A Survey of the Four Pillars for Small Object Detection: Multiscale Representation, Contextual Information, Super-Resolution, and Region Proposal

Guang Chen, *Member, IEEE*, Haitao Wang, Kai Chen, Zhijun Li, *Senior Member, IEEE*, Zida Song, Yinlong Liu, Wenkai Chen, and Alois Knoll, *Senior Member, IEEE*

***Abstract*—Although great progress has been made in generic object detection by advanced deep learning techniques, detecting small objects from images is still a difficult and challenging problem in the field of computer vision due to the limited size, less appearance, and geometry cues, and the lack of large-scale datasets of small targets. Improving the performance of small object detection has a wider significance in many real-world applications, such as self-driving cars, unmanned aerial vehicles, and robotics. In this article, the first-ever survey of recent studies in deep learning-based small object detection is presented. Our review begins with a brief introduction of the four pillars for small object detection, including *multiscale representation, contextual information, super-resolution, and region-proposal*. Then, the collection of state-of-the-art datasets for small object detection is listed. The performance of different methods on these datasets is reported later. Moreover, the state-of-the-art small object detection networks are investigated along with a special focus on the differences and modifications to improve the detection performance comparing to generic object detection architectures. Finally, several promising directions and tasks for future work in small object detection are provided. Researchers can track up-to-date studies on this webpage available at: https://github.com/tjtum-chenlab/SmallObjectDetectionList.**

Manuscript received January 1, 2020; revised April 26, 2020; accepted June 8, 2020. Date of publication July 17, 2020; date of current version January 17, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61906138; in part by the State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body Open Project under Grant 31815005; in part by the European Union's Horizon 2020 Research and Innovation Program under Grant 785907; and in part by the Shanghai AI Innovation Development Program 2018. This article was recommended by Associate Editor Z. Liu. *(Corresponding author: Zhijun Li.)*

Guang Chen is with the School of Automotive Studies, Tongji University, Shanghai 200092, China, also with the State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, Hunan University, Changsha 410082, China, and also with the Department of Informatics, Technical University of Munich, 80333 Munich, Germany.

Haitao Wang, Kai Chen, and Zida Song are with the School of Automotive Studies, Tongji University, Shanghai 200092, China.

Zhijun Li is with the Department of Automation, University of Science and Technology of China, Hefei 230000, China (e-mail: zjli@ieee.org).

Yinlong Liu and Alois Knoll are with the Department of Informatics, Technical University of Munich, 80333 Munich, Germany.

Wenkai Chen is with the School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSMC.2020.3005231.

Digital Object Identifier 10.1109/TSMC.2020.3005231

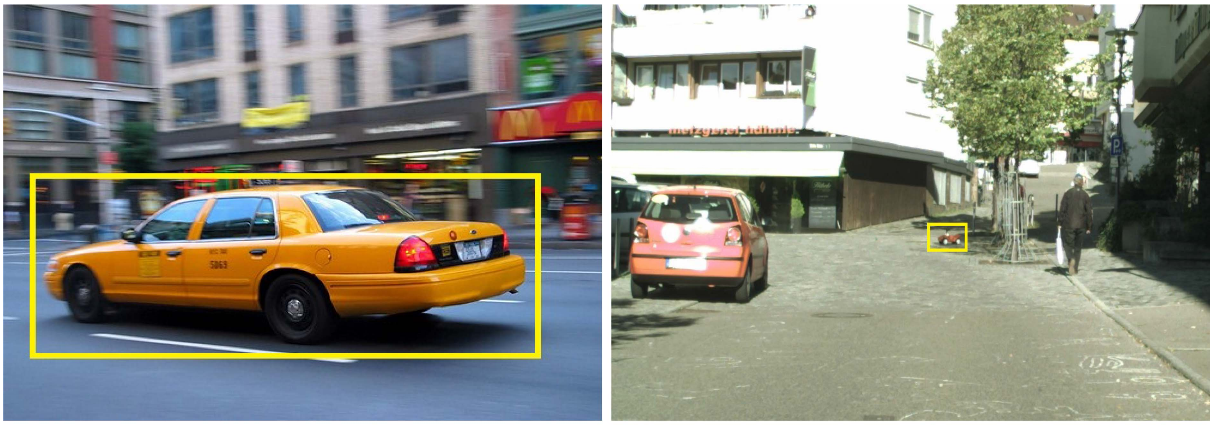
***Index Terms*—Contextual information, multiscale representation, region proposal, small object dataset, small object detection, super-resolution.**

I. Introduction

OBJECT detection consists of two subtasks, that is, localization and classification, indicating that not only all object instances need to be accurately located in an image but also their categories should be correctly recognized. As a promising technology related to computer vision, object detection has been applied to many application scenes, such as pedestrian detection [1], [2], face detection [3]–[5], autonomous driving [6]–[8], and robotic vision [9]–[12]. More and more object detection tasks are successfully implemented, because of the continuous development of deep learning techniques [13]. Current state-of-the-art detection models, such as mask R-CNN [14], cascade R-CNN [15], and hybrid task cascade [16], have achieved great performances on the large image datasets, such as MS COCO [17], PASCAL VOC [18], and ImageNet [19].

Before the advent of deep learning techniques, object detection task has been studied for several decades [20], [21]. Different methods, such as SIFT [22], histograms of oriented gradient (HOG) [23], SPM [24], DPM [25], and Selective Search [26], have been proposed by researchers to extract local handcrafted features. These methods have also achieved excellent detection performance on specific applications with a small-scale dataset. For example, HOG descriptor [23] could extract line features quickly by counting local gradient information. However, handcrafted features cannot capture multiple levels of representation for large-scale datasets, such as MS COCO [17]; therefore, these traditional handcrafted methods are less robust to intro-class variability due to the failure of representing the semantics of the data.

The deep convolution neural network (DCNN) [27] was proposed to autonomously learn features in order to overcome these drawbacks of traditional handcrafted features; this method exhibits powerful detection performance on generic object detection [14], [16], [28]–[31]. However, as a subcategory of object detection, small object detection has been



(a)

(b)

Fig. 1. Comparison between small object detection and generic object detection. (a) Common object image from the ImageNet dataset [19]. Target taxi covers about 34.4% in this figure. (b) Small object image from the Lost and Found dataset [32]. The small target just occupies approximately 0.3% of the whole image.

TABLE I

PRECISION COMPARISON OF SEVERAL LEADING GENERIC OBJECT DETECTION ALGORITHMS ON THE MS COCO DATASET [17]

Model	**Backbone**	**Dataset**	**Avg.Precision,IoU:**			**Avg.Precision,Area:**			**DOR**
			0.5:0.95	0.5	0.75	S	M	L	
YOLOv2 [33]	DarkNet-19	COCO 2015 test-dev	21.6	44	19.2	5	22.4	35.5	**30.5**
RetinaNet [34]	ResNet-101	COCO 2015 test-dev	39.1	59.1	42.3	21.8	42.7	50.2	**28.4**
SSD513 [35]	ResNet-101	COCO 2015 test-dev	31.2	50.4	33.3	10.2	34.5	49.8	**39.6**
DSSD513 [36]	Residual-101	COCO 2015 test-dev	33.2	53.3	35.2	13	35.4	51.1	**38.1**
Faster R-CNN [30]	VGG	COCO 2015 test-dev	26.9	44.3	27.8	8.3	28.2	41.1	**32.8**
FPN [37]	ResNet-50	COCO 2014 minival	33.9	56.9	-	17.8	37.7	45.8	**28.0**
Mask R-CNN [14]	ResNet-101	COCO 2015 test-dev	38.2	60.3	41.7	20.1	41.1	50.2	**30.1**
Double-Head R-CNN [38]	ResNet-101	COCO 2014 minival	41.9	62.4	45.9	23.9	45.2	55.8	**31.9**
Cascade R-CNN [15]	ResNet-101	COCO 2015 test-dev	42.8	62.1	46.3	23.7	45.5	55.2	**31.5**
Hybrid Task Cascade [16]	ResNet-101-FPN	COCO 2015 test-dev	47.1	63.9	44.7	22.8	43.9	54.6	**31.8**

grossly neglected. Different from the rapid progress of the generic object detection, small object detection has not been addressed very well. This motivates us to provide the first-ever survey of recent studies in vision-based small object detection.

A. *Motivation*

A detailed definition of small objects could be illustrated with different aspects. For example, [17] illustrates that the length and width pixels of small objects bounding box should be less than 32 and [39] states that the bounding box of small object should cover less than 1% of the original image. Small object detection suffers more difficulties than generic object detection due to lower image cover rate, fewer appearance cues, and large datasets. A clear description of small object detection and generic object detection is displayed in Fig. 1. Moreover, detection results of most good-performing generic object detectors based on the MS COCO [17] datatset could be found in Table I. In order to facilitate the readers to follow Table I, we explain the metric used by MS COCO [17] again here. Traditionally, AP is an averaged precision for each class while mAP is the averaged precision of all AP. However, MS COCO [17] makes no difference between them. So Avg.Precision,IoU 0.5:0.95 means the average AP for IoU from 0.5 to 0.95 with a step size of 0.05. Avg.Precision, IoU 0.5 corresponds to the AP with IoU $= 0.5$ and Avg.Precision,IoU 0.75 corresponds to the AP with IoU $= 0.75$. Besides, the metric for object size is: small objects (less than 32^2), medium objects (from 32^2 to 96^2), and large objects (larger than 96^2). In Table I, we also define an item called degrade of reduction (DOR), to illustrate the large gap of performance between large object detection and small object detection. It can be seen that the average precision (AP) of small objects is much lower comparing to medium or large objects. Nearly all generic object detectors trained in this dataset have a poor performance on the small objects as the number of medium and large objects is far more than that of small objects. Furthermore, current surveys of deep learning-based object detection mainly focus on generic object detectors, as shown in Table II. Therefore, we identify this work as a timely complement to the object detection community. It is noted that our work mainly focuses on nature scenes, such as traffic road scene, indoor scene [40], etc. The small objects detection from aerial perspective [41], [42] is not the main content of our survey.

B. *Challenges for Small Object Detection*

Compared with a medium and large object, the small object is more difficult to be detected and located. First, small object covers fewer pixels, indicating that features used for detection are insufficient and feature representation is weak. Otherwise,

TABLE II
SUMMARY OF RECENT OBJECT DETECTION SURVEYS WHICH ARE BASED ON DEEP LEARNING AND RGB IMAGES

Survey Title	Year	Published
Deep Learning for Generic Object Detection: A Survey [43]	2018	arxiv
Object Detection with Deep Learning: A Review [44]	2019	TNNLS
Recent progresses on object detection: a brief review [45]	2019	MULTIMED TOOLS APPL
A Survey of Deep Learning-Based Object Detection [46]	2019	IEEE Access
Object Detection in 20 Years: A Survey [47]	2019	arxiv
A Survey of the Four Pillars for Small Object Detection: Multi-scale Representation, Contextual Information, Super-resolution, and Region Proposal	**2019**	**Ours**

the larger anchor size in the region-proposal stage of generic object detectors causes small objects to receive less attention or even be ignored. Second, objects may appear in any position of input image such as the corner or overlapping area with other objects due to the smaller size. Besides, it is also difficult to distinguish small objects from noisy clutter in the background and accurately locate their boundaries. Third, *AP and mean AP (mAP)* that adopt the IoU threshold to determine true positive (TP) or false positive (FP) value are commonly regarded as the performance metrics of object detection. However, AP and mAP may not be suitable for evaluating the performance of small object detection because a large difference in IoU value would be caused by even a small shift of bounding box in the image. Therefore, a novel evaluation metric tailored for small object detection is absolutely necessary. Fourth, there are few authoritative datasets for small object detection. There are several simple datasets for small object detection existing, facilitating the comparison of different approaches and providing insight into the development of different approaches. However, it is not evident how to extrapolate those results obtained on simple datasets to more complex scenarios.

C. Four Pillars for Small Object Detection

With the development of object detection based on deep learning, many novel detection networks tailored for small objects are proposed. In this article, small object detection methods are mainly classified into four pillars. The basis for the division of the four pillars is based on the popular object detection frameworks such as the definition in mmdetection [48], which divides the detector into several modules, e.g., Backbone, Neck, AnchorHead, RoIExtractor, and RoIHead. The first two pillars about multiscale representation and contextual information belong to Neck component, which make refinements or reconfigurations on the raw feature map produced by the Backbone. The region-proposal pillow is mainly related to AnchorHead component. While the super-resolution is not exactly a component of the above, which adds two branch networks, e.g., generator network and discriminator network on the basis of baseline detectors. Considering that it has become an independent research direction of small object detection, we also describe it as a kind of pillar.

Multiscales Representation: On the one hand, detailed information in shallow conv layers is necessary for object location. On the other hand, semantic information in deep conv layers facilitates object classification a lot. Due to the tiny size and low resolution of small objects, location details are gradually lost in high-level feature maps. While most generic detectors only adopt the output of final layer for detection tasks, which contains rich segmentation information but lacks detailed information. Multiscales representation is a strategy of combining detailed location information from low-level feature maps and rich semantic information from high-level feature maps.

Contextual Information: Leveraging the relationship between an object and its coexisting environment in the real world, contextual information is another novel method to improve small object detection accuracy. The medium and large objects could provide sufficient ROI features in generic detectors. However, it is much necessary to extract more additional contextual information as the supplement of original ROI features because the ROI features extracted from the small objects are so few.

Super-Resolution: As mentioned above, fine details are critical for object instance localization. Super-resolution techniques attempt to recover or reconstruct raw low-resolution images to a higher resolution, which means more details of small objects could be obtained. For example, the core idea of GAN is the generator network and discriminator network. In this adversarial process, the ability of generator to generate real-like images and the ability of discriminator to distinguish between real and fake images is constantly improving at the same time.

Region-Proposal: Region-proposal is a strategy aiming at designing more suitable anchors for small objects. The anchors of current leading detectors mainly focus on generic objects, indicating that the anchor size, shape, and amount used in the generic detectors could not match well with small objects. Otherwise, extra noise information will cause a huge computational cost and reduce the detection accuracy if these anchor parameters of generic detectors are directly applied to the small objects.

According to these four pillars, related small object detection researches are described with more details in our paper, which is organized as follows. A summary of related small object detection datasets is provided in Section II. Then, small object detection methods are expanded specifically in Section III. Finally, a discussion of several promising directions is illustrated in Section IV and our conclusion is drawn in Section V.

II. SMALL OBJECT DETECTION DATASETS

A less bias benchmark is fundamental for deep learning research. Although some generic object detection datasets, such as PASCAL VOC and ImageNet, are accessible, there is no common accepted dataset for small objects. Most researchers have to perform and evaluate their small object detection networks on the datasets built by themselves or extracted from large datasets such as MS COCO. Based on

TABLE III
INFORMATION FOR SMALL OBJECT DATASETS. SOME EXAMPLE IMAGES ARE SHOWN IN FIG. 2

Dataset Name	Total Images	Annotated images	Categories	Image Size	Instances	Instances Size(pixels)
Lost and Found [32]	2,104	2,104	37	2,048x1,024	-	-
STS [49]	20,000	4000	7	1,280x960	3,488	3x5-263x248
Tsinghua-Tencent 100K [50]	100,000	100,000	45	2,048x2,048	30,000	80% smaller than 70x70
GTSDB [51]	900	900	43	1,360x800	1,206	16-128(longer edge)
CURE-TSD [52]	1,719,900	1,719,900	14	1,628x1,236	2,206,106	3x7-206x277
Small Object Dataset [53]	4,925	4,925	10	640x480 & 500x300	8,393	16x16-42x42
CURE-OR [54]	1,000,000	1,000,000	6	480x640 - 726x1,292	-	-
WIDER FACE [55]	32,203	32,203	1	-	393,703	50% 10-50, 43% 50-300
DeepScores [56]	300,000	300,000	123	1,894x2,668	80 millions	-

different application scenarios and data sources, some high-quality datasets about small objects are briefly introduced in the following sections. Detailed information on small object datasets is collected in Table III.

A. Datasets for Traffic Road Scene

Datasets for traffic road scene are mainly collected by camera fixed in the front of vehicle. These datasets could be divided into two major categories, including road obstacles and traffic signs.

Lost and Found [32]: Lost and Found is the first lost-cargo dataset for detecting small barriers on the road, which are collected from 13 different street scenarios and 37 different obstacle types. These selected objects vary in size, distance, color, and material. Besides, 112 video stereo sequences are included, corresponding with 2104 annotated frames.

Swedish Traffic Signs (STS) [49]: The STS dataset contains 3488 traffic signs which are captured on highways and cities from more than 350 km of Swedish roads in this dataset. It contains more than 20 000 images and 20% of images are labeled for training. The labeled objects contain sign types, such as pedestrian crossing, designated lane right, no standing or parking, priority road, give way, 50 kph, and 30 kph. Moreover, explicit visibility status (occluded, blurred, or visible) and road status are also included in the dataset.

Tsinghua-Tencent 100K [50]: Zhu *et al.* [50] built the Tsinghua-Tencent 100K dataset, which may be the largest and most challenging traffic sign dataset, including annotated 100 000 images in 45 classes and 30 000 traffic sign instances. All images in this dataset have a high resolution (2048 × 2048), and 80% of instances occupy less than 0.1% in the whole images.

GTSDB [51]: German traffic sign detection benchmark (GTSDB) is the successor of GTSRB [58], [59]. The recording of GTSDB is finished by a Prosilica GC1380CH camera with automatic exposure control. The images are selected from sequences recorded near Bochum, Germany, in different scenarios, such as urban, rural, and highway during daytime and dusk. Image samples are shown in Fig 2(e).

CURE-TSD [52]: CURE-TSD datasets consist of real-world data and synthetic virtual data, in which 49 challenge-free real-world video sequences are generated by combining 300 frames from BelgiumTs [60] and another 49 synthesized video sequences are generated with a game development tool Unreal Engine4. What is more, the Adobe After Effects are used to emulate weather and vision system challenges at post-production.

B. Datasets for Generic Small Objects

Small Object Dataset [53]: In [53], a small object dataset and validated classic R-CNN detection model on this dataset were first introduced. Some large image datasets such as MSCOCO also contain categories about small objects. However, the image number of small objects is fewer than that of medium and big objects, significantly causing the nonuniformity of experimental samples. Thus, this dataset extracted purely ten categories of small objects from MSCOCO and Scene Understanding database [61], containing approximately 8393 object instances and 4925 images. Moreover, different IoU thresholds were set according to different categories in order to avoid the problem that the commonly used 0.5 IoU value causes a low recall for small objects.

CURE-OR [54]: In Challenging Unreal and Real Environments for Object Recognition (CURE-OR), there are 1 000 000 images of 100 objects with varying size, color, and texture. These objects are grouped into six categories as toys, personal belongings, office supplies, household items, sports/entertainment items, and health/personal care items. The image resolution of the dataset includes: 648 × 968, 756 × 1008, 480 × 640, 460 × 816, and 726 × 1292.

C. Datasets for Single Category

WIDER FACE [55]: As shown in Fig. 2(d), WIDER FACE contains 32 203 images, which are extracted from the public WIDER dataset [62]. Images in WIDER FACE are categorized into 60 social event classes, which have much more diversities and are closer to the real-world scenario. Besides, they are also divided into three subsets, small, medium, and large, based on the heights of the ground-truth faces. The small/medium/large subsets contain faces with heights larger than 10/50/300 pixels, respectively. The small subset accounts for 50% of WIDER FACE while medium accounts for 43%.

DeepScores [56]: DeepScores [56] focuses on pages of written music, containing 300 000 pages of digitally rendered music scores and 123 different symbol classes. This dataset could execute tasks, such as object classification, semantic segmentation, and object detection.

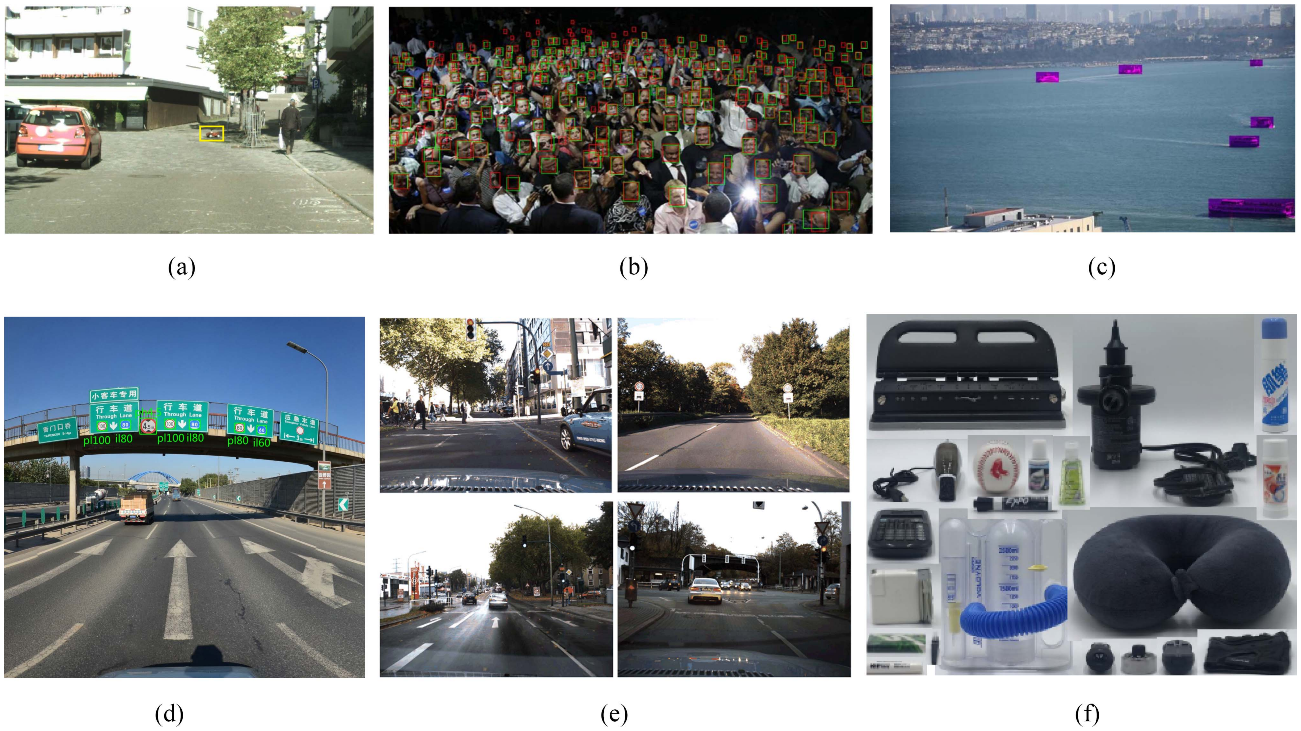


Fig. 2. Some image samples and annotations from several datasets focusing on small objects. (a) Lost and Found sample [32]. (b) WIDER face sample [55]. (c) Small-vessel sample [57]. (d) Tsinghua-Tencent 100K sample [50]. (e) GTSDB sample [51]. (f) CURE-OR sample [54].

III. Small Object Detection Networks

The frameworks of small object detection are mainly divided into two paradigms, that is, one leverages handcrafted features and shallow classifiers, detecting objects such as barriers or traffic signs on the road, which usually has poor performance because of the weak feature extraction method. The other adopts DCNN to extract image features and then modifies leading generic object detection networks to reach a good tradeoff of accuracy and computational cost. A vary of novel methods have been proposed to improve traditional small object detection performance significantly. In Fig. 3, an overview of small object detection research community is illustrated. Based on the core theories utilized in each method, the research works of small object detection are classified into five categories in this work, namely, multiscale representation, contextual information, super-resolution, region proposal, and other methods. The top performing models among each category are described in detail while other similar models are going to be stated briefly in order to give a clear explanation of each category.

A. Multiscale Representation

Weak feature representations of small object are the main reason for the poor detection performance. After repeated downsampling operations from CNN and pooling layers, fewer small object features exist in the final feature map. Moreover, with the increase of neural network layers, the inherent hierarchy generates feature maps with different spatial resolutions. Specifically, although deeper layers represent larger receptive field, stronger semantics, higher robustness to deformation, overlap, and illumination variances, the resolution of feature maps becomes lower and more detailed information is lost. In contrast, shallow layers have a smaller receptive field, leading to a higher resolution, while they lack semantic information.

1) Multiple Feature Maps Fusion: Some prevailing object detectors, such as R-CNN, Fast R-CNN, Faster R-CNN, and YOLO, only use the feature map of the last layer to localize objects and predict confidence scores, as simply displayed in Fig. 4(a). Due to lack of detailed information, these models often fail to detect small objects. Then, single-shot multibox detector (SSD) introduces the pyramidal hierarchy feature to assemble each feature map from bottom to the top network layer, as shown in Fig. 4(b), resulting in improving small object detection. However, much unnecessary representation noise and high computation complexity could be caused by taking all-levels features into consideration. To simplify network and improve detection, some researchers adopt the deconvolution layer and only choose several important feature maps that contain most detailed and semantic information.

MDSSD [63]: Deconvolution Fusion Block was proposed in [63], which adopted skip connection to fuse more contextual features. In this model, three high-level semantic feature maps from different scales (conv8_2, conv9_2, and conv10_2 from SSD layers) were first introduced into the deconvolution layers and then sum with three shallow layers by element (conv3_3, conv4_3, and conv7 from VGG16 layers). It should be noted that deconvolution layers are applied to upsample the high-level feature maps into the same resolution with corresponding low layers. SSD is the backbone of the whole model; the fusion process is finished in the Fusion Block. The basic idea is depicted in Fig. 4(c). The detection results on MS

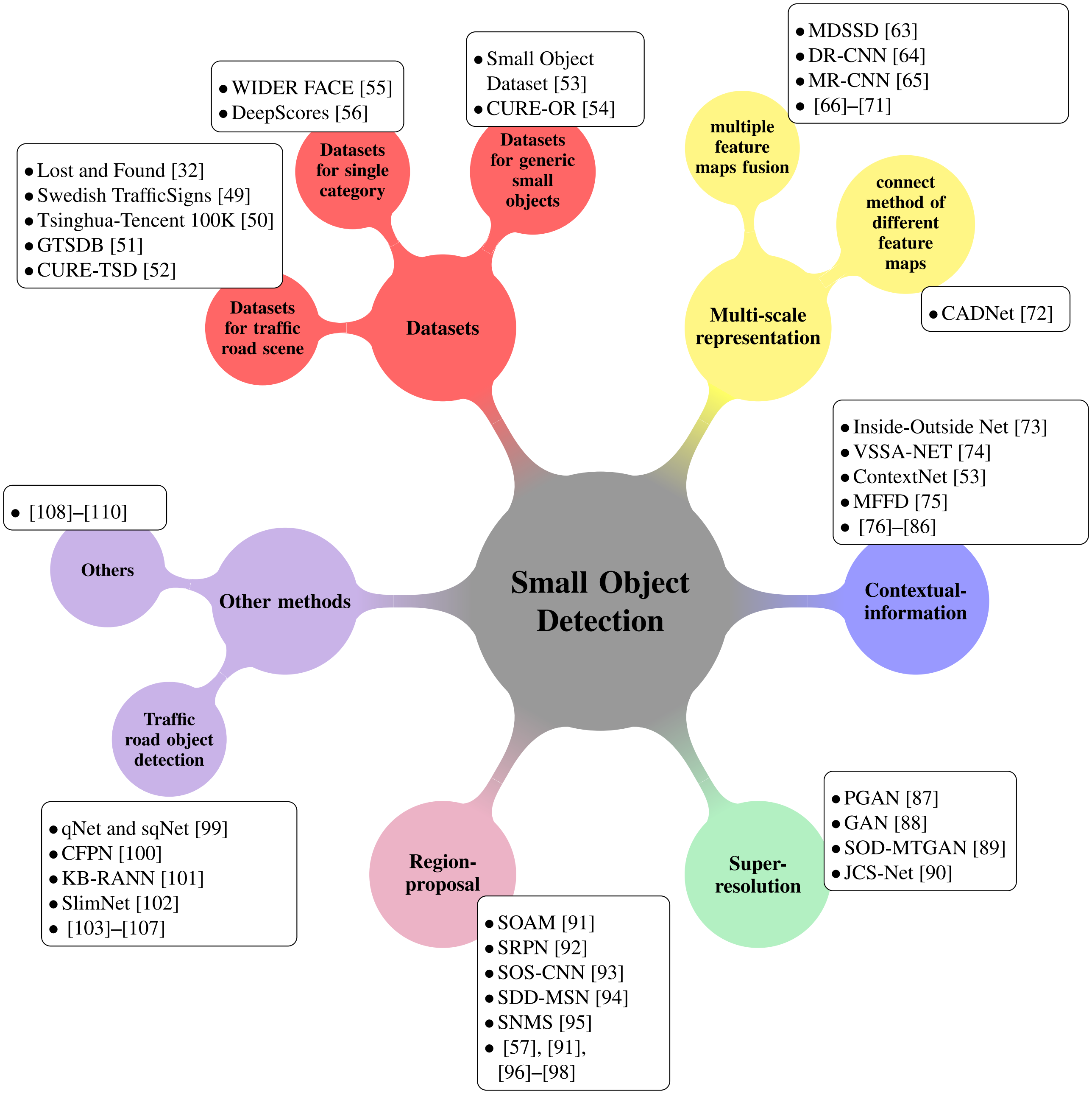


Fig. 3. Four main solutions for small object detection problem: multiscale representation, contextual information, super-resolution, and region-proposal. Relative small object datasets are also collected in this survey.

COCO [17] and PASCAL VOC2007 [111] are collected in Tables IX and X.

DR-CNN [64]: Different from the element-sum strategy taken by MDSSD, concatenation strategy was adopted in the deconvolution region-based convolutional neural network (DR-CNN) to fuse multiscale feature maps for small traffic sign detection. DR-CNN selects conv3, conv4, and conv5 from VGG16 to form a fusion feature map for followed RPN and detection. After each deconvolution module, the L2 normalization layer is also used to ensure the concatenated features on the same scale. Another innovation of this network is about loss function. Hard negative samples benefit the training phase a lot. However, it is hard for common cross entropy loss function to distinguish easy positive samples from hard negative samples. Therefore, the common cross entropy loss function is replaced with a novel two-stage classification adaptive loss function in the RPN and fully connected network in order to fully leverage hard negative samples for better performance. The result shows that DR-CNN achieves excellent performance on the MS COCO and Tsinghua-Tencent 100 K datasets. Detailed information are collected in Tables IX and XI.

MR-CNN [65]: The multiscale region-based convolutional neural network (MR-CNN) was proposed for small traffic sign recognition, where a multiscale deconvolution operation was used to upsample the features of deeper convolution layers

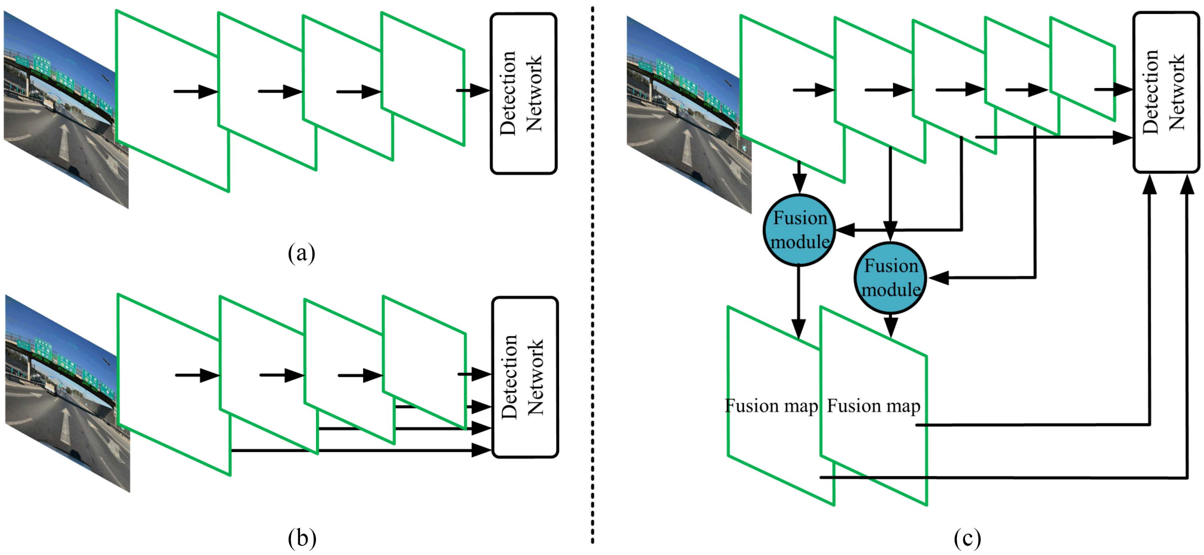


Fig. 4. (a) Traditional detector uses one top feature map for the detection network. (b) Detectors take all-level feature maps for detection network via pyramidal feature hierarchy. (c) Basic idea of multiscale feature maps fusion. Both original deep feature map and fusion maps are leveraged. Specifically, the fusion module usually takes elementwise product or concatenating operation.

that were concatenated with those of the shallow layer directly to construct fused feature map. Thus, the fused feature map could generate fewer region proposals and achieve a higher recall rate. Moreover, the test result indicates that this method can effectively enhance feature representation and boost the performance of small traffic sign detection.

Other Simply Introduced Methods: Sun [66] presented a multiple receptive field and small-object-focusing weakly supervised segmentation network to enhance the performance of small objects detection. In [67], an image block architecture was utilized to divide raw images into fixed size blocks; then, these blocks were sent to the VGG-16 network as input. Besides, feature map fusion and image pyramid were also adopted in order to solve the issue that details of small objects generally lost in deeper layers. Moreover, fused multiscale feature maps were applied to locate object position and used information from deep layers to execute object classification [68]. A backward feature enhancement network (BFEN) was designed to transport more semantic information from high layers to low layers [69]; then, fine-grained features were concatenated into a spatial layout preserving network (SLPN), preserving the spatial information of ROI pooling layer, and achieving better location accuracy. In [70], the feature maps of third, fourth, and fifth convolution layers were extracted and combined into a one-dimensional vector for classification and localization. Besides, an optimizing anchor size method and fused multilevel feature maps for road garbage detection were proposed [71]. Inspired by the Inception module, a novel feature fusion mechanism was putting forward [112]. They chose YOLOv3 as the basic framework and used multiscale convolution kernels to form various size receptive fields, which can make full use of low-level information. Furthermore, multiscale feature maps-based ResNet-50 and merged these feature maps by means of the feature pyramid network were generated by means of [113].

2) Connection Method of Different Feature Maps: Although many methods based on multiscale representation mentioned above have been proposed to improve small object detection, there is little relevant work focusing on how to fuse high-level feature map and low-level feature map.

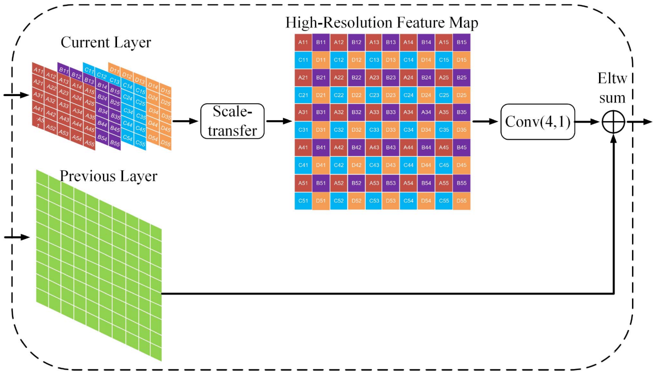

Current Layer
Previous Layer
Scale-transfer
High-Resolution Feature Map
A11 B11 A12 B12 A13 B13 A14 B14 A15 B15
C11 D11 C12 D12 C13 D13 C14 D14 C15 D15
A21 B21 A22 B22 A23 B23 A24 B24 A25 B25
C21 D21 C22 D22 C23 D23 C24 D24 C25 D25
A31 B31 A32 B32 A33 B33 A34 A34 A35 B35
C31 D31 C32 D32 C33 D33 C34 D34 C35 D35
A41 B41 A42 B42 A43 B43 A44 B44 A45 B45
C41 D41 C42 D42 C43 D43 C44 D44 C45 D45
A51 B51 A52 B52 A53 B53 A54 B54 A55 B55
C51 D51 C52 D52 C53 D53 C54 D54 C55 D55
Conv(4,1)
Eltw sum

Fig. 5. Channel-aware deconvolution block [72].

CADNet [72]: The channel-aware deconvolutional network (CADNet) was proposed to study the relationship of feature maps in different channels from deeper layers in order to avoid the simple stacking of feature maps. The recall rate of small objects could be improved at a low computational cost through exploiting the correlation between different scale characteristic. As shown in Fig. 5, the framework is divided into three steps, including the scale transfer layer, convolution layer, and elementwise-sum module. Particularly, the scale-transfer layer reorganizes four pixels of each four channels into the same position on a two-dimensional plane in order to obtain the location details and increase the resolution of the feature map. Then, more semantic information of the feature map is exploited by a convolution layer with a 4 × 4 kernel size; feature maps with the previous layer are connected by the elementwise method. Thus, the fusion layer contains both details in the low-level layer and semantic information in the high-level layer.

TABLE IV
SUMMARY OF MULTISCALES REPRESENTATION-BASED METHODS

Model	Fusion method
[63]	deconvolution layer + skip connections + element-wise sum
[64]	deconvolution layer + concatenating
[65]	deconvolution layer + concatenating
[67]	bilinear upsampling + element-wise max
[68]	maxpooling on low-level features + element-wise sum
[69]	deconvolution layer + element-wise sum
[70]	L2 normalized on selected feature maps + concatenating
[71]	bilinear upsampling + concatenating
[72]	Channel-Aware deconvolution
[112]	Inception [114] -like convolution kernel on low-level features
[113]	upsample(2x) + element-wise sum (FPN)

In general, multifeature map fusion helps to capture detailed information and rich semantic information, facilitating object location and classification, respectively. However, many multiscale representation methods increase the computational burden while improving the detection performance. Moreover, redundant information fusion design may lead to background noise, resulting in performance degradation. In Table IV, the fusion methods of primary models mentioned in this section are described.

B. Contextual Information

Since small objects only occupy a small portion of the image, the information that be directly obtained from fine-grained local areas is greatly limited. Generic object detectors usually ignore many contextual features outside those local regions. It is well known that every object always exists in particular environments or coexists with other objects. Then, some detection methods based on contextual information were proposed to leverage the relationship between small objects and other objects or background. Oliva and Torralba [115] illustrated that the around region of the small object could provide useful contextual information to help detect target object. Besides, the experimental results in [116] also demonstrate that detection accuracy could be significantly improved by adding a special context module. Next, several important network models using contextual information are described in detail.

ContextNet [53]: Augmented R-CNN [53] could be considered as the first detector focusing on small object detection. In this work, a novel region proposal network (RPN) is proposed to encode the context information around a small object proposal. First, according to the size of small objects, the RPN anchor size is scaled from the original 128^2, 256^2, 512^2 pixel^2 to 16^2, 40^2, 100^2 pixel^2 and small object proposal is extracted in conv4_3 feature map rather than the conv5_3 of VGG16. Second, a ContextNet module consisting of three subnetworks is designed to obtain the context information around the proposal object, as shown in Fig. 6. The same two front-end subnetworks are composed of a few convolutional layers followed by one fully connected layers; the back-end subnetworks consists of two fully connected layers. The proposal region extracted by a modified RPN and a larger context region with the same center point with proposal region is passed into the two front-end networks, respectively. Meanwhile, two 4096-D feature vectors obtained from front-end networks are concatenated before they are inputted into the back-end network. The experimental results show that this augmented R-CNN improves the mAP of small object detection by 29.8% over the original R-CNN model.

Inside–Outside Net [73]: Spatial recurrent neural networks (RNNs) are adopted in an Inside–Outside Net (ION) [73] to search for contextual information outside the target region; then, skip pooling is taken to obtain multilevel feature maps inside. Two consecutive four-directional spatial RNN units are employed to move through each column of the image. This model concatenates multiple scales and context information for detection. In the ION method, the context feature map is generated by mentioned IRNN modules at the top of the network. It is noted that IRNN is composed of rectified linear units (RELUs), which is initialized by Le *et al.* [117]. Besides, four copies of the conv5 layer of original VGG16 are taken as input of first four-directional RNN (left-to-right, right-to-left, top-to-bottom, and bottom-to-up) by a 1×1 convolution layer; then, the output of each direction is concatenated as input to next IRNN unit. Finally, context features are obtained.

VSSA-NET [74]: In [74], a multiresolution feature fusion network exploiting deconvolution layers with skip connecting and a vertical spatial sequence attention module was designed for traffic signs detection. This network is mainly divided into two stages. The first stage is a multiscale feature extracting module, which forms multiresolution feature maps through Mobile Net [118] and deconvolution layers. The second stage is constructing a vertical spatial sequence attention module. Particularly, each column of three feature maps is regarded as spatial sequence in order to fully exploit context information. The traditional encoder–decoder model based on the LSTM network is modified by introducing the attention mechanism at the decoding stage, which could encode the contextual feature disregarding the noise.

MFFD [75]: With the improvement of detection accuracy, the deeper detection network means high computation costs. A kind of modular lightweight network model that is called modular feature fusion detector (MFFD) was proposed in [75]; it not only has a great performance on small object detection but also could be embedded into the resource limited equipment such as advanced assistance systems (ADASs). Two novel modules are designed in this network. Among them, the front module uses small size filters in convolution layers to reduce information loss while the Tinier module changes the number of input channels with pointwise convolution layers (1 × 1 convolution) before entering in the convolution layer. The advantage is that the network fuses multiscale context information from available modules, instead from an individual layer directly, leading to efficient computation.

Other Simply Introduced Methods: The concatenation module or element-sum module is employed in a multilevel feature fusion module to introduce context information into SSD [76]. Meanwhile, a special layer called CSSD is designed to integrate multiscale context information [77]. This context layer adopts dilated convolution and deconvolution to extract context information from multiscale feature maps. In [78], a spatial

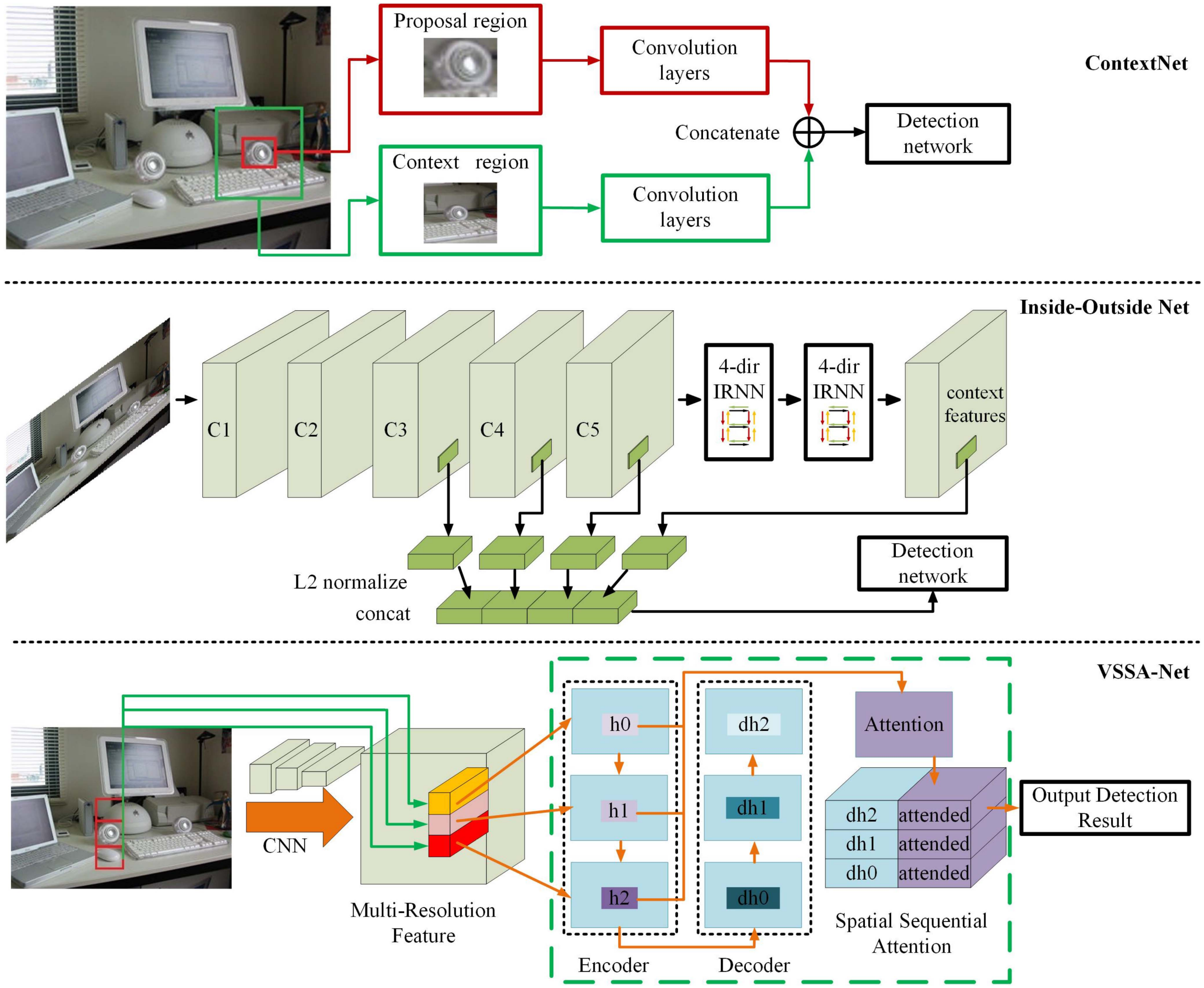


Fig. 6. Three kinds of contextual information-based small object detection network, ContextNet [53], Inside–Outside Net [73], and VSSA-net [74]. ContextNet shows the basic idea about how to leverage the around contextual clue; Inside–Outside Net adopts a spatial recurrent neural network; and VSSA-Net uses an LSTM-based spatial sequential attention module to obtain contextual information.

memory network was introduced to store semantic information and preserve conditional distributions on previous detections. Memory augmented scores are added to the Faster-RCNN score and then optimized to finish region classification. P-CNN [86] consists of three blocks, where global features are obtained from squeeze-and-excitation (SE) block and part features are extracted from the part localization network (PLN). Then, the second stream of part classification network (PCN) concatenates part local features and global image feature together into a joint feature for the final classification.

Besides, the TL-SSD network, where inception modules concatenate different size receptive fields, was presented [79]. The feature concatenation combines shallow and deep feature layers; the shallow one could provide accurate location and state information while the deep one makes the decision whether the object belongs to traffic light.

Multilevel context information through pyramid pooling was used to construct context-aware features [80]. The context fusion module focused on adding scales of context information into feature maps. Context-aware RoI pooling avoiding to harm the structure of small objects and keeping the contextual information is also designed, where a scale-intensive convolutional neural network was applied to vehicle detection scenes [81]. Leng *et al.* [82] integrated a U-V disparity algorithm with faster R-CNN that combines internal and contextual information.

Similar to multiscale representations, contextual information is also intended to provide more information to the final detection network. The difference is that the contextual information is mainly to obtain the information around the ROI area and improve object classification by learning the relationship between objects and surrounding information. Therefore, redundant context information also causes information noise. Most novel architectures about capturing contextual information are shown in Table V.

C. *Super-Resolution*

Super-resolution methods aim at recovering high resolution from corresponding low-resolution features. High-resolution

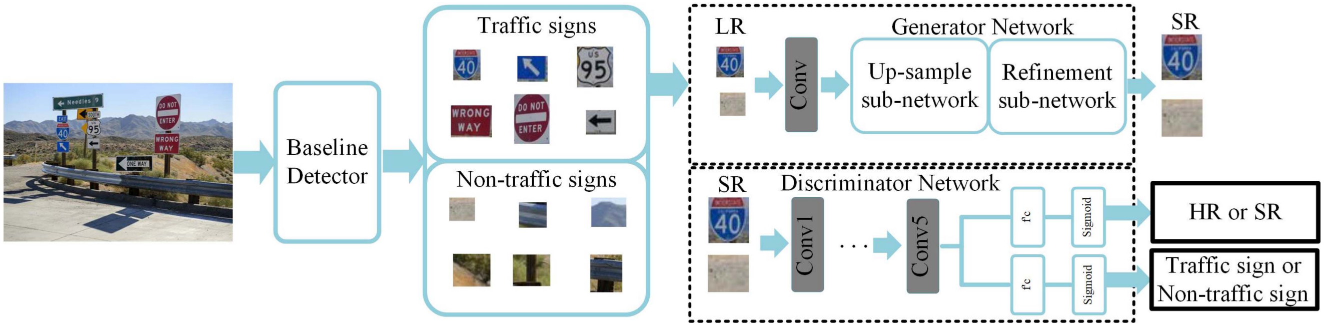


Fig. 7. Generic process of the GAN-based object detection method. The baseline detector generates region proposals for target objects from the input image and generator network.

TABLE V
SUMMARY OF CONTEXTUAL INFORMATION-BASED METHODS

Model	Method for obtaining contextual information
[73]	spatial recurrent neural network and skip pooling
[74]	vertical spatialsequence attention (VSSA) module
[75]	post-context fusion
[76]	Element_sum module / concatenating module
[77]	multi-scale dilated convolution deconvolution
[78]	spatial reasoning module (fully-convolution network)
[80]	context fusion module
[81]	context-aware RoI pooling
[82]	context-aware module

image offers more refined details about the original scene, which could be well applied to small object detection. GAN-based algorithms have been proposed to reconstruct high-resolution images. The generative adversarial network [119] has obtained tremendous stride in image super-resolving [120], which consists of two subnetworks, a generator network and a discriminator network. The generator produces super-resolved images to fool discriminator while discriminator tries to distinguish the real images from fake images generated by the generator. The common overflow of GAN-based methods is exhibited in Fig. 7.

Perceptual GAN [87]: In [87], the GAN method was first used in the small object detection task. A novel conditional generator was introduced; it took the low-level features as the input to obtain more details for super-resolved representation. The generator includes multiple residual blocks to learn the residual representation between small objects and similar large objects. The discriminator consists of two branches, namely, adversarial branch and perception branch. From one perspective, the adversarial branch distinguishes the generated super-resolved region of small objects from similar large objects. From another perspective, regular objection detection tasks are achieved in the perception branch and detection accuracy is justified from the generated super-resolved representation. Both branches try to obtain minimal loss while the generator is trained to maximize the probability of discriminator making a wrong judgement.

GAN [88]: However, the high-resolution images generated by GAN are still not clear enough. Thus, a refinement module [88] is added to recover some details for small face detection. First, MB-FCN [121] is selected as the baseline detector to generate regions containing faces or not, which are passed into the generator and discriminator separately. Second, the low-resolution faces pass through an upsample module and a refinement module, obtaining clear and super-resolution regions. Third, the nonfaces regions are treated as negative data for training discriminator that has two tasks simultaneously to distinguish super-resolution regions from the high-resolution ones; faces regions from nonface regions.

SOD-MTGAN [89]: A novel multitask generative adversarial network (MTGAN) was presented in [89]. In MTGAN, super-resolved images are produced by the generator network; the multitask discriminator network is introduced to distinguish real high-resolution images from fake ones, predict object categories, and refine bounding boxes, simultaneously. More importantly, the classification and regression loss are back-propagated to further guide the generator network to produce super-resolved images for easier classification and better localization. The loss function of the generator in MTGAN consists of adversarial loss (to the objective loss), pixelwise MSE loss, classification loss (to the overall objective), and bounding box regression loss, enforcing the reconstructed images to be similar with real high-resolution images containing high-frequency details. Compared to the previous GANs, the classification and regression losses of generated super-resolved images were added to the generator loss in order to ensure the super-resolved images recovered from the generator networks; they are more realistic than those optimized by only using the adversarial and MSE losses.

JCS-Net [90]: Focusing on small pedestrian detection, JCS-Net consists of a classification subnetwork and a super-resolution subnetwork [90]. The two subnetworks are integrated as a unified network by combining classification loss and super-resolution loss. Similar residual architecture such as VDSR [122] is adopted in the super-resolution subnetwork to explore the relationship between large-scale pedestrian and small-scale pedestrian for recovering details of small-scale pedestrians. Therefore, the reconstructed small-scale pedestrian contains both the original information of small-scale pedestrian and the output information of super-resolution subnetwork. In the training phase, multilayer channel features (MCFs) [123] are based on HOG + LUV [124] and JCS-Net are applied to train the detector. Furthermore,

TABLE VI
SUMMARY OF SUPER-RESOLUTION-BASED METHODS

Reference	Novel architecture
[39]	modified RPN(smaller anchor size) + super-resolution network
[87]	a generator (learns the additive residual representation between large and small objects) + a discriminator
[88]	Genenrator network(up-sample sub-network+refinement sub-network) + Discrimianator network(VGG19)
[89]	super-resolution network (generator) + multi-task network (discriminator)
[90]	super-resolution sub-network + multi-layer channel features(MCF)

multiscale representation is combined with MCF to enhance detection.

GAN-based method is effective to enhance the detail information of image, especially for the super-resolution application. There is no need to design specific architecture and it can be applied to any kind of generator network. However, it also faces two intractable problems. First, GAN is hard to train, which means it is difficult to achieve a good balance between generators and discriminators. Second, when the generator in the training process produces limited rewards of samples and the learning process stops, the phenomenon of model collapse is easily to emerge, resulting the increase of final detection error. Relative researches in this section are displayed in Table VI.

D. Region-Proposal

Before the appearance of deep learning techniques, the best-performing method of region proposal is the Selective Search algorithm [26]. However, computing efficiency in this method is highly limited. Faster R-CNN first introduced RPN [30] to identify region of interest; then, R-FCN was proposed to generate $k \times k \times (C + 1)$ feature maps rather than single feature map, where each map is responsible for each category detection. However, it is still difficult to accurately locate for the small object detection due to larger anchor sizes.

Based on FastMask, AttentionMask [91] was proposed to generate a tailored region proposal for small objects. An additional larger scale ($S8$) was added to the feature scale space at the early stage of the base network. Particularly, a scale-specific objectness attention mechanism (SOAM) was adopted to select most promising windows at each feature map with different scale in order to reduce the number of sampled windows. Although all scales are jointly justified according to their attention values to find the optimal locations for sampling windows, this strategy only prioritizes the most promising windows to sample and process, resulting in saving the memory and GPU source for adding the scale ($S8$) of small object detection. More precise locations of anchor boxes usually have lower confidence scores while they are more likely rejected by the post-process of NMS. Thus, a smooth NMS (SNMS) [95] was designed to utilize those anchor boxes and IoU-prediction was adopted to provide more classification evidences. Besides, several pixels of input image are shifted circularly in four directions in order to avoid missing small objects that located in the gap of near anchor box.

The underfitting problem often exists in the training model of RPN because part parameters in RPN are determined by prior knowledge. Therefore, a strengthened RPN (SRPN) [92] was designed by increasing the parameters. Besides, particle swarm optimization and bacterial foraging optimization are introduced to find the optimal parameter values; then, a high-quality detection proposal could be acquired. Oversampling images containing small objects and small object augmentation were also introduced to make the model focus more on the small objects [97]. It should be noted that small object augmentation is copy-pasting small object region several times in one image; the pasted objects do not overlap with the existing objects. This increased the number of positively matched anchors and region proposals containing small objects. The result on the MS COCO dataset indicated that the most gain was obtained by processing the image with 3× oversampling and copy-pasting strategies, with an increase of a 9.7% relative improvement for instance segmentation and 7.1% for small object detection compared to the original mask R-CNN.

It took a huge amount of time and memory to process background regions in neural networks. A cascade mask generation framework was proposed to reach a balance between computational speed and accuracy [98]. The raw image was first resized into multiscales. Then, each scale produced region proposal and mask through the mask generation module (MGM) inspired by RoI convolution [126]. Finally, the feature maps from each scale were concatenated for the ROI pooling and post-detection. A CNN-based cascaded architecture is also designed to detect far objects in the outdoor surveillance [57]. After trained in the SSD model, the feature maps of input images were divided into obscure object samples and prominent object samples according to their confidence scores. The details of the prominent object samples are enough to recognize, while the obscure object samples (mostly far small objects) are confirmed through verification of SSD detection, object-size confirmation, duplication-object removal, and off-scope object removal. This method is also suitable for other detection models without architecture modification. In [96], both region-proposal stage and classifier stage were investigated in detail for detecting company logo. Area proposal network is applied when regions containing at least one object in the raw image are cropped and then enlarged to the same input size [94]. That makes the original small objects become more similar to large objects and easier to be detected for ordinary SSD detector.

A well-designed region proposal strategy could take advantage of limited anchor size and anchor amount, reduce computational cost in generating interested region, and efficiently detect small targets. All novel architectures are classified in Table VII.

E. Others

In addition to these four pillars for small object detection, other related works are replenished according to their application scenarios in this section. Besides, our work mainly

TABLE VII
SUMMARY OF REGION-PROPOSAL-BASED METHODS

Reference	**Novel architecture in region-proposal stage**
[53]	a small region proposal generator(AlexNet or VGG)
[91]	class-agnostic object proposal generation and visual attention
[92]	strengthened RPN (PBLS) which increases parameters
[93]	cropping large images into small patches as input to SOS-CNN
[94]	area proposal network(APN)
[95]	conducting circular shifts input image + Smooth NMS + IoU-Prediction
[96]	modifying anchor size of RPN for small objects
[97]	oversampling those images containing small objects and copy-pasting small objects many times
[98]	mask generation module (MGM)+RPN
[57]	additional steps processed the obtained candidate regions
[125]	Atrous Region Proposal Network

TABLE VIII
SUMMARY OF OTHER METHODS

Reference	**Novel architecture**
[66]	proposed a small- object-focusing weakly-supervised segmentation module
[99]	lightweight and accurate network for fast detection
[100]	Concatenated blok
[101]	Attention mechanism and RNN (LSTM)
[102]	SlimNet_freeze
[103]	a novel IoU loss function
[105]	investigate optimal combination of color space and deep model for traffic light detection
[106]	a Comprehensive Feature Enhancement(CFE) module
[127]	somatic topological line localization and temporal feature aggregation

discussing traffic road small object detection and a few works in other scenarios. Related information is summarized in Table VIII.

Traffic road object detection could help the driver or self-driving car make decisions earlier and avoid danger. However, many objects on the road, such as traffic signs, small obstacles, and pedestrians, appear small in the detected images. There is only a short time to capture and detect them due to fast speed of vehicle. Hence, it is more difficult but critical to detect such objects at a fast speed and high accuracy. Some methods have been proposed to tackle this issue as follows.

Considering the limited device resource, convNet (qNet) and small ConvNet (sqNet) with uniform macro-architecture and depthwise separable convolution for fast traffic sign detection are proposed [99]. The network has only one fully connected layer after the average pooling layer leading a decrease in the number of parameters. Otherwise, standard convolution is replaced by the depthwise separable convolution to reduce the computations. ConvNet is compacted through reducing the number of channels in each convolutional layer by a fixed ratio, which significantly makes networks smaller. In [100], a shallow network called concatenated feature pyramid network (CFPN) was introduced with a novel concatenated block for real-time embedded traffic flow estimation system.

Based on the faster R-CNN, a new IoU loss function [103] to improve location accuracy and the bilinear interpolation was applied to reduce location deviation. KB-RANN [101] focuses on the detection of traffic signs, where a pretrained SqueezeNet generates feature maps and an RNN architecture (LSTM) with attention mechanism searches contextual information. Besides, the pool4 layer of VGG-16 is reduced and dilation for ResNet is adopted to extract the characteristics of small signs [104] because original region proposal generator from faster R-CNN is too large for traffic signs. Afterward, online Hart examples mining (OHEM) is combined to make the network more robust.

A topological line annotation method was adopted to detect pedestrians on the road, which obtained a better location accuracy than the traditional bounding box through IoU criteria [127]. Each line annotation was connected by two point, top and bottom vertex locations; they were modeled as a Gaussian peak and then can be solved by the Hungary algorithm [129]. Moreover, a Markov random field (MRF)-based post-processing method was proposed to solve the crowd scenes.

Six kinds of color space and different network models are exploited to build a traffic light recognition system [105]. It is found that the combination of RGB color space and the faster R-CNN model perform best. In [106], a comprehensive feature enhancement module was added into single shot detector for small object detection on the road at a high speed. Moreover, a modified faster R-CNN for vehicle detection that adopted a small anchor scale was designed in [107], which converted object detection problem into a binary object detection classification problem. There are also some methods for other scenarios. Simple VGG16 is modified to detect small insects such as spiders [109]; a kernelized correlation filter tracking algorithm based on faster R-CNN was used for detecting and tracking small sea objects [110].

IV. DISCUSSION

Deep learning has been widely adopted in the task of computer vision and object detection. For the medium and large object, the generic object detection model has reached high precision and short inference time. However, the practical requirement still could not be satisfied with the result of small object detection. Besides, there is still a huge gap between the current best performing small object detection networks and those generic detection models. Many open issues remain to be dealt with, which we discuss at the following aspects.

A. Novel Metric for Small Object Detection

In fact, the widely used AP metric has several drawbacks. First, AP value is the area under the recall–precision (RP) curve. It could not reflect the tendency from the RP curve, indicating that different RP curves, either low-recall–high-precision or high-recall–low-precision could obtain the same AP. Second, detailed tightness level information of bounding box detection could not be obtained from AP. However, the real tightness level of the bounding box is critical

TABLE IX
PERFORMANCE COMPARISON OF FOUR PILLOWS BASED ON MS COCO

Group	Model	Backbone	Dataset	Avg.Precision, IoU			Avg.Precision		
				0.5:0.95	0.5	0.75	S	M	L
Multi-scale representaion	MDSSD512 [63]	VGG16	COCO 2015 test-dev	30.1	50.5	31.4	13.9	-	-
	DR-CNN [64]	VGG16	COCO 2015 test-dev	36.5	59.7	38.2	18.6	39.3	47.7
	MRFSWSnet512 [66]	VGG16	COCO 2015 test-dev	33.1	53	34.7	17.5	37.1	47.9
	CADNet512 [72]	VGG16	COCO 2015 test-dev	30.5	50.8	32.1	11.4	35	44.8
Contextual information	ION [73]	VGG16	COCO 2015 test-dev	33.1	55.7	34.6	14.5	35.2	47.2
	DiCSSD300* [77]	VGG16	COCO 2015 test-dev	26.9	46.3	27.7	8.2	27.5	43.4
	SMN [78]	VGG16	COCO 2014 minval	31.6	52.2	33.2	14.4	35.7	45.8
Super-resolution	SOD-MTGAN [89]	ResNet101	COCO 2014 minval	41.5	62.5	45.4	25.1	44.6	54.1
Region-proposal	SSD-MCN [94]	VGG16	COCO 2015 test-dev	45.6	66.7	51.7	29.4	48.4	56.6
	PBLS-RPN [92]	VGG16	COCO 2015 test-dev	31.5	53.8	31.6	13.1	33.8	45.7

TABLE X
DETECTION RESULTS ON PASCAL VOC2007 TEST SET. EXCEPT FOR THE ION [76], THE OTHER METHODS ARE TRAINED ON VOC2007 AND VOC2012 TRAINVAL, AND TESTED ON VOC2007 TEST. ION ADDS SBD SEGMENTATION LABELS [128] DURING TRAINING

Model	Backbone	mAp	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
YOLO-inception [112]	Darknet53	78.37	88.32	85.88	76.67	69.07	66.63	83.76	87.08	87.96	61.08	79.51	73.71	86.84	85.87	85.22	82.4	52.16	79.11	74.78	86.42	74.86
MDSSD [63]	VGG16	80.3	**88.8**	**88.7**	**83.2**	73.7	58.3	88.2	**89.3**	87.4	62.4	85.1	75.1	84.7	89.7	**88.3**	83.2	56.7	**84.0**	77.4	83.9	77.6
CADNet512 [72]	VGG16	**80.6**	87.3	85.6	79.4	**74.8**	63.6	**88.3**	88.8	**88.3**	**65.5**	85.1	74.2	86	88.3	87.4	**84.2**	**58.8**	80.2	79.7	87.7	79.7
Multi-scale [68]	VGG16	76.6	78.8	82.4	75.6	67.2	64.9	85.3	88.2	87.3	59.8	83.2	73.6	85.2	86.3	77.9	79.3	48.9	76	72.8	83.6	75.4
ION [73]	VGG16	79.2	80.2	85.2	78.8	70.9	62.6	86.6	86.9	89.8	61.7	**86.9**	76.5	**88.4**	87.5	83.4	80.5	52.4	78.1	77.2	86.9	**83.5**
DICSSD300* [77]	VGG16	78.1	82.2	85.4	76.5	69.8	51.1	86.4	86.4	88	61.6	82.7	76.4	86.5	87.9	85.7	78.8	54.2	76.9	77.6	**88.9**	78.2
SSD+VSSA(Vertical) [74]	MobileNet	78.7	81.5	88.4	82.7	72.8	55.4	83.5	87.5	87.6	65.2	83.1	74.5	86.3	**90.0**	83.4	75.8	50.1	80.8	**82.4**	**88.9**	73.1
Feature-Fused-SSD [76]	VGG16	78.9	82	86.5	78	71.7	52.9	86.6	86.9	**88.3**	63.2	83	**76.8**	86.1	88.5	87.5	80.4	53.9	80.6	79.5	88.2	77.9
PBLS_SRPN [92]	VGG16	78.9	79.7	84.6	79.2	69.7	**68.9**	**88.3**	87.8	87.6	61.8	83.7	74.9	86.2	86.6	85.7	79.3	52.2	77.5	75.5	86.1	82.3

for small object detection because of its sensibility to localization accuracy. A new performance metric for small object detection is expected to guarantee high localization accuracy. Therefore, the pixel distance of center point between the predicted box and ground truth might be a new evaluation metric for small object detection.

B. Weakly Supervised Object Detection

Almost most previous generic object detectors train their model based on large public datasets, such as PASCAL VOC, ImageNet, and MS COCO. However, a few scenes and categories are only included in the small object datasets. It is difficult to train a generic network for small objects using a fully supervised learning method. Therefore, some researches adopt weakly supervised learning to detect small objects. For example, [130] only used image-level annotations to learn the object detectors by a dynamic curriculum learning strategy. Feng *et al.* [131] leveraged surrounding context information by a progressive contextual instance refinement (PCIR) method to avoid information loss of the whole object in the existing weakly supervised object detection. Cheng *et al.* [132] proposed a proposal generation method combining selective search and a gradient-weighted class activation mapping (Grad-CAM)-based technique, which can be widely applied to weakly supervised object detection.

C. Small Object Datasets

To date, there no large small object dataset such as COCO has existed. Many researchers adopted dataset constructed by themselves, which could not demonstrate obvious cross evaluation performance comparing with other methods. Moreover, most of the datasets focus on limited scenes, such as the face, pedestrian, and traffic signs. Therefore, a common small object dataset is vital which is accepted by most researchers and can provide a universal performance evaluation. However, building a small object dataset costs a lot of time and placing the bounding box properly for IOU evaluation is hard for the limited pixels of small objects.

D. Combination of Multiple Kinds of Methods

Contextual information, the fusion of multiscale feature maps, super-resolution images, and smaller anchor size in the region-proposal stage are four different methods currently used to improve the performance of small object detection. Generally, the current leading generic detection framework is selected as the backbone of the small object detection network and some other modules could be integrated into the backbone, like the above four. Moreover, these modules could be combined to improve the detection result.

E. Small Object Detection in Videos

With the increasing of video data, relative work about object detection in videos has attracted much attention because it provides more consequent and richer information compared to a still image. The task of object detection in the video requires objects of each frame to be located with bounding boxes. Otherwise, it is a kind of real-time object detection applied to the autonomous driving and monitoring systems. Although current deep learning-based methods have obtained impressive performance on object detection in still images, object detection in the video is facing many challenges. Exploring the

TABLE XI
DETECTION RESULTS ON TSINGHUA-TENCENT 100K

Model	**Metrics**	**i2**	**i4**	**i5**	**il100**	**il60**	**il80**	**io**	**ip**	**p10**	**p11**	**p12**	**p19**	**p23**	**p26**	**p27**
MR-CNN [65]	recall	80.7	88.6	92.4	93.1	90.3	95.5	89.6	84.1	90.4	84.3	88.6	94.5	89.7	88.2	92.3
	precision	82.2	91.8	94.5	92.2	94.8	87.7	80.3	87.4	74.7	91.2	90.3	95.1	93.4	83.5	87
	F1-measure	81.4	90.2	93.4	92.6	92.5	91.4	84.7	85.7	81.8	87.6	89.4	94.8	91.5	85.8	89.6
FPN* [113]	recall	87	97	96	97	98	100	94	88	92	95	95	91	94	95	98
	precision	90	92	94	93	98	94	86	90	89	90	94	75	93	89	98
PGAN [87]	recall	84	95	95	95	92	95	92	91	89	96	97	97	95	94	98
	accuracy	85	92	94	97	95	83	79	90	84	85	88	84	92	83	98
Model	**Metrics**	**p3**	**p5**	**p6**	**pg**	**ph4**	**ph4.5**	**ph5**	**pl100**	**pl120**	**pl20**	**pl30**	**pl40**	**pl5**	**pl50**	**pl60**
MR-CNN [65]	recall	88.4	92.1	88.9	91.5	78.7	88	75.9	93.9	94.2	85.3	91.7	91.4	85.3	92.2	83.7
	precision	76.6	93.6	76.7	93.2	80.5	84.2	82.8	94.7	91.4	90.6	90.8	90.5	87.6	86.5	91.8
	F1-measure	82.1	92.8	82.4	92.3	79.6	86.1	79.2	94.3	92.8	87.9	91.2	90.9	86.4	89.3	87.6
FPN* [113]	recall	96	98	97	98	86	90	90	100	97	98	97	97	94	97	98
	precision	81	91	90	93	94	80	78	98	99	90	92	91	92	90	95
PGAN [87]	recall	93	96	100	93	78	88	85	96	98	96	93	96	92	96	91
	accuracy	92	90	83	93	97	68	69	97	98	92	91	90	86	87	92
Model	**Metrics**	**pl70**	**pl80**	**pm20**	**pm30**	**pm55**	**pn**	**pne**	**po**	**pr40**	**w13**	**w32**	**w55**	**w57**	**w59**	**wo**
MR-CNN [65]	recall	88.6	92.3	88.4	91.8	93.5	88.2	92.5	70.6	92.8	83.2	68.6	63.1	84.6	74.5	42.8
	precision	84.5	86.3	92.7	88.9	78.1	90.4	89.2	77.5	93.3	82.5	82.3	82.5	89.3	75.1	41.6
	F1-measure	86.5	89.2	90.5	90.3	85.1	89.3	90.8	73.9	93	82.8	74.8	71.5	86.9	74.8	42.2
FPN* [113]	recall	93	99	94	96	97	96	96	82	100	90	91	95	94	93	42
	precision	98	92	98	97	86	90	97	81	97	90	95	95	90	68	50
PGAN [87]	recall	91	99	88	94	100	96	97	83	97	94	85	95	94	95	53
	accuracy	97	86	90	77	81	89	93	78	92	66	83	88	93	71	54

spatial and temporal correlation by methods, such as optical flow and LSTM, may be a breakthrough point.

F. High Precision or Real-Time Detection Framework

Similar to generic object detection, small object detection also faces the problem that how to make a balance between precision and inference time. As the classic region-proposal network, faster R-CNN is famous for its high detection accuracy. However, it has a longer inference time compared with YOLO and SSD. In fact, the balance of detection accuracy and inference time is decided by different application scenarios. For example, high accuracy may be the key point when personal identity is verified by detecting faces in the bank. From another perspective, high detection speed would be the first choice when small object detection techniques are applied to intelligent transportation, military monitoring, and unmanned aerial vehicle. Thus, lightweight networks need to be designed to suit the resource-limited devices.

V. CONCLUSION

Generic object detection has achieved great success due to powerful learning ability of deep learning methods. These neural networks even perform better than humans to some extent. In recent years, small object detection has also drawn more attention in the computer vision field. Deep learning-based studies make significant contributions to the development of the self-driving car, UAV, and other vision systems, which can detect objects fast in a far distance. Based on the current research status of small object detection, lots of related excellent deep learning-based works are collected in this article, providing a detailed classification summary. Moreover, the most representative networks at five major categories are described in detail and specific datasets of small object are summarized. This survey is meaningful for the development of small object detection, which could provide guidance for further research in this field.

REFERENCES

[1] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 304–311.

[2] Y. Xu, D. Xu, S. Lin, T. X. Han, X. Cao, and X. Li, "Detection of sudden pedestrian crossings for driving assistance systems," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 729–739, Jun. 2012.

[3] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[4] J. Chen, R. Wang, S. Yan, S. Shan, X. Chen, and W. Gao, "Enhancing human face detection by resampling examples through manifolds," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 37, no. 6, pp. 1017–1028, Nov. 2007.

[5] G. Chen, L. Hong, J. Dong, P. Liu, J. Conradt, and A. Knoll, "EDDD: Event-based drowsiness driving detection through facial motion analysis with neuromorphic vision sensor," *IEEE Sens. J.*, vol. 20, no. 11, pp. 6170–6181, Jun. 2020.

[6] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1907–1915.

[7] S. Yang, W. Wang, C. Liu, and W. Deng, "Scene understanding in deep learning-based end-to-end controllers for autonomous vehicles," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 1, pp. 53–63, Jan. 2018.

[8] G. Chen, H. Cao, J. Conradt, H. Tang, F. Röhrbein, and A. Knoll, "Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception," *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 34–49, Jul. 2020.

[9] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati, "Active vision for sociable robots," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 31, no. 5, pp. 443–453, Sep. 2001.

[10] Z. Li, J. Li, S. Zhao, Y. Yuan, Y. Kang, and C. L. P. Chen, "Adaptive neural control of a kinematically redundant exoskeleton robot using brain–machine interfaces," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3558–3571, Dec. 2019.

[11] Y. Liu *et al.*, "Motor-imagery-based teleoperation of a dual-arm robot performing manipulation tasks," *IEEE Trans. Cogn. Develop. Syst.*, vol. 11, no. 3, pp. 414–424, Sep. 2018.

[12] J. R.-D. Solar, P. Loncomilla, and N. Soto, "A survey on deep learning methods for robot vision," 2018. [Online]. Available: arXiv:1803.10862.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, p. 436, May 2015.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[15] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.

[16] K. Chen *et al.*, “Hybrid task cascade for instance segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4974–4983.

[17] T.-Y. Lin *et al.*, “Microsoft COCO: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VoC) challenge,” *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[20] P. A. Viola and M. J. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proc. CVPR*, vol. 1, 2001, pp. 511–518.

[21] D. K. Prasad, “Survey of the problem of object detection in real images,” *Int. J. Image Process.*, vol. 6, no. 6, p. 441, 2012.

[22] D. G. Lowe *et al.*, “Object recognition from local scale-invariant features,” in *Proc. ICCV*, vol. 99, 1999, pp. 1150–1157.

[23] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[24] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, 2006, pp. 2169–2178.

[25] P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in *Proc. CVPR*, vol. 2, 2008, p. 7.

[26] K. E. A. Van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, “Segmentation as selective search for object recognition,” in *Proc. ICCV*, vol. 1, 2011, p. 7.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[28] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[29] R. Girshick, “Fast R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[30] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[31] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[32] P. Pinggera, S. Ramos, S. Gehrig, U. Franke, C. Rother, and R. Mester, “Lost and found: Detecting small road hazards for self-driving vehicles,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2016, pp. 1099–1106.

[33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[35] W. Liu *et al.*, “SSD: Single shot multibox detector,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[36] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, “DSSD: Deconvolutional single shot detector,” 2017. [Online]. Available: arXiv:1701.06659.

[37] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[38] Y. Wu *et al.*, “Rethinking classification and localization in R-CNN,” 2019. [Online]. Available: arXiv:1904.06493.

[39] H. Krishna and C. V. Jawahar, “Improving small object detection,” in *Proc. 4th IAPR Asian Conf. Pattern Recognit. (ACPR)*, 2017, pp. 340–345.

[40] G. Chen *et al.*, “A novel visible light positioning system with event-base neuromorphic vision sensor,” *IEEE Sens. J.*, early access, Apr. 27, 2020, doi: 10.1109/JSEN.2020.2990752.

[41] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, “Semantic annotation of high-resolution satellite images via weakly supervised learning,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[42] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, “Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[43] L. Liu *et al.*, “Deep learning for generic object detection: A survey,” 2018. [Online]. Available: arXiv:1809.02165.

[44] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, “Object detection with deep learning: A review,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[45] H. Zhang and X. Hong, “Recent progresses on object detection: A brief review,” *Multimedia Tools Appl.*, vol. 78, pp. 27809–27847, Jun. 2019.

[46] L. Jiao *et al.*, “A survey of deep learning-based object detection,” *IEEE Access*, vol. 7, pp. 128837–128868, 2019.

[47] Z. Zou, Z. Shi, Y. Guo, and J. Ye, “Object detection in 20 years: A survey,” 2019. [Online]. Available: arXiv:1905.05055.

[48] K. Chen *et al.*, “MmDetection: Open MMLAB detection toolbox and benchmark,” 2019. [Online]. Available: arXiv:1906.07155.

[49] F. Larsson and M. Felsberg, “Using Fourier descriptors and spatial models for traffic signal recognition,” in *Proc. Scandinavian Conf. Image Anal.*, 2011, pp. 238–249.

[50] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, “Traffic-sign detection and classification in the wild,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2110–2118.

[51] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, “Detection of traffic signs in real-world images: The German traffic sign detection benchmark,” in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 2013, pp. 1–8.

[52] D. Temel, T. Alshawi, M.-H. Chen, and G. AlRegib, “Challenging environments for traffic sign detection: Reliability assessment under inclement conditions,” 2019. [Online]. Available: arXiv:1902.06857.

[53] C. Chen, M.-Y. Liu, O. Tuzel, and J. Xiao, “R-CNN for small object detection,” in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 214–230.

[54] D. Temel, J. Lee, and G. AlRegib, “Cure-OR: Challenging unreal and real environments for object recognition,” in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, 2018, pp. 137–144.

[55] S. Yang, P. Luo, C.-C. Loy, and X. Tang, “Wider face: A face detection benchmark,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5525–5533.

[56] L. Tuggener, I. Elezi, J. Schmidhuber, M. Pelillo, and T. Stadelmann, “DeepScores—A dataset for segmentation, detection and classification of tiny objects,” in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, 2018, pp. 3704–3709.

[57] A. Ghahremani, E. Bondarev, and P. H. N. De With, “Cascaded CNN method for far object detection in outdoor surveillance,” in *Proc. 14th Int. Conf. Signal Image Technol. Internet Based Syst. (SITIS)*, 2018, pp. 40–47.

[58] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, “The German traffic sign recognition benchmark: A multi-class classification competition,” in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2011, pp. 1453–1460.

[59] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, “Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Netw.*, vol. 32, pp. 323–332, Aug. 2012.

[60] R. Timofte, K. Zimmermann, and L. Van Gool, “Multi-view traffic sign detection, recognition, and 3D localisation,” *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 633–647, 2014.

[61] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva, “Sun database: Exploring a large collection of scene categories,” *Int. J. Comput. Vis.*, vol. 119, pp. 3–22, Aug. 2016.

[62] Y. Xiong, K. Zhu, D. Lin, and X. Tang, “Recognize complex events from static images by fusing deep channels,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1600–1609.

[63] M. Xu *et al.*, “MDSSD: Multi-scale deconvolutional single shot detector for small objects,” 2018. [Online]. Available: arXiv:1805.07009.

[64] Z. Liu, D. Li, S. S. Ge, and F. Tian, “Small traffic sign detection from large image,” *Appl. Intell.*, vol. 50, pp. 1–13, Jan. 2020.

[65] Z. Liu, J. Du, F. Tian, and J. Wen, “MR-CNN: A multi-scale region-based convolutional neural network for small traffic sign recognition,” *IEEE Access*, vol. 7, pp. 57120–57128, 2019.

[66] S. Sun, “Multiple receptive fields and small-object-focusing weakly-supervise segmentation network for fast object detection,” 2019. [Online]. Available: arXiv:1904.12619.

[67] Q. Meng, H. Song, G. Li, Y. Zhang, and X. Zhang, "A block object detection method based on feature fusion networks for autonomous vehicles," *Complexity*, vol. 2019, pp. 1–14, Feb. 2019.

[68] D. W. Ma, X. J. Wu, and H. Yang, "Efficient small object detection with an improved region proposal networks," in *Proc. IOP Conf. Mater. Sci. Eng.*, vol. 533, 2019, Art. no. 012062.

[69] W. Liu, S. Liao, W. Hu, X. Liang, and Y. Zhang, "Improving tiny vehicle detection in complex scenes," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2018, pp. 1–6.

[70] G. X. Hu, Z. Yang, L. Hu, L. Huang, and J. M. Han, "Small object detection with multiscale features," *Int. J. Digit. Multimedia Broadcast.*, vol. 2018, pp. 1–10, Sep. 2018.

[71] R. Zhang *et al.*, "A detection method for low-pixel ratio object," *Multimedia Tools Appl.*, vol. 78, pp. 11655–11674, Oct. 2018.

[72] K. Duan, D. Du, H. Qi, and Q. Huang, "Detecting small objects using a channel-aware deconvolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1639–1652, Jun. 2020.

[73] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2874–2883.

[74] Y. Yuan, Z. Xiong, and Q. Wang, "VSSA-NET: Vertical spatial sequence attention network for traffic sign detection," 2019. [Online]. Available: arXiv:1905.01583.

[75] Y. Liu, S. Cao, P. Lasang, and S. Shen, "Modular lightweight network for road object detection using a feature fusion approach," *IEEE Trans. Syst., Man, Cybern., Syst.*, early access, Oct. 16, 2019, doi: 10.1109/TSMC.2019.2945053.

[76] G. Cao, X. Xie, W. Yang, Q. Liao, G. Shi, and J. Wu, "Feature-fused SSD: Fast detection for small objects," in *Proc. 9th Int. Conf. Graph. Image Process. (ICGIP)*, vol. 10615, 2018, Art. no. 106151E.

[77] W. Xiang, D.-Q. Zhang, H. Yu, and V. Athitsos, "Context-aware single-shot detector," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018, pp. 1784–1793.

[78] X. Chen and A. Gupta, "Spatial memory for context reasoning in object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 4106–4116.

[79] J. Müller and K. Dietmayer, "Detecting traffic lights by single shot detection," 2018. [Online]. Available: arXiv:1805.02523.

[80] L. Guan, Y. Wu, and J. Zhao, "SCAN: Semantic context aware network for accurate small object detection," *Int. J. Comput. Intell. Syst.*, vol. 11, no. 1, p. 936, 2018.

[81] X. Hu *et al.*, "SINet: A scale-insensitive convolutional neural network for fast vehicle detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 1010–1019, Apr. 2019.

[82] J. Leng, Y. Liu, D. Du, T. Zhang, and P. Quan, "Robust obstacle detection and recognition for driver assistance systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1560–1571, Apr. 2020.

[83] P. Fang and Y. Shi, "Small object detection using context information fusion in faster R-CNN," in *Proc. IEEE 4th Int. Conf. Comput. Commun. (ICCC)*, 2018, pp. 1537–1540.

[84] H. Long, Y. Chung, Z. Liu, and S. Bu, "Object detection in aerial images using feature fusion deep networks," *IEEE Access*, vol. 7, pp. 30980–30990, 2019.

[85] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian, "Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1758–1770, Jun. 2020.

[86] J. Han, X. Yao, G. Cheng, X. Feng, and D. Xu, "P-CNN: Part-based convolutional neural networks for fine-grained visual categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 6, 2019, doi: 10.1109/TPAMI.2019.2933510.

[87] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1951–1959.

[88] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Finding tiny faces in the wild with generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 21–30.

[89] Y. Zhang, Y. Bai, M. Ding, and B. Ghanem, "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 206–221.

[90] Y. Pang, J. Cao, J. Wang, and J. Han, "JCS-Net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 12, pp. 3322–3331, Dec. 2019.

[91] C. Wilms and S. Frintrop, "AttentionMask: Attentive, efficient object proposal generation focusing on small objects," 2018. [Online]. Available: arXiv:1811.08728.

[92] G. Wang, J. Guo, Y. Chen, Y. Li, and Q. Xu, "A PSO and BFO-based learning strategy applied to faster R-CNN for object detection in autonomous driving," *IEEE Access*, vol. 7, pp. 18840–18859, 2019.

[93] Z. Meng, X. Fan, X. Chen, M. Chen, and Y. Tong, "Detecting small signs from large images," in *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, Aug. 2017, pp. 217–224.

[94] Z. Chen, K. Wu, Y. Li, M. Wang, and W. Li, "SSD-MSN: An improved multi-scale object detection network based on SSD," *IEEE Access*, vol. 7, pp. 80622–80632, 2019.

[95] L. Fang, X. Zhao, and S. Zhang, "Small-objectness sensitive detection based on shifted single shot detector," *Multimedia Tools Appl.*, vol. 78, no. 4, pp. 13227–13245, 2018.

[96] C. Eggert, D. Zecha, S. Brehm, and R. Lienhart, "Improving small object proposals for company logo detection," in *Proc. ACM Int. Conf. Multimedia Retrieval (ICMR)*, 2017, pp. 167–174.

[97] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," 2019. [Online]. Available: arXiv:1902.07296.

[98] G. Wang, Z. Xiong, D. Liu, and C. Luo, "Cascade mask generation framework for fast small object detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2018, pp. 1–6.

[99] X. Luo, J. Zhu, and Q. Yu, "Efficient ConvNets for fast traffic sign recognition," *IET Intell. Transp. Syst.*, vol. 13, no. 6, pp. 1011–1015, 2019.

[100] P.-Y. Chen, J.-W. Hsieh, M. Gochoo, C.-Y. Wang, H.-Yuan, and M. Liao, "Smaller object detection for real-time embedded traffic flow estimation using fish-eye cameras," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2019, pp. 2956–2960.

[101] K. Yi, Z. Jian, S. Chen, Y. Yang, and N. Zheng, "Knowledge-based recurrent attentive neural network for small object detection," 2018. [Online]. Available: arXiv:1803.05263.

[102] Z. Yang, Y. Liu, L. Liu, X. Tang, J. Xie, and X. Gao, "Detecting small objects in urban settings using SlimNet model," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8445–8457, Nov. 2019.

[103] C. Cao *et al.*, "An improved faster R-CNN for small object detection," *IEEE Access*, vol. 7, pp. 106838–106846, 2019.

[104] C. Han, G. Gao, and Y. Zhang, "Real-time small traffic sign detection with revised faster-RCNN," *Multimedia Tools Appl.*, vol. 78, pp. 13263–13278, Aug. 2018.

[105] H.-K. Kim, J. H. Park, and H.-Y. Jung, "An efficient color space for deep-learning based traffic light recognition," *J. Adv. Transp.*, vol. 2018, pp. 1–12, Dec. 2018.

[106] Q. Zhao, T. Sheng, Y. Wang, F. Ni, and L. Cai, "CFENet: An accurate and efficient single-shot object detector for autonomous driving," 2018. [Online]. Available: arXiv:1806.09790.

[107] Q. Zhang, C. Wan, and S. Bian, "Research on vehicle object detection method based on convolutional neural network," in *Proc. 11th Int. Symp. Comput. Intell. Design (ISCID)*, 2018, pp. 271–274.

[108] P. Pham, D. Nguyen, T. Do, T. D. Ngo, and D.-D. Le, "Evaluation of deep models for real-time small object detection," *Neural Inf. Process.*, vol. 10636, pp. 516–526, Oct. 2017.

[109] M. Menikdiwela, C. Nguyen, H. Li, and M. Shaw, "CNN-based small object detection and visualization with feature activation mapping," in *Proc. Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, 2017, pp. 1–5.

[110] F. Xu *et al.*, "Real-time detecting method of marine small object with underwater robot vision," in *Proc. OCEANS MTS/IEEE Kobe Techno-Oceans (OTO)*, 2018, pp. 1–4.

[111] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[112] P. Du, X. Qu, T. Wei, C. Peng, X. Zhong, and C. Chen, "Research on small size object detection in complex background," in *Proc. Chin. Autom. Congr. (CAC)*, 2018, pp. 4216–4220.

[113] Z. Liang, J. Shao, D. Zhang, and L. Gao, "Small object detection using deep feature pyramid networks," in *Proc. Adv. Multimedia Inf. Process. (PCM)*, 2018, pp. 554–564.

[114] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[115] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends Cogn. Sci.*, vol. 11, no. 12, pp. 520–527, 2007.

[116] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015. [Online]. Available: arXiv:1511.07122.

[117] Q. V. Le, N. Jaitly, and G. E. Hinton, "A simple way to initialize recurrent networks of rectified linear units," 2015. [Online]. Available: arXiv:1504.00941.

[118] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available: arXiv:1704.04861.

[119] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[120] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.

[121] Y. Bai and B. Ghanem, "Multi-branch fully convolutional network for face detection," 2017. [Online]. Available: arXiv:1707.06330.

[122] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.

[123] J. Cao, Y. Pang, and X. Li, "Learning multilayer channel features for pedestrian detection," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3210–3220, Jul. 2017.

[124] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Conf.*, 2009, pp. 1–11.

[125] T. Guan and H. Zhu, "Atrous faster R-CNN for small scale object detection," in *Proc. 2nd Int. Conf. Multimedia Image Process. (ICMIP)*, 2017, pp. 16–21.

[126] D. Chen, G. Hua, F. Wen, and J. Sun, "Supervised transformer network for efficient face detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 122–138.

[127] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on topological LIN localization and temporal feature aggregation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 536–551.

[128] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 991–998.

[129] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logist. Quart.*, vol. 2, nos. 1–2, pp. 83–97, 1955.

[130] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo, "Automatic weakly supervised object detection from high spatia resolution remote sensing images via dynamic curriculum learning," *IEEE Trans. Geosci. Remote Sens.*, early access, May 18, 2020, doi: 10.1109/TGRS.2020.2991407.

[131] X. Feng, J. Han, X. Yao, and G. Cheng, "Progressive contextual instance refinement for weakly supervise, object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Apr. 27, 2020, doi: 10.1109/TGRS.2020.2985989.

[132] G. Cheng, J. Yang, D. Gao, L. Guo, and J. Han, "High-quality proposals for weakly supervised object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 5794–5804, Apr. 2020.

Guang Chen (Member, IEEE) received the B.S. degree from the College of Mechanical and Vehicle Engineering, Hunan University, Changsha, China, in 2008, the M.Eng. degree from Hunan University in 2011, and the Ph.D. degree in computer science from the Technical University of Munich, Munich, Germany, in 2016.

He is a Research Professor with Tongji University, Shanghai, China, and a Senior Research Associate (Guest) with the Technical University of Munich. He is a Visiting Researcher with the State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, Hunan University, from 2019 to 2020. He is leading the Intelligent Perception and Intelligent Computation Group with Tongji University. He was a Research Scientist with fortiss GmbH, Munich, from 2012 to 2016, and a Senior Researcher with the Chair of Robotics, Artificial Intelligence and Real-Time Systems, Technical University of Munich from 2016 to 2017. His research interests include computer vision, machine learning, and bio-inspired vision with applications in robotics and autonomous vehicle.

Dr. Chen was awarded the Program of Tongji Hundred Talent Research Professor 2018.

Haitao Wang received the B.S. degree in vehicle engineering from Jilin University, Changchun, China, in 2019. He is currently pursuing the M.S. degree in vehicle engineering with Tongji University, Shanghai, China.

His current research interests include computer vision and autonomous vehicle.

Kai Chen received the B.S. degree in vehicle engineering from Tongji University, Shanghai, China, in 2019, where he is currently pursuing the Ph.D. degree in vehicle engineering.

His current research interests include computer vision and autonomous vehicle.

Zhijun Li (Senior Member, IEEE) received the Ph.D. degree in mechatronics from Shanghai Jiao Tong University, Shanghai, China, in 2002.

From 2003 to 2005, he was a Postdoctoral Fellow with the Department of Mechanical Engineering and Intelligent Systems, University of Electro-Communications, Tokyo, Japan. From 2005 to 2006, he was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and Nanyang Technological University, Singapore. Since 2017, he has been a Professor with the Department of Automation, University of Science and Technology, Hefei, China, where he has been the Vice Dean with the School of Information Science and Technology since 2019. His current research interests include wearable robotics, teleoperation systems, nonlinear control, and neural network optimization.

Prof. Li has been the Co-Chair of IEEE SMC Technical Committee on Bio-Mechatronics and Bio-Robotics Systems and IEEE RAS Technical Committee on Neuro-Robotics Systems since 2016. He is serving as an Editor-at-Large for the *Journal of Intelligent & Robotic Systems*, and an Associate Editor of several IEEE TRANSACTIONS.

Zida Song received the B.S. degree in vehicle engineering from Jilin University, Changchun, China, in 2018. He is currently pursuing the M.S. degree in vehicle engineering with Tongji University, Shanghai, China.

His current research interests include deep learning and computer vision.

Yinlong Liu received the M.S. degree in biomedical engineering from the Digital Medical Research Center, School of Basic Medical Science, Fudan University, Shanghai, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Informatics, Technical University of Munich, Munich, Germany.

His research interests include medical robot and geometric computer vision.

Wenkai Chen received the B.S. degree in mechanical engineering from Wuhan University, Wuhan, China, in 2017. He is currently pursuing the M.S. degree in mechanical engineering with Shanghai Jiao Tong University, Shanghai, China.

From 2019 to 2020, he worked as a Research Assistant with the Intelligent Perception and Computing Group, Tongji University, Shanghai. His research interest includes perception and localization of autonomous robot, computer vision based on deep learning, and cross-modal learning.

Alois Knoll (Senior Member, IEEE) received the Diploma (M.Sc.) degree in electrical/communications engineering from the University of Stuttgart, Stuttgart, Germany, in 1985, and the Ph.D. degree *(summa cum laude)* in computer science from the Technical University of Berlin, Berlin, Germany, in 1988.

He served on the faculty of the Computer Science Department, Technical University of Berlin until 1993. He joined the University of Bielefeld, Bielefeld, Germany, as a Full Professor and the Director of the Research Group Technical Informatics until 2001. Since 2001, he has been a Professor with the Department of Informatics, Technical University of Munich, Munich, Germany, where he was the Executive Director of the Institute of Computer Science from 2004 to 2006.

Prof. Knoll was the Program Chairman of IEEE Humanoids2000, the General Chair of IEEE Humanoids2003, the Program Chair of IEEE-IROS 2015, and the Editor-in-Chief of *Frontiers in Neurorobotics*.