Image and Vision Computing 97 (2020) 103910

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Review article

Recent advances in small object detection based on deep learning: A review☆

Kang Tong, Yiquan Wu *, Fei Zhou

College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China

ARTICLE INFO

Article history:
Received 8 March 2020
Accepted 19 March 2020
Available online 23 March 2020

Keywords:
Small object detection
Deep learning
Computer vision
Convolutional neural networks

ABSTRACT

Small object detection is a challenging problem in computer vision. It has been widely applied in defense military, transportation, industry, etc. To facilitate in-depth understanding of small object detection, we comprehensively review the existing small object detection methods based on deep learning from five aspects, including multi-scale feature learning, data augmentation, training strategy, context-based detection and GAN-based detection. Then, we thoroughly analyze the performance of some typical small object detection algorithms on popular datasets, such as MS-COCO, PASCAL-VOC. Finally, the possible research directions in the future are pointed out from five perspectives: emerging small object detection datasets and benchmarks, multi-task joint learning and optimization, information transmission, weakly supervised small object detection methods and framework for small object detection task.

Contents

☆ This paper has been recommended for acceptance by Sinisa Todorovic.

* Corresponding author.
E-mail addresses: tkangcv@nuaa.edu.cn (K. Tong), mltd2099@163.com (Y. Wu), F.zhouip@nuaa.edu.cn (F. Zhou).

https://doi.org/10.1016/j.imavis.2020.103910
0262-8856/

1. Introduction

Small object detection is a fundamental computer technology related to image understanding and computer vision that deals with detecting instances of small objects of a certain class in digital images and videos. As an indispensable and challenging problem in computer vision, small object detection forms the basis of many other computer vision tasks, such as object tracking [1], instance segmentation [2,3], image captioning [4], action recognition [5], scene understanding [6], etc. In recent years, the compelling success of deep learning techniques has brought new blood into small object detection, pushing it forward to a research highlight. Small object detection has been widely used in academia and real world applications, such as robot vision, autonomous driving, intelligent transportation, drone scene analysis, military reconnaissance and surveillance.

There are mainly two definitions of small objects. One refers to objects with smaller physical sizes in the real world. Another definition of small objects is mentioned in MS-COCO [7] metric evaluation. Objects occupying areas less than and equal to 32×32 pixels come under "small objects" category and this size threshold is generally accepted within the community for datasets related to common objects. Some instances of small objects (e.g. "baseball", "tennis" and traffic sign "pg") are shown in Fig. 1. Although many object detectors perform well on medium and large objects, they perform poorly on the task of detecting small objects. This is because that there are three difficulties in small object detection. First, small objects lack appearance information needed to distinguish them from background or similar categories. Then the locations of small objects have much more possibilities. That is to say, the precision requirement for accurate localization is higher. Furthermore, the experiences and knowledge of small object detection are very limited because the majority of prior efforts are tuned for the large object detection problem.

In this paper we provide a comprehensive and in-depth survey on small object detection in the deep learning era. Our survey aims to cover thoroughly five respects of small object detection algorithms, including multi-scale feature learning, data augmentation, training strategy, context-based detection and GAN-based detection. Aside from taxonomically reviewing the existing small object detection methods, we investigate datasets and evaluation metrics of small object detection. Meanwhile, we thoroughly analyze the performance of small object detection methods and present several promising directions for future work.

1.1. History and scope

Compared with other computer vision tasks, the history of small object detection is relatively short. Earlier work on small object detection is mostly about detecting vehicles utilizing hand-engineered features and shallow classifiers in aerial images [8,9]. Before the prevalent of deep learning, color and shape-based features are also used to address traffic sign detection problems [10]. With the rapid advancement of convolutional neural networks (CNNs) in deep learning, some deep learning-based small object detection methods have sprung up. However, there are relatively few surveys and researches focusing only on small object detection. Most of the state-of-the-art methods are based on existing object detection algorithms with some modifications so as to improve the detection performance of small objects. To the best of our knowledge, Chen et al. [11] are perhaps the first to introduce a small object detection (SOD) dataset, an evaluation metric, and provide a baseline score in order to explore small object detection. Later, Krishna and Jawahar [12] build upon their ideas and suggest an effective upsampling-based technique that performs better results on small object detection. Different from the R-CNN (regions with CNN features) used in [11,12], Zhang et al. [13] use deconvolution R-CNN [14] for small object detection on remote sensing images. Faster R-CNN [15] and single shot detector (SSD) [16] are two major approaches in object detection. Based on Faster R-CNN or SSD, some small object detection methods [17–21] are proposed. Furthermore, multi-scale techniques [22,23], data augmentation techniques [24], training strategies [25,26], contextual information [27,28] and generative adversarial networks (GAN) [29,30] are also used for detecting small objects. A brief chronology is exhibited in Table 1.

We carefully and thoroughly select significant or influential papers published in prestigious journals and conferences. This review mainly focuses on the major progress of small object detection in the last from three to five years. But some other related works are also included in order to completeness and better readability. It is worth noting that we restrict this review to image-level small object detection methods. Other work for small object detection, such as 3D small object detection and video small object detection will not be included in our discussion.

1.2. Comparison with previous reviews

A number of notable object detection reviews have been published. These include many reviews on the detection under specific application scenarios, such as face detection [54–56], text detection [57,58], vehicle detection [59], traffic sign detection [60], pedestrian detection [61,62], and remote sensing target detection [63,64]. In addition to these category specific object detection surveys, there are many generic object detection surveys [65–70]. Among these works, Agarwal et al. [65] systematically review recent papers on object detection with deep CNN, including typical architectures, current challenges and public datasets. Zhao et al. [66] provide a thoroughly survey on deep learning based object detection frameworks. Furthermore, Liu et al. [67] comprehensively review many aspects of generic object detection, covering detection frameworks, fundamental subproblems, evaluation issues and state-of-the-art performance. Similarly, Zou et al. [69] extensively review many topics of object detection in 20 years, including milestone detectors in history, detection methods, detection metrics, datasets, speed up techniques, detection applications and challenges.

From the perspective of small object detection, these surveys rarely or even do not elaborate small object detection. Zou et al. [69] and Zhao et al. [66] both mention small object detection, but they only show it in the future directions. In addition, Agarwal et al. [65] briefly introduce several methods for detecting small objects in the major challenges of the object detection task. However, they do not systematically analyze it in depth. Unlike these previous object detection surveys, we present a systematic and comprehensive review of deep learning-based algorithms that handle small object detection problems. Our survey is featured by in-depth analysis of small object detection. We summarize existing small object detection algorithms based on five different perspectives: multi-scale feature learning, data augmentation, training strategy, context-based detection and GAN-based detection. Furthermore, the performance of some typical algorithms on popular object detection datasets is carefully analyzed. Last the possible future research directions are discussed. We hope that our survey can provide researchers and practitioners with timely reviews and novel inspirations to facilitate understanding of small object detection and further catalyze research on detection systems.

1.3. Our contributions

Our contributions in this paper are summarized as follows:

1) Systematic review of small object detection methods. We categorize and summarize the existing detection methods of small objects from five respects, including multi-scale feature learning, data augmentation, training strategy, context-based detection and GAN-based detection. The proposed taxonomies aim to help researchers with deeper and comprehensive understanding of small object detection in the deep learning era.

Fig. 1. Some instances of small objects.

2) In-depth summary and analysis of small object detection performance. Based on our taxonomy of small object detection algorithms, we assess and analyze the detection results of these state-of-the-art algorithms on different datasets.
3) Overview of future directions. According to the taxonomy methods and performance analysis of small object detection, we shed light on potential directions for future research from five perspectives: emerging small object detection datasets and benchmarks, multi-task joint learning and optimization, information transmission, weakly supervised small object detection methods, framework for small object detection task.

The rest of the paper is organized as follows. The details and methods of small object detection are listed in Section 2. Then datasets and the detection performance of small objects are presented in Section 3. Finally, we conclude and discuss future directions in Section 4.

2. Methods for small object detection

In this section, we will extensively review the methods of detecting small objects from five aspects, including multi-scale feature learning, data augmentation, training strategy, context-based detection and GAN-based detection. A taxonomy of small object detection methods is shown in Table 2.

2.1. Multi-scale feature learning

Handling feature scale issues is of crucial importance for small object detection. There are seven main paradigms addressing multi-scale feature learning problem: featurized image pyramids, single feature map, pyramidal feature hierarchy, integrated features, feature pyramid network, feature fusion and feature pyramid generation, and multi-scaled fusion module. These are briefly illustrated in Fig. 2.

Table 1
A brief chronology of small object detection.

Year	Methods (chronological order from left to right)
2014–2016	R-CNN [14]➔ SPPNet [31]➔ Context-SVM [32]➔ DeepIDNet [33]➔ Fast R-CNN [34]➔ MRCNN [35]➔ Faster R-CNN [15]➔ SegDeepM [36]
2016–2017	HyperNet [37]➔ R-FCN [38]➔ MSCNN [39]➔ MPNet [40]➔ GBDNet [41]➔ CPF [42]➔ R-CNN-SOD [11]➔ ION [27]➔ SSD [16]
2017–2018	DSSD [19]➔ SMN [43]➔ FPN [22]➔ ACCNN [44]➔ Perceptual-GAN [29]➔ CoupleNet [45]➔ SOD-Faster-R-CNN [17] ➔ FFSSD [20]➔ ISOD [12]➔ FSSD [46]
2018–2019	RefineDet [47]➔ SCAN [28]➔ R-FCN++ [48]➔ SNIP [25]➔ ORN [49] ➔ SIN [50]➔ Deconv-R-CNN [13]➔ SAN [51]➔ MTGAN [30]➔ DFPN [23]➔ CasMaskGF [52]➔ SNIPER [26]
2019–2020	M2DNet [53]➔ MDSSD [21]➔ Augmentation [24]➔ Improved-Faster-R-CNN-SOD [18]

A simple operation is to resize input images into many different scales and to learn multiple detectors, each of which is in charge of a certain range of scales. Most of previous object detectors are based on hand-engineered features [71,72] utilizing featurized image pyramids (see Fig. 2(a)) to detect objects at various scales. However, the hand-

Table 2
A taxonomy of small object detection methods.

Type		Method
Multi-scale feature learning (see Section 2.1)	Featurized image pyramids	SNIP [25]
		CasMaskGF [52]
	Single feature map	Fast R-CNN [34]
		Faster R-CNN [15]
		SPPNet [31]
		R-FCN [38]
	Pyramidal feature hierarchy	SSD [16]
		MSCNN [39]
	Integrated features	ION [27]
		HyperNet [37]
	Feature pyramid network	FPN [22]
	Variants of FPN (including feature fusion and feature pyramid generation, multi-scaled fusion module, etc.)	RefineDet [47]
		M2Det [53]
		DFPN [23]
		FFSSD [20]
		FSSD [46]
		DSSD [19]
		MDSSD [21]
Data augmentation (see Section 2.2)		Augmentation [24]
Training strategy (see Section 2.3)		SNIP [25]
		SNIPER [26]
		SAN [51]
Context-based detection (see Section 2.4)	Local context	MRCNN [35]
		MPNet [40]
		GBDNet [41]
		ACCNN [44]
		CoupleNet [45]
		SCAN [28]
	Global context	ION [27]
		R-FCN++ [48]
		DeepIDNet [33]
		SegDeepM [36]
		CPF [42]
	Context interactive	SMN [43]
		ORN [49]
		Context-SVM [32]
		SIN [50]
GAN-based detection (see Section 2.5)		Perceptual-GAN [29]
		MTGAN [30]

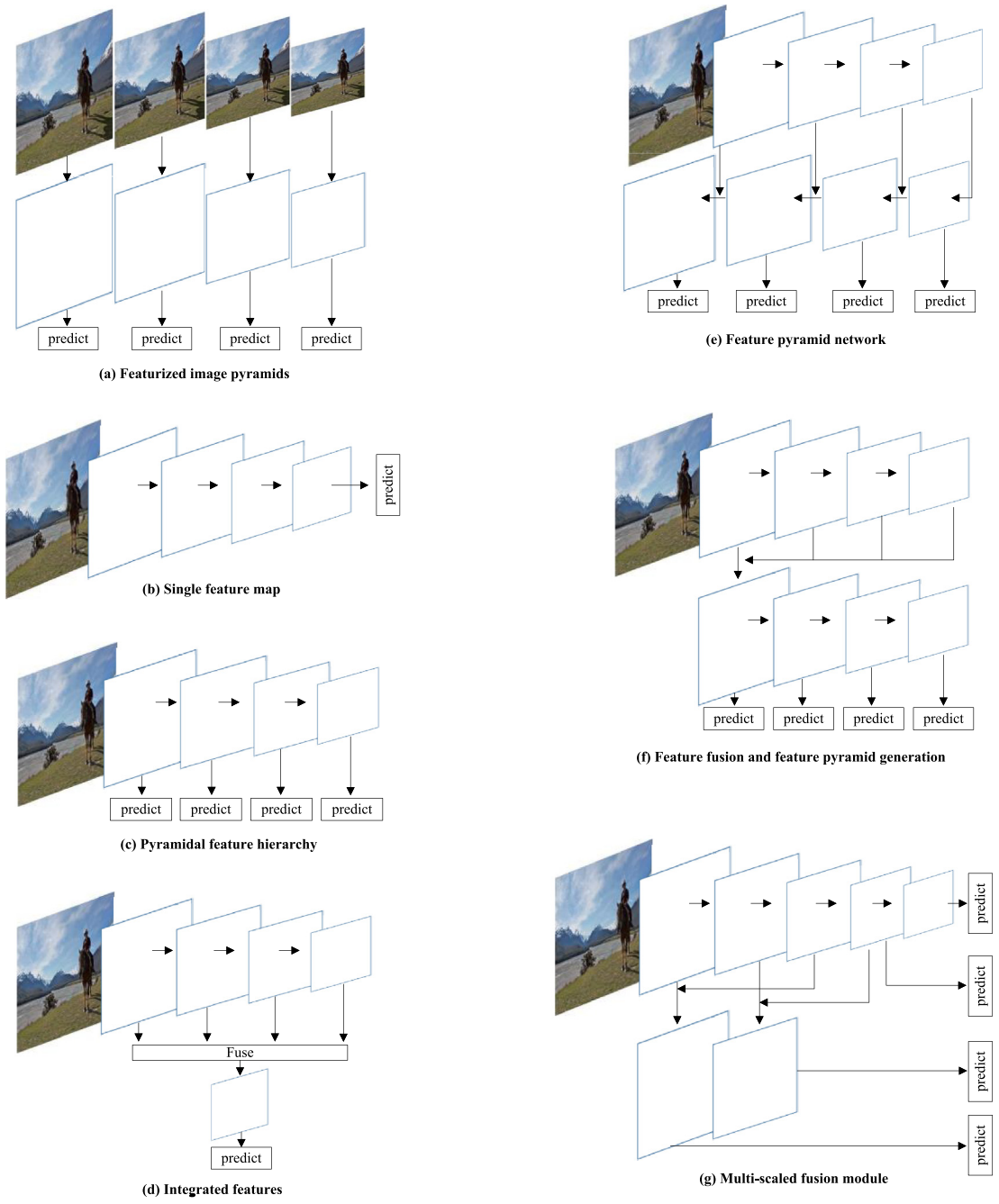

Fig. 2. Seven paradigms for multi-scale feature learning. (a) It is slow to use an image pyramid to build a feature pyramid because features are computed on each of the image scales independently. (b) Detection systems like Faster RCNN use only single scale features (the outputs of the last conv layer) for faster detection. (c) An alternative to the featurized image pyramid is to predict each of the pyramidal feature hierarchy from a CNN. (d) It refers to predict on single feature map generated from multiple features. (e) Feature pyramid network (FPN) integrates the structure of (b), (c) and (d). (f) Features from different layers with different scales are concatenated together first and used to produce a set of pyramid features later. (g) Multi-scale fusion module with skip connections.

crafted features have been replaced with features computed by CNNs because deep CNNs have improved significantly the performance of object detection. Liu et al. [73] proposed a scale-aware network to resize images so that all objects were in a similar scale and then trained a single scale detector. Through conducting comprehensive experimental study on small object detection, Singh and Davis [25] held that training a single scale-robust detector to cope with all scale objects was harder than training scale-dependent detectors with image pyramids. Thus, they designed a new framework called scale normalization for image pyramids (SNIP). SNIP trained multiple scale-dependent detectors and each of which was in charge of a specific scale objects. Wang et al. [52] presented a cascade mask generation framework, which efficiently combined multi-scale inputs for fastly detecting small objects. Nevertheless, these works are computationally expensive considering rapid increase of memory consumption and inference time.

Recent detection algorithms such as Fast R-CNN [34], Faster R-CNN [15], SPPNet [31] and R-FCN [38] utilize the top-most feature maps computed by CNNs on a single input scale to predict candidate bounding boxes with different aspect ratios and scales (see Fig. 2(b)). Nevertheless, the top-most feature maps conflict with objects at different scales in images due to their fixed receptive field. There is little information left on the top-most features especially for small objects, so it may

compromise detection performance of small objects. Deep CNNs learn hierarchical features in different layers which capture information from different scale objects. Specifically, spatial-rich features in shallow layers have higher resolutions and smaller receptive fields and are more beneficial to the detection of small objects. Deep layer features with semantic-rich information have coarse resolutions but larger receptive fields, and thus are more suitable for detecting large objects. The in-network feature hierarchy in a deep CNN produces feature maps of different spatial resolutions while introduces large semantic gaps caused by different depths (see Fig. 2(c)). Liu et al.'s SSD [16] detected objects with different scales and aspect ratios from multiple layers. Afterwards, predictions were made from multiple layers, where each layer was in charge of a certain scale of objects. SSD used the features from the shallower layers to detect smaller objects, while exploited the features from the deeper layers for bigger objects detection. Later, multi-scale deep convolutional neural network (MSCNN) [39] was proposed by Cai et al. MSCNN used deconvolutional layers on multiple feature maps to improve their resolutions and then predictions were made by using these refined feature maps.

Integrated features refer to build a single feature map via integrating features in multiple layers and making final predictions based on the new built feature map (see Fig. 2(d)). Bell et al. [27] presented inside-outside network (ION), which cropped region features from different layers through region of interest (ROI) pooling, and integrated these multi-scale region features for the final prediction. Kong et al. put forward a deep hierarchical network, namely HyperNet [37], which followed a similar idea as ION. In order to generate proposals and detect objects, they carefully devised high resolution hyper feature maps via integrating shallow and intermediate layer features. Combined feature map is better suitable for localization and classification because it has features from different levels of abstraction of the input image. However, the combined feature map increases the memory of a model and decreases the speed of a model.

To combine the advantage among single feature map, pyramidal feature hierarchy and integrated features, Lin et al. put forward feature pyramid network (FPN) [22]. FPN constructed a top-down architecture with lateral connections to produce a series of scale-invariant feature maps, and learned multiple scale-dependent classifiers on these feature pyramids (see Fig. 2(e)). Specifically, the shallow spatially-rich features were strengthened by the deep semantic-rich features. These lateral and top-down features were integrated by concatenation or element-wise summation. Subsequently, many variants of FPN are presented [19–21,23,46,47,53]. Compared with conventional detectors, these methods show dramatic improvements in detection accuracy with some modifications to the feature pyramid block. Zhang et al. [47] and Kong et al. [74] constructed scale invariant feature maps with lateral connections. Unlike original FPN which produced region proposals followed by categorical classifiers, their methods were more efficient than FPN because of omitting proposal generation. Liang et al. developed a new detector called deep feature pyramid networks (DFPN) [23]. Employing the feature pyramid architecture with lateral connections in DFPN made the semantic feature of small objects more sensitive. In addition, they devised specialized anchors to detect the small objects from large resolution image, and then trained the network with focal loss. In order to enhance the detection accuracy of small objects, Cao et al. [20] presented a multi-level feature fusion algorithm for introducing contextual information in SSD, namely feature-fused SSD (FFSSD). Moreover, they used concatenation module and element-sum module in fusion stage. It is hard to fuse the features from different scales in SSD's feature pyramid detection method. Thus, Li and Zhou presented feature fusion single shot multibox detector (FSSD) [46], an improved SSD with a lightweight feature fusion module which can enhance the small object detection performance over SSD with just a little speed drop. Features from different layers with different scales are concatenated together in feature fusion stage, followed by some down-sampling blocks to generate new feature pyramid, which will be fed to multibox detectors to achieve the detection results eventually (see Fig. 2(f)). Fu et al. [19] developed deconvolutional single shot detector (DSSD). By leveraging skip connections and deconvolution layers, they added more semantic information in dense feature maps, which in turn helped detect small objects. DSSD [19] and FPN [22] leverage the deconvolution layer from the top-most feature maps which have lost the majority of fine details for small objects. And the following deconvolution fusion modules completely rely on the last convolution layer, increasing a heavy burden on the top-most layer. In addition, these systems based on fusion features carry out connections for every prediction layer, which means more additional layers bring about more computational cost at the same time. Unlike these architectures, Xu et al. [21] proposed the multi-scale deconvolutional single shot detector (MDSSD), which took SSD as the base framework. Afterwards, they added the high-level features with semantic information to the low-level features by multi-scale deconvolution fusion module to achieve the feature maps with rich information (see Fig. 2(g)). Especially, they carried out deconvolution layers on multi-scale features before the top-most layer and merged them with some of the bottom features to produce more semantic feature maps. Moreover, they deliberately added conv3_3 output via the backbone network for prediction so as to enhance the detection performance of CNNs for small objects. Based on the above analysis, we can find that multi-scale feature learning is useful for detecting small objects.

2.2. Data augmentation

Data augmentation refers to perturbing an image through transformations, including flipping, cropping, rotating, scaling etc. The purpose of data augmentation is to produce additional samples of the class under the underlying category unchanged. Data augmentation can be used in training, testing, or both. In a sense, the performance of deep learning is improved via using a large amount of data. Similarly, the detection performance of small objects can also be boosted by increasing the types and numbers of small objects samples in the dataset. Kisantal et al. [24] investigated the problem of small object detection task and carefully analyzed the state-of-the-art model called Mask-RCNN on MS-COCO dataset. They demonstrate that one of the factors behind the poor detection performance for small objects is lack of representation of small objects in a training set. That is to say, only a few images contain small objects, and small objects do not appear enough even within each image containing small objects. Especially, existing object detectors require the presence of enough objects for predicted anchors to match during training. Thus Kisantal et al. [24] put forward two approaches for augmenting the original MS-COCO dataset to deal with this problem. During training, they display that the detection performance of small objects can easily enhance via oversampling images containing small objects. Second, they present an augmentation method by using the segmented mask to extract small objects in images and then copy-pasting small objects. Although data augmentation is good for the final detection performance, the increasing of computational complexity in training and testing limits its usage in real applications.

2.3. Training strategy

Inspired by the intuitive understanding that large and small objects are difficult to detect at larger and smaller scales respectively, Singh and Davis proposed an efficient training strategy named scale normalization for image pyramids (SNIP) [25] which selectively back-propagates the gradients of object instances of different sizes. Specifically, all ground truth boxes are utilized to assign labels to proposals in order to train the classifier. During training, they only select ground truth boxes and proposals which fall in a specified size range at a particular resolution. Similarly, all ground truth boxes are also used to assign labels to anchors for the training of RPN. These anchors that have an overlap higher than 0.3 with an invalid ground truth box are excluded. In testing stage, they

generate proposals utilizing RPN for each resolution and classify them independently at each resolution. Meanwhile, they merely select detections which fall in a specified range at each resolution. The final detection results are obtained by employing NMS [75] to integrate detections from multiple resolutions after bounding box regression and classification. Later, Singh et al. proposed SNIPER [26], another approach for efficient multi-scale training. It only processed context regions around ground truth instances at the appropriate scale instead of processing a whole image pyramid. SNIPER sampled low resolution regions from a multi-scale image pyramid to accelerate multi-scale training. Besides, Kim et al. [51] designed a scale-aware network, namely SAN, and also built a novel learning approach which considered purely the relationship between channels without the spatial information. In order to make current detectors based on CNN more robust to the scale variation, SAN maps the convolutional features obtained from the different scales onto a scale-invariant subspace. SAN first needs to extract convolutional features from scale normalized patches, then SAN and detection network are trained simultaneously by utilizing these extracted features. These training strategies mentioned above are help for detecting small objects to some extent.

2.4. Context-based detection

Context plays an essential role in object detection. Visual objects usually occur in specific environments and sometimes coexist with other related objects. A typical example is that birds commonly fly in the sky. CNNs have been implicitly already learned contextual information from hierarchical feature representations with multiple levels of abstraction. However, there is still value in exploring contextual information explicitly in small object detection based on deep learning. It is recognized that effectively utilizing contextual information can help promote object detection performance, especially for detecting small objects with insufficient cues. In this section, we survey context-based methods for small object detection from three perspectives: local context, global context, and context interactives.

Local context refers to the visual contextual information in the area that surrounds the object to detect. In order to yield richer object representations, Gidaris and Komodakis proposed an object detection system named multi-region CNN (MRCNN) [35] to extract features from many different subregions of the object proposals, including border regions, central regions, half regions, contextual regions and semantically segmented regions. And these features were integrated simply via concatenation. Zagoruyko et al. [40] proposed a multipath network (MPNet) closely related to MRCNN. Different from MRCNN, MPNet just utilized four contextual regions organized in a foveal structure, where the classifier was trained jointly end to end. Furthermore, Zeng et al. put forward gated bi-directional CNN (GBDNet) [41] which extracted features from multi-scale contextualized subregions surrounding an object proposal to facilitate detection performance. Unlike MRCNN, GBDNet passed messages among features from different contextual regions, implemented by convolution. Notably, GBDNet adopted gated functions to control message transmission, such as long short term memory (LSTM) networks, because not all contextual information was helpful for detection. Concurrent with GBDNet, a novel attention to context CNN (ACCNN) was proposed by Li et al. [44]. ACCNN used global and local context information to promote object detection performance. Meanwhile, ACCNN encoded local surroundings context in the similar way as MRCNN. Zhu et al. proposed a fully convolutional network named CoupleNet [45] with two branches. One branch captured the local part information of the object, while the other encoded the global contextual information with ROI pooling. These methods based on deep learning improved the detection performance of small objects to a certain degree. Guan et al. [28] proposed a novel semantic context aware network (SCAN) that contained two modules: local fusion module and context fusion module. The location fusion module built fine-grained and expressive feature maps utilizing a top-down flow and lateral connections, while the context fusion module built context-aware features adopting multiple pooling operations. Finally, these features were used to perform region proposal and classification. SCAN combined context information and precise location information to facilitate object detection performance for occluded and small objects. Moreover, Chen et al. [11] augmented the R-CNN with the context patch in parallel to the proposal patch produced from region proposal network in order to enhance accuracy of detecting small objects.

Global context refers to learning from image or scene level context in the whole image. For example, detecting a baseball ball from an image becomes easier to recognize the baseball ball object when the contextual information from the rest of the image is leveraged (e.g. baseball field, bat, and players). A number of detectors based on deep learning integrate global context to improve object detection performance. Li et al. [48] introduced global context module to promote the classification score maps via using large and separable convolutional kernels. Meanwhile, they also proposed a new pooling method to better extract scores from the score maps, through adopting column-wise or row-wise max pooling. Bell et al. presented the inside-outside network (ION) [27] that used information both inside and outside the region of interest. ION exploited spatial recurrent neural networks to encode contextual information across the entire image from four directions in order to improve small object detection. Ouyang et al.'s DeepIDNet [33] took the image classification scores as contextual features and concatenated with the object detection scores to facilitate small object detection results. Zhu et al. presented segmentation deep model (SegDeepM) [36] that exploited both segmentation and context to promote small object detection performance. Specifically, SegDeepM put forward a Markov random field (MRF) model that scored appearance as well as context for each detection, and allowed each candidate bounding box to select a segment and scored the agreement between them. Moreover, semantic segmentation was considered as a detection method in contextual priming and feedback (CPF) [42] as well.

Context interactive conveys contextual information through the interactions of visual elements, like constraints and dependencies. Recently, some works have shown that object detectors can be improved by considering context interactives. Some recent efforts can be divided into two classes: exploring the relationship between individual objects [43] [49] [32], and modeling the dependencies between objects and scenes [50]. Liu et al. presented structure inference network (SIN) [50]. SIN formulated object detection as a problem of graph structure inference via considering scene contextual information and object relationships within a single image. In SIN, objects were regarded as nodes in a graph and relationships between different objects were modeled as graph edges. They improved the detection performance of small objects by using contextual information in SIN. Furthermore, Song et al. [32] proposed a contextualized learning method by contextualized support vector machine (Context-SVM) with two characteristics: adaptive contextualization and configurable model complexity. Meanwhile, they put forward an iterative contextualization procedure based on the Context-SVM, such that object classification and detection performance can be iteratively and mutually boosted. Chen and Gupta designed spatial memory network (SMN) [43]. In SMN, the spatial memory module obtained instance-level contextual information via assembling object instances back into a pseudo image representation. Hu et al. [49] put forward a light-weight object relation network (ORN), which formulated the interaction of different objects between their appearance feature and geometry. Moreover, the ORN did not require additional supervision and showed the improvements of performance in small object detection. It is worth noting that the above three context-based methods facilitate the detection accuracy of small objects.

2.5. GAN-based detection

The generative adversarial networks (GAN) introduced by Goodfellow et al. [76] in 2014, has received great attention in recent

years. GAN is structurally inspired by the two-person zero-sum game in game theory. A typical GAN consists of a generator network and a discriminator network, contesting with each other in a minimax optimization framework. The generator learns to capture the potential distribution of true data samples and generates new data samples, while the discriminator aims to discriminate between instances from the true data distribution and those produced by the generator. To the best of our knowledge, Li et al. [29] put forward a novel perceptual GAN model that made the first attempt to accommodate GAN on object detection task to boost small object detection performance via generating super-resolved representations for small objects to narrow representation difference of small objects from the large ones. Its generator learns to enhance the poor representations of the small objects to super-resolved ones that are similar enough to real large objects to fool its discriminator, while its discriminator competes with its generator to recognize the generated representation. Meanwhile, on the generator, its discriminator imposes an additional perceptual requirement that the generated representations must be beneficial for detecting small objects. Bai et al. [30] presented a multi-task generative adversarial network, namely MTGAN, in order to handle the detection problem of small objects. The generator in the MTGAN is a super-resolution network which upsamples the small blurred images into fine-scale clear images. Unlike the generator, the discriminator in the MTGAN is a multi-task network. In the discriminator, each super-resolved image patch is described by a real or fake score, object category scores and regression offsets. Moreover, the bounding box regression and classification losses in the discriminator are back-propagated to the generator during training in order to make the generator obtain more details for more accurate detection. Extensive experiments show the superiority of the above two GAN-based detection methods in detecting small objects such as traffic signs, over state-of-the-art algorithms.

3. Datasets and performance evaluation

3.1. Datasets

Datasets have played a critical role in object detection, because they are able to draw a standard comparison between different competing algorithms and set goals for solutions. A number of well-known datasets have been released in the past years, including PASCAL-VOC [77], MS-COCO [7], FlickrLogos [78,79], KITTI [80], SUN [81], Tsinghua-Tencent 100 K (TT100K) [82], Caltech [83], etc. MS-COCO and PASCAL-VOC are for generic object detection. Caltech and KITTI are for pedestrian detection. Moreover, FlickrLogos, TT100K and SUN are for logo detection, traffic signs detection and scene detection respectively. Especially, the small object dataset (SOD) [11] is designed to solve small object detection problems. For more details, see Table 3.

3.2. Evaluation criteria

In addition to the dataset, the evaluation criteria are equally important. At the time of evaluation, a metric called the overlap ratio intersection over union (IoU) between objects and predictions is used to evaluate the quality of localization:

$$IoU(b_{pred}, b_{gt}) = \frac{Area(b_{pred} \cap b_{gt})}{Area(b_{pred} \cup b_{gt})} \quad (1)$$

b_{pred} and b_{gt} refers to the predict bounding box, and the ground truth bounding box or mask respectively. An IoU threshold α is set to determine whether a prediction tightly covers the object or not. A typical value is 0.5. For object detection, a prediction with correct categorical label as well as successful localization prediction is regarded as positive, otherwise it's a negative prediction.

The most commonly used metric is Average Precision (AP). AP is defined as the average detection precision under different recalls, and is usually evaluated in a category specific manner. The mean Average Precision (mAP) refers to average score of AP across all classes, which is used as evaluation metric for many object detection datasets. The

Table 3
An overview of some popular detection datasets.

Dataset	Description	Published
PASCAL-VOC [77]	It is the most iconic object detection dataset. Two versions of PASCAL-VOC are commonly used in papers: VOC2007 and VOC2012. VOC2007 consists of 2501 training images, 2510 validation images and 5011 testing images. However, VOC2012 consists of 5717 training images, 5823 validation images and 10,991 testing images. They are both mid-scale datasets for object detection with 20 categories. The dataset can be downloaded at http://host.robots.ox.ac.uk/pascal/VOC/.	IJCV
MS-COCO [7]	It is one of the most popular and challenging object detection datasets today. It contains about 164,000 images and 897,000 annotated objects from 80 categories. There are three image splits in it: training, validation and testing set. 118,287 images, 5000 images and 40,670 images are for training, validation and testing set respectively. The objects distribution in MS-COCO is closer to real world scenarios. The annotation information of MS-COCO testing set is not available. The URL of the dataset is http://cocodataset.org.	ECCV
KITTI [80]	It is well-known dataset for traffic scene analysis. It contains 7481 labeled images and another 7518 images for testing. There are 100,000 instances of pedestrians. With around 6000 identities and one person in average per image. The person class in KITTI is divided into two subclasses: pedestrian and cyclist. The object labels are grouped into easy, moderate and hard levels, based on the extent to which the objects are occluded and truncated. The dataset can be obtained at http://www.cvlibs.net/datasets/kitti/index.php.	CVPR
Caltech [83]	It is one of the most popular and challenging pedestrian detection datasets. The training set and testing set contains 192,000 and 155,000 pedestrian instances respectively. The URL of this dataset is http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/.	PAMI
FlickrLogos [78,79]	Two versions of FlickrLogos are commonly used in papers: FlickrLogos-32 and FlickrLogos-47. FlickrLogos-47 uses the same image corpus as FlickrLogos-32 but new classes are introduced and missing object instances have been annotated. FlickrLogos-47 contains 833 training images and 1402 testing images. The dataset can be downloaded at http://www.multimedia-computing.de/flickrlogos/.	ICMR
SUN [81]	It contains about 132,000 images from 908 distinct scene categories. Two versions of SUN are mostly used in the literature, namely SUN397 and SUN2012. The former is for scene recognition and the latter is for object detection. The URL of the dataset is http://groups.csail.mit.edu/vision/SUN/.	IJCV
TT100K [82]	It is the largest traffic sign detection dataset so far, with 100,000 images and 30,000 traffic sign instances of 128 classes. The resolution of the images is as large as 2048 × 2048, but the typical traffic sign instances are less than 32 × 32 pixels. Each instance is annotated with class label, bounding box and pixel mask. It has small objects in abundance, huge illumination and scale variations. There are 45 classes with at least 100 instances present. The URL of the dataset is http://cg.cs.tsinghua.edu.cn/traffic%2Dsign/.	CVPR
SOD [11]	Small object dataset (SOD) is composed by utilizing a subset of images from both the MS-COCO dataset and SUN dataset. SOD contains about 8393 object instances in 4925 images from 10 categories. The selected object categories are “mouse”, “telephone”, “switch”, “outlet”, “clock”, “toilet paper”, “tissue box”, “faucet”, “plate”, and “jar”. All the object instances in SOD are small. For more related information, please refer to the website: http://www.merl.com.	ACCV

threshold in PASCAL-VOC is set to a single value, 0.5, but is belong to [0.5,0.95] with an interval 0.05 that is 10 values to calculate the mAP in MS-COCO. Also the special AP for small, medium and large objects is calculated separately in MS-COCO. For KITTI, standard mAP is used as evaluation metric with 0.5 IoU threshold. For Caltech, the log-average miss rate over 9 points ranging from $1e^{-2}$ to 100 FPPI (false positive per image) is used to evaluate the performance of the detectors (lower is better). The details of evaluation metrics are shown in Table 4.

3.3. Performance analysis

We have comprehensively summarized the methods of detecting small objects from five aspects. They are multi-scale feature learning, data augmentation, training strategy, context-based detection and GAN-based detection. Based on our taxonomy of small object detection algorithms, we show the detection results of these state-of-the-art algorithms on the MS-COCO and PASCAL-VOC dataset in Tables 5 and 6 respectively. By comparing Tables 5 and 6, we can see that the detection difficulty of the MS-COCO dataset is much greater than that of the PASCAL-VOC dataset. This is maybe because that the number of objects per image in the MS-COCO dataset is more than the PASCAL-VOC dataset. Compared with the PASCAL-VOC dataset, the majority objects in MSCOCO dataset are small objects with large scale ranges, which leads to poor detection results.

As shown in Table 5, the detection performance of large objects is the best among large, medium and small objects (countdown columns 1, 2 and 3), while that of small objects is the worst. This also indicates the difficulty of small object detection. SNIPER [26] improves the SNIP [25] from 45.7% AP to 46.1% AP, and also enhances the SNIP from 29.3% to 29.6% on small objects. SNIP with multi-scale feature learning and training strategy achieves 45.7% AP on the test-devset, which outperforms other algorithms (except SNIPER) with a large margin. Meanwhile, SNIP's AP on small objects is still higher than other methods (except SNIPER). Moreover, variants of FPN such as RefineDet512++ [47] and M2Det800++ [53], achieve more than 25% AP on small object detection. R-FCN++ [48] and MTGAN [30] utilize global context information and multi-task GANs to detect small objects respectively, which both achieve over 24% AP on small objects.

In Table 6, we summarize some typical detectors from the VOC2007 and VOC2012 challenges. Compared with other methods, RefineDet512++ [47], FSSD512 [46], DSSD513 [19], CoupleNet [45] and R-FCN++ [48] obtain excellent detection results on the PASCAL-VOC challenge. Furthermore, ION [27] used VGG16 as the backbone network achieves 79.2% mAP and 76.4% mAP on VOC2007 and VOC2012 respectively, which has better performance compared with other most algorithms with using VGG16. It is worth noting that ION obtains good performance via utilizing integrated features and global context information. In addition, we calculate AP of each of the 20 categories in the PASCAL-VOC dataset and their mAP. Comparative results of PASCAL-VOC 2007 and PASCAL-VOC 2012 are exhibited in Tables 7 and 8 respectively.

Based on the above comparative analysis from Tables 5–8, the following remarks can be obtained.

(1) Multi-scale feature learning, data augmentation, training strategy, context-based detection and GAN-based detection methods can achieve better detection performance of small objects on MS-COCO and PASCAL-VOC dataset.
(2) Detection performance of small objects is also affected by adopting different input resolutions, such as SSD300 [16] and SSD512 [16], DSSD321 [19] and DSSD513 [19]. A higher input

Table 4
Summary of common evaluation metrics for object detection.

Alias	Meaning	Definition and description	
α	IoU threshold	The IoU threshold to assess localization.	
λ	Confidence threshold	A confidence threshold for computing P_λ and R_λ.	
D_λ	All Predictions	Top λ predictions returned by the detectors with highest confidence score.	
TP_λ	True Positive	Correct predictions from sampled predictions.	
FP_λ	False Positive	False predictions from sampled predictions.	
P_λ	Precision	The fraction of TP_λ out of D_λ.	
R_λ	Recall	The fraction of TP_λ out of all positive samples.	
F1	F1-measure	$\frac{2P_\lambda R_\lambda}{P_\lambda + R_\lambda}$	
AP	Average Precision	Computed over the different levels of recall by varying the confidence λ.	
mAP	mean AP	Average score of AP across all categories.	
AR	Average Recall	The maximum recall given a fixed number of detections per image, averaged over all classes and IoU thresholds.	
FPPI	False Positive Per Image	The fraction of false positive for each image.	
LaMR	Log-average Missing Rate	Average miss rate over different FPPI rates evenly spaced in log-space.	
Object detection			
mAP	Mean average precision	PASCAL-VOC	mAP at 0.5 IoU threshold over all 20 categories.
		MS-COCO	AP: mAP averaged over ten IoUs: [0.5:0.05:0.95]; AP_{50}: mAP at 0.5 IoU threshold; AP_{75}: mAP at 0.75 IoU threshold; AP_S: AP for small objects of area smaller than 32^2; AP_M: AP for objects of area between 32^2 and 96^2; AP_L: AP for large objects of area bigger than 96^2
		KITTI	mAP (easy): mAP for easy level pedestrians; mAP (mid): mAP for midlevel pedestrians; mAP (hard): mAP for hard level pedestrians
		TT100K	the same detection metrics as for MS-COCO.
		SOD	mAP at 0.5 IoU threshold over all 10 classes.
		SUN	A detection is correct if the predicted label matches the human label. mAP (SUN): the average precision over all categories.
LaMR	Log-average Miss Rate	Caltech	LaMR: log-average Miss Rate on FPPI in [0.01, 1].
AR	Average recall	MS-COCO	AR(max = 1): AR given 1 detection per image; AR(max = 10): AR given 10 detection per image; AR(max = 100): AR given 100 detection per image; AR_S: AR for small objects of area smaller than 32^2; AR_M: AR for objects of area between 32^2 and 96^2; AR_L: AR for large objects of area bigger than 96^2

Table 5
Detection results on the MS-COCO test-dev dataset of some typical methods. "++" denotes applying inference strategy such as multi scale test, horizontal flip, etc. (in %).

Type		Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Multi-scale feature learning	Featurized image pyramids	SNIP [25]	DPN98 [84]	45.7	67.3	51.1	29.3	48.8	57.1
	Single feature map	Faster R-CNN [15]	VGG16	21.9	42.7	–	–	–	–
	Pyramidal feature hierarchy	SSD300 [16]	VGG16	23.2	41.2	23.4	5.3	23.2	39.6
		SSD512 [16]	VGG16	26.8	46.5	27.8	9.0	28.9	41.9
	Integrated features	ION [27]	VGG16	24.6	46.3	23.3	7.4	26.2	38.8
	Feature pyramid network	FPN [22]	ResNet101	36.2	59.1	39.0	18.2	39.0	48.2
	Variants of FPN	RefineDet512 [47]	ResNet101	36.4	57.5	39.5	16.6	39.9	51.4
	(including feature fusion and feature pyramid generation, multi-scaled fusion module, etc.)	RefineDet512++ [47]	ResNet101	41.8	62.9	45.7	25.6	45.1	54.1
		M2Det800 [53]	VGG16	41.0	59.7	45.0	22.1	46.5	53.8
		M2Det800++ [53]	VGG16	44.2	64.6	49.3	29.2	47.9	55.1
		FSSD300 [46]	VGG16	27.1	47.7	27.8	8.7	29.2	42.2
		FSSD512 [46]	VGG16	31.8	52.8	33.5	14.2	35.1	45.0
		DSSD321 [19]	ResNet101	28.0	46.1	29.2	7.4	28.1	47.6
		DSSD513 [19]	ResNet101	33.2	53.3	35.2	13.0	35.4	51.1
		MDSSD300 [21]	VGG16	26.8	46.0	27.7	10.8	–	–
		MDSSD512 [21]	VGG16	30.1	50.5	31.4	13.9	–	–
Data augmentation		Augmentation [24]	ResNet50	30.4	–	–	17.9	32.9	38.6
Training strategy		SNIP [25]	DPN98 [84]	45.7	67.3	51.1	29.3	48.8	57.1
		SNIPER [26]	ResNet101	46.1	67.0	51.6	29.6	48.9	58.1
		SAN [51]	R-FCN	36.3	59.6	–	16.7	40.5	55.5
Context-based detection	Local context	MPNet [40]	ResNet	33.2	51.9	36.3	13.6	37.2	47.8
		CoupleNet [45]	ResNet101	33.1	53.5	35.4	11.6	36.3	50.1
		CoupleNet++ [45]	ResNet101	34.4	54.8	37.2	13.4	38.1	50.8
	Global context	ION [27]	VGG16	24.6	46.3	23.3	7.4	26.2	38.8
		R-FCN++ [48]	R-FCN	42.3	63.8	–	25.2	46.1	54.2
		ORN [49]	ResNet50	30.5	50.2	32.4	–	–	–
		SIN [50]	VGG16	23.2	44.5	22.0	7.3	24.5	36.3
GAN-based detection		MTGAN [30]	ResNet101	41.4	63.2	45.4	24.7	44.2	52.6

resolution maybe obtain better detection results than a lower input resolution.

(3) If combined in a proper way, more powerful backbone CNN models can better promote small object detection performance.

(4) The detection performance of SNIP on MS-COCO dataset and ION on PASCAL-VOC dataset shows that combining multiple small object detection methods can definitely improve detection performance of small objects.

(5) In order to obtain robust features, data augmentation is significant for deep learning based detection models (SSD512 with '07', '07 + 12' and '07 + 12 + COCO').

Since the pedestrian instances on the Caltech dataset are often of small scales, the overall performance on it can be used to evaluate the capability of an algorithm in detecting small objects. We compare the detection results of SPPNet [31], MSCNN [39] and Perceptual-GAN

Table 6
Detection performance on PASCAL-VOC dataset. For VOC2007, the models are trained on VOC2007 and VOC2012 trainval sets and tested on VOC2007 test set. For VOC2012, the models are trained on VOC2007 and VOC2012 trainval sets plus VOC2007 test set and tested on VOC2012 test set by default.

Type		Method	Backbone	mAP (%)	
				VOC2007	VOC2012
Multi-scale feature learning	Single feature map	Fast R-CNN [34]	VGG16	70.0	68.4
		Faster R-CNN [15]	VGG16	73.2	70.4
		R-FCN [38]	ResNet101	80.5	77.6
	Pyramidal feature hierarchy	SSD300 [16]	VGG16	74.3	72.4
		SSD512 [16]	VGG16	76.8	74.9
	Integrated features	ION [27]	VGG16	79.2	76.4
		HyperNet [37]	VGG16	76.3	71.4
	Variants of FPN	RefineDet512 [47]	VGG16	81.8	80.1
	(including feature fusion and feature pyramid generation, multi-scaled fusion module, etc.)	RefineDet512++ [47]	VGG16	83.8	83.5
		FFSSD (Elt_sum) [20]	VGG16	78.9	–
		FFSSD (Concat) [20]	VGG16	78.8	–
		FSSD300 [46]	VGG16	78.8	82.0[a]
		FSSD512 [46]	VGG16	80.9	84.2[a]
		DSSD321 [19]	ResNet101	78.6	76.3
		DSSD513 [19]	ResNet101	81.5	80.0
		MDSSD300 [21]	VGG16	78.6	–
		MDSSD512 [21]	VGG16	80.3	–
Training strategy		SAN [51]	R-FCN	80.6	–
Context-based detection	Local context	MRCNN [35]	VGG16	78.2	73.9
		ACCNN [44]	VGG16	72.0	70.6
		CoupleNet [45]	ResNet101	82.7	80.4
	Global context	ION [27]	VGG16	79.2	76.4
		R-FCN++ [48]	R-FCN	81.2	79.7
		CPF [42]	VGG16	76.4	72.6
	Context interactive	SIN [50]	VGG16	76.0	73.1

[a] This entry reports the model is trained with VOC2007 trainval, VOC2007 test, VOC2012 trainval and MS-COCO sets (in %).

Table 7
Comparative results on PASCAL-VOC2007 test set. All methods use VGG16 as the backbone network. Legend: '07': VOC2007 trainval, '07 + 12': union of VOC2007 and VOC2012 trainval, '07 + 12 + COCO': trained on COCO trainval35k at first and then fine-tuned on 07 + 12, The S in ION '07 + 12 + S' denotes SBD segmentation labels (in %).

Method	Trained on	mAP	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
Fast R-CNN [34]	07 + 12	70.0	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8
Faster R-CNN [15]	07 + 12	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9
ION [27]	07 + 12 + S	79.2	80.2	85.2	78.8	70.9	62.6	86.6	86.9	89.8	61.7	86.9
HyperNet [37]	07 + 12	76.3	77.4	83.3	75.0	69.1	62.4	83.1	87.4	87.4	57.1	79.8
RefineDet512 [47]	07 + 12	81.8	88.7	87.0	83.2	76.5	68.0	88.5	88.7	89.2	66.5	87.9
RefineDet512++ [47]	07 + 12 + COCO	85.8	90.4	89.6	88.2	84.9	78.3	89.8	89.9	90.0	75.9	90.0
MRCNN [35]	07 + 12	78.2	80.3	84.1	78.5	70.8	68.5	88.0	85.9	87.8	60.3	85.2
ACCNN [44]	07 + 12	72.0	79.3	79.4	72.5	61.0	43.5	80.1	81.5	87.0	48.5	81.9
SSD300 [16]	07 + 12 + COCO	79.6	80.9	86.3	79.0	76.2	57.6	87.3	88.2	88.6	60.5	85.4
SSD512 [16]	07	71.6	75.1	81.4	69.8	60.8	46.3	82.6	84.7	84.1	48.5	75.0
SSD512 [16]	07 + 12	76.8	82.4	84.7	78.4	73.8	53.2	86.2	87.5	86.0	57.8	83.1
SSD512 [16]	07 + 12 + COCO	81.6	86.6	88.3	82.4	76.0	66.3	88.6	88.9	89.1	65.1	88.4
MDSSD300 [21]	07 + 12	78.6	86.5	87.6	78.9	70.6	55.0	86.9	87.0	88.1	58.5	84.8
MDSSD512 [21]	07 + 12	80.3	88.8	88.7	83.2	73.7	58.3	88.2	89.3	87.4	62.4	85.1
FFSSD-Eltsum [20]	07 + 12	78.9	82.0	86.5	78.0	71.7	52.9	86.6	86.9	88.3	63.2	83.0
FFSSD-Concat [20]	07 + 12	78.8	82.4	85.7	77.8	73.8	52.3	87.5	86.8	87.6	62.6	82.1
SAN [51]	07 + 12	80.6	82.0	84.3	79.7	72.5	70.2	87.3	87.7	89.5	68.7	87.5
SIN [50]	07 + 12	76.0	77.5	80.1	75.0	67.1	62.2	83.2	86.9	88.6	57.7	84.5

Method	Trained on	mAP	Table	Dog	Horse	Mbike	Person	Plant	Sheep	Sofa	Train	tv
Fast R-CNN [34]	07 + 12	70.0	68.9	84.7	82.0	76.6	69.9	31.8	70.1	74.8	80.4	70.4
Faster R-CNN [15]	07 + 12	73.2	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
ION [27]	07 + 12 + S	79.2	76.5	88.4	87.5	83.4	80.5	52.4	78.1	77.2	86.9	83.5
HyperNet [37]	07 + 12	76.3	71.4	85.1	85.1	80.0	79.1	51.2	79.1	75.7	80.9	76.5
RefineDet512 [47]	07 + 12	81.8	75.0	86.8	89.2	87.8	84.7	56.2	83.2	78.7	88.1	82.3
RefineDet512++ [47]	07 + 12 + COCO	85.8	80.0	89.8	90.3	89.6	88.3	66.2	87.8	83.5	89.3	85.2
MRCNN [35]	07 + 12	78.2	73.7	87.2	86.5	85.0	76.4	48.5	76.3	75.5	85.0	81.0
ACCNN [44]	07 + 12	72.0	70.7	83.5	85.6	78.4	71.6	34.9	72.0	71.4	84.3	73.5
SSD300 [16]	07 + 12 + COCO	79.6	76.7	87.5	89.2	84.5	81.4	55.0	81.9	81.5	85.9	78.9
SSD512 [16]	07	71.6	67.4	82.3	83.9	79.4	76.6	44.9	69.9	69.1	78.1	71.8
SSD512 [16]	07 + 12	76.8	70.2	84.9	85.2	83.9	79.7	50.3	77.9	73.9	82.5	75.3
SSD512 [16]	07 + 12 + COCO	81.6	73.6	86.5	88.9	85.3	84.6	59.1	85.0	80.4	87.4	81.2
MDSSD300 [21]	07 + 12	78.6	73.4	84.8	89.2	88.1	78.0	52.3	78.6	74.5	86.8	80.7
MDSSD512 [21]	07 + 12	80.3	75.1	84.7	89.7	88.3	83.2	56.7	84.0	77.4	83.9	77.6
FFSSD-Eltsum [20]	07 + 12	78.9	76.8	86.1	88.5	87.5	80.4	53.9	80.6	79.5	88.2	77.9
FFSSD-Concat [20]	07 + 12	78.8	76.6	86.1	88.2	86.6	80.3	53.7	78.0	80.1	87.3	78.0
SAN [51]	07 + 12	80.6	75.7	88.4	88.2	83.8	81.1	53.7	81.8	81.0	87.2	81.1
SIN [50]	07 + 12	76.0	70.5	86.6	85.6	77.7	78.3	46.6	77.6	74.7	82.3	77.1

[29]. As shown in Table 9, SPPNet outperforms the other two methods and achieves the lowest LaMR of 8.56%. Similarly, a comparison on KITTI challenge between MSCNN [39] and SCAN [28] is also exhibited in Table 9. The column "mAP" of SCAN show substantial gains (3.44 points) over MSCNN. Specifically, the columns "Pedestrians" and "Cyclists" obtain much better performance than the MSCNN. Particularly, the column "Cyclists-Hard" achieves about 14 points higher than MSCNN.

We compare Perceptual-GAN [29], DFPN [23] and CasMaskGF [52] these three approaches on TT100K dataset. According to [82], this dataset is divided into three categories based on the area size, including small objects (area smaller than 32×32 pixels), medium objects (area between 32×32 and 96×96 pixels) and large objects (area bigger than 96×96 pixels). In Table 10, a three-stage detector with input resolutions of 128×128, 512×512 and 2048×2048 is denoted by 128-512-2048. Two-stage and other three-stage detectors utilize the same representation. From Table 10, we can observe that "CasMaskGF 512-1024-2048" offers a substantial improvement in terms of F1-measure on three subsets of different object sizes. More concretely, the columns "F1-measure-small", "F1-measure-medium" and "F1-measure-large" show substantial gains (1.74, 3.45, and 4.79 points respectively) over Perceptual-GAN. Moreover, the detection performance of small objects is the lowest among three object sizes. As shown in Table 10, the column "F1-measure-small" of DFPN achieves the highest F1-measure of 88.27% among three methods. We also exhibit precision and recall of most commonly used traffic signs in Table 11. Perceptual-GAN achieves excellent performance in most categories, with some classes reach 100% recall, such as "p6" and "pm55". Furthermore, DFPN obtains better performance in most categories than Perceptual-GAN. Some classes reach 100% recall, including "il80", "pl100" and "pr40". In particular, category "pl120" achieves the highest precision of 99% among all categories. In addition to the previous remarks, from the analysis in Tables 9–11, it can be also seen that GAN-based detection and multi-scale feature learning methods can obtain good detection performance of small objects on the Caltech and TT100K dataset.

4. Conclusions and future directions

Small object detection is an extremely challenging problem in computer vision. This paper comprehensively reviews the small object detection methods based on deep learning from five dimensions, compares and analyzes the current classic detection algorithms of small objects on some popular object detection datasets, such as PASCAL-VOC, MS-COCO, KITTI, TT100K. The results on these datasets show that multi-scale feature learning, data augmentation, training strategy, context-based detection and GAN-based detection methods can achieve better performance for detecting small objects. Although the promising progress of this domain has been achieved recently, there remains a huge gap between the state-of-the-art and human-level performance. Much work remains to be done, which we see focused on the following aspects:

Table 8
Comparative results on PASCAL-VOC2012 test set. All methods use VGG16 as the backbone network. Legend: '07++12': union of VOC2007 trainval and test and VOC2012 trainval, '07++12 + COCO': trained on COCO trainval35k at first then fine-tuned on 07++12. '07 + 12': union of VOC2007 and VOC2012 trainval, The S in ION '07 + 12 + S' denotes SBD segmentation labels (in %).

Method	Trained on	mAP	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
Fast R-CNN [34]	07++12	68.4	82.3	78.4	70.8	52.3	38.7	77.8	71.6	89.3	44.2	73.0
Faster R-CNN [15]	07++12	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1
ION [27]	07 + 12 + S	76.4	87.5	84.7	76.8	63.8	58.3	82.6	79.0	90.9	57.8	82.0
MRCNN [35]	07++12	73.9	85.5	82.9	76.6	57.8	62.7	79.4	77.2	86.6	55.0	79.1
ACCNN [44]	07++12	70.6	83.2	80.8	70.8	54.9	42.1	79.1	73.4	89.7	47.0	75.9
SSD300 [16]	07++12	70.3	84.2	76.3	69.6	53.2	40.8	78.5	73.6	88.0	50.5	73.5
SSD300 [16]	07++12 + COCO	79.3	91.0	86.0	78.1	65.0	55.4	84.9	84.0	93.4	62.1	83.6
SSD512 [16]	07++12 + COCO	82.2	91.4	88.6	82.6	71.4	63.1	87.4	88.1	93.9	66.9	86.6
HyperNet [37]	07++12	71.4	84.2	78.5	73.6	55.6	53.7	78.7	79.8	87.7	49.6	74.9
FSSD300 [46]	07++12 + COCO	82.0	92.2	89.2	81.8	72.3	59.7	87.4	84.4	93.5	66.8	87.7
FSSD512 [46]	07++12 + COCO	84.2	92.8	90.0	86.2	75.9	67.7	88.9	89.0	95.0	68.8	90.9
SIN [50]	07++12	73.1	84.8	79.5	74.5	59.7	55.7	79.5	78.8	89.9	51.9	76.8
RefineDet512 [47]	07++12	80.1	90.2	86.8	81.8	68.0	65.6	84.9	85.0	92.2	62.0	84.4
RefineDet512++ [47]	07++12 + COCO	86.8	94.7	91.5	88.8	80.4	77.6	90.4	92.3	95.6	72.5	91.6

Method	Trained on	mAP	Table	Dog	Horse	Mbike	Person	Plant	Sheep	Sofa	Train	tv
Fast R-CNN [34]	07++12	68.4	55.0	87.5	80.5	80.8	72.0	35.1	68.3	65.7	80.4	64.2
Faster R-CNN [15]	07++12	70.4	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
ION [27]	07 + 12 + S	76.4	64.7	88.9	86.5	84.7	82.3	51.4	78.2	69.2	85.2	73.5
MRCNN [35]	07++12	73.9	62.2	87.0	83.4	84.7	78.9	45.3	73.4	65.8	80.3	74.0
ACCNN [44]	07++12	70.6	61.8	87.8	80.9	81.8	74.4	37.8	71.6	67.7	83.1	67.4
SSD300 [16]	07++12	70.3	61.7	85.8	80.6	81.2	77.5	44.3	73.2	66.7	81.1	65.8
SSD300 [16]	07++12 + COCO	79.3	67.3	91.3	88.9	88.6	85.6	54.7	83.8	77.3	88.3	76.5
SSD512 [16]	07++12 + COCO	82.2	66.3	92.0	91.7	90.8	88.5	60.9	87.0	75.4	90.2	80.4
HyperNet [37]	07++12	71.4	52.1	86.0	81.7	83.3	81.8	48.6	73.5	59.4	79.9	65.7
FSSD300 [46]	07++12 + COCO	82.0	70.4	92.1	90.9	89.6	87.7	56.9	86.8	79.0	90.7	81.3
FSSD512 [46]	07++12 + COCO	84.2	68.7	92.8	92.1	91.4	90.2	63.1	90.1	76.9	91.5	82.7
SIN [50]	07++12	73.1	58.2	87.8	82.9	81.8	81.6	51.2	75.2	63.9	81.8	67.8
RefineDet512 [47]	07++12	80.1	64.9	90.6	88.3	87.2	87.8	58.0	86.3	72.5	88.7	76.6
RefineDet512++ [47]	07++12 + COCO	86.8	69.9	93.9	93.5	92.4	92.6	68.8	92.4	78.5	93.6	85.2

(1) Emerging small object detection datasets and benchmarks: Despite popular datasets such as the MS-COCO contains several "small" object classes, many instances of the objects in the "small" object classes occupy a large part of an image. Actually, we do not have much understanding on how difficult the small object detection task is or how well existing object detectors work. To better evaluate the performance of small object detection algorithms, we need large-scale datasets specifically for small object detection, just like ImageNet [85] dataset in image classification, Kinetics [86] dataset in action recognition. So establishing large-scale small object datasets and corresponding benchmarks is a research direction for small object detection domain.

(2) Multi-task joint learning and optimization: As we all known, combining multiple small object detection methods [25,27] can facilitate detection performance of small objects. Moreover,

Table 9
Summary of detection results of different models on KITTI and Caltech challenge (in %).

Type	Method	KITTI									
		mAP	Cars (AP)			Pedestrians (AP)			Cyclists (AP)		
			Easy	Mid	Hard	Easy	Mid	Hard	Easy	Mid	Hard
Multiscale feature learning	MSCNN [39]	78.52	90.03	89.02	76.11	83.92	73.70	68.31	84.06	75.46	66.07
Context-based detection	SCAN [28]	81.96	96.68	80.65	70.42	87.16	78.22	70.28	87.88	86.65	79.70

Type		Method	Caltech (LamR)
Multi-scale feature learning	Single feature map	SPPNet [31]	8.56
	Pyramidal feature hierarchy	MSCNN [39]	10.0
GAN-based detection		Perceptual-GAN [29]	9.48

Table 10
Detection performance of different methods for different sizes of traffic signs in TT100K dataset (in %).

Type	Method	F1-measure-small	F1-measure-medium	F1-measure-large
GAN-based detection	Perceptual-GAN [29]	86.43	93.43	89.99
Multi-scale feature learning	DFPN [23]	88.27	95.99	93.69
	CasMaskGF 400-800-1600 [52]	82.20	95.89	93.91
	CasMaskGF 512-2048 [52]	86.86	96.52	93.85
	CasMaskGF 128-512-2048 [52]	86.57	96.49	93.78
	CasMaskGF 512-1024-2048 [52]	88.17	96.88	94.78

Table 11
Comparisons of detection performance for some commonly used classes in TT100K. "P" and "R" refers to precision and recall respectively (in %).

Class	i2	i4	i5	il100	il60	il80	io	ip	p10	p11	p12	p19	p23	p26	p27
PercepGAN(R) [29]	84	95	95	95	92	95	92	91	89	96	97	97	95	94	98
PercepGAN(P) [29]	85	92	94	97	95	83	79	90	84	85	88	84	92	83	98
DFPN(R) [23]	87	97	96	97	98	100	94	88	92	95	95	91	94	95	98
DFPN(P) [23]	90	92	94	93	98	94	86	90	89	90	94	75	93	89	98

Class	p3	p5	p6	pg	ph 4	ph 4.5	ph 5	pl100	pl120	pl20	pl30	pl40	pl5	pl50	pl60
PercepGAN(R) [29]	93	96	100	93	78	88	85	96	98	96	93	96	92	96	91
PercepGAN(P) [29]	92	90	83	93	97	68	69	97	98	92	91	90	86	87	92
DFPN(R) [23]	96	98	97	98	86	90	90	100	97	98	97	97	94	97	98
DFPN(P) [23]	81	91	90	93	94	80	78	98	99	90	92	91	92	90	95

Class	pl70	pl80	pm20	pm30	pm55	pn	pne	po	pr40	w13	w32	w55	w57	w59	wo
PercepGAN(R) [29]	91	99	88	94	100	96	97	83	97	94	85	95	94	95	53
PercepGAN(P) [29]	97	86	90	77	81	89	93	78	92	66	83	88	93	71	54
DFPN(R) [23]	93	99	94	96	97	96	96	82	100	90	91	95	94	93	42
DFPN(P) [23]	98	92	98	97	86	90	97	81	97	90	95	95	90	68	50

adopting multiple computer vision tasks (such as object detection, semantic segmentation, instance segmentation, etc.) simultaneously can improve the performance of separate task with a large margin because of richer information. Wang et al.'s RDSNet [87] designs a two-stream structure to learn features on both the object level (i.e., bounding boxes) and the pixel level (i.e., instance masks) jointly. The RDSNet that combines object detection and instance segmentation task has been achieved good results on the MS-COCO dataset. That is to say, performing multi-task learning and optimization is a good way to aggregate multiple tasks in a network. Thus how to effectively utilize multi-task joint learning and optimization to improve small object detection performance is also the focus for future research.

(3) Information transmission: Making use of context surrounding the small object instances play a vital role in small object detection. Nevertheless, contextual information is not always useful for small object detection. GBDNet [41] uses LSTM to control different region information transmission which may avoid the introduction of useless background noises. Therefore how to effectively control contextual information transmission will be the further work to research.

(4) Weakly supervised small object detection methods: Current small object detectors based on deep learning make use of fully-supervised models learned from well-annotated images with bounding boxes or segmentation masks. But the annotation process of fully-supervised learning is extremely time-consuming and inefficient. Weakly supervised learning refers to use a few fully-annotated images to detect considerable non-fully-annotated ones, while fully-supervised learning is not scalable in the absence of fully-labeled training data. It is easy and high efficient that small object detectors are only trained with object class annotations but not the bounding box images. Thus developing small object detection algorithms based on weakly-supervised learning is an important issue for further study.

(5) Framework for small object detection task: It has become a paradigm to use weights of models pretrained on large-scale image classification datasets into object detection problem. Nevertheless, it is may not an optimal solution because of existing conflicts between detection and classification tasks. A majority of detection algorithms are based on or modified from classification backbone networks, and only a few of them try different selections (such as CornerNet [88] and ExtremeNet [89] based on Hourglass network [90]). Therefore how to develop a novel framework that directly handles the detection task of small objects is also a significant research direction in the future.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61573183 and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) under Grant 201900029.

References

[1] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, W. Ouyang, T-CNN: tubelets with convolutional neural networks for object detection from videos, IEEE Trans. Circ. Syst. Video Tech. 28 (10) (2018) 2896–2907.
[2] J. Dai, K. He, J. Sun, Instance-aware semantic segmentation via multi-task network cascades, Computer Vision and Pattern Recognition 2016, pp. 3150–3158.
[3] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2020) 386–397.
[4] Qi Wu, Chunhua Shen, Peng Wang, Anthony R. Dick, Anton van den Hengel, Image captioning and visual question answering based on attributes and external knowledge, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1367–1381.
[5] S. Herath, M. Harandi, F. Porikli, Going deeper into action recognition: a survey, Image Vis. Comput. 60 (2017) 4–21.
[6] B. Zhou, Z. Hang, X. Puig, T. Xiao, S. Fidler, A. Barriuso, A. Torralba, Semantic understanding of scenes through the ADE20K dataset, Int. J. Comput. Vis. 127 (3) (2016) 302–321.
[7] T.Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár, Microsoft COCO: common objects in context, European Conference on Computer Vision 2014, pp. 740–755.
[8] A. Kembhavi, D. Harwood, L.S. Davis, Vehicle detection using partial least squares, IEEE Trans. Pattern Anal. Mach. Intell. 33 (6) (2011) 1250–1265.
[9] V.I. Morariu, E. Ahmed, V. Santhanam, D. Harwood, L.S. Davis, Composite discriminant factor analysis, IEEE Winter Conference on Applications of Computer Vision 2014, pp. 564–571.
[10] T.T. Le, S.T. Tran, S. Mita, T.D. Nguyen, Real time traffic sign detection using color and shape-based features, Asian Conference on Intelligent Information and Database Systems 2010, pp. 268–278.
[11] C. Chen, M.-Y. Liu, O. Tuzel, J. Xiao, R-CNN for small object detection, Asian Conference on Computer Vision 2016, pp. 214–230.
[12] H. Krishna, C.V. Jawahar, Improving small object detection, Asian Conference on Pattern Recognition 2017, pp. 340–345.
[13] W. Zhang, S. Wang, S. Thachan, J. Chen, Y. Qian, Deconv R-CNN for small object detection on remote sensing images, IEEE International Geoscience and Remote Sensing Symposium 2018, pp. 2483–2486.
[14] R.B. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, IEEE Conference on Computer Vision and Pattern Recognition 2014, pp. 580–587.
[15] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Adv. Neural Inf. Proces. Syst. (2015) 91–99.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S.E. Reed, C.-Y. Fu, A.C. Berg, SSD: single shot multibox detector, European Conference on Computer Vision 2016, pp. 21–37.
[17] C. Eggert, S. Brehm, A. Winschel, D. Zecha, R. Lienhart, A closer look: small object detection in faster R-CNN, International Conference on Multimedia and Expo 2017, pp. 421–426.
[18] C. Cao, B. Wang, W. Zhang, X. Zeng, X. Yan, Z. Feng, Y. Liu, Z. Wu, An improved faster R-CNN for small object detection, IEEE Access 7 (2019) 106838–106846.
[19] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, A.C. Berg, DSSD: Deconvolutional Single Shot Detector, CoRR abs/1701.06659 2017.
[20] G. Cao, X. Xie, W. Yang, Q. Liao, G. Shi, J. Wu, Feature-Fused SSD: Fast Detection for Small Objects, *CoRR abs/1709.05054* 2017.
[21] M. Xu, L. Cui, P. Lv, X. Jiang, J. Niu, B. Zhou, M. Wang, MDSSD: multi-scale deconvolutional single shot detector for small objects, SCIENCE CHINA Inf. Sci. 63 (2) (2020) 120113.
[22] T.-Y. Lin, P. Dollár, R.B. Girshick, K. He, B. Hariharan, S.J. Belongie, Feature pyramid networks for object detection, IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 936–944.
[23] Z. Liang, J. Shao, D. Zhang, L. Gao, Small object detection using deep feature pyramid networks, Pacific-Rim Conference on Multimedia 2018, pp. 554–564.
[24] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, K. Cho, Augmentation for Small Object Detection, *CoRR abs/1902.07296* 2019.
[25] B. Singh, L.S. Davis, An analysis of scale invariance in object detection SNIP, IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 3578–3587.
[26] B. Singh, M. Najibi, L.S. Davis, SNIPER: efficient multi-scale training, Neural Information Processing Systems 2018, pp. 9333–9343.
[27] S. Bell, C.L. Zitnick, K. Bala, R. Girshick, Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks, IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 2874–2883.
[28] L. Guan, Y. Wu, J. Zhao, SCAN: semantic context aware network for accurate small object detection, Int. J. Comput. Int. Sys. 11 (1) (2018) 936–950.
[29] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, S. Yan, Perceptual generative adversarial networks for small object detection, IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 1951–1959.
[30] Y. Bai, Y. Zhang, M. Ding, B. Ghanem, SOD-MTGAN: small object detection via multi-task generative adversarial network, European Conference on Computer Vision 2018, pp. 210–226.
[31] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.
[32] Q. Chen, Z. Song, J. Dong, Z. Huang, Y. Hua, S. Yan, Contextualizing object detection and classification, IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2015) 13–27.
[33] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy, X. Tang, DeepID-Net: deformable deep convolutional neural networks for object detection, IEEE Conference on Computer Vision and Pattern Recognition 2015, pp. 2403–2412.
[34] R. Girshick, Fast R-CNN, IEEE International Conference on Computer Vision 2015, pp. 1440–1448.
[35] S. Gidaris, N. Komodakis, Object detection via a multi-region & semantic segmentation-aware CNN model, International Conference on Computer Vision 2015, pp. 1134–1142.
[36] Y. Zhu, R. Urtasun, R. Salakhutdinov, S. Fidler, segDeepM: exploiting segmentation and context in deep neural networks for object detection, IEEE Conference on Computer Vision and Pattern Recognition 2015, pp. 4703–4711.
[37] T. Kong, A. Yao, Y. Chen, F. Sun, Hypernet: towards accurate region proposal generation and joint object detection, IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 845–853.
[38] J. Dai, Y. Li, K. He, J. Sun, R-FCN: object detection via region-based fully convolutional networks, Advances in Neural Information Processing Systems 2016, pp. 379–387.
[39] Z. Cai, Q. Fan, R.S. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, European Conference on Computer Vision 2016, pp. 354–370.
[40] S. Zagoruyko, A. Lerer, T.-Y. Lin, P.O. Pinheiro, S. Gross, S. Chintala, P. Dollár, A MultiPath network for object detection, British Machine Vision Conference, 2016.
[41] X. Zeng, W. Ouyang, B. Yang, J. Yan, X. Wang, Gated bi-directional CNN for object detection, European Conference on Computer Vision 2016, pp. 354–369.
[42] A. Shrivastava, A. Gupta, Contextual priming and feedback for faster R-CNN, European Conference on Computer Vision 2016, pp. 330–348.
[43] X. Chen, A. Gupta, Spatial memory for context reasoning in object detection, International Conference on Computer Vision 2017, pp. 4106–4116.
[44] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, S. Yan, Attentive contexts for object detection, IEEE Trans. Multimed. 19 (5) (2017) 944–954.
[45] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, H. Lu, CoupleNet: coupling global structure with local parts for object detection, International Conference on Computer Vision 2017, pp. 4146–4154.
[46] Z. Li, F. Zhou, FSSD: Feature Fusion Single Shot Multibox Detector, *CoRR abs/1712.00960* 2017.
[47] S. Zhang, L. Wen, X. Bian, Z. Lei, Single-shot refinement neural network for object detection, IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 4203–4212.
[48] Z. Li, Y. Chen, G. Yu, Y. Deng, R-FCN++: towards accurate region-based fully convolutional networks for object detection, The Association for the Advance of Artificial Intelligence 2018, pp. 7073–7080.
[49] H. Hu, J. Gu, Z. Zhang, J. Dai, Y. Wei, Relation networks for object detection, IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 3588–3597.
[50] Y. Liu, R. Wang, S. Shan, X. Chen, Structure inference net: object detection using scene-level context and instance-level relationships, IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 6985–6994.
[51] Y. Kim, B.-N. Kang, D. Kim, SAN: learning relationship between convolutional features for multi-scale object detection, European Conference on Computer Vision 2018, pp. 328–343.
[52] G. Wang, Z. Xiong, D. Liu, C. Luo, Cascade mask generation framework for fast small object detection, IEEE International Conference on Multimedia and Expo 2018, pp. 1–6.
[53] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, H. Ling, M2Det: a single-shot object detector based on multi-level feature pyramid network, AAAI 2019, pp. 9259–9266.
[54] S. Zafeiriou, C. Zhang, Z. Zhang, A survey on face detection in the wild: past, present and future, Comput. Vis. Image Und. 138 (9) (2015) 1–24.
[55] N. Wang, X. Gao, D. Tao, H. Yang, X. Li, Facial feature point detection: a comprehensive survey, Neurocomputing 275 (1) (2017) 50–65.
[56] W. Yue, J. Qiang, Facial landmark detection: a literature survey, Int. J. Comput. Vis. 127 (2) (2019) 115–142.
[57] Q. Ye, D.S. Doermann, Text detection and recognition in imagery: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 37 (7) (2015) 1480–1500.
[58] X.C. Yin, Z.Y. Zuo, S. Tian, C.L. Liu, Text detection, tracking and recognition in video: a comprehensive survey, IEEE Trans. Image Process. 25 (6) (2016) 2752–2773.
[59] S. Sivaraman, M.M. Trivedi, Looking at vehicles on the road: a survey of vision-based vehicle detection, tracking, and behavior analysis, IEEE Trans. Intell. Transp. Syst. 14 (4) (2013) 1773–1795.
[60] A. Mogelmose, M.M. Trivedi, T.B. Moeslund, Vision-based traffic sign detection and analysis for intelligent driver assistance systems: perspectives and survey, IEEE Trans. Intell. Transp. Syst. 13 (4) (2012) 1484–1497.
[61] G. David, A.M. López, A.D. Sappa, G. Thorsten, Survey of pedestrian detection for advanced driver assistance systems, IEEE Trans. Pattern Anal. Mach. Intell. 32 (7) (2010) 1239–1258.
[62] A. Brunetti, D. Buongiorno, G.F. Trotta, V. Bevilacqua, Computer vision and deep learning techniques for pedestrian detection and tracking: a survey, Neurocomputing 300 (1) (2018) 17–33.
[63] L. Wei, D. Qian, A survey on representation-based classification and detection in hyperspectral remote sensing imagery, Pattern Recogn. Lett. 83 (2016) 115–123.
[64] G. Cheng, J. Han, A survey on object detection in optical remote sensing images, ISPRS J. Photogramm. Remote Sens. 117 (2016) 11–28.
[65] S. Agarwal, J.O.d. Terrail, F.e.e. Jurie, Recent Advances in Object Detection in the Age of Deep Convolutional Neural Networks, CoRR abs/1809.03193 2018.
[66] Z.Q. Zhao, P. Zheng, S.T. Xu, X. Wu, Object detection with deep learning: a review, IEEE Trans. Neural Netw. Learn. Syst. 30 (11) (2019) 3212–3232.
[67] L. Liu, W. Ouyang, X. Wang, P.W. Fieguth, J. Chen, X. Liu, M. Pietikäinen, Deep learning for generic object detection: a survey, Int. J. Comput. Vis. 128 (2) (2020) 261–318.
[68] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, R. Qu, A survey of deep learning-based object detection, IEEE Access 7 (2019) 128837–128868.
[69] Z. Zou, Z. Shi, Y. Guo, J. Ye, Object Detection in 20 years: A Survey, *arXiv: 1905.05055v2* 2019 1–39.
[70] X. Wu, D. Sahoo, S.C.H. Hoi, Recent Advances in Deep Learning for Object Detection, arXiv:1908.03673v1 2019 1–40.
[71] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, Computer Vision and Pattern Recognition 2005, pp. 886–893.
[72] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[73] Y. Liu, H. Li, J. Yan, F. Wei, X. Wang, X. Tang, Recurrent scale approximation for object detection in CNN, IEEE International Conference on Computer Vision 2017, pp. 571–579.
[74] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, Y. Chen, RON: reverse connection with objectness prior networks for object detection, IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 5244–5252.
[75] N. Bodla, B. Singh, R. Chellappa, L.S. Davis, Soft-NMS - improving object detection with one line of code, IEEE International Conference on Computer Vision 2017, pp. 5562–5570.
[76] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial nets, Neural Information Processing Systems 2014, pp. 2672–2680.
[77] M. Everingham, L. Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[78] S. Romberg, L.G. Pueyo, R. Lienhart, R.V. Zwol, Scalable logo recognition in real-world images, International Conference on Multimedia Retrieval 2011, pp. 25–33.
[79] C. Eggert, D. Zecha, S. Brehm, R. Lienhart, Improving small object proposals for company logo detection, International Conference on Multimedia Retrieval 2017, pp. 167–174.
[80] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, Computer Vision and Pattern Recognition 2012, pp. 3354–3361.
[81] J. Xiao, K.A. Ehinger, J. Hays, A. Torralba, A. Oliva, SUN database: exploring a large collection of scene categories, Int. J. Comput. Vis. 119 (1) (2010) 3–22.
[82] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, S. Hu, Traffic-sign detection and classification in the wild, Computer Vision and Pattern Recognition, 2016.
[83] C. Wojek, P. Dollar, B. Schiele, P. Perona, Pedestrian detection: an evaluation of the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 743–761.
[84] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, J. Feng, Dual path networks, Neural Information Processing Systems 2017, pp. 4467–4475.
[85] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, F.F. Li, ImageNet: a large-scale hierarchical image database, Computer Vision and Pattern Recognition 2009, pp. 248–255.

[86] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, The Kinetics Human Action Video Dataset, CoRR abs/1705.06950 2017.
[87] Shaoru Wang, Yongchao Gong, Junliang Xing, Lichao Huang, Chang Huang, Weiming Hu, RDSNet: a new deep architecture for reciprocal object detection and instance segmentation, The Association for the Advance of Artificial Intelligence, 2020 , (pp. CoRR abs/1912.05070).
[88] H. Law, J. Deng, CornerNet: detecting objects as paired keypoints, European Conference on Computer Vision 2018, pp. 765–781.
[89] X. Zhou, J. Zhuo, P. Krähenbühl, Bottom-up object detection by grouping extreme and center points, Computer Vision and Pattern Recognition 2019, pp. 850–859.
[90] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, European Conference on Computer Vision 2016, pp. 483–499.