction attention," in *ECCV*, 2018, pp. 136–152.

[214] J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *CVPR*, 2020, pp. 122–132.

Jianxin Wu is currently a professor in the Department of Computer Science and Technology at Nanjing University, China, and is associated with the State Key Laboratory for Novel Software Technology, China.

Yuxin Peng is currently the Boya Distinguished Professor with Wangxuan Institute of Computer Technology, Peking University, China.

Xiu-Shen Wei is a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. He was a Program Chair for the workshops associated with ICCV, IJCAI, ACM Multimedia, etc. He has also served as a Guest Editor of Pattern Recognition Journal, and a Tutorial Chair for Asian Conference on Computer Vision (ACCV) 2022.

Jinhui Tang is a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China.

Yi-Zhe Song is a Chair Professor of Computer Vision and Machine Learning, and Director of SketchX Lab at the Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey, UK. He is an Associate Editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), and a Program Chair for British Machine Vision Conference (BMVC) 2021.

Jian Yang is a Chang-Jiang professor in the School of Computer Science and Technology of Nanjing University of Science and Technology, China.

Oisin Mac Aodha is a Lecturer in Machine Learning at the University of Edinburgh, UK. He is a Fellow of the Alan Turing Institute and a European Laboratory for Learning and Intelligent Systems (ELLIS) Scholar. He was a Program Chair for the British Machine Vision Conference (BMVC) 2020.

Serge Belongie is a professor of Computer Science at the University of Copenhagen (DIKU) and the director of the Pioneer Centre for AI, Denmark.