Machine Intelligence Research
www.mi-research.net
21(3), June 2024, 431-451
DOI: 10.1007/s11633-022-1404-6

Parsing Objects at a Finer Granularity: A Survey

Yifan Zhao[1] Jia Li[2] Yonghong Tian[1]

[1]School of Computer Science, Peking University, Beijing 100871, China

[2]State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University, Beijing 100191, China

Abstract: Fine-grained visual parsing, including fine-grained part segmentation and fine-grained object recognition, has attracted considerable critical attention due to its importance in many real-world applications, e.g., agriculture, remote sensing, and space technologies. Predominant research efforts tackle these fine-grained sub-tasks following different paradigms, while the inherent relations between these tasks are neglected. Moreover, given most of the research remains fragmented, we conduct an in-depth study of the advanced work from a new perspective of learning the part relationship. In this perspective, we first consolidate recent research and benchmark syntheses with new taxonomies. Based on this consolidation, we revisit the universal challenges in fine-grained part segmentation and recognition tasks and propose new solutions by part relationship learning for these important challenges. Furthermore, we conclude several promising lines of research in fine-grained visual parsing for future research.

Keywords: Finer granularity, visual parsing, part segmentation, fine-grained object recognition, part relationship.

Citation: Y. Zhao, J. Li, Y. Tian. Parsing objects at a finer granularity: A survey. *Machine Intelligence Research*, vol.21, no.3, pp.431–451, 2024. http://doi.org/10.1007/s11633-022-1404-6

1 Introduction

Fine-grained visual parsing of image objects is a basic and crucial task in the computer vision community, which is fundamentally difficult, owing that there are usually subtle visual cues for distinguishing different objects or part regions. Recent advances in deep learning have significantly boosted the image understanding abilities of machine systems, e.g., its performance on the large-scale ImageNet dataset[1] surpasses the human-level recognition, but it is still a great challenge facing the fine-grained visual tasks. In particular, we consider two representative fine-grained visual parsing tasks in this paper, i.e., semantic part segmentation and fine-grained object recognition.

In contrast with coarse-grained object segmentation and base-level classification, fine-grained parsing is meant to segment or distinguish visually similar objects that belong to different fine-grained concepts, for example, decomposing objects into parts and dividing the base category into subcategories. A tremendous amount of research efforts[2–10] has been proposed to solve this important problem, which can also be applied for downstream applications[11–13]. Conventional machine learning techniques build explicit structures for parsing and understanding these fine-grained objects, e.g., graph and tree structures for part segmentation[14–17], and part learning in fine-grained recognition[18, 19]. In the era of deep learning, fine-grained segmentation and recognition approaches follow different paradigms, which achieve huge success compared to conventional models. Although there are more than 100 research papers each year to investigate this important problem, these papers seem to be disorganized, owing to various sundry research focuses including new task settings, benchmarking, and learning strategies. In particular, there are few survey papers that summarize the recent advances in fine-grained part segmentation. Thus, the relationships among different fine-grained sub-tasks are still under-explored, and these subtasks are developed independently by regarding them as less-relevant tasks.

In this paper, we make a comprehensive study of advances in fine-grained visual parsing tasks in the last decade. Besides analyzing recent deep learning works, we seek to explain the differences between non-deep learning and deep models, since these works often share similar intuitions and observations, and some of the previous studies could inspire further research. For consolidating these recent advances, we propose a new taxonomy for fine-grained part segmentation and recognition tasks, and also provide a collection of predominant benchmark datasets following our taxonomy. Besides these improvements compared to other survey papers[20, 21], in this paper, we start from the novel view of part relationship learning and regard it as the correspondences of different fine-grained sub-tasks. In this view, we revisit both the indi-

Review
Manuscript received on June 26, 2022; accepted on December 13, 2022; published online on January 12, 2024
Recommended by Associate Editor Jingyi Yu
Colored figures are available in the online version at https://link.springer.com/journal/11633

vidual and universal challenges of part segmentation and fine-grained recognition and make an attempted solution using the guidance of part relationship learning. In addition to these insights, we finish by discussing the future directions of fine-grained visual parsing tasks.

To summarize, the main contributions of this survey are as follows:

1) We present a comprehensive survey of fine-grained visual parsing tasks by collecting recent advances of two representative tasks, i.e., semantic part segmentation and fine-grained recognition.

2) We revisit these fine-grained visual tasks from a novel perspective of part relationship learning, by revealing the connections of these fine-grained tasks and providing a promising solution to tackle challenges in fine-grained tasks.

3) We consolidate recent fine-grained research by reorganizing these works with new taxonomy, providing a collection of prevailing benchmark datasets, and make comprehensive discussions to inspire future works.

4) We provide promising future directions of fine-grained visual parsing tasks to inform further studies.

The remainder of this paper is organized as follows: Section 2 provides new taxonomies, benchmark settings, and recent research on the problem of fine-grained part segmentation. Section 3 consolidates benchmarks, challenges, and advanced research on fine-grained object recognition. In Section 4, we delve into the connections of different fine-grained visual tasks from the perspective of part relationship learning and provide new solutions to improve existing challenges. We then highlight the future directions in Section 5 and then conclude this paper in Section 6.

2 Fine-grained object segmentation: A part-level perspective

2.1 Taxonomy

In this section, we construct a new taxonomy for the semantic part segmentation task and revisit the fine-grained object segmentation from the part-level perspective. As in Fig. 1, we conclude and re-organize the methods to solve semantic part segmentation tasks from two different views, i.e., the problem setting and learning strategies.

Problem setting. Considering target objects to segment, we categorize these methods into two lines, i.e., single-class and multi-class parsing. The single-class part segmentation methods only tackle one specific category while objects of other categories should be taken as backgrounds. The multi-class setting aims to segment multiple classes that appear in the visual stream simultaneously. Regarding data collection in segmentation tasks, we further divide them into strongly-supervised, weakly-supervised and unsupervised learning. In similar ways, we consider the instance-level semantic-level, and video-based parsing problems. Given these terminologies defined here, some sub-areas show linguistic crosses with other ones, e.g., unsupervised learning with multi-class part parsing. However, as these sub-areas have not yet been explored, here we discuss the main branches that attract major research attention in Section 2.2.

Strategy. Learning to segment object parts has attracted a wide variety of research attention. In previous decades, several successful hand-crafted models have achieved success in segmenting objects with clear fore-

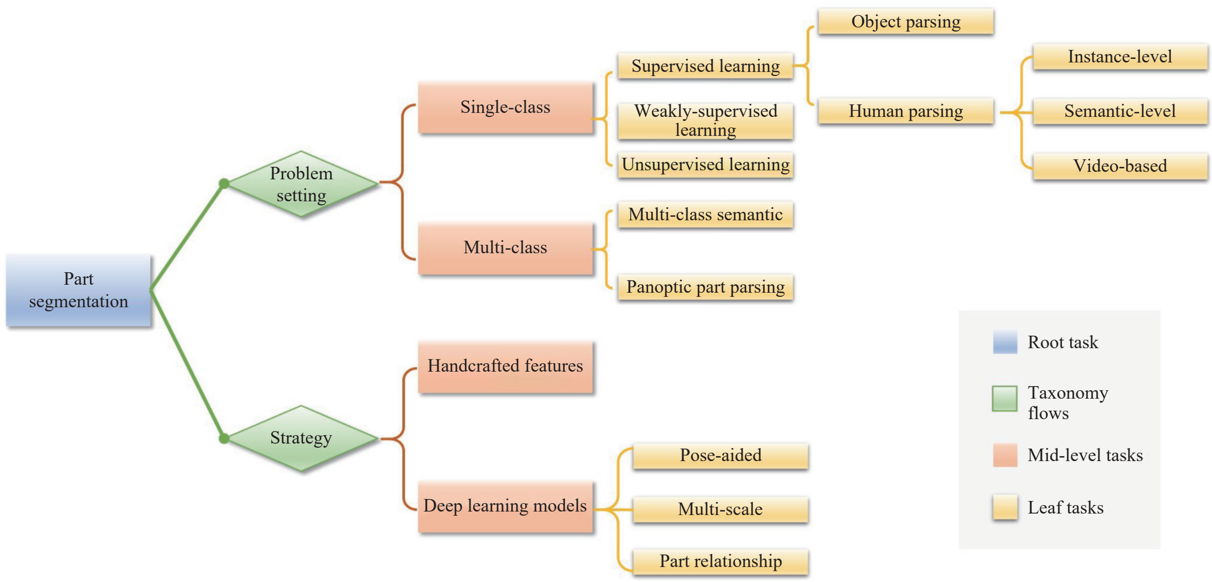


Fig. 1 Landscape of semantic part segmentation tasks in our taxonomy. We summarize the recent advances from two different aspects: problem setting and learning strategy.

Table 1 Summarization and comparisons of 12 widely-used part segmentation benchmark datasets. Note that PPS[22] re-organizes two new datasets based on existing data annotations.

Dataset	Pub.	Year	Task	Image No.	Category	Description
Fashionista[23]	CVPR	2012	Human parsing	685	56	Human clothes parsing
PASCAL-Part[24]	CVPR	2014	Detection & segmentation	10 103	NA	First large-scale part segmentation dataset
Horse-Cow dataset[25]	CVPR	2015	Single-class part segmentation	521	5	Quadruped animal parsing, reorganized from [24]
ATR[26]	ICCV	2015	Human parsing	17 700	18	Human clothes parsing
PASCAL-Person-Part[27]	CVPR	2016	Human parsing	3 533	7	Human body parsing, reorganized from [24]
MHP[28]	arXiv	2017	Human parsing	4 980	18	Multiple human clothes parsing
LIP[29]	T-PAMI	2018	Human parsing	50 462	20	Clothes parsing with human poses
VIP[30]	ACM MM	2018	Video-based human parsing	404 videos	19	Video-based human clothes parsing
CIHP[31]	ECCV	2018	Instance-level human parsing	38 280	20	Human clothes parsing
PASCAL-Part-58[32]	CVPR	2019	Multi-class part segmentation	10 103	58	First multi-class dataset, reorganized from [24]
PPS[22]	CVPR	2021	Part-aware panoptic segmentation	10 103/3 475	194/23	Derived from VOC-2010/Cityscape dataset
UDA-Part[33]	CVPR	2022	Single-class part segmentation	200	5	Unsupervised domain adaptation from synthetic vehicles
ADE-20K-Part[34]	IJCV	2022	Multi-class part segmentation	10 103	544	Large scale multi-class dataset, reorganized from [35]

ground representations, i.e., salient objects. We will briefly introduce several pioneer works in the following section. Besides these hand-crafted models, deep learning techniques have substantially improved the accuracy of segmentation models. We thus roughly group these techniques into three lines, i.e., pose-aided, multi-scale techniques, and using part relationships. Note that similar ideas could also be proposed in the non-deep learning methods. We will elaborate on their relations and differences in Section 2.3.

2.2 Task settings in part segmentation

Following the taxonomy in Section 2.1, we first summarize these datasets according to the task settings and annotation labels. We then elaborate on the detailed task settings and popular methods to solve these problems. According to the segmentation targets, here we summarize the popular datasets for semantic part segmentation tasks, which span the publications, image numbers, segmentation categories, and detailed descriptions.

2.2.1 Single-class part segmentation

Human parsing. As in Table 1, earlier works first tend to solve the specific categories of part segmentation, i.e., human parsing[1]. Representative datasets including Fashionista[23] focus on the human clothes parsing, which segments human objects into typical classes including shorts, shoeS, boots and sweaters. However, this dataset contains over 56 categories with a limited number of 685 images, which is not applicable to large machine-learning systems. With the development of deep learning techniques, large datasets are proposed to train and benchmark these deep models, e.g., ATR[26] and LIP[29], which consist of over 50 000 images of 20 categories for training and testing. These large benchmarks, as well as the accompanying baseline, have achieved great success in parsing humans into dressing clothes. Nevertheless, decomposing human objects with different clothing parts would lead to semantic inconsistencies on certain occasions. Hence, the other line of works proposes to segment human bodies into semantic parts following the morphological rules, which share the same definitions with human poses. For example, Chen et al.[27] propose to organize the PASCAL-Person-Part dataset to segment human bodies into 7 semantic parts, including lower/upper-arms, torsos, lower/upper-legs, heads, and backgrounds. Leading by this trend, dozens of works[2–5, 29, 36–46] propose to address this critical issue using deep learning techniques, which build well-established parsing baselines for understanding human structures.

The MHP dataset[28, 47] is presented to address multiple human parsing challenges that involve multiple human identities in one image, in addition to the conventional human parsing tasks. Beyond this challenging task, CIHP[31] is established to solve the instance-level human parsing task, e.g., [3, 48–54] which not only requires the semantic information of foreground objects, but also disentangles these parts into different human identities. Moreover, other works offer to investigate the part seg-

[1] Also noted as human part segmentation in some works.

mentation problem from a video-based perspective, i.e., video-based human parsing, with the proposal of VIP[30]. Video-based part segmentation requires a semantic consistency of temporal sequences. In summary, the tasks of these new trends continue to be founded on the segmentation of human clothing, while ignoring human body structure exploration and leaving space for future research.

Object part parsing. Except for human parsing tasks, here we review the object part segmentation task and divide the existing literature into two different lines: 1) Rigid object part segmentation[17, 33, 55]: including cars, aeroplanes, motorbikes, and other vehicles; 2) Non-rigid object part segmentation[25, 56, 57]: including birds, horses, cows, and other living creatures. Although these two lines of segmentation tasks can be uniformly solved by the popular deep learning schemes, challenges remain due to ambiguous semantic representations, blurriness around part boundaries, and anti-topological predictions. However, as the rigid objects usually consist of stable structures, Song et al.[55] and Liu et al.[33] propose to embed the canonical 3D models into learnable 2D part segmentation tasks, while in [17], the static geometric relationships are calculated for deducting the related part regions. However, when extending these strategies to non-rigid objects, namely articulated objects, the connections between object parts show a significant variance because of the various part shapes. Hence, the dynamic part relationships[25, 56] or pose-aided strategies[57] are proposed to solve this problem, which will be elaborated on in the next subsection.

Weakly-supervised part parsing. Several recent works have proposed exploiting weak supervision to generate dense semantic part segmentation masks. Unlike conventional weakly-supervised semantic segmentation methods, part segmentation using conventional image-level or box-level supervision would inevitably lead to highly repetitive semantic meanings (e.g., every human image contains the semantic label of torso and legs) and ambiguous annotations (bounding box overlap) respectively. Thus, several recent works propose to use human pose information as a weak supervision, which also shows high relationships with the part segmentation masks. We would like to elaborate on this in Section 2.3.2. Wu et al.[58] and Yang et al.[59] propose to generate accurate part segmentation masks using keypoint annotations. Moreover, Zhao et al.[60] propose a pose-to-part framework that gradually transfers weak pose annotations to the accurate segmentation masks and then use the image-level boundaries to correct the ambiguous regions.

Unsupervised learning. The aforementioned part segmentation tasks require accurate pixel-level annotations. It is an extremely labor-consuming work[61], especially performing annotations in the fine-grained part levels. Hence another trend of works[62−67] proposes to explore the semantic information through unsupervised manners. In [62], the automatic discovery of semantic parts and the relationships between the linguistical definition and activations discovered by convolutional neural networks (CNNs) are first explored. Leading by this thought, several works[63, 64] are proposed to leverage the advantages of deep representations. One specific feature is that the part representation after geometric transformation should be invariant over all instances of a category. Beyond this idea, Choudhury et al.[67] propose to discover the object part by using the contrastive loss among local regions. In addition, as all object instances from one category share the same part compositions, Gao et al.[65] propose to leverage the consistency of specific parts from different object instances, for example, different wings of birds share similar shapes and localization with respect to holistic objects.

The above works undoubtedly demonstrate the strong ability of automatic part discovery by adding constraints to deep neural networks. Without any prior guidance, these localized parts show a strong relation to the morphological structures of object categories. Thus, a natural concern arises: Can we use this object compositional information to guide the learning of other tasks? Furthermore, with weak supervision (e.g., image-level class labels), can we localize the accurate part regions that are most helpful for the learning objective, and what are the relations among these parts in recognition tasks? Keeping these concerns in mind, we will explain and discuss these details in Section 3.

2.2.2 Multi-class part segmentation

When revisiting the single-class part segmentation task in Fig. 2(b), there remain challenging problems for understanding the image content. Focusing on a single class, e.g., person class, and ignoring other meaningful classes, e.g., cars, and horses, would lead to severe parsing issues for understanding the context. In Fig. 2, only parsing human bodies into parts leads to a lack of object interaction with the context, for example, what is the human doing and where is the human sitting?

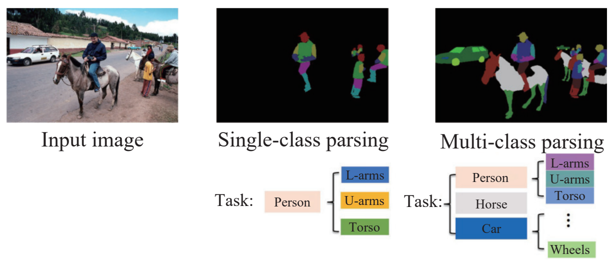


Fig. 2 Task settings of single-class and multi-class part segmentation. Single-class part segmentation only focuses on segmenting the objects of one specific class, while multi-class part segmentation aims to segment multiple objects that occurred in one scenario.

Multi-class semantic part segmentation. In response to these above challenges, the multi-class part segmentation tasks[32] are naturally proposed, as in Fig. 2(c).

The multi-class part segmentation tasks aim to segment objects of multiple classes into parts. In [32], a re-organized benchmark of the PASCAL-Part dataset is first proposed to solve this task, resulting in 58 semantic part classes. This new benchmark setting introduces additional challenges compared to the conventional single-class setting: 1) Semantic ambiguity: Parts of different object categories could share similar appearances, e.g., horse and cow legs; 2) Boundary ambiguity: Part boundaries of different objects are usually hard to disentangle. Toward this end, pioneer work[32] proposes a joint boundary-semantic awareness framework with auxiliary supervision. In [68], a graph-based matching network is proposed to construct the complex relationships between different parts, handling the part-level ambiguity and localization problems, which achieves success in handling part segmentation tasks of large scales, i.e., 108 part classes, while its journal version[34] focuses on the improvements of edge localization and extends the ADE-20K dataset[35] with part parsing labels, namely ADE 20K-Part. Besides, Tan et al.[69] propose a semantic ranking loss to re-rank these semantic parts by their predicted confidence. Singh et al.[70] develop a new learning framework that increases scalability and reduces task complexity compared to the monolithic label space counterpart. Additionally, this new research[70] introduces more complex part challenges, i.e., distinguishing left and right part localizations with more than 200 semantic classes.

Panoptic part segmentation. Motivated by the panoptic segmentation proposed by [71], parsing objects into disjoint parts along with the background regions seems to construct a comprehensive interaction with the environmental context. Geus et al.[22] establish the part-aware panoptic segmentation (PPS) task to understand a scene at multiple levels of abstraction. This PPS benchmark is founded on two representative datasets, PASCAL-VOC[72] for daily images and Cityscapes[73] for autonomous driving. In [22], a two-stage semantic parsing framework is proposed and the evaluation criteria for this new task are founded.

2.3 Strategies in part segmentation

Beyond the specific challenges of task setting in Section 2.2, part segmentation methods are designed following certain basic principles. Even the deep learning models and the non-deep ones share similar thoughts for constraining the optimization process. In this subsection, we will first introduce their commonalities and contrasts and then discuss and explore the promising future directions.

2.3.1 Non-deep learning models: Hand-crafted priors

Object part parsing in the past decade does not strictly follow the current definition of semantic segmentation, but decompose a holistic object into basic compositional units shares the same concerns. In [74], the deformable part models are proposed to localize and understand the whole object, which constructs a feature pyramid with respective deformative locations. Eslami and Williams[75] propose a generative model to jointly learn the appearances and part shapes and use block-Gibbs Markov chain Monte Carlo (MCMC) for fast inference. Following this trend, Liu et al.[76] adopt the Markov random fields to model the color and appearance similarities, deciding the part belongings. Meng et al.[77] propose to initialize part seed proposals and then develop a seed propagation strategy to combine other potential regions. Some other researches[6, 78, 79] segment object parts as an intermediate result to help the downstream tasks, including object detection, pose estimation, and action recognition.

Besides these works, the other line of works proposed to build trees[14–16, 80, 81] or graph models[17, 24, 25], depicting the relationships of different object parts. In [14], a joint bottom-up and top-down procedure is proposed to hierarchically decompose the holistic object into coarse parts, fine-grained parts, and basic lines/keypoints. Wang et al.[16] introduce hierarchical poselets, which decompose the human bodies into poselets (e.g., torso + left arm). Moreover, Several studies[80, 81] construct "And-Or" graphs to assemble the outputs of parts. e.g., Dong et al.[80] build a deformable mixture parsing model to simultaneously handle the deformation and multi-modalities of Parselets. Other works resort to graph structures, which are relatively flexible compared to hierarchical trees. For example, Chen et al.[24] construct a relational graph by the part attributes itself and pair-wise relationships. Wang and Yuille[25] propose to learn the part compositional model under multiple viewpoints and poses, constructing a robust transformation of different conditions.

Revisiting non-deep learning models. With the development of deep learning techniques[82–84], there is no doubt that the deep part segmentation models occupy the predominant places, benefited from their significant leading performance. Following the end-to-end training framework[85] in semantic segmentation, recent part segmentation models achieve more success than the conventional hand-crafted feature extractors, e.g., histogram of oriented gradient (HOG) or scale-invariant feature transform (SIFT) features. However, these deep learning models neglect the consideration of hierarchical body structures and would face great challenges in understanding unseen data and generating unreasonable segmentation masks. For example, in human parsing tasks, deep models always follow the statistical rules that deservedly take the round-shaped objects as human heads, which leads to incorrect parsing results for car wheels. In some error parsing cases, the lower legs could be connected with the upper bodies which breaks the basic topological rules. Interestingly, these phenomena are usually rare in conventional non-deep learning models, which follow the strict constraints of topological or morphological compositional principles, e.g., the human bodies are hierarchically decomposed into basic structures thus adjacent body parts show strong correlations. Moreover, the non-deep learning models require very little training data, showing great application potential in handling extreme circumstances in real-world applications.

2.3.2 Deep learning strategies

In addition to these aforementioned differences, the non-deep learning and deep learning models share similar designs and basic foundations to solve the fine-grained part parsing tasks. Whether hand-crafted feature extractors or deep feature extractors are employed, the basic challenges still remain for parsing reasonable and clear segmentation results. Three important characteristics of deep learning-based models are discussed in this subsection.

Pose-aided learning. The pose estimation and part segmentation are dual problems. Compared to the dense pixel prediction task of part segmentation, pose estimation is a more lightweight estimation task with significantly less annotation consumption. Conventional non-deep learning methods[23, 81, 86, 87] have proposed the importance of joint learning of these two related tasks. In the era of deep learning, major research efforts[5, 29, 36, 38] focus on the joint optimization of human part parsing and pose estimation with the proposals of large datasets. In [38], a mutual learning framework is proposed by embedding the dynamic kernel of pose estimation to part segmentations. Fang et al.[36] propose to transfer the human pose estimation knowledge as a coarse parsing prior and then to refine these coarse masks in the subsequent stages. Besides, other weakly-supervised methods[60] using keypoints information also achieves notable successes. Methods of this category verify that the accurate localizations of pose key points, including animals and human beings, show strong benefits to the tasks of part segmentations.

Multi-scale zooming. Different from the object segmentation tasks, part segmentation demonstrates a great demand in parsing detailed regions inside objects. Chen et al.[88] propose the atrous convolutional network to enhance the receptive field while [89] introduces an improved structure of atrous spatial pyramid pooling (ASPP), which incorporates multi-scale features in one single layer. Besides these general improvements, in [27], a two-stream CNN is proposed to fuse the global features and local features. Xia et al.[37] propose a stage-wise framework to detect and segment object parts from image-levels to object-levels and then part-levels.

Part relationship guidance. Several recent works[4, 40−43, 50, 53, 56, 90] propose to embed the part-level relationship as learning priors to guide the segmentation process. For example, Wang et al.[56] propose a joint conditional random field (CRF) to model the object-part and part-level relationships after the encoding of image features. Zhao et al.[4] decouple the part segmentation learning as multiple independent tasks while using the part-level learning order to constrain the recurrent learning process. Gong et al.[41] adopt a universal graph learning strategy to model the part relationship across multiple datasets. Wang et al.[43] propose a hierarchical part parsing network to gradually decompose the object from the coarse level to the finer level. In addition, in [40], a tree structure is constructed based on the CNN architectures and models the part-level relationship for understanding. Methods of this category successfully incorporate relationship learning to promote the segmentation process, while also using the accurate feature extraction of CNNs. The key challenge in fine-grained visual parsing is to understand the compositional relationships. Here we summarize these relationships as follows: 1) object-part relationship; 2) part-level relationship within one object; and 3) part-level across different images/objects. By understanding these relationships, deep models can further promote the learning of action recognition, fine-grained object recognition, and re-identification tasks.

3 Fine-grained object recognition: Understanding local structures

3.1 Definition and challenges

Definition. Image object classification has achieved great success benefiting from the development of deep learning systems and proposals of large datasets. Here we conclude the tasks of image object classification as base-level recognition, for example, classifying horses and aeroplanes, as in Fig. 3. Objects in base-level categories can be easily distinguished by image-level global features and usually has large margins in semantic definitions. For fine-grained object recognition, deep learning systems are required to distinguish the subtle differences among sub-categories that have similar appearances and semantic definitions. In this problem, methods developed for base-level recognition usually face great challenges for classifying fine-grained classes as in Fig. 3.

The formulation of fine-grained object recognition is similar to the common base-level ones, by learning using a much more "compact" semantic label space. Moreover, the generalized definition of fine-grained object recognition problems consists of two different levels of recognition: 1) Subcategory-level: Recognizing different fine-grained sub-categories that consist of multiple identities; 2) Instance-level: Distinguishing and identifying two different instances, e.g., person re-identification, vehicle re-identification, and face recognition. In this survey, we mainly focus on the first level of research but it should be noted that these two sub-field share many common techniques, which will be discussed in Sections 3.3 and 4.2.

Challenges. Here, we summarize three typical challenges of the fine-grained object recognition task in Fig. 4:

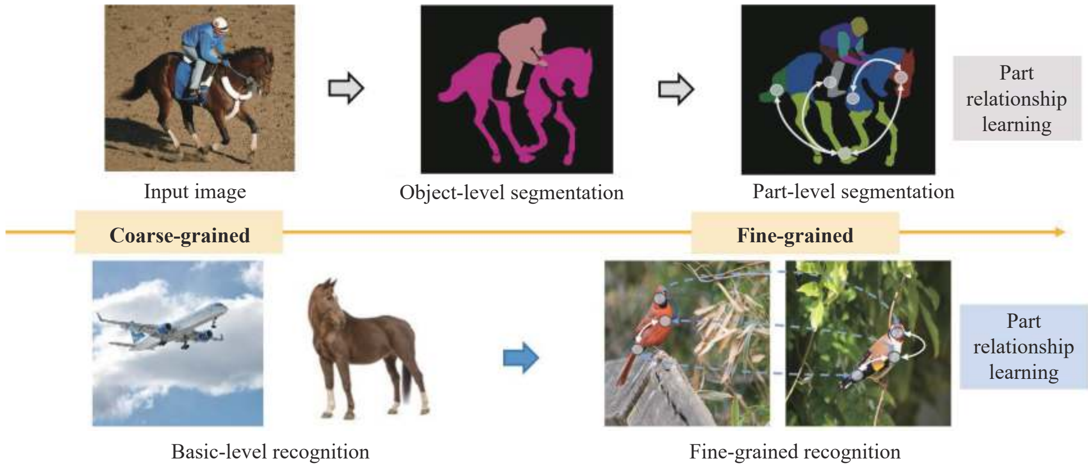


Fig. 3 Comparisons of coarse-grained learning and fine-grained learning. Representative fine-grained learning tasks, i.e., semantic part segmentation and fine-grained recognition, rely on the part relationship learning to build robust local features, while the coarse-grained tasks can be achieved by image-level global features.

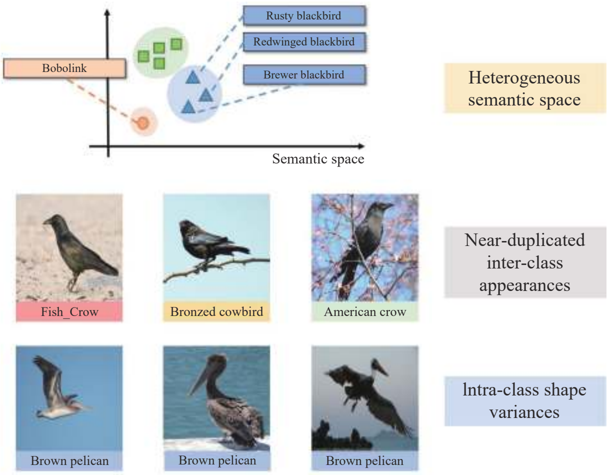


Fig. 4 Three typical challenges in fine-grained recognition tasks (images from CUB dataset[91]). 1) Heterogeneous semantic spaces: The semantic definitions of fine-grained text labels are usually cluster distributed. 2) Near-duplicated inter-class appearances: Objects of different categories present visually similar appearances. 3) Inter-class shape variances: The shape structures of objects in the same category can be inconsistent.

1) Heterogeneous semantic space: Although fine-grained labels are distributed in a compact space compared to the base-level category labels, their semantic definitions are still heterogeneous. For example, there are three types of blackbirds but only one bobolink in the semantic space. This phenomenon is still less-explored in the field of fine-grained recognition which leaves challenges for learning appropriate decision boundaries. 2) Near-duplicated inter-class appearances: In the middle of Fig. 4, we present three images from different fine-grained categories while sharing much common ground in visual appearances. Thus, deep learning models need to clearly distinguish their differences by observing local details. 3) Intra-class shape variances: Image objects that belong to the same categories can present in various shapes and structures. As for the bird classification task, the flying attitude shares less intuitive visual cues with that sitting one, bringing challenges to deducting these images with various shapes in the same categories. In most cases, these three challenges show mutual effects on each other, and a good learning model needs to have the ability to handle semantic imbalance, inter-class similarities, and intra-class diversities simultaneously.

3.2 Benchmark datasets

In this subsection, we summarize the prevailing benchmark datasets in the field of fine-grained recognition. In Table 2, with the development of machine learning systems, earlier works have established benchmark datasets with more than 100 categories for the classification of common daily objects, including Oxford 102 flowers[92] for plants, CUB-200-2011[91] for more than 200 bird categories, and Stanford-Dogs[93] for the classification of 120 dog sub-categories.

Pioneer machine learning methods, including support vector machine (SVM), and dictionary learning face great challenges in tackling these problems with less than 30% accuracies[18], indicating that these works can not be directly used in real-world industrial applications. Thus to solve these problems, these datasets provide bounding box information for localizing the main objects and providing the box or segmentation masks for part learning, e.g., bird heads and torsos for the CUB dataset[93]. Integrating this fore-ground information or part localization priors significantly helps the learning of fine-grained objects, especially the subtle differences of near-duplicated objects.

With the development of deep learning systems, especially CNNs, the representation ability for fine-grained objects has been significantly improved, e.g., from 28% to 75% accuracy on the CUB-200-2011 benchmark in [6]. Despite their effectiveness, deep learning models usually rely on the acquisition of a large number of training data

Table 2 Summarization and comparisons of 13 widely-used part segmentation benchmark datasets. Bbox and Part in annotation items indicate that the dataset provides object bounding box labels and part-level localization labels, respectively.

Dataset	Pub.	Year	Image number	Category	Annotation	Description
Oxford 102 flowers[92]	ICCVGIP	2008	8 189	102	–	Flower classification
CUB-200-2011[91]	–	2011	11 788	200	Bbox & Part	Birds, best-known fine-grained benchmark
Stanford dogs[93]	CVPRW	2011	20 580	120	Bbox	Dog classification
Stanford cars[94]	ICCVW	2013	16 185	196	Bbox	Car classification
FGVC aircraft[95]	Arxiv	2013	10 000	100	Bbox	Aircraft classification
Food 101[96]	ECCV	2014	101 000	101	–	Food classification
BirdSnap[97]	CVPR	2014	49 829	500	Bbox & Part	Large bird datasets
NAbirds[98]	CVPR	2015	48 562	555	Bbox & Part	Large bird datasets
CompCars[99]	ECCV	2018	136 727	431	Part images	Cars from web-nature and surveillance-nature
DeepFashion[100]	CVPR	2016	800 000	1 050	Bbox & Part	Clothes classification
iNat2017[101]	CVPR	2018	857 877	5 089	Bbox	Large-scale species classification
Dogs-in-the-Wild[102]	ECCV	2018	299 458	362	–	Large-scale dog classitication
iNat2021[103]	CVPR	2021	3 286 843	10 000	–	Improved version of iNat2017[101]

with similar distributions. Thus many new datasets with plentiful annotations are proposed, including Food 101[96] with more than 101 K images, NAbirds[98] and BirdSnap[97] for nearly 50 K images. These datasets not only provide high-quality annotations but also introduce new challenges for complicated semantic definitions, intra-class diversities, and inter-class similarities.

Beyond these earlier deep learning benchmarks, recent advanced research proposes large-scale annotations with a huge number of fine-grained categories. For example, iNat2017[101] provides more than 0.85 M images of 5 K categories, while its improved version iNat2021[103] provides more than 3 M images of 10 K categories. In addition, several other large-scale benchmarks[104, 105] have been proposed for fish recognition and landmark recognition. Beyond the aforementioned challenges, these datasets span the new dilemmas: 1) Imbalanced/Long-tailed data distributions: Objects of some rare categories usually consist of few annotations, while other main classes consist of thousands of images; 2) Noisy labels: Images of large scale datasets are usually collected webly and would introduce many noisy ambiguous labels. Thus, the classification model needs to further purify these noisy factors by learning from predominant clean annotations.

3.3 Strategies in fine-grained recognition

Recognizing fine-grained objects has attracted much research attention in the last two decades. In this subsection, we first conclude the non-deep learning techniques including the hand-crafted features and human-in-loop learning frameworks in Section 3.3.1. We then conclude the recent advanced research using deep learning techniques from two aspects, i.e., the part-guided learning in Section 3.3.2 and learning with feature representation constraints in Section 3.3.3.

3.3.1 Non-deep learning models

Hand-crafted feature extraction. Pioneer fine-grained works[18, 106–108] propose to use hand-crafted features to recognize objects. For example, Zhang et al.[18] propose to incorporate the SVM into understanding pose structures, learning with SIFT and bag of words (BoW) features. Yao et al.[106] propose dense sampling strategies with random forests to extract local features. Other researches including [107] adopt the codebook learning strategy for encoded dictionaries. Methods of this category can benefit from learning with local descriptors or part features while still facing difficulties in understanding fine-grained semantics.

Human in the loop. Conventional machine learning methods usually lead to relatively low performance, e.g., 28% for CUB-200-2011 classification tasks, which have difficulties for applying in realistic applications. Hence earlier works propose to incorporate human expert knowledge into the learning process. For example, Wah et al.[109] leverage computer vision techniques and analyze the user responses to gradually enhance the final learning accuracies. Branson et al.[19] propose an interactive scheme with deformable part models to distinguish the subtle differences between similar objects.

3.3.2 Part-guided learning

Supervised part learning. With the development of deep learning techniques, recognizing common base-level objects has made significant progress. Although the performance of fine-grained recognition has been improved in many applications[6], considering the challenges mentioned above, distinguishing subtle differences among near-duplicated objects usually faces serious dilemmas. Thus dozens of works propose to employ the part-level features[6–10, 110–112] to amplify these differences in a local perspective. Zhang et al.[6] propose a part-based R-CNN to locate the part features and then build pose-normalized features as the enhancement for global features. Krause et al.[113] propose to use keypoint annotations for fine-grained recognition, which leverages the co-segmentation techniques to align different views of images. Huang et al.[7] propose a dual-stream part-stacked CNN to jointly learn discriminative features from high-resolution part features and low-resolution global ones. In addition to these techniques using part detection methods, Wei et al.[9] propose to use part segmentation masks to regularize the local descriptor learning process. Although segmentation masks provide more accurate learning guidance, learning with all feasible part proposals would lead to a globally homogeneous amplification of every pixel. To summarize, supervised part-learning techniques adopt the part detectors or segmentation masks as local feature selection guidance, and then fuse these local features with the image-level global ones. In this manner, not only the global features but the local details are taken into account for final feature distance measurements. However, these part excavation methods still rely on accurate part segmentation or detection annotations, requiring enormous labor consumption. In addition, considering that the accurate part annotations of test data are usually infeasible, transferring this expert part knowledge into the testing environment would also lead to an inductive bias, which would further limit the effectiveness in real-world applications.

Unsupervised part learning.[2] Considering the labour-intensive computation and unstable generalization ability in the inference stage, recent ideas[114–123] propose to use unsupervised part attention techniques. Gonzalez-Garcia et al.[62] prove that during the back-propagation process, neural networks have the potential to discover semantic parts automatically. Leading by this trend, Simon and Rodner[114] propose neural activation constellations to localize semantic parts without any supervision. Different from this work, Fu et al.[116] propose a multi-stage zooming strategy to automatically locate and rescale the attention regions, by learning the confidence scores of different zooming proposals. Similar to this work, Recasens et al.[117] develop a saliency-based sampling layer for neural networks after finding the activated regions. Ge et al.[119] incorporate the weakly-supervised detection and segmentation models for localizing the discriminative features for fine-grained distinguishing. Besides, Wang et al.[118] propose a Gaussian mixture model for investigating the object parts with an auxiliary branch for supervision. In [123], a graph-propagation correlation learning method is proposed to model and propagate the discriminative part features to other parts. Nevertheless, these methods have shortcomings in two aspects: 1) introducing auxiliary learning branches or stages for optimization; and 2) the number of part proposals can sometimes be large whereas only a few are useful for recognition.

To solve this issue as well as reduce computational costs, Lam et al.[124] propose an HSNet searching architecture to explore the most discriminative parts, while other work[8] builds a weakly-supervised part selection mechanism based on their response scores. Zhao et al.[125] propose a Transformer architecture to build inter-part relationships and adopt multiple auxiliary branches for part-awareness learning, while in the inference stage, these auxiliary branches are not used for computational consideration. Methods of unsupervised part learning[125–128] lead the prevailing trend in fine-grained recognition, which benefits from its strong ability in understanding local differences and discovering object parts. Furthermore, selecting and modeling the part relationships becomes an emerging topic in fine-grained recognition.

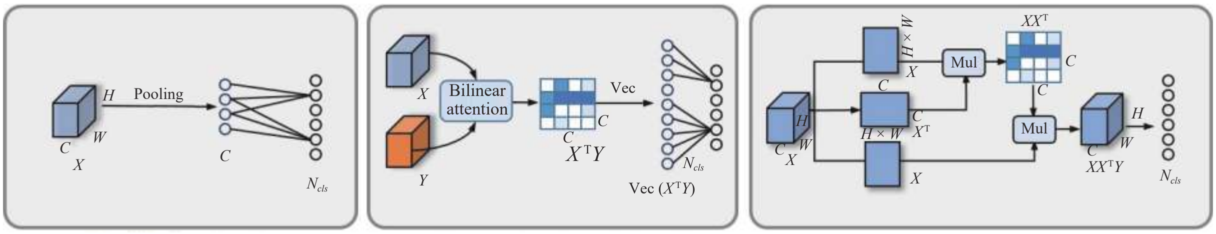
Different from the unsupervised part learning in semantic segmentation, part attention in fine-grained recognition aims to discover the discriminative features and exploits these local features as an enhancement for distinguishing near-duplicated objects. Thus the semantic information of the unsupervised part in recognition is usually not strictly aligned with the natural common definitions.

3.3.3 Feature representation learning

Besides the methods using part-level features for enhancing the local details, the other crucial problem in fine-grained recognition is feature representation learning. There is intuitive thinking that well-represented features can provide a more robust and generalization ability for downstream tasks, including segmentation, detection, and also fine-grained recognition. Despite the experimental evidence, enhancing the detailed representation ability helps the measurement of local subtle differences hidden among different features, which may be vital factors for discrimination. When features of various images are distributed in one generalized and robust fine-grained feature space, these subtle differences would be easy to discover. Guided by the theory, this line of methods tends to regularize the feature learning process[129–134] or generate rich feature representations[135–144] without using additional annotations.

High-order representations. As in Fig. 5(a), given an input image $\mathcal{I}$, the conventional classification model can be represented as $\boldsymbol{X} = \Phi(\mathcal{I})$. Thus, $\boldsymbol{X} \in \mathbf{R}^{W\times H\times C}$ denotes the C-dimensional with $H \times W$ feature maps and the final classification vector would be $\mathrm{Pool}(\boldsymbol{X}) \in \mathbf{R}^{1\times 1\times C}$. Considering the spatial feature relationship is neglected during the pooling operations, high-order interactions are proposed in advanced works. In Fig. 5(b), as the pioneer work using the second-order relationship, Lin et al.[136] propose a bilinear model to extract shape and appearances by two different CNNs and then construct a bilinear pooling operation to generate rich second-order representations, i.e., $\boldsymbol{X} = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}\mathtt{vec}(\Phi_1(\mathcal{I})_{i,j}^{\mathrm{T}}\times \Phi_2(\mathcal{I})_{i,j})$. Although this bilinear pooling operation en-

[2] Also noted as weakly-supervised fine-grained recognition.



(a) Vanilla classification (b) Second order relationship (c) Trilinear attention/Third order

Fig. 5 Three typical high-order relations as in [135]. Vanilla classification: Encoded features are pooled into vectors for classification, used in most of the works. Second-order relationship[135–139]: Learning rich second-order features by keeping the spatial dimension. Trilinear attention[140–145]: Preserving the same size as input features for learning spatial-wise or channel-wise attention matrix.

riches the fine-grained representation and amplifies the differences of similar embedding, it also introduces high computational costs, i.e., $C\times C\times N_{cls}$ for optimization, and N_{cls} denotes the category numbers. To solve this, Gao et al.[137] propose a compact bilinear pooling model that uses the network itself to build second-order relationships, i.e., $\boldsymbol{X}\equiv\boldsymbol{Y}$. Other works propose to use matrix factorization[138], Grassmann constraints[139], and low-rank learning[146] to reduce computation costs. Besides these works, Yu et al.[143] propose a hierarchical feature interaction operation to build heterogeneous second-order relationships. Zhao et al.[135] propose a graph-based high-order relationship learning to reduce the high-dimension space into discriminative low dimensions.

However, the second-order feature learning still introduces the curse of dimensionality for optimization, thus the other line of works proposes to use the third-order relationship in Fig. 5(c), namely trilinear attention[140, 141] or non-local mechanisms[145]. For example, Zheng et al.[140] propose the trilinear attention in the channel-dimension with a distillation mechanism, which can be formulated as $\mathrm{softmax}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}^{\mathrm{T}}\in\mathbf{R}^{WH\times C}$. While Gao et al.[141] propose a contrastive loss to learn the channel-wise relationship of inter-images and intra-images. Methods using the third-order relationship maintain the output size and can be embedded into different network stages to enhance the representations.

Feature interactions and regularization. Besides building high-order rich features, the other line of works proposes to use feature interactions[102, 129, 147–152] or using additional constraints[130–132, 153–155]. Wang et al.[129] construct a discriminative feature bank of convolutional filters that captures class-specific discriminative patches. Sun et al.[102] propose a multi-attention multi-constraint network to regularize the feature distributions based on the selected anchors. Luo et al.[148] propose to learn cross-level and cross-images relationships for building interactive feature representations. For robust feature learning, Chen et al.[147] incorporate an additional destruction and construction branch as an additional learning task. These works rely on additional blocks or feature interaction networks, which may introduce additional computation costs.

Besides works using additional parameters to feature enhancement, several works propose to use auxiliary constraints in addition to the basic cross-entropy constraints. In [130], pair-wise confusion is proposed among Siamese networks to alleviate the overfitting issues. While Dubey et al.[132] propose an entropy maximizing approach to regularize the final classification confidence. Aodha et al.[131] propose geographically guided loss functions that deduct the fine-grained features using temporal and geographical spatial priors. Besides, other works[153, 154, 156] follow a self-supervised learning trend for fine-grained recognition. Wu et al.[153] propose to solve the dilemma between self-supervised learning and fine-grained recognition by enhancing the salient foreground regions.

To summarize, 1) methods using high-order relations modules mainly focus on rich representations at the feature-level, and enhancing these representations would amplify the subtle differences among different object features, and 2) methods using additional constraints make fine-grained features to be distributed in compact and precise spaces, while alleviating the overfitting issue and concentrating more on object regions. This overfitting issue is further studied in existing works by generating accurate class activation maps while preventing only focusing on the local part regions. In the next section, we will discuss why we need local details and why only local details cannot perform accurate fine-grained recognition.

4 Part relationship in segmentation and recognition

Fine-grained visual parsing, including recognition, segmentation, detection and other high-level image understanding tasks, leaves us with challenges in its detailed and complex "fine-grained" parsing requirements. Understanding images with fine-grained objects can be substantially different from common "coarse-grained" ones. In this section, we investigate two representative fine-grained visual tasks, i.e., segmentation and recognition with the following natural concerns:

1) What are the key challenges in fine-grained recognition, or what are the unique problems in this subfield?

2) Why does part relationship learning help the understanding of these fine-grained tasks? What are the relations among them?

4.1 Problems in fine-grained parsing

Fine-grained visual parsing is a relatively-defined concept compared to the common daily categories. Considering the specific tasks of fine-grained recognition and semantic part segmentation, understanding image objects would face the following challenges.

Non-salient/Less-prominent in image-level. The fine-grained visual features are imperceptible using existing learning systems or not easily understood by human visual systems. In fine-grained recognition, objects of different semantic classes usually share visually similar appearances but still show imperceptible discrepancies. This means that objects of these categories are recognizable by detailed local differences while understanding them using coarse global features is impracticable. In the task of fine-grained part segmentation, distinguishing different parts relies on subtle visual cues, including imperceptible part boundaries and near-duplicated local visual patterns. The term imperceptible here means that these part boundaries can be relatively non-significant compared to object silhouettes and even contextual noisy information. In addition, in some extreme cases, due to the small scale of object parts (in Fig. 6(a)), it is difficult to recognize them by only observing small objects themselves. And in some cases, these discriminative cues in both parsing tasks are relatively non-significant, and are suppressed in the image-level feature representations.

Locally distinguishable and indistinguishable. As mentioned in the first challenge, the fine-grained objects are only recognizable in local details, e.g., Fig. 4 in Section 3. Thus, enhancing the representation of these local parts is beneficial to learn discriminative embedding. However, these local part regions are not always distinguishable, for example, birds of two different categories may exhibit similar appearances in torso, tail, and wings while only differing in their heads. Analogous to the fine-grained recognition task, we present a similar segmentation problem in Fig. 6(b). The legs of the horse and cow are locally indistinguishable whereas the holistic object categories are easy to recognize and segment, since their head regions are salient for distinguishing. To sum up, the local regions of an image may be distinguishable or indistinguishable when compared with different images.

Ambiguous semantic definition. The last challenge in fine-grained parsing is the ambiguous semantic definition, which is less explored in recent works. Conventional part parsing works[14, 15] propose to build hierarchical structures, e.g., from the holistic body to object parts, and then to line segments. However, considering the goal of segmentation and recognition, the semantic definitions of "fine-grained" tasks become an increasingly critical problem. For segmentation tasks, LIP dataset[29] defines the human bodies with different fashion clothes, e.g., skirts and coats, while other datasets[24] tend to segment bodies with the morphological rules, i.e., upper and lower bodies. Similarly, if we define two different categories of objects, e.g., bulldogs and poodles, the bulldogs can be subdivided into English bulldogs and American bulldogs with fewer differences. Thus the definitions of fine-

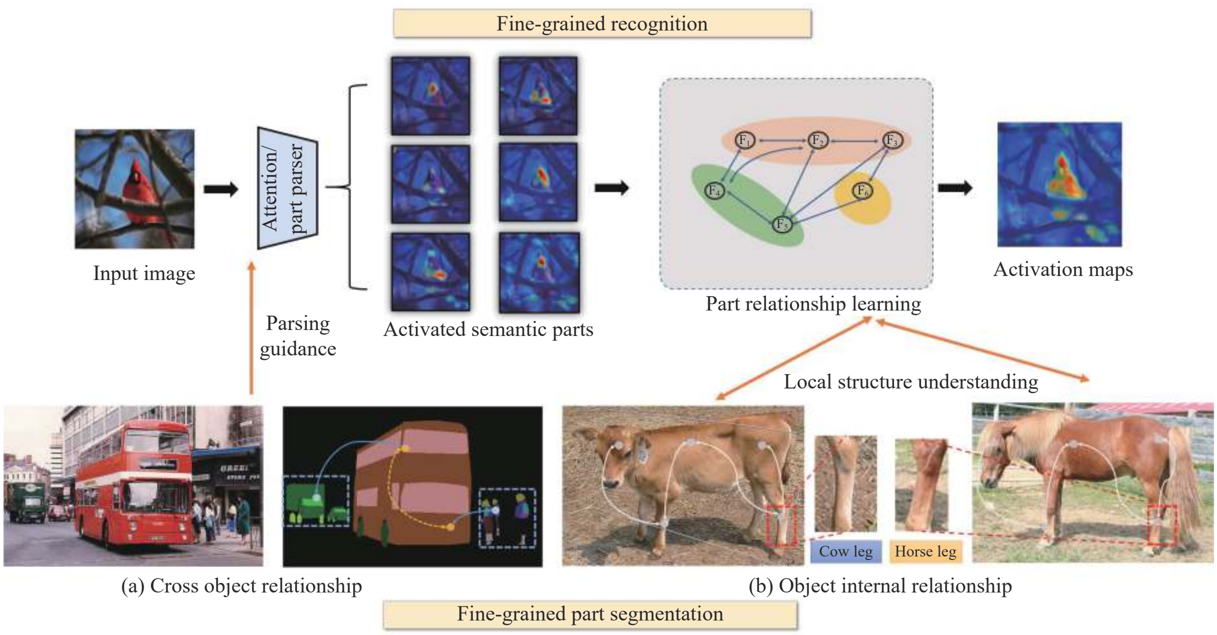


Fig. 6 Part relationship learning in two representative fine-grained visual tasks, fine-grained part segmentation[32] and fine-grained recognition[135]. First row: Understanding complex fine-grained images requires the accurate parsing of local part relationships. Second row: The cross-object relationship uses contextual information to help the understanding of small parts, while the object internal relationship with other parts helps the distinguishing of locally similar regions. Besides, the segmentation results can serve as parsing guidance and relationship learning in both tasks constructing the robust local structure understanding.

grained semantics leave us with severe challenges, or in other words, "how fine is fine-grained parsing"?

4.2 Part relationship learning: A solution

Towards the aforementioned challenges, in this survey, we argue that building part relationships in fine-grained visual parsing would be one reliable and promising solution. Here we elaborate on its effectiveness in solving these challenges: 1) Considering that the fine-grained cues are usually hidden in local details and cannot be distinguished using the image-level features. Hence enhancing the learning process with part relations helps a dynamic understanding, as illustrated in the first line of Fig. 6 by [135]. It should be mentioned that most of the prevailing works in fine-grained parsing and even re-identification tend to amplify fixed part regions of the same image, which could be solved by introducing part relationships. 2) Given one image, we usually do not understand which part should network focus on and what are discriminative features for recognition and segmentation. Thus, understanding part relationships helps this problem in many ways, e.g., cross-part relationships within each object helps the understanding of geometric structures, object-part relationship helps the distinguishing of locally similar visual patterns, and part relationships across different images and objects help to learn the semantic consensus. Besides, considering the comparison of visually similar images, these different part-level relationships help the dynamic enhancement of feature extractions. 3) For the semantic ambiguities of fine-grained definition, here we advocate the learning with hierarchical structures, which are still less explored in deep learning works. For example, the human body can be decomposed into heads and bodies, while the head regions can be further subdivided into faces, eyes, and other organs. Building hierarchical trees or graphs helps the holistic geometric structures be more reasonable and is also beneficial for handling the heterogeneous semantic definitions as mentioned in Fig. 4.

Relations to fine-grained tasks. What roles does the part relationship learning play in different fine-grained understanding tasks? Here, we present a schematic in Fig. 7 with fine-grained part segmentation, fine-grained recognition and part relationship representations. Part segmentation is the subset of the part relationship representations, while the latter consists of other structural parsing and detection sub-tasks. The intersections of part segmentation and fine-grained recognition are also illustrated in Section 3.3.2. Methods of this category tend to utilize the part priors to guide the feature extraction, including the fine-grained classification[6–10, 112] and re-identifications[110, 157–160]. Understanding part relationships without segmenting them as explicit masks or boxes also considerably facilitates the distinguishing of subtle differences, considering the joint region of part relation-

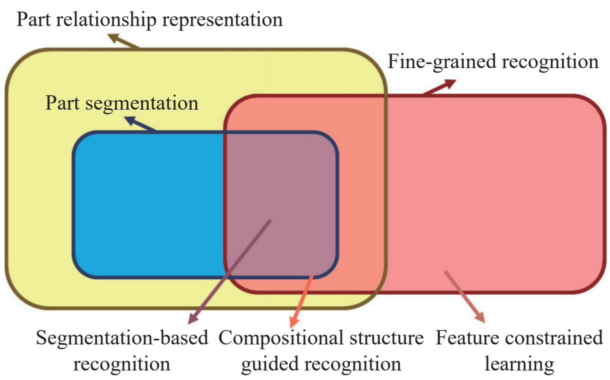


Fig. 7 Correlations schematic of three relevant tasks, i.e., part relationship representation, part segmentation and fine-grained recognition.

ship representations and fine-grained recognition. Methods using this idea usually adopt the attention mechanisms[8, 62, 114, 116–119, 124, 125] or graph-based structures[126, 127] to guide the learning process. Besides, the other line of works[129–132, 135–141, 143] does not rely on part relationships and focuses on the feature representation learning process as mentioned above. It should be noted here we extend the concept of part relationship learning, by incorporating learning with implicit and explicit part localizations, and those methods build explicit part-level relationships. Moreover, fine-grained part segmentation is one of the explicit ways to understand the part relationship procedure, and learning in hierarchical manners or tree/graph structures also leads in promising directions.

5 Future directions

Despite the significant progress made by existing works, there are still many unsolved problems in fine-grained recognition. Here we propose several promising future directions for discussion.

Dynamic part relationship learning. As discussed in Section 4.2, prevailing part relationship learning works focus on building static connections and responses. For example, in fine-grained recognition tasks, networks tend to amplify fixed regions of the same image. Although this helps the discriminative embedding in training data distributions, it faces great challenges when compared to unknown novel testing examples. Besides, the dynamic relationships help the understanding of semantic parsing when objects face occlusions or abnormal gestures.

Few-shot fine-grained learning. The learning of fine-grained classification is based on sufficient training data, while few-shot learning[161–163] is nowadays an attractive trend for understanding novel concepts with only a few annotated images. As for the few-shot fine-grained classification, there are several advances[164–168] in this field. However, these works usually follow the N-way K-shot trend, and N is usually set as 5 for the number of categories, indicating the huge gap compared to popular datasets with 500 to 5 000 categories. Besides, few-shot learning also can be regarded as a long-tailed setting where several categories are labeled with sufficient images but a few categories with limited annotations. This few-shot/long-tailed setting is a more realistic challenge that can be taken as a promising direction.

Hierarchical structures of fine-grained learning. In Fig. 4 and Section 4.1, it is noted that the definition of fine-grained settings still exists ambiguous and some subcategories can be further divided into finer levels. Thus to solve this imbalance in semantic space and visual space, we argue for developing hierarchical structures to gradually subdivide these components into meaningful leaf nodes. These leaf nodes can be presented using basic units including pixels and line segments but also basic structures, i.e., organs of human beings. In other words, the hierarchical structures help to maintain similar concepts in semantic space aligned with that of visual spaces. Several pioneer works[155, 169, 170] have explored the tree structures or hyper classes for fine-grained semantic structures. However, how to unify the semantic language embedding with the image-level visual features is still an under-explored problem. One promising direction is to unify the language and visual spaces using contrastive learning and mask modeling, including contrastive language-image pretraining (CLIP)[171], grounded language-image pre-training (GLIP)[172], and other multi-modality learning methods.

3D-aware fine-grained learning. In addition to the aforementioned 2D-based learning mechanisms, the other promising direction is 3D-aware fine-grained learning. In the semantic parsing of 3D models, many research efforts[173–176] have been proposed to parse 3D objects with point cloud, mesh and voxel representations. Thus an interesting question arises here: what is the relationship between 3D parsing models and the real-world 2D images? Earlier works collected in [177] propose to use 3D models to aid the recognition of human faces by handcrafted filters or template learning techniques. In the era of deep learning, several works[178] propose to embed the 3D canonical model with learnable warping parameters to represent diverse 2D images. The emerging field can be further boosted with learnable mechanisms including the neural rendering field[179].

6 Conclusions

In this paper, we present a comprehensive survey of fine-grained visual parsing tasks from the novel perspective of part relationship learning. In this view, we delve into the connections of two representative fine-grained tasks, i.e., fine-grained recognition and part segmentation, and propose a new taxonomy to reorganize recent research advances including the conventional methods and deep learning methods. By consolidating these works and popular benchmarks, we propose the universal challenges left in fine-grained visual parsing and make an attempted solution from the view of part relationship learning. Besides, we also point out several promising research directions that can be further explored. We hope these contributions will provide new inspiration to inform future research in the field of fine-grained visual parsing.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (Nos. 62132002, 61825101 and 62202010), the Key-Area Research and Development Program of Guangdong Province, China (No. 2021B010140 0002), and the China Postdoctoral Science Foundation (No. 2022M710212).

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

References

[1] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, F. F. Li. ImageNet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Miami, USA, pp. 248–255, 2009. DOI: 10.1109/CVPR.2009.5206848.

[2] X. D. Liang, X. H. Shen, D. L. Xiang, J. S. Feng, L. Lin, S. C. Yan. Semantic object parsing with local-global long short-term memory. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 3185–3193, 2016. DOI: 10.1109/CVPR.2016.347.

[3] T. Ruan, T. Liu, Z. L. Huang, Y. C. Wei, S. K. Wei, Y. Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Honolulu, USA, pp. 4814–4821, 2019. DOI: 10.1609/aaai.v33i01.33014814.

[4] Y. F. Zhao, J. Li, Y. Zhang, Y. F. Song, Y. H. Tian. Ordinal multi-task part segmentation with recurrent prior generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1636–1648, 2021. DOI: 10.1109/TPAMI.2019.2953854.

[5] F. T. Xia, P. Wang, X. J. Chen, A. L. Yuille. Joint multi-person pose estimation and semantic part segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 6780–6789, 2017. DOI: 10.1109/CVPR.2017.644.

[6] N. Zhang, J. Donahue, R. Girshick, T. Darrell. Part-based R-CNNs for fine-grained category detection. In *Proceedings of the 13th European Conference on Computer Vision*, Springer, Zurich, Switzerland, pp. 834–849, 2014. DOI: 10.1007/978-3-319-10590-1_54.

[7] S. L. Huang, Z. Xu, D. C. Tao, Y. Zhang. Part-stacked CNN for fine-grained visual categorization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1173–1182, 2016. DOI: 10.1109/CVPR.2016.132.

[8] X. T. He, Y. X. Peng. Weakly supervised learning of part

selection model with spatial constraints for fine-grained image classification. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, San Francisco, USA, pp. 4075–4081, 2017.

[9] X. S. Wei, C. W. Xie, J. X. Wu, C. H. Shen. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition*, vol. 76, pp. 704–714, 2018. DOI: 10.1016/j.patcog.2017.10.002.

[10] Z. X. Huang, Y. Li. Interpretable and accurate fine-grained recognition via region grouping. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 8659–8669, 2020. DOI: 10.1109/CVPR42600.2020.00869.

[11] V. T. Bickel, J. Aaron, A. Manconi, S. Loew, U. Mall. Impacts drive lunar rockfalls over billions of years. *Nature Communications*, vol. 11, no. 1, Article number 2862, 2020. DOI: 10.1038/s41467-020-16653-3.

[12] X. Sun, P. J. Wang, Z. Y. Yan, F. Xu, R. P. Wang, W. H. Diao, J. Chen, J. H. Li, Y. C. Feng, T. Xu, M. Weinmann, S. Hinz, C. Wang, K. Fu. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 184, pp. 116–130, 2022. DOI: 10.1016/j.isprsjprs.2021.12.004.

[13] D. Pakhomov, V. Premachandran, M. Allan, M. Azizian, N. Navab. Deep residual learning for instrument segmentation in robotic surgery. In *Proceedings of the 10th International Workshop on Machine Learning in Medical Imaging*, Springer, Shenzhen, China, pp. 566–573, 2019. DOI: 10.1007/978-3-030-32692-0_65.

[14] L. Zhu, C. X. Lin, H. D. Huang, Y. H. Chen, A. Yuille. Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion. In *Proceedings of the 10th European Conference on Computer Vision*, Springer, Marseille, France, pp. 759–773, 2008. DOI: 10.1007/978-3-540-88688-4_56.

[15] J. W. Hsieh, C. H. Chuang, S. Y. Chen, C. C. Chen, K. C. Fan. Segmentation of human body parts using deformable triangulation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, vol. 40, no. 3, pp. 596–610, 2010. DOI: 10.1109/TSMCA.2010.2040272.

[16] Y. Wang, D. Tran, Z. C. Liao. Learning hierarchical poselets for human parsing. In *Proceedings of IEEE/CVF Canference on Computer Vision and Pattern Recognition*, IEEE, Colorado Springs, USA, pp. 1705–1712, 2011. DOI: 10.1109/CVPR.2011.5995519.

[17] W. H. Lu, X. C. Lian, A. Yuille. Parsing semantic parts of cars using graphical models and segment appearance consistency. In *Proceedings of British Machine Vision Conference*, Nottingham, UK, 2014. DOI: 10.5244/C.28.118.

[18] N. Zhang, R. Farrell, T. Darrell. Pose pooling kernels for sub-category recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Providence, USA, pp. 3665–3672, 2012. DOI: 10.1109/CVPR.2012.6248364.

[19] S. Branson, P. Perona, S. Belongie. Strong supervision from weak annotation: Interactive training of deformable part models. In *Proceedings of International Conference on Computer Vision*, IEEE, Barcelona, Spain, pp. 1832–1839, 2011. DOI: 10.1109/ICCV.2011.6126450.

[20] B. Zhao, J. S. Feng, X. Wu, S. C. Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, vol. 14, no. 2, pp. 119–135, 2017. DOI: 10.1007/s11633-017-1053-3.

[21] X. S. Wei, Y. Z. Song, O. M. Aodha, J. X. Wu, Y. X. Peng, J. H. Tang, J. Yang, S. Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8927–8948, 2022. DOI: 10.1109/TPAMI.2021.3126648.

[22] D. De Geus, P. Meletis, C. Y. Lu, X. X. Wen, G. Dubbelman. Part-aware panoptic segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 5481–5490, 2021. DOI: 10.1109/CVPR46437.2021.00544.

[23] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, T. L. Berg. Parsing clothing in fashion photographs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Providence, USA, pp. 3570–3577, 2012. DOI: 10.1109/CVPR.2012.6248101.

[24] X. J. Chen, R. Mottaghi, X. B. Liu, S. Fidler, R. Urtasun, A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, pp. 1979–1986, 2014. DOI: 10.1109/CVPR.2014.254.

[25] J. Y. Wang, A. Yuille. Semantic part segmentation using compositional model combining shape and appearance. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 1788–1797, 2015. DOI: 10.1109/CVPR.2015.7298788.

[26] X. D. Liang, C. Y. Xu, X. H. Shen, J. C. Yang, S. Liu, J. H. Tang, L. Lin, S. C. Yan. Human parsing with contextualized convolutional neural network. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 1386–1394, 2015. DOI: 10.1109/ICCV.2015.163.

[27] L. C. Chen, Y. Yang, J. Wang, W. Xu, A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 3640–3649, 2016. DOI: 10.1109/CVPR.2016.396.

[28] J. S. Li, J. Zhao, Y. C. Wei, C. Y. Lang, Y. D. Li, T. Sim, S. C. Yan, J. S. Feng. Multiple-human parsing in the wild, [Online], Available: https://arxiv.org/abs/1705.07206, 2017.

[29] X. D. Liang, K. Gong, X. H. Shen, L. Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 871–885, 2019. DOI: 10.1109/TPAMI.2018.2820063.

[30] Q. X. Zhou, X. D. Liang, K. Gong, L. Lin. Adaptive temporal encoding network for video instance-level human parsing. In *Proceedings of the 26th ACM International Conference on Multimedia*, ACM, Seoul, Republic of Korea, pp. 1527–1535, 2018. DOI: 10.1145/3240508.3240660.

[31] K. Gong, X. D. Liang, Y. C. Li, Y. M. Chen, M. Yang, L. Lin. Instance-level human parsing via part grouping network. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 805–822, 2018. DOI: 10.1007/978-3-030-01225-0_47.

[32] Y. F. Zhao, J. Li, Y. Zhang, Y. H. Tian. Multi-class part parsing with joint boundary-semantic awareness. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, 2019. DOI: 10.1109/ICCV.2019.00927.

[33] Q. Liu, A. Kortylewski, Z. S. Zhang, Z. Z. Li, M. Q. Guo, Q. H. Liu, X. D. Yuan, J. T. Mu, W. C. Qiu, A. Yuille. Learning part segmentation through unsupervised domain adaptation from synthetic vehicles. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 19118–19129, 2022. DOI: 10.1109/CVPR52688.2022.01855.

[34] U. Michieli, P. Zanuttigh. Edge-aware graph matching network for part-based semantic segmentation. *International Journal of Computer Vision*, vol. 130, no. 11, pp. 2797–2821, 2022. DOI: 10.1007/s11263-022-01671-z.

[35] B. L. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba. Scene parsing through ADE20K dataset. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 5122–5130, 2017. DOI: 10.1109/CVPR.2017.544.

[36] H. S. Fang, G. S. Lu, X. L. Fang, J. W. Xie, Y. W. Tai, C. W. Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 70–78, 2018. DOI: 10.1109/CVPR.2018.00015.

[37] F. T. Xia, P. Wang, L. C. Chen, A. L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 648–663, 2016. DOI: 10.1007/978-3-319-46454-1_39.

[38] X. C. Nie, J. S. Feng, S. C. Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 519–534, 2018. DOI: 10.1007/978-3-030-01228-1_31.

[39] S. S. Li, J. Zhao, C. Y. Lang, Y. D. Li, Y. C. Wei, G. D. Guo, T. Sim, S. C. Yan, J. S. Feng. Multi-human parsing with a graph-based generative adversarial model. *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, no. 1, Article number 29, 2021. DOI: 10.1145/3418217.

[40] W. G. Wang, Z. J. Zhang, S. Y. Qi, J. B. Shen, Y. W. Pang, L. Shao. Learning compositional neural information fusion for human parsing. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 5702–5712, 2019. DOI: 10.1109/ICCV.2019.00580.

[41] K. Gong, Y. M. Gao, X. D. Liang, X. H. Shen, M. Wang, L. Lin. Graphonomy: Universal human parsing via graph transfer learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 7442–7451, 2019. DOI: 10.1109/CVPR.2019.00763.

[42] X. C. Liu, M. Zhang, W. Liu, J. K. Song, T. Mei. Braid-Net: Braiding semantics and details for accurate human parsing. In *Proceedings of the 27th ACM International Conference on Multimedia*, Nice, France, pp. 338–346, 2019. DOI: 10.1145/3343031.3350857.

[43] W. G. Wang, H. L. Zhu, J. F. Dai, Y. W. Pang, J. B. Shen, L. Shao. Hierarchical human parsing with typed part-relation reasoning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 8926–8936, 2020. DOI: 10.1109/CVPR42600.2020.00895.

[44] T. F. Zhou, W. G. Wang, S. Liu, Y. Yang, L. Van Gool. Differentiable multi-granularity human representation learning for instance-aware human semantic parsing. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 1622–1631, 2021. DOI: 10.1109/CVPR46437.2021.00167.

[45] D. Zeng, Y. H. Huang, Q. Bao, J. J. Zhang, C. Su, W. Liu. Neural architecture search for joint human parsing and pose estimation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 11365–11374, 2021. DOI: 10.1109/ICCV48922.2021.01119.

[46] Y. N. Liu, S. S. Zhang, J. Yang, P. C. Yuen. Hierarchical information passing based noise-tolerant hybrid learning for semi-supervised human parsing. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pp. 2207–2215, 2021. DOI: 10.1609/aaai.v35i3.16319.

[47] J. Zhao, J. S. Li, H. Z. Liu, S. C. Yan, J. S. Feng. Fine-grained multi-human parsing. *International Journal of Computer Vision*, vol. 128, no. 8, pp. 2185–2203, 2020. DOI: 10.1007/s11263-019-01181-5.

[48] L. Yang, Q. Song, Z. H. Wang, M. Jiang. Parsing R-CNN for instance-level human analysis. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 364–373, 2019. DOI: 10.1109/CVPR.2019.00045.

[49] P. K. Li, Y. Q. Xu, Y. C. Wei, Y. Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3260–3271, 2022. DOI: 10.1109/TPAMI.2020.3048039.

[50] H. Y. He, J. Zhang, Q. M. Zhang, D. C. Tao. Grapy-ML: Graph pyramid mutual learning for cross-dataset human parsing. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, USA, pp. 10949–10956, 2020. DOI: 10.1609/aaai.v34i07.6728.

[51] R. Y. Ji, D. W. Du, L. B. Zhang, L. Y. Wen, Y. J. Wu, C. Zhao, F. Y. Huang, S. W. Lyu. Learning semantic neural tree for human parsing. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 205–221, 2020. DOI: 10.1007/978-3-030-58601-0_13.

[52] S. Y. Zhang, G. J. Qi, X. C. Cao, Z. J. Song, J. Zhou. Human parsing with pyramidical gather-excite context. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1016–1030, 2021. DOI: 10.1109/TCSVT.2020.2990531.

[53] X. M. Zhang, Y. Y. Chen, B. K. Zhu, J. Q. Wang, M. Tang. Part-aware context network for human parsing. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 8968–8977, 2020. DOI: 10.1109/CVPR42600.2020.00899.

[54] A. Loesch, R. Audigier. Describe me if you can! characterized instance-level human parsing. In *Proceedings of IEEE International Conference on Image Processing*, Anchorage, USA, pp. 2528–2532, 2021. DOI: 10.1109/

ICIP42928.2021.9506509.

[55] Y. F. Song, X. W. Chen, J. Li, Q. P. Zhao. Embedding 3D geometric features for rigid object part segmentation. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 580–588, 2017. DOI: 10.1109/ICCV.2017.70.

[56] P. Wang, X. H. Shen, Z. Lin, S. Cohen, B. Price, A. Yuille. Joint object and part segmentation using deep learned potentials. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 1573–1581, 2015. DOI: 10.1109/ICCV.2015.184.

[57] S. Naha, Q. Y. Xiao, P. Banik, A. Reza, D. J. Crandall. Part segmentation of unseen objects using keypoint guidance. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, USA, pp. 1741–1749, 2021. DOI: 10.1109/WACV48630.2021.00178.

[58] Z. H. Wu, G. S. Lin, J. F. Cai. Keypoint based weakly supervised human parsing. *Image and Vision Computing*, vol. 91, Article number 103801, 2019. DOI: 10.1016/j.imavis.2019.08.005.

[59] Z. Y. Yang, Y. C. Li, L. J. Yang, N. Zhang, J. B. Luo. Weakly supervised body part segmentation with pose based part priors. In *Proceedings of the 25th International Conference on Pattern Recognition*, IEEE, Milan, Italy, pp. 286–293, 2021. DOI: 10.1109/ICPR48806.2021.9412887.

[60] Y. F. Zhao, J. Li, Y. Zhang, Y. H. Tian. From pose to part: Weakly-supervised pose evolution for human part segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3107–3120, 2023. DOI: 10.1109/TPAMI.2022.3174529.

[61] Y. Yang, X. T. Cheng, H. Bilen, X. Y. Ji. Learning to annotate part segmentation with gradient matching. In *Proceedings of the 10th International Conference on Learning Representations*, 2021.

[62] A. Gonzalez-Garcia, D. Modolo, V. Ferrari. Do semantic parts emerge in convolutional neural networks? *International Journal of Computer Vision*, vol. 126, no. 5, pp. 476–494, 2018. DOI: 10.1007/s11263-017-1048-0.

[63] D. Lorenz, L. Bereska, T. Milbich, B. Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 10947–10956, 2019. DOI: 10.1109/CVPR.2019.01121.

[64] W. C. Hung, V. Jampani, S. F. Liu, P. Molchanov, M. H. Yang, J. Kautz. SCOPS: Self-supervised co-part segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 869–878, 2019. DOI: 10.1109/CVPR.2019.00096.

[65] Q. Z. Gao, B. Wang, L. B. Liu, B. Q. Chen. Unsupervised co-part segmentation through assembly. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 3576–3586, 2021.

[66] S. L. Liu, L. Zhang, X. Yang, H. Su, J. Zhu. Unsupervised part segmentation through disentangling appearance and shape. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 8351–8360, 2021. DOI: 10.1109/CVPR46437.2021.00825.

[67] S. Choudhury, I. Laina, C. Rupprecht, A. Vedaldi. Unsupervised part discovery from contrastive reconstruction. In *Proceedings of the 35th Neural Information Processing Systems*, pp. 28104–28118, 2021.

[68] U. Michieli, E. Borsato, L. Rossi, P. Zanuttigh. GMNet: Graph matching network for large scale part semantic segmentation in the wild. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 397–414, 2020. DOI: 10.1007/978-3-030-58598-3_24.

[69] Xin Tan, J. C. Xu, Z. Ye, J. K. Hao, L. Z. Ma. Confident semantic ranking loss for part parsing. In *Proceedings of IEEE International Conference on Multimedia and Expo*, Shenzhen, China, 2021. DOI: 10.1109/ICME51207.2021.9428332.

[70] R. Singh, P. Gupta, P. Shenoy, R. Sarvadevabhatla. Float: Factorized learning of object attributes for improved multi-object multi-part scene parsing. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 1435–1445, 2022. DOI: 10.1109/CVPR52688.2022.00150.

[71] A. Kirillov, K. M. He, R. Girshick, C. Rother, P. Dollár. Panoptic segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 9396–9405, 2019. DOI: 10.1109/CVPR.2019.00963.

[72] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015. DOI: 10.1007/s11263-014-0733-5.

[73] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 3213–3223, 2016. DOI: 10.1109/CVPR.2016.350.

[74] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010. DOI: 10.1109/TPAMI.2009.167.

[75] S. M. A. Eslami, C. K. I. Williams. A generative model for parts-based object segmentation. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Lake Tahoe, USA, pp. 100–107, 2012.

[76] S. Liu, J. S. Feng, C. Domokos, H. Xu, J. S. Huang, Z. Z. Hu, S. C. Yan. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 253–265, 2014. DOI: 10.1109/TMM.2013.2285526.

[77] F. M. Meng, H. L. Li, Q. B. Wu, K. N. Ngan, J. F. Cai. Seeds-based part segmentation by seeds propagation and region convexity decomposition. *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 310–322, 2018. DOI: 10.1109/TMM.2017.2739919.

[78] C. Desai, D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *Proceedings of the 12th European Conference on Computer Vision*, Springer, Florence, Italy, pp. 158–172, 2012. DOI: 10.1007/978-3-642-33765-9_12.

[79] H. Azizpour, I Laptev. Object detection using strongly-

supervised deformable part models. In *Proceedings of the 12th European Conference on Computer Vision*, Springer, Florence, Italy, pp. 836–849, 2012. DOI: 10.1007/978-3-642-33718-5_60.

[80] J. Dong, Q. Chen, Z. Y. Huang, J. C. Yang, S. C. Yan. Parsing based on Parselets: A unified deformable mixture model for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 88–101, 2016. DOI: 10.1109/TPAMI.2015.2420563.

[81] F. T. Xia, J. Zhu, P. Wang, A. L. Yuille. Pose-guided human parsing by an and/or graph using pose-context features. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Phoenix, USA, pp. 3632–3640, 2016.

[82] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[83] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA, 2015.

[84] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 1026–1034, 2015. DOI: 10.1109/ICCV.2015.123.

[85] J. Long, E. Shelhamer, T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 3431–3440, 2015. DOI: 10.1109/CVPR.2015.7298965.

[86] Y. Yang, D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of IEEE/CVF Conferenceon Computer Vision and Pattern Recognition*, IEEE, Colorado Springs, USA, pp. 1385–1392, 2011. DOI: 10.1109/CVPR.2011.5995741.

[87] J. Dong, Q. Chen, X. H. Shen, J. C. Yang, S. C. Yan. Towards unified human parsing and pose estimation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, pp. 843–850, 2014. DOI: 10.1109/CVPR.2014.113.

[88] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018. DOI: 10.1109/TPAMI.2017.2699184.

[89] L. C. Chen, G. Papandreou, F. Schroff, H. Adam. Rethinking atrous convolution for semantic image segmentation, [Online], Available: https://arxiv.org/abs/1706.05587, 2017.

[90] X. D. Liang, X. H. Shen, J. S. Feng, L. Lin, S. C. Yan. Semantic object parsing with graph LSTM. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 125–143, 2016. DOI: 10.1007/978-3-319-46448-0_8.

[91] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset, Technical Report CNS-TR-2011-001, California Institute of Technology, Pasadena, USA, 2011.

[92] M. E. Nilsback, A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the 6th Indian Conference on Computer Vision, Graphics & Image Processing*, IEEE, Bhubaneswar, India, pp. 722–729, 2008. 10.1109/ICVGIP.2008.47.

[93] A. Khosla, N. Jayadevaprakash, B. Yao, F. F. Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proceedings of CVPR Workshop on Fine-grained Visual Categorization*, vol. 2, Article number 1, 2011.

[94] J. Krause, M. Stark, J. Deng, L. Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, Sydney, Australia, pp. 554–561, 2013. DOI: 10.1109/ICCVW.2013.77.

[95] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, A. Vedaldi. Fine-grained visual classification of aircraft, [Online], Available: https://arxiv.org/abs/1306.5151, 2013.

[96] L. Bossard, M. Guillaumin, L. Van Gool. Food-101-mining discriminative components with random forests. In *Proceedings of the 13th European Conference on Computer Vision*, Springer, Zurich, Switzerland, pp. 446–461, 2014. DOI: 10.1007/978-3-319-10599-4_29.

[97] T. Berg, J. X. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, P. N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, pp. 2019–2026, 2014. DOI: 10.1109/CVPR.2014.259.

[98] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, S. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 595–604, 2015. DOI: 10.1109/CVPR.2015.7298658.

[99] L. J. Yang, P. Luo, C. C. Loy, X. O. Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 3973–3981, 2015. DOI: 10.1109/CVPR.2015.7299023.

[100] Z. W. Liu, P. Luo, S. Qiu, X. G. Wang, X. O. Tang. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1096–1104, 2016. DOI: 10.1109/CVPR.2016.124.

[101] G. Van Horn, O. M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, S. Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 8769–8778, 2018. DOI: 10.1109/CVPR.2018.00914.

[102] M. Sun, Y. C. Yuan, F. Zhou, E. R. Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 834–850, 2018. DOI: 10.1007/978-3-030-01270-0_49.

[103] G. Van Horn, E. Cole, S. Beery, K. Wilber, S. Belongie, O. MacAodha. Benchmarking representation learning for natural world image collections. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 12897–12888,

2021. DOI: 10.1109/CVPR46437.2021.01269.

[104] P. Q. Zhuang, Y. L. Wang, Y. Qiao. WildFish: A large benchmark for fish recognition in the wild. In *Proceedings of the 26th ACM International Conference on Multimedia*, Seoul, Republic of Korea, pp. 1301–1309, 2018. DOI: 10.1145/3240508.3240616.

[105] T. Weyand, A. Araujo, B. Y. Cao, J. Sim. Google landmarks dataset v2 – A large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 2572-2581, 2020. DOI: 10.1109/CVPR42600.2020.00265.

[106] B. P. Yao, A. Khosla, F. F. Li. Combining randomization and discrimination for fine-grained image categorization. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recogntion*, IEEE, Colorado Springs, USA, pp. 1577–1584, 2011. DOI: 10.1109/CVPR.2011.5995368.

[107] B. P. Yao, G. Bradski, F. F. Li. A codebook-free and annotation-free approach for fine-grained image categorization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Providence, USA, pp. 3466–3473. IEEE, 2012. DOI: 10.1109/CVPR.2012.6248088.

[108] C. Göering, E. Rodner, A. Freytag, J. Denzler. Nonparametric part transfer for fine-grained recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, pp. 2489–2496, 2014. DOI: 10.1109/CVPR.2014.319.

[109] C. Wah, S. Branson, P. Perona, S. Belongie. Multiclass recognition and part localization with humans in the loop. In *Proceedings of International Conference on Computer Vision*, IEEE, Barcelona, Spain, pp. 2524–2531, 2011. DOI: 10.1109/ICCV.2011.6126539.

[110] B. He, J. Li, Y. F. Zhao, Y. H. Tian. Part-regularized near-duplicate vehicle re-identification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 3992–4000, 2019. DOI: 10.1109/CVPR.2019.00412.

[111] Y. X. Peng, X. T. He, J. J. Zhao. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1487–1500, 2018. DOI: 10.1109/TIP.2017.2774041.

[112] D. Q. Wang, Z. Q. Shen, J. Shao, W. Zhang, X. Y. Xue, Z. Zhang. Multiple granularity descriptors for fine-grained categorization. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 2399–2406, 2015. DOI: 10.1109/ICCV.2015.276.

[113] J. Krause, H. L. Jin, J. C. Yang, F. F. Li. Fine-grained recognition without part annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 5546–5555, 2015. DOI: 10.1109/CVPR.2015.7299194.

[114] M. Simon, E. Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 1143–1151, 2015. DOI: 10.1109/ICCV.2015.136.

[115] Y. Zhang, X. S. Wei, J. X. Wu, J. F. Cai, J. B. Lu, V. A. Nguyen, M. N. Do. Weakly supervised fine-grained categorization with part-based image representation. *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1713–1725, 2016. DOI: 10.1109/TIP.2016.2531289.

[116] J. L. Fu, H. L. Zheng, T. Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 4476–4484, 2017. DOI: 10.1109/CVPR.2017.476.

[117] A. Recasens, P. Kellnhofer, S. Stent, W. Matusik, A. Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 52–67, 2018. DOI: 10.1007/978-3-030-01240-3_4.

[118] Z. H. Wang, S. J. Wang, S. H. Yang, H. J. Li, J. J. Li, Z. Z. Li. Weakly supervised fine-grained image classification via Guassian mixture model oriented discriminative learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 9746–9755, 2020. DOI: 10.1109/CVPR42600.2020.00977.

[119] W. F. Ge, X. R. Lin, Y. Z. Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 3029–3038, 2019. DOI: 10.1109/CVPR.2019.00315.

[120] G. L. Sun, H. Cholakkal, S. Khan, F. Khan, L. Shao. Fine-grained recognition: Accounting for subtle differences between similar classes. In *Proceedings of AAAI Conference on Artificial Intelligence*, New York, USA, pp. 12047–12054, 2020. DOI: 10.1609/aaai.v34i07.6882.

[121] H. L. Zheng, J. L. Fu, Z. J. Zha, J. B. Luo, T. Mei. Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. *IEEE Transactions on Image Processing*, vol. 29, pp. 476–488, 2020. DOI: 10.1109/TIP.2019.2921876.

[122] Y. Ding, Y. Z. Zhou, Y. Zhu, Q. X. Ye, J. B. Jiao. Selective sparse sampling for fine-grained image recognition. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 6598–6607, 2019. DOI: 10.1109/ICCV.2019.00670.

[123] Z. H. Wang, S. J. Wang, H. J. Li, Z. Dou, J. J. Li. Graph-propagation based correlation learning for weakly supervised fine-grained image classification. In *Proceedings of AAAI Conference on Artificial Intelligence*, New York, USA, pp. 12289–12296, 2020. DOI: 10.1609/aaai.v34i07.6912.

[124] M. Lam, B. Mahasseni, S. Todorovic. Fine-grained recognition as hsnet search for informative image parts. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 6497–6506, 2017. DOI: 10.1109/CVPR.2017.688.

[125] Y. F. Zhao, J. Li, X. W. Chen, Y. H. Tian. Part-guided relational transformers for fine-grained visual recognition. *IEEE Transactions on Image Processing*, vol. 30, pp. 9470–9481, 2021. DOI: 10.1109/TIP.2021.3126490.

[126] R. Y. Ji, L. Y. Wen, L. B. Zhang, D. W. Du, Y. J. Wu, C. Zhao, X. L. Liu, F. Y. Huang. Attention convolutional binary neural tree for fine-grained visual categorization. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 10465–10474, 2020. DOI: 10.1109/CVPR42600.2020.

01048.

[127] M. Nauta, R. Van Bree, C. Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 14928–14938, 2021. DOI: 10.1109/CVPR46437.2021.01469.

[128] Z. Yang, T. G. Luo, D. Wang, Z. Q. Hu, J. Gao, L. W. Wang. Learning to navigate for fine-grained classification. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 438–454, 2018. DOI: 10.1007/978-3-030-01264-9_26.

[129] Y. M. Wang, V. I. Morariu, L. S. Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 4148–4157, 2018. DOI: 10.1109/CVPR.2018.00436.

[130] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, N. Naik. Pairwise confusion for fine-grained visual classification. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 71–88, 2018. DOI: 10.1007/978-3-030-01258-8_5.

[131] O. M. Aodha, E. Cole, P. Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 9595–9605, 2019. DOI: 10.1109/ICCV.2019.00969.

[132] A. Dubey, O. Gupta, R. Raskar, N. Naik. Maximum entropy fine-grained classification. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 635–645, 2018.

[133] Y. Cui, Y. Song, C. Sun, A. Howard, S. Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 4109–4118, 2018. DOI: 10.1109/CVPR.2018.00432.

[134] X. W. Zheng, R. R. Ji, X. H. Sun, B. C. Zhang, Y. J. Wu, F. Y. Huang. Towards optimal fine grained retrieval via decorrelated centralized loss with normalize-scale layer. In *Proceedings of the 33th AAAI Conference on Artificial Intelligence*, Honolulu, USA, pp. 9291–9298, 2019. DOI: 10.1609/aaai.v33i01.33019291.

[135] Y. F. Zhao, K. Yan, F. Y. Huang, J. Li. Graph-based high-order relation discovery for fine-grained recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 15074–15083, 2021. DOI: 10.1109/CVPR46437.2021.01483.

[136] T. Y. Lin, A. RoyChowdhury, S. Maji. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 1449–1457, 2015. DOI: 10.1109/ICCV.2015.170.

[137] Y. Gao, O. Beijbom, N. Zhang, T. Darrell. Compact bilinear pooling. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 317–326, 2016. DOI: 10.1109/CVPR.2016.41.

[138] Y. H. Li, N. Y. Wang, J. Y. Liu, X. D. Hou. Factorized bilinear models for image recognition. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 2098–2106, 2017. DOI: 10.1109/ICCV.2017.229.

[139] X. Wei, Y. Zhang, Y. H. Gong, J. W. Zhang, N. N. Zheng. Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 365–380, 2018. DOI: 10.1007/978-3-030-01219-9_22.

[140] H. L. Zheng, J. L. Fu, Z. J. Zha, J. B. Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 5007–5016, 2019. DOI: 10.1109/CVPR.2019.00515.

[141] Y. Gao, X. T. Han, X. Wang, W. L. Huang, M. Scott. Channel interaction networks for fine-grained image categorization. In *Proceedings of 34th AAAI Conference on Artificial Intelligence*, New York, USA, pp. 10818–10825, 2020. DOI: 10.1609/aaai.v34i07.6712.

[142] X. S. Wei, J. H. Luo, J. X. Wu, Z. H. Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2868–2881, 2017. DOI: 10.1109/TIP.2017.2688133.

[143] C. J. Yu, X. Y. Zhao, Q. Zheng, P. Zhang, X. G. You. Hierarchical bilinear pooling for fine-grained visual recognition. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 595–610, 2018. DOI: 10.1007/978-3-030-01270-0_35.

[144] L. B. Zhang, S. L. Huang, W. Liu, D. C. Tao. Learning a mixture of granularity-specific experts for fine-grained categorization. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 8330–8339, 2019. DOI: 10.1109/ICCV.2019.00842.

[145] X. L. Wang, R. Girshick, A. Gupta, K. M. He. Non-local neural networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 7794–7803, 2018. DOI: 10.1109/CVPR.2018.00813.

[146] S. Kong, C. Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 7025–7034, 2017. DOI: 10.1109/CVPR.2017.743.

[147] Y. Chen, Y. L. Bai, W. Zhang, T. Mei. Destruction and construction learning for fine-grained image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 5152–5161, 2019. DOI: 10.1109/CVPR.2019.00530.

[148] W. Luo, X. T. Yang, X. J. Mo, Y. H. Lu, L. Davis, J. Li, J. Yang, S. N. Lim. Cross-x learning for fine-grained visual categorization. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 8241–8250, 2019. DOI: 10.1109/ICCV.2019.00833.

[149] P. Q. Zhuang, Y. L. Wang, Y. Qiao. Learning attentive pairwise interaction for fine-grained classification. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, USA, pp. 13130–13137, 2020. DOI: 10.

1609/aaai.v34i07.7016.

[150] C. B. Liu, H. T. Xie, Z. J. Zha, L. Y. Yu, Z. N. Chen, Y. D. Zhang. Bidirectional attention-recognition model for fine-grained object classification. *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1785–1795, 2019. DOI: 10.1109/TMM.2019.2954747.

[151] P. Rodríguez, J. M. Gonfaus, G. Cucurull, F. X. Roca, J. Gonzàlez. Attend and rectify: A gated attention mechanism for fine-grained recovery. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 357–372, 2018. DOI: 10.1007/978-3-030-01237-3_22.

[152] C. B. Liu, H. T. Xie, Z. J. Zha, L. F. Ma, L. Y. Yu, Y. D. Zhang. Filtration and distillation: Enhancing region attention for fine-grained visual categorization. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, USA, pp. 11555–11562, 2020. DOI: 10.1609/aaai.v34i07.6822.

[153] D. Wu, S. Y. Li, Z. L. Zang, K. Wang, L. Shang, B. G. Sun, H. Li, S. Z. Li. Align yourself: Self-supervised pre-training for fine-grained recognition via saliency alignment, [Online], Available: https://arxiv.org/abs/2106.15788, 2021.

[154] J. B. Wang, Y. Li, X. S. Wei, H. Li, Z. Miao, R. Zhang. Bridge the gap between supervised and unsupervised learning for fine-grained classification, [Online], Available: https://arxiv.org/abs/2203.00441, 2022.

[155] D. L. Chang, K. Y. Pang, Y. X. Zheng, Z. Y. Ma, Y. Z. Song, J. Guo. Your "flamingo" is my "bird": Fine-grained, or not. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 11471–11480, 2021. DOI: 10.1109/CVPR46437.2021.01131.

[156] M. H. Zhou, Y. L. Bai, W. Zhang, T. J. Zhao, T. Mei. Look-into-object: Self-supervised structure modeling for object recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 11771–11780, 2020. DOI: 10.1109/CVPR42600.2020.01179.

[157] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, M. Shah. Human semantic parsing for person re-identification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 1062–1071, 2018. DOI: 10.1109/CVPR.2018.00117.

[158] D. C. Meng, L. Li, X. J. Liu, Y. D. Li, S. J. Yang, Z. J. Zha, X. Y. Gao, S. H. Wang, Q. M. Huang. Parsing-based view-aware embedding network for vehicle re-identification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 7101–7110, 2020. DOI: 10.1109/CVPR42600.2020.00713.

[159] J. J. Zhao, Y. F. Zhao, J. Li, K. Yan, Y. H. Tian. Heterogeneous relational complement for vehicle re-identification. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 205–214, 2021. DOI: 10.1109/ICCV48922.2021.00027.

[160] W. C. Chen, X. Y. Yu, L. L. Ou. Pedestrian attribute recognition in video surveillance scenarios based on view-attribute attention localization. *Machine Intelligence Research*, vol. 19, no. 2, pp. 153–168, 2022. DOI: 10.1007/s11633-022-1321-8.

[161] B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015. DOI: 10.1126/science.aab3050.

[162] F. F. Li, R. Fergus, P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006. DOI: 10.1109/TPAMI.2006.79.

[163] E. G. Miller, N. E. Matsakis, P. A. Viola. Learning from one example through shared densities on transforms. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, USA, pp. 464–471, 2000. DOI: 10.1109/CVPR.2000.855856.

[164] L. M. Tang, D. Wertheimer, B. Hariharan. Revisiting pose-normalization for fine-grained few-shot recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 14340–14349, 2020. DOI: 10.1109/CVPR42600.2020.01436.

[165] P. Koniusz, H. G. Zhang. Power normalizations in fine-grained image, few-shot image and graph classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 591–609, 2022. DOI: 10.1109/TPAMI.2021.3107164.

[166] H. X. Huang, J. J. Zhang, J. Zhang, J. S. Xu, Q. Wu. Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. *IEEE Transactions on Multimedia*, vol. 23, pp. 1666–1680, 2021. DOI: 10.1109/TMM.2020.3001510.

[167] Y. H. Zhu, C. L. Liu, S. Q. Jiang. Multi-attention meta learning for few-shot fine-grained image recognition. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, Yokohama, Japan, pp. 1090–1096, 2020.

[168] X. Li, J. Wu, Z., Sun, Z. Ma, J. Cao, J. H. Xue. BSNet: Bi-similarity network for few-shot fine-grained image classification. *IEEE Transactions on Image Processing*, vol. 30, pp. 1318–1331, 2020.

[169] X. F. Zhang, F. Zhou, Y. Q. Lin, S. T. Zhang. Embedding label structures for fine-grained feature representation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1114–1123, 2016. DOI: 10.1109/CVPR.2016.126.

[170] S. N. Xie, T. B. Yang, X. Y. Wang, Y. Q. Lin. Hyper-class augmented and regularized deep learning for fine-grained image classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 2645–2654, 2015. DOI: 10.1109/CVPR.2015.7298880.

[171] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

[172] L. H. Li, P. C. Zhang, H. T. Zhang, J. W. Yang, C. Y. Li, Y. W. Zhong, L. J. Wang, L. Yuan, L. Zhang, J. N. Hwang, K. W. Chang, J. F. Gao. Grounded language-image pre-training. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 10955–10965, 2022. DOI:

10.1109/CVPR52688.2022.01069.

[173] E. Kalogerakis, M. Averkiou, S. Maji, S. Chaudhuri. 3D shape segmentation with projective convolutional networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 6630–6639, 2017. DOI: 10.1109/CVPR.2017.702.

[174] F. G. Yu, K. Liu, Y. Zhang, C. Y. Zhu, K. Xu. PartNet: A recursive part decomposition network for fine-grained and hierarchical shape segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 9483–9492, 2019. DOI: 10.1109/CVPR.2019.00972.

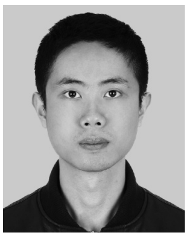
[175] K. C. Mo, S. L. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, H. Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 909–918, 2019. DOI: 10.1109/CVPR.2019.00100.

[176] R. Q. Charles, H. Su, K. C. Mo, L. J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 77–85, 2017. DOI: 10.1109/CVPR.2017.16.

[177] J. Kittler, A. Hilton, M. Hamouz, J. Illingworth. 3D assisted face recognition: A survey of 3d imaging, modelling and recognition approachest. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, USA, pp. 114–114, 2005. DOI: 10.1109/CVPR.2005.377.

[178] S. Joung, S. Kim, M. Kim, I. J. Kim, K. Sohn. Learning canonical 3D object representation for fine-grained recognition. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 1015–1025, 2021. DOI: 10.1109/ICCV48922.2021.00107.

[179] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021. DOI: 10.1145/3503250.



Yifan Zhao received the B. Eng. degree in computer science from Harbin Institute of Technology, China in 2016, and the Ph. D. degree in computer science from School of Computer Science and Engineering, Beihang University, China in 2021. He is currently a postdoctoral researcher with School of Computer Science, Peking University, China.

His research interests include computer vision and image/video understanding.

E-mail: zhaoyf@pku.edu.cn
ORCID: 0000-0002-5691-013X

Jia Li received the B. Eng. degree in computer science from Tsinghua University, China in 2005, and the Ph. D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences, China in 2011. He is currently a full professor with State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, China. Before he joined Beihang University in 2014, he used to work at Nanyang Technological University, Singapore, Shanda Innovations, China and Peking University, China. He has co-authored more than 110 articles in peer-reviewed top-tier journals and conferences. He also has one Monograph published by Springer and more than 60 patents issued from USA and China. He is a fellow of IET, and senior members of IEEE/ ACM/CCF/CIE.

His research interests include computer vision, multimedia and artificial intelligence, especially the visual computing in extreme environments.

E-mail: jiali@buaa.edu.cn (Corresponding author)
ORCID: 0000-0002-4346-8696

Yonghong Tian received the Ph. D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences, China in 2005. He is currently a Boya distinguished professor with School of Computer Science, Peking University, China, and is also the deputy director of Artificial Intelligence Research Center, Peng Cheng Laboratory, China. He is the author or coauthor of over 280 technical articles in refereed journals and conferences. He was/is an Associate Editor of IEEE TCSVT (2018.1–2021.12), IEEE TMM (2014.8–2018.8), *IEEE Multimedia Magazine* (2018.1–), and *IEEE Access* (2017.1–). He Co-initiated *IEEE International Conference on Multimedia Big Data* (BigMM) and served as the TPC Co-chair of BigMM 2015, and aslo served as the Technical Program Co-chair of IEEE ICME 2015, IEEE ISM 2015 and IEEE MIPR 2018/2019, and General Co-chair of IEEE MIPR 2020 and ICME 2021. He is the steering member of IEEE ICME (2018–2020) and IEEE BigMM (2015–), and is a TPC Member of more than ten conferences such as CVPR, ICCV, ACM KDD, AAAI, ACM MM and ECCV. He was the recipient of the Chinese National Science Foundation for Distinguished Young Scholars in 2018, two National Science and Technology Awards and three ministerial-level awards in China, and obtained the 2015 EURASIP Best Paper Award for *Journal on Image and Video Processing*, and the best paper award of IEEE BigMM 2018. He is a Fellow of IEEE, a Senior Member of CIE and CCF, a Member of ACM.

His research interests include neuromorphic vision, distributed machine learning and multimedia big data.

E-mail: yhtian@pku.edu.cn