Image and Vision Computing 123 (2022) 104471

Contents lists available at ScienceDirect

Image and Vision Computing

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

Review article

Deep learning-based detection from the perspective of small or tiny objects: A survey

Kang Tong, Yiquan Wu *

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

ARTICLE INFO

Article history:
Received 6 December 2021
Received in revised form 19 March 2022
Accepted 4 May 2022
Available online 10 May 2022

Keywords:
Object detection
Small or tiny objects
Deep learning
Datasets
Convolutional neural networks

ABSTRACT

Detecting small or tiny objects is always a difficult and challenging issue in computer vision. In this paper, we provide a latest and comprehensive survey of deep learning-based detection approaches from the perspective of small or tiny objects. Our survey is featured by thorough and exhaustive analysis of small or tiny object detection. We comprehensively introduce 30 existing datasets about small or tiny objects, and summarize different definitions of small or tiny objects based on different application scenarios, such as pedestrian detection, traffic signs detection, face detection, remote sensing target detection and object detection in common life. Then small or tiny object detection techniques are overviewed systematically from seven aspects, including super-resolution techniques, context-based information, multi-scale representation learning, anchor mechanism, training strategy, data augmentation, and schemes based on loss function. Finally, the detection performance of small or tiny objects on 12 popular datasets is analyzed in depth. Based on performance analysis, we also discuss the promising research directions in the future. We hope this survey could provide researchers guidance to catalyze understanding of small or tiny object detection and further facilitate research on small or tiny object detection systems.

Contents

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

https://doi.org/10.1016/j.imavis.2022.104471
0262-8856/

1. Introduction

Object detection is a fundamental task in computer vision. When given an image, object detection aims at finding where and what each object instance is. From the application perspective, object detection can be grouped into two types: generic object detection and domain-specific detection. The first type aims at detecting different types of visual objects under a unified framework, while the purpose of the second type is to the detection under specific application scenarios, such as face detection [1,2], traffic sign detection [3,4], pedestrian detection [5,6], remote sensing target detection [7–10] and so on. Object detection has been widely used in many applications, such as robot vision, autonomous driving, intelligent transportation surveillance, human-computer interaction, content based image retrieval, drone scene analysis, consumer electronics, and augmented reality.

Even though impressive results have been achieved on large and medium objects in large-scale detection benchmarks, the performance on small or tiny objects is far from satisfactory. As shown in detection leaderboard of MS-COCO challenge[1], the detection accuracy of small objects is much lower than that of large objects. Nowadays, small or tiny object detection [5–9,11–26] has become an extremely challenging problem because of low-resolution, insufficient appearance information, limited prior knowledge, etc. In this paper, we mainly focuses on the major progress of deep learning-based small or tiny object detection methods in recent three years. Some other related works are also included in order to completeness and better readability.

1.1. Comparison with Previous Reviews

As summarized in Table 1, many object detection reviews have been published in recent years. These include generic object detection [27–31]. Zou et al. [28] and Zhao et al. [27] just show small object detection in the future directions. These works provide a comprehensive, systematic, and thorough survey. However, they focus on general-size objects, not small or tiny objects. Meanwhile, they do not analyze small or tiny object detection in depth. In addition to these generic object detection surveys, there are some recent reviews on the small object detection [32–35]. Nguyen et al. [32] mainly focuses on small object performance evaluation on four models for deep learning, including Fast R-CNN [36], Faster R-CNN [37], RetinaNet [38], and YOLOv3 [39]. They also provide a profound assessment of the advantages and limitations of these models. In our previous work [33], we comprehensively reviews 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. Besides, we introduce evaluation criteria in detail and analyze experimental results of different algorithms on five datasets, including MS-COCO [40], PASCAL-VOC [41], Caltech [42], KITTI [43] and TT100K [44]. Finally, we point out five promising research directions in the future. Chen et al. [34] survey the four pillars for deep learning-based small object detection: multiscale representation, contextual information, super-resolution, and region-proposal. Then, they list some small object detection datasets, and report the performance of different methods on three datasets, such as MS-COCO, PASCAL-VOC and TT100K. The six possible future directions are also provided in their survey. Liu et al. [35] summarize the four challenges of small object detection: 1) individual feature layers do not contain sufficient information of small objects; 2) limited context information of small objects; 3) class imbalance for small objects; 4) insufficient positive examples for small objects. Then they offer the corresponding solutions as follows: 1) combine multiple feature maps; 2) add context information 3) balance category examples; 4) increase sufficient number of positive examples. They also compare the performance of some approaches for small object detection, such as YOLOv3, Faster R-CNN, and SSD [45], based on three benchmark datasets. Although these reviews focus on the detection methods and performance evaluation of small objects, they just only cover literatures before 2020 and do not involve the tiny object detection as well. Moreover, these works lack a systematic and comprehensive summary of the definitions and datasets for small or tiny objects. Last but not least, these papers are not thorough, comprehensive and in-depth for the analysis and comparison of different small object detection approaches on different datasets.

Unlike these previous surveys, we present a latest survey of deep learning-based methods that focus on detecting small or tiny objects. Our review is featured by thorough and in-depth analysis of small or tiny object detection. We comprehensively introduce the existing datasets about small or tiny object detection, and summarize different definitions of small or tiny objects based on different application scenarios. Then small or tiny object detection techniques are overviewed systematically. Last the performance for detecting small or tiny objects is analyzed and discussed in depth.

1.2. Our contributions

Our contributions in this paper are summarized as follows:

1) Provide the latest survey of deep learning-based detection algorithms from the perspective of small or tiny objects.

2) Comprehensively summarize the 30 datasets about small or tiny objects, and offer different definitions for small or tiny objects based on different application scenarios, such as pedestrian detection, traffic signs detection, face detection, remote sensing target detection and object detection in common life.

3) Systematically review small or tiny object detection techniques from seven aspects: super-resolution techniques, context-based information, multi-scale representation learning, anchor mechanism, training strategy, data augmentation, and schemes based on loss function.

4) In-depth analysis the detection performance of small or tiny objects on 12 datasets, including DOTA [46], UAVDT [47], AI-TOD [48], DIOR [49], KITTI, TinyPerson [50], TT100K, WIDER FACE [51], PASCAL-VOC, MS-COCO, SOD [52] and USC-GRAD-STDdb [53]. Based on performance analysis, we discuss the possible future research directions.

These contributions altogether bring an up-to-date, thorough, and exhaustive survey, and differentiate it from previous review works significantly. We identify this paper as a timely complement to the small object detection community. Also, we hope this survey will provide researchers with novel inspirations to facilitate understanding of small or tiny object detection and further catalyze research on detection systems.

The rest of the paper is organized as follows. In Section 2, we summary the datasets and definitions for small or tiny objects. Then we review small object detection techniques based on deep learning in Section 3. Performance analysis and discussion are given in Section 4. Finally, our conclusion is drawn in Section 5.

[1] https://cocodataset.org/#detection-leaderboard

Table 1
Summarization of some related reviews since 2019

Title	Venue/Year	Description
Object Detection with Deep Learning: A Review [27]	TNNLS/2019	It provides a detailed review on deep learning based object detection frameworks.
Object Detection in 20 Years: A Survey [28]	arXiv/2019	Surveys extensively 400 + papers of object detection in the light of its technical evolution in 20 years.
Deep Learning for Generic Object Detection: A Survey [29]	IJCV/2020	A comprehensive survey of deep learning for generic object detection.
Recent Advances in Deep Learning for Object Detection [30]	Neurocomputing /2020	Reviews systematically the existing object detection methods from three parts: detection components, learning strategies, applications and benchmarks.
Imbalance problems of object detection: A review [31]	PAMI/2020	Presents a comprehensive review of the imbalance problems in object detection.
An evaluation of deep learning methods for small object detection [32]	JECE/2020	Focuses on small object performance evaluation on four models for deep learning: Fast R-CNN, Faster R-CNN, RetinaNet, and YOLOv3. Also, pros and cons of these models are introduced.
Recent advances in small object detection based on deep learning: A review [33]	IVC/2020	Reviews the existing deep learning-based small object detection methods from five aspects, analyses experimental results on five datasets, and points out five promising directions.
A survey of the four pillars for small object detection: multiscale representation, contextual information, super-resolution, and region proposal [34]	TSMC/2020	Surveys the four pillars for deep learning-based small object detection, lists some small object detection datasets and reports the performance of different methods on three datasets. Six possible directions in the future are also provided.
A survey and performance evaluation of deep learning methods for small object detection [35]	ESWA/2021	Aiming at the four challenges of small object detection, it summarizes the corresponding solutions and offers some experimental analysis.
Deep learning-based detection from the perspective of small or tiny objects: A survey	Ours	This survey comprehensively discusses small or tiny object datasets, the definitions of small or tiny objects, techniques for small or tiny object detection, detection performance analysis of small or tiny objects, and promising directions for small/tiny object detection.

These papers provide a comprehensive and thorough review for object detection. Nevertheless, they only focus on general-size objects, not small or tiny objects.

These just cover literatures before 2020 and can't involve tiny object detection. Lack a systematic summary of the definitions and datasets for small objects. The performance analysis of small object detection is not thorough.

2. Datasets and definitions for small or tiny objects

In this section, we first introduce some popular datasets about small or tiny objects in chronological order. Then we summarize different definitions for small or tiny objects based on different application scenarios.

2.1. Datasets about small or tiny objects

Datasets have played a critical role in object detection. They not only provide the data for data-driven algorithms, but also enable comparison between different object detection algorithms. However, there are few generally accepted datasets for small or tiny objects. Most researchers have to evaluate their small or tiny object detection methods on the datasets built by themselves or extracted from large datasets such as MS-COCO, WIDER FACE, etc. Table 2 summarizes some popular datasets about small or tiny objects in chronological order.

2.2. Definitions for small or tiny objects

The definition of small or tiny objects refers to clarify how small scales or sizes of objects are or how many pixels they occupy in an image. There are two main ways to define small objects. One is the relative size. According to the definition of SPIE, if the object size is less than 0.12% of the original image, it is regarded as a small object. Krishna and Jawahar [72] show that an object is considered small if it occupies only a tiny portion of the image (less than 1% of the image area). Namely, the bounding box of small object should cover less than 1% of the original image. The other is the absolute size, such as the small object whose size is less than 32×32 pixels defined in MS-COCO dataset, 16×16 pixels defined in USC-GRAD-STDdb [53]. To facilitate in-depth understanding of small object detection, researchers offer different definitions for small or tiny objects based on different application scenarios, as shown in Table 3.

3. Techniques for small or tiny object detection

3.1. An overview of small/tiny object detection

As depicted in Fig. 1, we summarize techniques for small or tiny object detection. We attempt to understand small/tiny object detection from two perspectives: definition and difficulty. The definition of small/tiny object detection refers to determine whether there are any instances of small/tiny objects from given categories in an image and, if present, to return the spatial location and extent of each small/tiny object instance (e.g., via a bounding box). In short, small/tiny object detection needs to complete two steps: localization and classification. On the one hand, rich semantic information is beneficial for the classification of small objects. Semantic context-based information and deep features of CNN cover semantic-rich information, which can facilitate small object classification a lot. On the other hand, rich spatial details are vital to small object localization. Shallow features of CNN and super-resolution technique can capture more details of small objects, which improve the localization accuracy of small objects. Moreover, spatial context-based information and anchor mechanism are also of importance for small object localization.

Compared with large and medium objects, the small/tiny objects are more difficult to be detected accurately. This is because that there are four difficulties in small/tiny object detection. Firstly, small/tiny objects have low resolution and insufficient features. Secondly, the span of object-scale is large and multiple scales coexist. Thirdly, the examples of small/tiny objects are scarce. Finally, the categories for small/tiny objects are imbalance. There are six methods to deal with the above four difficulties, as shown in Fig. 1. Specifically, it is much necessary to capture more additional contextual information as the supplement of the small objects, since the effective features extracted from the small objects are very limited. Multi-scale representation learning can not only provide more effective information for small objects, but also alleviate the problem of large span of object-scale to a certain extent. Besides, training strategy is also used to deal with object-scale problem. Anchor

Table 2
An overview of some popular detection datasets about small or tiny objects

Dataset	Description	Published/Year
SDOTA [54]	This dataset for small object detection based on DOTA-v1.5 [46]. The size of images ranges from 800×800 to 4000×4000 pixels, including 227656 instances covering four classes, and most of them are small objects less than 50 pixels and a few large objects. The four categories of the dataset are small-vehicle, storage tank, ship, and large-vehicle. Besides, the small-vehicle class contains a large number of small objects under 10 pixels.	J-STARS/2021
SDD [54]	This dataset comes from DOTA-v1.5 dataset and DIOR dataset [49], including five classes, 12628 aerial images, and 343961 labeled examples. The size of imagery in this dataset is from 800×800 to 1024×1024 pixels. The five categories of this dataset are vehicle, airplane, ship, windmill, and swimming-pool.	J-STARS/2021
Small Object Dataset [55]	It contains about 2200 images with the labeling manually. The students' heads are considered as the small objects for detection. There are 550 images for training, 1100 images for testing, and 550 images for verification. These images of the small objects are taken from a university classroom video record.	IEEE Access/2021
Small Target Detection database (USC-GRAD-STDdb) [53]	It contains 115 video segments with more than 25000 annotated frames of HD 720p resolution with small objects of interest from 16 to 256 as pixel area. The total number of labeled small objects is over 56000. The test subset holds 11337 objects, where almost 90% of them (10136 objects) correspond to the very small subset. The videos in this database contain the three main landscapes with five object classes, namely: air (drone, bird), 57 videos with 12139 frames; sea (boat), 28 videos with 7099 frames; and land (vehicle, person), 30 videos with 6619 frames. 80% of the videos of the database are used for training (92 videos), while the remaining 20% are used for testing (23 videos) [53].	EAAI/2020
DIOR [49]	This dataset is a large publicly available dataset for evaluating object detectors in the field of remote sensing. It includes 23463 images and 190288 objects, covering 20 objects classes: basketball-court, baseball-field, airplane, windmill, ground-track-field, expressway-toll-station, storage-tank, train station, expressway-service-area, overpass, dam, harbor, golf-field, bridge, chimney, ship, stadium, TC, airport and vehicle. Each instance in the dataset is labeled by experts utilizing horizontal bounding boxes, and the size of the imagery is 800×800 pixels.	ISPRS-JPRS/2020
TinyPerson [50]	It contains 5 categories, including sea person, earth person, uncertain sea person, uncertain earth person, and ignore region. There are 1610 images, with 794 images for training and 816 images for training. The total number of instances is 72651. The objects' relative size of TinyPerson is smaller than that of CityPersons [56].	WACV/2020
TinyCityPersons [50]	For a tiny object dataset, extreme small size is one of the main challenges. TinyCityPersons is constructed through down-sampling CityPersons by 4×4, where mean of objects' absolute size is same as that of TinyPerson. It is used to quantify the effect of absolute size reduction on performance.	WACV/2020
Tiny Object Detection in Aerial Images (AI-TOD) [48]	AI-TOD is a new dataset for advancing tiny object detection in aerial images. It comes with 700621 object instances for 8 categories across 28036 aerial images with sizes of 800×800 pixels. Compared to existing object detection datasets in aerial images, the mean size of objects in AI-TOD is about 12.8 pixels, which is much smaller than others. For dataset splits, 2/5, 1/10 and 1/2 of the images are used to form training set, validation set and test set.	ICPR/2020
Small and Dense Object Dataset of Milk Tea (SDOD-MT) [57]	The dataset contains 16919 images collected in different scenes and 392969 instances within 13 categories. It is divided into two subsets according to the popular principle of dividing datasets: trainval set and test set, where the proportion is about 4:1. This dataset could be applied into multiple scenes through combining with the related hard-ware system, such as warehouse management and physical retail.	ACM MM/2019
VisDrone [58,59]	It provides a dataset of 10209 images for detection task, with 3190 images used for testing, 6471 images for training, and 548 images for validation. The images of the three subsets are taken at different locations, but share similar environments and attributes. This dataset mainly focuses on vehicles and human in our daily life, and defines 10 object classes of interest including car, truck, bus, van, motor, bicycle, awning-tricycle, tricycle, pedestrian and person.	ICCV Workshops/2019
UAVDT [47]	Objects in this dataset are usually tiny or small due to high altitude of UAV views, resulting in difficulties to detect them. Three levels are annotated, including low altitude (low-alt), medium altitude (medium-alt) and high altitude (high-alt). When shooting in low-alt, more details of objects are captured. However, shooting in much higher altitude, plentiful vehicles are of less clarity. For example, most tiny objects just contain 0.005% pixels of a frame, yet object numbers can be more than a hundred. It contains 375884 test objects, where 76215 are considered within the very small subset (20.3%) and 281532 within the small subset (74.9%).	ECCV/2018
DOTA [46]	DOTA is a large-scale dataset for object detection in aerial images. It includes 2086 images, 188282 instances, and 15 common categories, such as plane, ship, storage tank, harbor, bridge, baseball diamond, tennis court, basketball court, ground track field, small vehicle, large vehicle, helicopter, roundabout, basketball court and soccer ball field. Each image is of the size about 4000×4000 pixels and contains objects of different shapes, orientations and scales. The proportions of the testing set, training set and validation set in this dataset are 1/3, 1/2 and 1/6 respectively.	CVPR/2018
EuroCity Persons [60]	The images are collected on-board a moving vehicle in 31 cities of 12 European countries. This dataset includes around 47300 images with over 238200 person instances manually labeled. Moreover, it contains over 211200 person orientation annotations. For each city the recordings are split into chunks with a duration of at least 20 minutes. The recorded images of each chunk are separated into validation, training, and test by 10%, 60%, and 30% respectively.	arXiv/2018
DeepScores [61]	It includes high quality images of musical scores, partitioned into 300000 sheets of written music that contain 123 different symbol categories. This dataset is the largest public dataset that covers around a hundred million small objects. Also, it can be used for different visual tasks, such as object detection and semantic segmentation.	ICPR/2018
Small Object Dataset (from MS-COCO) [62]	They pick out the smaller object categories (knife, fork, etc.) from MS-COCO dataset that are really small and easy to cluster from the complex scene, such as kitchen, to compose the subset. It contains the training set, the validation set, and the test set. There are 10 categories, including knife, bottle, wine glass, cup, spoon, fork, bowl, sports ball, orange, vase.	ICCC/2018
CityPersons [56]	It is built upon the Cityscapes dataset [63] for pedestrian detection. It consists of images recorded across 27 cities, 3 seasons, various weather conditions and more common crowds. Train-val-test split (%): 60-10-30. Fine pixel-level annotations of 30 categories are provided for 5000 images. The fine annotations include instance labels for vehicles and persons. Besides, 20000 images from 23 other cities are annotated with coarse semantic labels, without instance labels.	CVPR/2017
Bosch Small Traffic Lights [64]	This database contains 13427 images of size 1280×720 pixels with about 24000 annotated traffic lights, annotated with bounding boxes and states (active light). It is the largest publicly available labeled traffic light dataset and includes labels down to the size of only 1 pixel in width.	ICRA/2017
Tsinghua-Tencent 100K (TT100K) [44]	It is the largest traffic sign detection dataset so far, with 100000 images and 30000 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	CVPR/2016

Table 2 (*continued*)

Dataset	Description	Published/Year
	abundance, huge illumination and scale variations. There are 45 classes with at least 100 instances present [33].	
Small Object Dataset (SOD) [52]	The dataset is composed by using a subset of images from both the SUN [65] dataset and MS-COCO dataset. It includes about 8393 object instances in 4925 images from 10 categories. The selected object categories are "clock", "telephone", "switch", "outlet", "mouse", "toilet paper", "tissue box", "faucet", "plate", and "jar". The "tissue box" category has the smallest number of instances: 103 instances in 100 images. The "mouse" category has the largest number of object instances: 2137 instances in 1739 images. All the object instances in this dataset are smaller than 30 centimeters.	ACCV/2016
WIDER FACE [51]	Images in this dataset are categorized into 60 event classes, which have much more diversities and are closer to the real-world scenario. This dataset consists of 393703 labeled face bounding boxes in 32203 images. Based on the heights of the ground-truth faces, they are also divided into three subsets, small, medium, and large. 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%.	CVPR/2016
Lost and Found [66]	This dataset mainly focuses on the detection of small hazards and lost cargo 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. In addition, 112 stereo video sequences are included, corresponding with 2104 annotated frames.	IROS/2016
Stanford Drone Dataset (SDD) [67]	The dataset is the first large-scale dataset that has images and videos of various categories of objects that are moving and interacting on a real-world university campus. It is very challenging due to the tiny size of three kinds of objects (pedestrian, biker and car). The training and validation set contain 69673 images, with 53224 images for testing.	ECCV/2016
DLR 3k Munich Dataset [68]	It is common used dataset in the small vehicle detection literature with 20 extra large images. 10 training images with up to 3500 cars and 70 trucks and 10 test images with 5800 cars 90 trucks.	GRSL/2015
PASCAL FACE [69]	The small/medium/large subsets contain faces with heights larger than 10/50/300 pixels, respectively. The small subset accounts for 41% of this dataset while medium accounts for 57%.	IVC/2014
MS-COCO [40]	The objects distribution in this dataset is closer to real world scenarios. It contains about 164000 images and 897000 annotated objects from 80 categories. 118287 images, 5000 images and 40670 images are for training, validation and testing set respectively.	ECCV/2014
German Traffic Sign Detection Benchmark (GTSDB) [70]	It provides a total of 900 images with 1206 traffic signs. Images capture different scenarios during the day-time and dusk featuring various weather conditions. The image resolution is 1360×800. The traffic sign sizes vary between 16 and 128 pixels w.r.t. the longest edge.	IJCNN/2013
KITTI [43]	The dataset contains 7481 labeled images and another 7518 images for testing. There are 100000 instances of pedestrians. With around 6000 identities and one person in average per image. The person class in this dataset is divided into two subclasses: pedestrian and cyclist. The object labels are grouped into hard, moderate and easy levels, based on the extent to which the objects are occluded and truncated [33]. Train-val-test split (%): 50-0-50.	CVPR/2012
Caltech [42]	It includes 350000 pedestrian bounding boxes labeled in 250000 frames. Also, it is one of the most popular pedestrian detection datasets. The training set contains 192000 pedestrian instances, and testing set contains 155000 pedestrian instances.	PAMI/2012
LISA Traffic Sign Dataset [71]	The dataset contains 47 US signs and 7855 annotations on 6610 video frames. Sign sizes vary from 6×6 to 167×168 pixels. Each sign is annotated with sign size, type, position, on side road (yes/no), occluded (yes/no).	TITS/2012
PASCAL-VOC [41]	It has two versions, including VOC2007 and VOC2012. They are both mid-scale datasets with 20 categories. VOC07 consists of 2501 training images, 2510 validation images and 5011 testing images. However, VOC12 contains 5717 training images, 5823 validation images and 10991 testing images.	IJCV/2010

mechanism can help more anchors match to ground truths of small objects by adaptively setting anchor scales and ratios, which improve scarce examples of small/tiny objects. Data augmentation is another effective strategy, which can not only alleviate the insufficient examples of small objects, but also improve the category imbalance of small objects. Furthermore, the usage of loss function also helps to balance the category of small objects. Last but not least, datasets and definitions for small or tiny objects shown in Section 2 are also briefly presented in Fig. 1.

3.2. Techniques for small or tiny object detection

Based on the above overview of small or tiny object detection, we will analyze techniques of small or tiny objects from seven respects: super-resolution techniques, context-based information, multi-scale representation learning, anchor mechanism, training strategy, data augmentation, and schemes based on loss function. To clearly explain the techniques of each aspect for small/tiny object detection, we first comprehensively describe the existing approaches among each aspect. Then we summarize the advantages and disadvantages of each aspect in detail.

3.2.1. Super-resolution techniques

Super-resolution techniques aim at recovering high resolution from corresponding low-resolution features. High-resolution image can be applied to small/tiny object detection because it provides more refined details about the original scene. The generative adversarial networks (GAN) [74] can be used to rebuild high-resolution images. It has achieved great progress in image super-resolution [75], which contains a generator and a discriminator. The generator yields super-resolved images to fool the discriminator while the discriminator attempts to distinguish the real images from fake images produced via the generator.

To the best of our knowledge, Li et al. [76] use the GAN method in small/tiny object detection task for the first time. The proposed perceptual GAN model improves small traffic sign detection by generating super-resolved representations for small objects to narrow representation difference of small objects from the large ones. The details of the perceptual GAN are shown in Fig. 2. Specifically, its generator is a deep residual network which takes the low-level features as the input to capture more details for super-resolved representation. Multiple residual blocks in the generator are employed to learn the residual representation between small objects and similar large objects. Its discriminator takes the features of large object and the super-resolved representation of small object as inputs and splits into adversarial branch and perception branch. The adversarial branch contains three fully connected layers followed via sigmoid, which is utilized to distinguish the generated super-resolved region of small objects from similar large objects. The perception branch contains two fully connected layers followed by two output sibling layers, which are employed for bounding box regression and classification respectively to justify the detection accuracy from the generated super-resolved representation.

To alleviate the influence of down-sampling operation on the loss of small traffic sign details, Yang et al. [77] propose a small object detection method in a coarse-to-fine manner. Specifically, some rough regions of

Table 3
The definition of small or tiny objects on different datasets

Application scenarios	Dataset	Definition or description of small or tiny objects
Remote sensing	SDOTA [54]	The four classes of this dataset are small-vehicle, ship, ST, and large-vehicle. Most of them are small objects less than 50 pixels. Moreover, the small-vehicle category contains many small objects under 10 pixels.
	SDD [54]	This dataset views the objects with a width or height of less than 50 pixels as small objects.
	DIOR [49,54]	View the objects with a width or height of less than 50 pixels as small objects.
	AI-TOD [48]	Consider objects in the range 2 to 8 pixels as very tiny, 8 to 16 pixels as tiny, 16 to 32 as small, 32 to 64 as medium, and no large objects. The percentages of very tiny, tiny, small and medium objects in AI-TOD are 13.3%, 72.3%, 12.3% and 2.1%, respectively.
	UAVDT [47]	Objects are usually small or tiny due to high altitude of UAV views. When shooting in much higher altitude (more than 70m), plentiful vehicles are of less clarity. For example, most tiny objects just contain 0.005% pixels of a frame, yet object numbers can be more than a hundred.
	DOTA [46]	They divide all the instances in DOTA into three splits according to the height of horizontal bounding box (call pixel size for short): small for range from 10 to 50, middle for range from 50 to 300, and large for range above 300.
	SDD [67]	All object instances have a size not larger than 0.2% of image size, and a considerable percentage of them are between 0.1% and 0.15%.
	DLR [68]	The vehicle detection is a challenging problem due to the small size of the vehicles (a car might be only 30×12 pixels).
Pedestrian detection or person detection	TinyPerson [50]	The size range is divided into 2 intervals: tiny objects [2,20], small objects [20,32]. And for tiny objects [2,20], it is partitioned into 3 sub-intervals: tiny1 [2,8], tiny2 [8,12], tiny3 [12,20].
	TinyCityPersons [50]	The mean of objects' absolute size is same as that of TinyPerson.
	EuroCity Persons [60]	Persons with a height between 30 pixels and 60 pixels which are occluded/truncated less than 40%.
	CityPersons [56]	The sizes of small-scale person vary between 30 and 80 pixels.
	Caltech [42]	They group pedestrians by their image size (height in pixels) into three scales: near (80 or more pixels), medium (between 30-80 pixels), and far (30 pixels or less). Most pedestrians are observed at heights of 30 to 80 pixels.
	Bosch [64]	The different traffic light sizes within the training set vary between approximately 1 and 85 pixels. The average width of the traffic lights in this dataset in only 10 pixels.
Traffic signs or lights detection	TT100K [44]	Small objects are objects whose sizes are filling 20% of an image. If the traffic sign has its square size, it is a small object when the width of the bounding box is less than 20% of an image and the height of the bounding box is less than the height of an image. Besides, the typical traffic sign instances are less than 32×32 pixels.
	GTSDB [70]	The traffic sign sizes vary between 16 and 128 pixels w.r.t. the longest edge.
	LISA [71]	Sign sizes vary from 6×6 to 167×168 pixels.
Face detection	WIDER FACE [51]	They group the faces by their image size (height in pixels) into three scales: small (between10-50 pixels), medium (between 50-300 pixels), large (over 300 pixels).
	PASCAL FACE [69]	The small/medium/large subsets contain faces with heights larger than 10/50/300 pixels, respectively. The small subset accounts for 41% of this dataset.
Object detection in common life	SDOD-MT [57]	They divide all the objects in SDOD-MT into three parts based on the height of a horizontal bounding box: small for range from 10 to 50, middle for range from 50 to 300, and large for range over 300.
	Small Object Dataset [62]	They pick out the smaller object categories (knife, fork, etc.) that are really small and easy to cluster from the MS-COCO dataset to compose the subset. The categories selected are knife, fork, bottle, wine glass, cup, spoon, bowl, sports ball, orange and vase.
	SOD [52]	Median of relative areas (the ratio of the bounding box area over the image area) of all the object instances in the same category is between 0.08% to 0.58%. This corresponds to 16×16 to 42×42 pixels areas in a VGA image.
	MS-COCO [40]	Objects occupying areas less than or equal to 32×32 pixels come under small objects category.
	PASCAL-VOC [41]	By counting the median of relative areas (the ratio of the bounding box area over the image area) of all the object instances in the same category in the PASCAL-VOC dataset, we define small objects as follows: the median of relative areas of object categories in the PASCAL-VOC dataset is less than 5%. Namely, bottle, car, potted plant, sheep, and boat are treated as small objects.
Others	Small Object Dataset [55]	According to the MS-COCO dataset, pixels smaller than 32×32 are defined as small objects.
	USC-GRAD-STDdb [53,73]	Objects occupying areas under 16×16 pixels are defined as small targets category.
	DeepScores [61]	There is a big variability in object size ranging from less than hundred to many thousands of pixels in area. It has large (more than 4 times larger than the average) images containing many very small (down to a few pixels, but vary by several orders of magnitude) objects that change their category belonging depending on the visual context.
	Lost and Found [66]	Small obstacles down to the height of 5cm can successfully be detected at 20m distance at low false positive rates.

interest (ROI) are computed from low-resolution images at first. The prior knowledge of the positions of objects is utilized to guide the generation of ROIs. Then the features of small ROIs are recomputed from high-resolution images, and the features of large ROIs are gained from the feature maps employed to generate ROIs. Furthermore, Pang et al. [78] propose a JSC-Net for small-scale pedestrian detection by jointly classification sub-network and super-resolution sub-network in a unified framework. The super-resolution sub-network aims to explore the relationship between large-scale pedestrian and small-scale pedestrian for recovering the detailed information of small-scale pedestrians from their large-scale counterparts. Based on HOG + LUV [79] and JCS-Net, multi-layer channel features (MCF) [80] are constructed to train the detector. Finally, multiscale representation is combined with MCF to enhance the detection performance further.

In addition to small traffic signs and small-scale pedestrian detection, some researchers also pay attention to small or tiny face detection. Bai et al. [81] take multi-branch fully convolutional network as baseline detector to crop regions containing faces or not, which are passed into the generator and discriminator respectively. The generator is trained to reconstruct a clear super-resolution face from the low-resolution input face, which includes the up-sample and refinement module. The non-faces regions are treated as negative data for training the discriminator that includes two fully connected layers to distinguish the real images or the generated super-resolution regions and to classify the faces

or non-faces, respectively. Subsequently, Liu et al. [82] put forward an algorithm to directly generate a clear high-resolution face from a blurry small one through using a GAN. Moreover, a prior information estimation network is devised to extract the facial image features, and estimate landmark heatmaps respectively. Via combining these two networks, the proposed end-to-end framework can both improve face resolution and detect the tiny faces. Focusing on the detection of small faces or common small objects, Zhang et al. [83] propose a new multi-task generative adversarial network, namely MTGAN. The generator is a super-resolution network which up-samples the small blurred images into fine-scale ones and recovers detailed information for more accurate detection. Unlike the generator, the discriminator is a multi-task network. It describes each super-resolved image patch with a real/fake score, object category scores, and regression offsets. The classification and regression losses in the discriminator are back-propagated into the generator during training process in order to make the generator gain more details for easier detection. Furthermore, Noh et al. [84] inspect existing small object detection approaches about feature-level super-resolution and discover the performance is significantly enhanced through using high-resolution target features as supervision signals and matching the relative receptive fields of input and target features. Thus, they put forward a new feature-level super-resolution (SR) model. As a GAN-based model, the SR feature generator learns to generate high-resolution features under the guidance of the SR feature discriminator utilizing the features from the SR target extractor as targets. During the inference, a large proposal is directly passed to the large predictor for localization and classification, while a small proposal is first super-resolved via the SR feature generator and then passed to the small predictor.

Focusing on small objects in common life, Chen et al. [52] compose a small object dataset by using a subset of images from both MS-COCO and SUN datasets. Krishna and Jawahar [72] train a CNN on the super-resolved train images like in [52]. The proposals in region proposal network are up-sampled and flow by a super-resolution network after which they are classified. By analyzing the factors that small object detection relies on and the trade-off between performance and efficiency, Liu et al. [85] propose a high-resolution detection network (HRDNet). The main idea of the HRDNet is to adopt a shallow backbone to deal with high-resolution images while employing a deep backbone to handle low-resolution images. The advantage of extracting features from high-resolution images with the tiny and shallow network has been demonstrated in [86]. HRDNet can not only gain more details for a small object in high-resolution, but also guarantee the effectiveness and efficiency via integrating multi-depth and multi-scale deep networks. Aiming at the limitation of bounding box prediction for small objects, Gu et al. [87] design a detection framework by generative and discriminative learning (GDL). A reconstruction generator network is first devised to reconstruct the mapping from low frequency to high frequency for anchor box prediction. Then a detector module extracts the ROIs from generated result and implements a RoI-Head to predict object class and refine bounding box. To guide the reconstructed image related to the corresponding one, a discriminator module is employed to tell from the generated result and the original image. Inspired by the positive impacts of super-resolution for object detection, Ji et al. [88] present a framework that can be incorporated with detection networks to improve small object detection performance, in which the low-resolution image is super-resolved by GAN in an unsupervised manner. The super-resolution network and the detection network are trained jointly. Particularly, the detection loss is back-propagated into the super-resolution network during training to facilitate detection. To accelerate the inference speed of feature-pyramid based object detectors, Yang et al. [89] propose a novel query mechanism, namely QueryDet. The locations (query keys) where small objects might exist are first predicted in the low-resolution features, and a sparse feature map (query values) is built employing high-resolution features in those locations. Lastly, a sparse detection head is utilized to output the detected boxes. This scheme is applied in a cascaded manner, enabling fast and accurate detection of small objects.

In addition to focusing on small objects in images, some researchers also explore small object detection in videos. Bosquet et al. [53] construct a video small object dataset USC-GRAD-STDdb. Meanwhile, they introduce a STDnet focused on the detection of small objects (under 16×16 pixels). The high performance of STDnet is rested on a new early visual attention mechanism, namely region context network (RCN), to select the most promising regions, while throwing away the rest of the input image. Handling only specific areas allows STDnet to keep high resolution feature maps in deeper layers providing low memory overhead and higher frame rates. High resolution feature maps are proved to be key to improving localization accuracy in such small objects. Based on STDnet [53], Bosquet et al. [90,91] propose a spatio-temporal neural network, namely STDnet-ST, to improve small object detection further. STDnet-ST can detect small objects over time and correlate pairs of the top-ranked regions with the highest likelihood of containing those small objects, which permits to link the small objects across the time as tubelets. To provide high quality tubelets and increase the precision, they also introduce a strategy to dismiss unprofitable object links. Moreover, Wang et al. [55] construct a small object dataset taken from a university classroom video record. Then, they present a new detection approach based on image super-resolution to improve the detection speed and accuracy of small objects. Specifically, they add a feature texture transfer module at the input end to increase the image resolution at this end as well as to remove the noise in the image. To reduce the number of network parameters, they adopt dense blocks to replace residual blocks. The neck integrates SPPnet [92] and PANet [93] to complete multi-scale feature fusion in order to make full use of the features of small objects in the image. Via adding the foreground and background balance loss function to the YOLOv4 [94] loss function part, the problem of image background and foreground imbalance is solved. The strengths and weaknesses of super-resolution techniques are shown in Table 4.

3.2.2. Context-based information

Since small objects themselves contain limited information, contextual cues play an essential role in small object detection. It is well known that visual objects usually occur in particular environments and sometimes coexist with other related objects. A typical example is that birds commonly fly in the sky. Before the prevalent of deep learning, Oliva and Torralba [95] demonstrate that the surrounding region of the small object can provide useful contextual information to help detect the target object. Although CNNs have been implicitly already learned contextual information from hierarchical feature representations with multiple levels of abstraction, there is still value in explicitly exploring contextual information (the relationship between small objects and other objects or background) in small object detection based on deep learning. Next, some deep learning-based detection methods using contextual information are described in detail.

Hu et al. [96] explore three facets of the problem in the context of finding small faces: the role of scale invariance, image resolution, and contextual reasoning. They make use of large local context in a scale-variant way. Namely, they define templates that utilize massively-large receptive fields (where 99% of the template extends beyond the object of interest). Meanwhile, they reveal that context is mostly useful for finding low-resolution faces. To this end, Zhang et al. [97] design an aggregation connection network (ACNet) which includes two significant parts: aggregation connection module and context module. Aggregation connection module can reduce the disappearance of features caused via image scaling. Context module can fully utilize the rich contextual cues without adding extra parameters. To address the hard face detection issue, Tang et al. [98] propose Pyramid-Box, a novel context-assisted single shot face detector. They improve the utilization of contextual information from three aspects by designing Pyramid-Anchors, introducing low-level FPN and building context-sensitive structure,

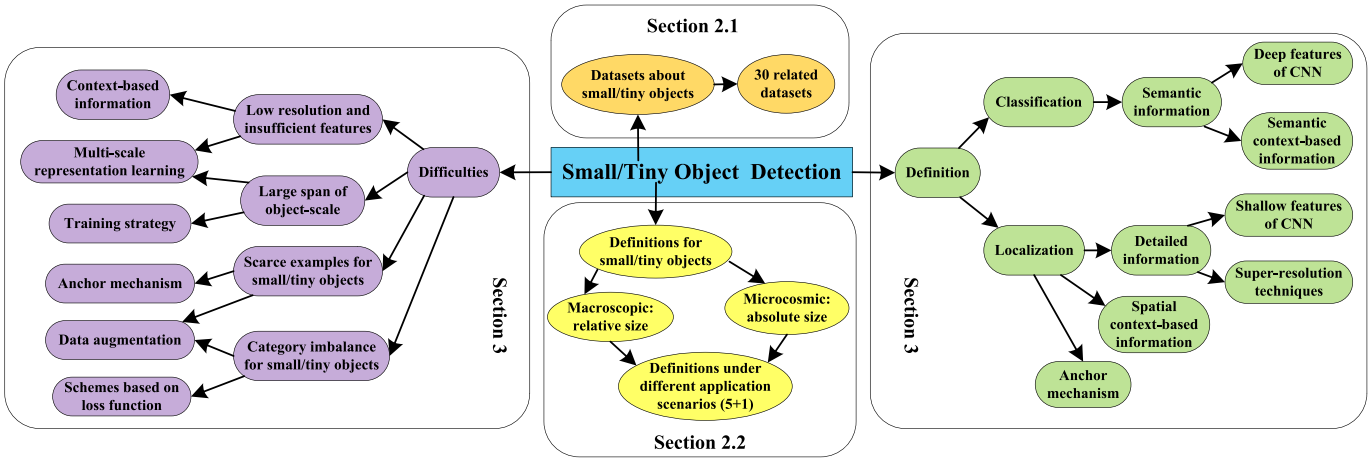

Fig. 1. An overview of small/tiny object detection. Datasets and definitions for small/tiny objects shown in Section 2 are also presented in this diagram.

respectively. Pyramid-Anchors can supervise high-level contextual feature learning via a semi-supervised method. The low-level FPN combines sufficient high-level context semantic features with low-level facial features, which allows the Pyramid-Box to predict faces of all scales in a single shot. Context-sensitive structure can increase the capacity of prediction network to improve the accuracy of final output. Subsequently, Li et al. [99] optimize each aspect in the Pyramid-Box [98] to further boost the detection performance of tiny faces, including balanced-data-anchor-sampling, dual-pyramid anchors and dense context module. Specifically, balanced-data-anchor-sampling gains more uniform sampling of faces with different sizes. Dual-pyramid anchors promote feature learning through introducing progressive anchor loss. Dense context module with dense connection not only enlarges receptive filed, but also efficiently transmits information. Unlike the above methods, Xi et al. [100] try to exploit the semantic similarity among all predicted objects in each image to promote current small face detectors. For this purpose, they propose a new framework to model semantic similarity as pairwise constraints within the metric learning strategy,

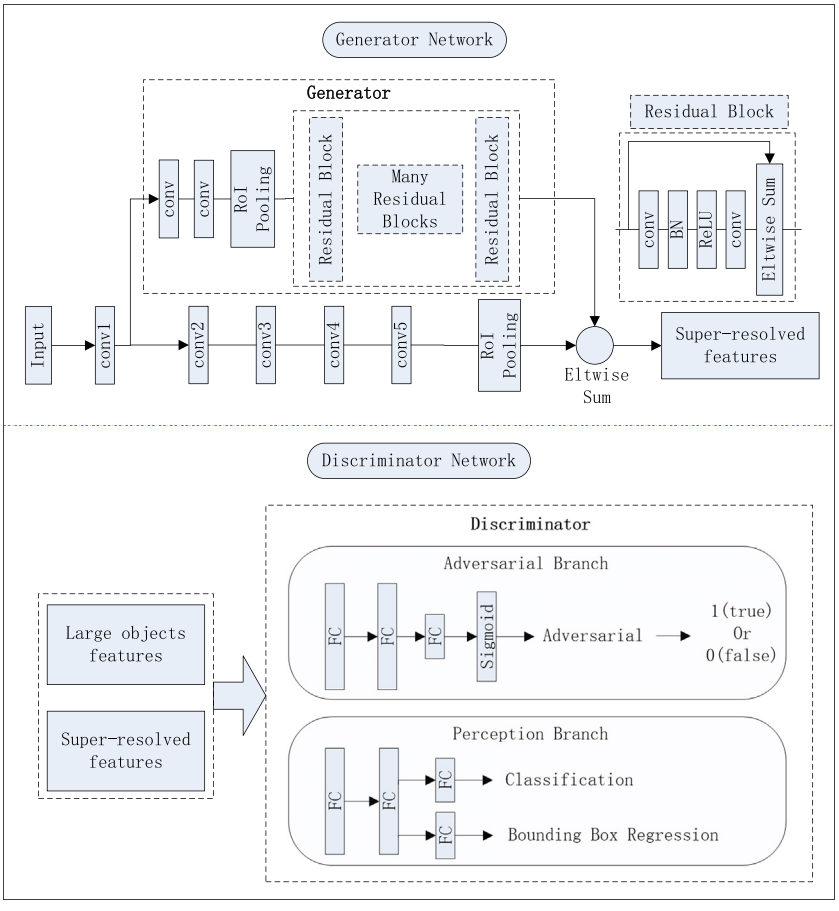

Fig. 2. Details of the perceptual GAN

and then improve their predictions with the semantic similarity via using the graph cut techniques. Later, Xi et al. [101] construct a pairwise constraint to depict the semantic similarity and develop a novel framework based on discriminative learning and graph-cut techniques. Experimental results on three widely used benchmark datasets reveal the effectiveness of the proposed method.

In order to take into account the small faces and small objects in common life at the same time, Leng et al. [102] develop a new internal-external network (IENet), which exploits both the appearance and context information of the object for robust detection. Aiming at feature extraction, proposal location and classification of small objects, they design three customized modules respectively: bidirectional feature fusion module (Bi-FFM), context reasoning module (CRM), and context feature augmentation module (CFAM). Specifically, Bi-FFM captures the internal feature of objects via transferring the semantic feature of deeper-level layers to lower-level layers and the detailed feature of lower-level layers to deeper-level layers in CNNs, which not only uses the hierarchy of convolutional features but also promotes its prediction by context relationships. CRM improves the quality of region proposals through context reasoning that utilizes easily detected objects to help understand hard ones. CFAM learns pair-wise relations between region proposals generated through CRM, and such relations are utilized to generate global feature information associated with the region proposals for accurate classification. Focusing on small object detection in common life, Fang et al. [62] employ a subset of the MS-COCO dataset to build a small object database and propose a more flexible context information integration approach based on Faster R-CNN. They crop eight corresponding context region enclosing the proposal region, including top-left, top-middle, top-right, middle-left, middle-right, bottom-left, bottom-middle and bottom-right. These eight contextual child-windows need to be discriminated whether they cross the boundary of the feature map. Namely, they add the valid context windows to the region proposals and only add the classification information rather than full information with both classification and regression information. In this way, the accuracy of small object detection can be improved to a certain extent. Fu et al. [103] present a novel context reasoning method for small object detection which models and infers the intrinsic semantic and spatial layout relationships between objects. Based on the initial regional features, they first design a semantic module to model the sparse semantic relationships. Also, they devise a spatial layout module to model the sparse spatial layout relationships based on their position and shape information. Both of them are then fed into a context reasoning module for integrating the contextual information, which is further fused with the original regional features for regression and classification. Experimental results show that the proposed scheme can effectively boost the performance of detecting small objects. Different from the above three schemes, Lim et al. [104] present a novel approach, which utilizes additional features from different layers as context via concatenating multi-scale features. Also, they introduce a detection method with attention mechanism which can focus on the object in image, and it can contain contextual information from target layer. Later, Yan et al. [105] propose a one-stage detector called LocalNet, which pays more attention to the detailed information modeling. The purpose of LocalNet is to preserve more detailed information in the early stage to enhance the representation of small objects. Besides, they devise a local detail-context module in order to heighten the semantics in the detection layers, which reintroduces the details lost in the network and exploits the local context within a restricted receptive field range.

Focusing on remote sensing small target detection, Liu et al. [106] construct a multi-component fusion network (MCFN) to improve the accuracy of small object detection in remote sensing imageries. They first design a dual pyramid fusion network, which densely concatenates spatial and semantic cues to extract features of small objects by encoding and decoding operations. Then they adopt a relative region proposal network to fully capture the features of small objects samples and parts of objects. Lastly they add contextual information to the proposal regions before final detection in order to obtain robustness against background disturbance. The inter feature maps and the feature map in different scales have different contribution to the network. To further strengthen the effective weights for detecting small objects in remote sensing images, Yang et al. [107] develop an inception parallel attention network, namely IPAN. It contains three parallel modules: multiscale attention module, contextual attention module, and channel attention module. The contextual attention module encodes a wide range of contextual cues into the local features, thus enhancing the representation capability. IPAN can extract not only rich multiscale, contextual features and the interdependencies of global features in different channels but also the long-range dependencies of the object to another based on the attention mechanism, which contributes to small object detection. Unlike the MCFN [106] and IPAN [107], Liang et al. [108] put forward a feature fusion and scaling-based SSD (FS-SSD) for small object detection in the UAV images. In addition to the deep features learned by the FS-SSD, spatial context analysis is proposed to further increase the detection precision via incorporating the object spatial relationships into object redetection. The interclass and intra-class distances between different object instances are computed as a spatial context, which proves effective for the detection of multiclass small objects. Subsequently, Cheng et al. [109] devise context feature enhancement module to exploit global context cues and selectively strengthen category-aware features via using image-level contextual information that indicates the presence or absence of an object class. Also, they leverage context encoding loss to regularize the model training which promotes the object detector to understand the scene better and narrows the probable object categories in prediction.

In addition to small face detection, the detection of small objects in daily life, and small target detection in remote sensing, context-based information is also used for small object detection in other domains. Bosquet et al. [73] develop a fully convolutional network focused on small targets in the videos. This network includes an early visual attention mechanism, named region context network (RCN), to select the most promising regions with one or more small objects together with their context and return them as a set of disjoint regions. The filtered feature maps, which only contain the most likely regions with small objects, are forwarded through the network up to an ending region proposal network (RPN) that feeds a final classification stage. RCN is key to improve localization precision via finer spatial resolution because of finer global effective strides, higher frame rates and low memory overhead. Fine-grained features in CNN are crucial for the network to precisely locate small or partially occluded objects. To alleviate the problem that deep CNN is hard to extract sufficient distinguishing

Table 4
Summarization of strengths and weaknesses of super-resolution techniques

Method	Advantage	Drawback
Perceptual GAN [76] PKG [77] JSC-Net [78] FaceGAN [81] TFPGAN [82] MTGAN [83] Feature SR [84] ISOD [72] HRDNet [85] GDL [87] Simultaneous SR [88] QueryDet [89] STDnet [53] STDnet-ST [91] ImageSR [55]	GAN-based approach effectively enhances the detail information of image, especially for the super-resolution application. In principle, it can be applied to any kind of generator without devising specific architecture. To some extent, it promotes small or tiny object detection.	Compared with general CNNs, GAN-based method is difficult to train. Namely, it is hard to keep a good balance between the generators and discriminators. Besides, the rewards of samples produced by the generator during the training process are limited, which will affect the further improvement of detection performance to a certain extent.

fine-grained features for high-level feature maps, Guan et al. [110] first build larger and more meaningful feature maps in top-down order and concatenate them and subsequently fuse multilevel contextual information via pyramid pooling to construct context aware features. Thus, a unified framework called the semantic context aware network (SCAN) is realized to enhance object detection accuracy. Via constructing high-resolution and strong semantic feature maps, Cui et al. [111] design a context-aware block network (CABNet) to improve the detection performance of small objects, such as traffic signs and small heads. To internally enhance the representation capability of feature maps with high spatial resolution, they devise the context-aware block (CAB) which leverages pyramidal dilated convolutions to incorporate multilevel contextual information without losing the original resolution of feature maps. Then, they assemble CAB to the end of the truncated backbone network with a relatively small down-sampling factor and cast off all following layers. Focusing on tiny person detection, Hong et al. [112] propose a scale selection pyramid network (SSPNet) that consists of three modules, including context attention module (CAM), scale enhancement module (SEM), and scale selection module (SSM). CAM considers context information to generate hierarchical attention heat-maps. SEM highlights features of specific scales at different layers, leading the detector to focus on objects of specific scales rather than vast backgrounds. SSM exploits adjacent layers' relationships to achieve suitable feature sharing between shallow and deep layers, thus avoiding the inconsistency in gradient computation across different layers. In Table 5, we summarize the advantages and disadvantages of the above context-based schemes.

3.2.3. Multi-scale representation learning

Multi-scale representation learning is also a significant and effective strategy for small or tiny object detection. First, we briefly show four classic methods in Fig. 3, including featurized image pyramids, single feature map, pyramidal feature hierarchy and feature pyramid network. Then, some improved approaches based on them are also illustrated in detail.

As shown in Fig. 3(a), previous object detectors usually adopt featurized image pyramids to detect objects at various scales. Namely, they resize input images into many different scales and to learn multiple detectors, each of which is in charge of a certain range of scales. With the development of deep learning, Liu et al. [113] propose an image pyramid guidance network (IPGNet) to make sure both the spatial and semantic information are abundant for each layer. It consists of two main modules: the IPG transformation and fusion module. Even in the deepest stage of the backbone, the IPG transformation module remains enough spatial information for bounding box regression and classification. Besides, IPG fusion module fuses the features from the image pyramid and backbone features. The main idea of proposed scheme is to introduce the IPG into the backbone network to handle the information imbalance problem, which alleviates the disappearance of the small object features. However, this method is computationally expensive due to rapid increase of memory consumption and inference time. Liu et al. [85] analyze the factors that small object detection relies on and the trade-off between efficiency and performance, as well as put forward a high-resolution detection network (HRDNet). It includes an important part, namely multi-depth image pyramid network (MD-IPN). MD-IPN maintains multiple position information utilizing multiple depth backbones. Through extracting various features from high to low resolutions, it can improve the small object detection performance as well as maintaining the detection performance of middle and large objects. To reduce the information imbalance between these features, a multi-scale feature pyramid network is also proposed to align and fuse multi-scale feature groups produced by MD-IPN. HRDNet can not only gain more details for a small object in high-resolution, but also guarantee the effectiveness and efficiency via integrating multi-depth and multi-scale networks.

Some detection schemes (such as Faster R-CNN [37] and R-FCN [114]) use 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. 3(b)). However, there is little information left for small objects on the top-most layers due to continuous down-sampling. This may compromise the detection performance of small objects. To this end, Liu et al. [45] propose a single shot detector named SSD, which detects objects with different scales and aspect ratios from multiple layers. Then, predictions are made from multiple layers, where each layer is in charge of a certain scale of objects. Actually, the in-network feature hierarchy in a deep CNN generates feature maps of different spatial resolutions while introduces large semantic gaps caused via different depths (see Fig. 3(c)). Namely, deep CNNs learn hierarchical features in different layers which capture information from different scale objects. Thus, SSD utilizes the features from the shallower layers to detect smaller objects, while exploits the features from the deeper layers for bigger objects detection. Afterwards, some SSD-based algorithms have been proposed.

Cao et al. [115] put forward a multi-level feature fusion approach for introducing contextual clues in SSD, namely feature-fused SSD (FFSSD). Furthermore, they adopt concatenation and element-sum module in fusion stage. To further improve the detection precision of small objects, Xu et al. [116] propose a multi-scale deconvolutional SSD named MDSSD. They inject the high-level features with semantic information to the low-level features through multi-scale deconvolution fusion module to obtain the feature maps with rich information. Especially, they implement deconvolution layers on multi-scale features before the top-most layer and merge them with some of the bottom features to generate more semantic feature maps. Also, they add conv3_3 output by the backbone network for prediction so as to improve the detection performance of CNNs for small objects. Different from these two architectures, Sun et al. [117] propose a mask-guided SSD to improve the performance of SSD for detecting small objects by boosting features with contextual information and introducing a segmentation mask to eliminate background regions. It consists of detection branch and segmentation branch. They construct a feature fusion module to allow the detection branch to exploit contextual information for feature maps with large resolution. The segmentation branch mainly built with dilated convolution to provide additional contextual cues to the detection branch. Moreover, segmentation features are applied to produce the mask, which is used to guide the detection branch to find objects in potential foreground regions.

Table 5
Summary of strengths and weaknesses of context-based schemes

Method	Advantage	Shortcoming
Pyramid-Box [98] Pyramid-Box++ [99] HR [96] ACNet [97] Xi et al. [100] BC-ESS [101] IENet [102] Fang et al. [62] IR RCNN [103] FA-SSD [104] LocalNet [105] MCFN [106] IPAN [107] FS-SSD [108] RCN [73] SCAN [110] CABNet [111] SSPNet [112]	Contextual information is intended to provide more information to the final detection network. These approaches based on contextual information make full use of the information related to the small or tiny object in the image, and can effectively improve the performance of detecting small objects.	These methods are mainly to gain the cues around the ROI area and improve object classification through learning the relationship between objects and surrounding information. Nevertheless, not all context-based information is helpful for detection. Redundant contextual information also causes information noise, which hurt the performance on small objects. On the other hand, these approaches do not consider the possible lack of contextual information in the scene, nor do they make targeted use of the easy-to-detect results in the scene to assist the detection of small objects.

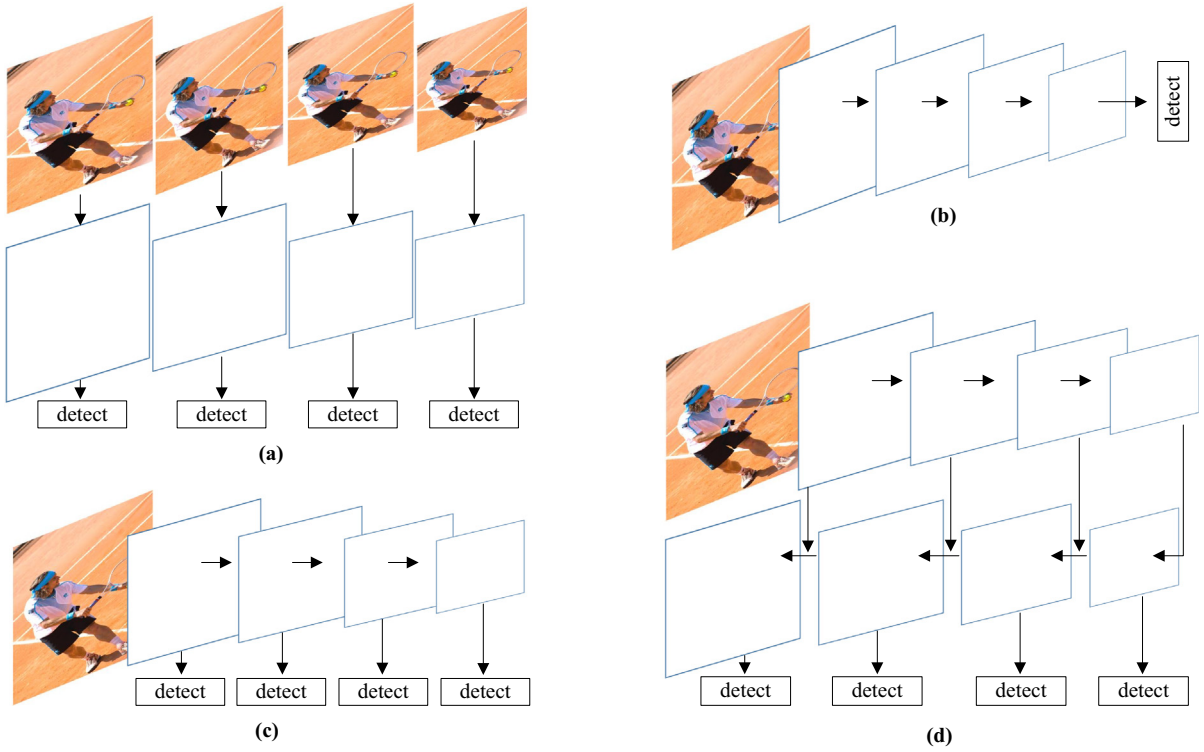

Fig. 3. Four approaches for multi-scale representation learning. (a) Featurized image pyramids (b) Single feature map (c) Pyramidal feature hierarchy (d) Feature pyramid network

Inspired by the structure of receptive fields (RFs) in human visual systems, Liu et al. [118] present a novel RF block (RFB), which takes the relationship between the size and eccentricity of RFs into account, to enhance the feature discriminability and robustness. By assembling RFB to the top of SSD, the RFBNet detector is constructed. RFBNet can reach the performance of advanced very deep detectors while keeping the real-time speed. Similar to RFBNet, Li et al. [119] develop a novel trident network (TridentNet) aiming to produce scale-specific feature maps with a uniform representational power. They build a parallel multi-branch structure in which each branch shares the same parameters but with different receptive fields. To improve the detection of small objects in images captured from a UAV, Razaak et al. [120] present a multi-scale approach of low-level feature combinations with deconvolutional modules on SSD object detector. Later, Han et al. [121] introduce a D-RFB module which can enhance the representation ability of the feature map. D-RFBNet integrated with D-RFB module can detect small objects in the UAV imagery more accurately.

To combine the advantage between single feature map and pyramidal feature hierarchy, Lin et al. [122] present feature pyramid network (FPN). It constructs a top-down architecture with lateral connections to yield a series of scale-invariant feature maps, and learns multiple scale-dependent classifiers on these feature pyramids (see Fig. 3(d)). Specifically, the shallow spatially-rich features are strengthened via the deep semantic-rich features. These lateral and top-down features are integrated through concatenation or element-wise summation. Subsequently, many variants of FPN are proposed. Compared with conventional detectors, these approaches reveal dramatic improvements in detection accuracy with some modifications to the feature pyramid block.

To remedy the shortcomings of current methods in detecting remote sensing small targets, Qu et al. [123] propose a dilated convolution and feature fusion detector named DFSSD. This detector exploits the structure of FPN to fuse the low-level feature map with high resolution and the high-level feature map with rich semantic information. It also boosts the receptive field of the third-level feature map of the DFSSD network by adopting dilated convolution. Later, Liang et al. [108] present a feature fusion and scaling-based SSD detector named FS-SSD for small object detection in the UAV images. They add an extra scaling branch in the deconvolution module with an average pooling operation to form a feature pyramid. Then the original feature fusion branch is adjusted to be suitable for the small object detection task better. Finally, these two feature pyramids produced by the deconvolution module and feature fusion module are used to make predictions together. Unlike the DFSSD and FS-SSD, Liu et al. [106] build a multi-component fusion network (MCFN) which contains three main parts: dual pyramid fusion network (DPFN), relative region proposal network (RRPN) and contextual information network (CIN). Specifically, DPFN densely concatenates spatial and semantic information to extract features of small objects through encoding and decoding operations. RRPN can fully capture the features of small objects samples and parts of objects. CIN is able to achieve robustness against background disturbance. Different from the multi-component fusion network (MCFN), Liu et al. [124] construct a multi-branch parallel feature pyramid network (MPFPN) to extract more abundant feature information of the objects with a small size in UAV-captured images. The parallel branch aims at recovering the features that missed in the deeper layers. To weaken the impact of background noise inference and focus object information, the supervised spatial attention module is applied in MPFPN. Besides, they use cascade architecture in the Fast R-CNN stage for a more powerful localization capability. In addition to the above methods, there are also some approaches [54,125] that try to use a bi-directional strategy. Zheng et al. [125] propose a bi-directional stepped concatenation feature pyramid approach. The stepped concatenation tactic is conductive to avoid the loss of information at the current layer during the pyramid construction process, and the bi-directional scheme ensures the fusion features contain both details and semantic information. Additionally, an attentional interaction module is devised to better aggregate dual-stream features for improving network performance. Similarly, Li et al. [54] put forward

a cross-layer attention network (CANet). They first devise an up-sampling and down-sampling feature pyramid to gain richer context information through bi-directionally fusing deep and shallow features, as well as skipping connections. Then, a cross-layer attention module is proposed to capture the non-local association of small objects in each layer, and further enhance its representation ability by cross-layer integration and balance. In order to solve the problem that the top-down operation in the feature pyramid would cause the mutual influence of different levels of features, Cheng et al. [126] propose aware feature pyramid network (AFPN). AFPN learns a vector for the higher level features in the feature pyramid to obtain clean features. Moreover, positive and negative labels can be better assigned by utilizing the new group assignment strategy.

In addition to the detection of small targets in remote sensing images, small object detection in common life has also attracted much attention from researchers. Liang et al. [127] develop a novel detector named deep feature pyramid networks (DFPN). Adopting the feature pyramid architecture with lateral connections in DFPN makes the semantic feature of small objects more sensitive. Besides, they design specialized anchors to detect the small objects from large resolution image, and then train the network with focal loss. To mitigate the imbalance at sample level, feature level and objective level in training for object detection, Pang et al. [128] propose an effective framework dubbed Libra R-CNN which integrates three new components: IoU-balanced sampling, balanced feature pyramid, and balanced L1 loss, respectively. Benefitted from the overall balanced design, Libra R-CNN significantly improves the detection performance of small objects. Unlike these two approaches, Zheng et al. [129] propose an interactive multi-scale feature representation enhancement (IMFRE) strategy, which includes two parts: multi-scale auxiliary enhancement network (MAEN) and adaptive interaction module (AIM). MAEN is proposed for feature interaction under multiple inputs. They scale the input to multiple scales corresponding to the prediction layers, and only pass through the lightweight module to extract more detailed features for enhancing the original features. AIM is devised to aggregate the features of adjacent layers. This method offers flexibility in achieving the improvement of small objects detection without changing the original network structure. Considering the detection of small objects on real-time embedded devices, Chen et al. [130] put forward RHF-Net, a novel recursive hybrid fusion pyramid network. The proposed RHFNet has two novelties: namely the bidirectional fusion module, the recursive concatenation and reshaping module. The former can fuse feature maps with both the top-down and bottom-up directions to produce flexible feature pyramids for small object detection. The latter can recursively concatenate not only high-level semantic features from deep layers but also reshape spatially richer features from shallower layers to prevent small objects from vanishing. RHFNet employs computationally low-cost and feature preserving operations in the fusion, hence it is efficient and accurate on embedded devices.

To detect tiny persons in a broad horizon and massive background, Liu et al. [131] devise a feature rescaling and fusion (SFRF) approach. By designing a nonparametric adaptive dense perceiving algorithm, it can automatically choose and produce a new resized feature map with the high density distribution of tiny objects. To enhance the feature representation, they also use a many-for-one strategy for feature fusion of the FPN layers. Gong et al. [132] argue that the top-down connections between adjacent layers in FPN bring a negative impact for tiny object detection. They introduce a novel concept called fusion factor to control the information transmission from deep layers to shallow layers, for adapting FPN to tiny object detection. Also, they explore how to estimate an effective value of fusion factor for a particular dataset via a statistical strategy. After series of experiments and analysis, they find that the estimation is dependent on the number of objects distributed in each layer. When configuring FPN with a proper fusion factor, the network can achieve significant detection performance on tiny objects. Although most existing approaches use FPN to enrich shallow layers' features via combing deep layers' contextual features, the shallow layers in FPN are not fully exploited to detect tiny objects due to the limitation of the inconsistency in gradient computation across different layers. To this end, Hong et al. [112] propose a scale selection pyramid network (SSPNet) for tiny person detection. It consists of three parts, including context attention module (CAM), scale enhancement module (SEM), and scale selection module (SSM). CAM considers context information to generate hierarchical attention heat-maps. SEM highlights features of specific scales at different layers, leading the detector to focus on objects of specific scales rather than vast backgrounds. SSM exploits adjacent layers' relationships to achieve suitable feature sharing between shallow and deep layers, thus avoiding the inconsistency in gradient computation across different layers. The strengths and weaknesses of the above multi-scale representation learning methods are presented in Table 6.

3.2.4. Anchor mechanism

In this section, we discuss anchor mechanism for small object detection. It is mainly divided into anchor-based mechanism and anchor-free mechanism.

The anchor is widely adopted by most of the existing detectors. Faster R-CNN [37] introduces the Region Proposal Network (RPN) to generate proposals. The RPN is based on anchors, which are predefined regions of different sizes and aspect ratios to handle multiple scales. The RPN produces the coordinates of the bounding boxes and their corresponding categories, namely object and background. Finally, given the output of the RPN and the last feature map of the feature extractor, the bounding box and category of the object are determined by a fully-connected classification network. Later, Dai et al. [114] propose a

Table 6
Summarization of strengths and weaknesses of multi-scale representation learning methods

Method		Strength	Weakness
IPG RCNN [113] HRDNet [85]	Featurized image pyramids	These methods can transform objects of all scales equally.	These approaches are inefficient.
Faster R-CNN [37] R-FCN [114]	Single feature map	These works use only the outputs of the last conv layer for faster detection.	There is little information left on the top-most layers for small objects.
SSD [45] FFSSD [115] MDSSD [116] Mask-guided SSD [117] RFBNet [118] Razaak et al. [120] D-RFBNet [121] TridentNet [119]	Pyramidal feature hierarchy (such as SSD, and some SSD-based approaches)	Deep CNNs learn hierarchical features in different layers which capture information from different scale objects. To a certain extent, these algorithms are conductive to small or tiny object detection.	The in-network feature hierarchy in a deep CNN introduces large semantic gaps caused by different depths. These works adopt some strategies to fuse semantic and contextual information. However, it is far from enough to explore the correlation of features at different scales.
FPN [122] MCFN [106] DFSSD [123] MPFPN [124] FS-SSD [108] DFPN [127] Libra R-CNN [128] RHFNet [130] IMFRE [129] SFRF [131] S-α [132] BSCF [125] CANet [54] SSPNet [112]	Feature pyramid network (In addition to FPN, some variants of FPN are includEd.)	Compared with conventional detectors, these approaches reveal dramatic improvements in small/tiny object detection accuracy with some modifications to the feature pyramid block.	The feature pyramid produces multi-level features thus sacrificing the feature consistency across different scales. This will lead to a decrease in effective training data and a higher risk of overfitting for each scale. Besides, the shallow layers in FPN are not fully exploited to detect tiny objects due to the limitation of the inconsistency in gradient computation across different layers.

region-based fully convolutional network (R-FCN) to generate $k \times k \times (C+1)$ feature maps instead of single feature map, where each map is responsible for each category detection. Nevertheless, both Faster R-CNN and R-FCN are not adequate for small object detection because of the larger sizes of the predefined anchors and the coarse global effective stride. To this end, Krishna and Jawahar [72] formulate finding that appropriate sizes of the anchor boxes mathematically and perform detailed experiments to reveal the effectiveness in their choice. By introducing the expected max overlapping (EMO) score, the authors in [133] calculate the expected max intersection over union (IoU) between anchor and object. They find the smaller stride of the anchor (SA) is, the higher EMO score achieves, statistically leading to improved average max IoU of all objects. It is noting that a smaller SA can sample more high-quality samples well capturing the small objects, which is of help for both detector training and testing. Based on the above analysis, Yang et al. [134] design a finer sampling and feature fusion network (SF-Net). In the anchor-based detection framework, the value of SA is equal to the reduction factor of the feature map relative to the original image. Subsequently, Cascade R-CNN [135] extends the Faster R-CNN to address the problems of over-fitting and quality mismatch. Wang et al. [136] devise a cascade mask generation framework, which takes in multi-scale images as input and processes them in ascending order of the scale. Then, each scale generates region proposal and mask via the mask generation module inspired by RoI convolution. Last, the feature maps from each scale are concatenated for the RoI pooling and post-detection.

Moreover, the anchor strategy is also used in small company logo detection [137], small-scale pedestrian detection [138], and small face detection [139–141]. Eggert et al. [137] derive a relationship which describes the minimum object size which can reasonably be proposed and provide a heuristic to select appropriate anchor scales for small company logos. Zhang et al. [138] propose an asymmetric multi-stage network (AMS-Net), which considers the asymmetry of a pedestrian's body shape in small-scale pedestrian detection. The rectangular anchors are utilized to produce various rectangular proposals that have a height greater than the width. Besides, asymmetric rectangular convolution kernels are adopted to capture the compact features for the pedestrian body. Focusing on small face detection, Zhang et al. [139] propose a new anchor densification strategy to make different kinds of anchors have the same density on the image, which significantly improves the recall rate of small faces. Later, Zhang et al. [140] tile anchors on a wide range of layers to ensure that all scales of faces have enough features for detection. Based on the effective receptive field and an equal proportion interval principle, they devise anchor scales. They also improve the recall rate of small faces via a scale compensation anchor matching strategy and reduce the false positive rate of small faces by a max-out background label. While anchor-based face detectors have achieved promising performance, they treat all faces equally and ignore the imbalance between easy faces and hard faces. Zhang et al. [141] develop an anchor-based face detector, which only outputs a single high-resolution feature map with small anchors, to specifically learn small faces and train it via a new hard image mining scheme which automatically adjusts training weights on images according to their difficulties.

It is necessary to tile massive dense anchors on high-resolution feature map for pursing high recall. However, it results in an extreme imbalance of category that drastically impacts the classification task in detection framework. An adaptive anchor tiling strategy, like MetaAnchor [142] and Guided Anchor [143], is proposed to shrink search space efficiently. Specifically, Yang et al. [142] present a novel anchor mechanism called MetaAnchor for object detection. Unlike many previous detectors model anchors by a predefined manner, anchor functions in the MetaAnchor could be dynamically produced from the arbitrary customized prior boxes. In this way, they empirically find that MetaAnchor is more robust to anchor settings and bounding box distributions. Wang et al. [143] present a scheme named Guided Anchoring, which uses semantic features to guide the anchoring. The proposed approach jointly predicts the locations where the center of objects of interest are likely to exist as well as the scales and aspect ratios at different locations. On top of predicted anchor shapes, they mitigate the feature inconsistency with a feature adaption module. This anchoring method can be seamlessly integrated into proposal methods and detectors. To perform more robust anchor matching, Duan et al. [144] propose a novel CPT-Matching strategy based on the center point translation of anchors to select more extending anchors as positive samples in the training phase. They first match the predicted boxes of the multi-RPN to any ground-truth boxes with an IoU higher than 0.5 and label their corresponding anchors. Then they further choose nearer anchors based on the threshold that represents the ratio of the center point distance between an anchor and a ground-truth box to the scale of the anchor. Meanwhile, the most negative anchors are removed. Finally, they translate the center point of the rest of anchors to the center of the nearest ground-truth boxes. If their IoUs are higher than the threshold, they are considered as positive examples. This strategy can help for accurate small object detection.

In addition to the anchor-based methods, some researchers discard the prior anchors and they utilize anchor-free approaches for object detection. Law et al. [145] propose CornerNet, a new method for object detection. They detect an object bounding box as a paired keypoints, namely the top-left corner and the bottom-right corner, using a single CNN. They also introduce corner pooling to help the network localize corners better. Unlike CornerNet, Lu et al. [146] introduce a new detection framework called Grid R-CNN, which employs a grid guided localization mechanism for accurate object detection. Instead of utilizing only two independent points, they devise a multi-point supervision formulation to encode more clues so as to reduce the influence of inaccurate prediction of specific points. To fully exploit the correlation of points in a grid, they present a two-stage fusion strategy to fuse feature maps of neighbor grid points. Grid R-CNN can lead to high quality object localization. Keypoint-based CornerNet achieves high accuracy among single-stage detectors yet comes at high processing cost. Law et al. [147] tackle this problem and introduce CornerNet-Lite. CornerNet-Lite is a combination of two variants based on CornerNet: CornerNet-Saccade and CornerNet-Squeeze. By combing these two efficient variants, it makes a balance between real-time efficiency and high accuracy. Based on CornerNet, Duan et al. [148] construct a framework named CenterNet. It detects each object as a triplet rather than a pair of keypoints, which improves both precision and recall. To this end, they devise two customized modules: cascade corner pooling and center pooling. These modules enrich information captured by both the top-left and bottom-right corners and provide more recognizable information from the central regions. Later, Zhou et al. [149] detect four extreme points (top-most, left-most, bottom-most, right-most) and one center point of objects utilizing a standard keypoint estimation network. The proposed ExtremeNet group the five keypoints into a bounding box if they are geometrically aligned.

Different from the above methods, Zhou et al. [150] model an object as a single point, namely the center point of its bounding box. The proposed detector CenterNet adopts keypoint estimation to find center points and regresses to all other object properties, such as size, location and orientation. Later, Wang et al. [48] present an detector named M-CenterNet, which uses multiple center points to locate accurate object center for improving the localization performance of tiny object detection in aerial images. Yang et al. [151] propose RepPoints (representative points), a novel finer representation of objects as a set of sample points useful for both localization and recognition. Given ground truth localization and recognition targets for training, RepPoints learn to automatically arrange themselves in a manner that bounds the spatial extent of an object and indicates semantically significant local areas. Besides, they do not use anchors to sample a space of bounding boxes. Unlike CenterNet [150] and RepPoints [151], Tian et al. [152] put forward a fully convolutional one-stage detector (FCOS) for object detection in a per-pixel prediction way. Through eliminating the predefined

set of anchor boxes, FCOS completely avoids the complicated computation related to anchor boxes such as calculating overlapping during training. Also, it avoids all hyper-parameters about anchor boxes, which are often very sensitive to the final detection performance. Similarly, Kong et al. [153] present FoveaBox for object detection, which directly learns the object existing possibility and the bounding box coordinates without anchor reference. Specifically, FoveaBox is achieved through predicting category-sensitive semantic maps for the object existing possibility, and generating category-agnostic bounding box for each position that potentially contains an object. Besides, an instance is assigned to adjacent feature levels to make the model more accurate. The merits and demerits of anchor-based and anchor-free mechanism are exhibited in Table 7.

3.2.5. Training strategy

Nowadays, object detectors based on deep CNN can benefit from multi-scale training and testing, where images are randomly resized into different resolutions. While training detectors for large objects is straightforward, the crucial challenge remains training detectors for small objects. Hu et al. [96] train separate face detectors for different scales. To keep the efficiency, face detectors are trained in a multi-task way: they utilize features extracted from multiple layers of single (deep) feature hierarchy. To detect small faces better, Luo et al. [154] propose a novel small faces attention (SFA) detector. Specifically, multi-branch face detection architecture is first devised to pay more attention for faces with small scale. Then, the feature maps of adjacent branches are merged, so that the features from the large scale can assist in the detection of hard faces with small scale. Finally, they simultaneously use multi-scale training and testing to make SFA robust to various scales. Thus, features learned in this way are more robust to scale variation. Multi-scale training strategy is not only adopted in small-scale face detection, but also employed in detecting tiny objects. Gao et al. [155] use multi-scale training method to improve the performance of tiny object detection. Specifically, the scale of shorter side is randomly sampled from 832, 896, 960, 1024, 1088, 1152, 1216, 1280, 1344, 1408, 1472, 1500 and the longer edge is fixed to 2000 in PaddleDetection[2]. In MMDetection [156], the scale of shorter side is randomly sampled from 480, 528, 576, 624, 672, 720, 768, 816, 912, 960 and the longer edge is fixed to 1280. During training, the number of proposals before NMS [157] is changed from 2000 to 12000. And the data is changed to 6000 in testing stage. Feng et al. [158] randomly rescale input images to 0.5, 1, 1.5 times of the original size while training to help address the scale variance issue. Besides, considering the objects in TinyPerson are extremely small, based on the multi-scale training strategy, they up-sample the training images for higher resolution and obtain better results. Nevertheless, this approach has its bottleneck. Up-sampling training images to larger scales helps detector extract more detailed features, but introduces more distortion at the same time. The model implemented by 1x training schedule does not converge well, thus they choose 2x training schedule to train the model and obtain better results. To a certain extent, the above training strategies are conductive to small face detection and tiny object detection.

To normalize the scales of objects in multi-scale training, Singh and Davis put forward a novel training strategy called scale normalization for image pyramids (SNIP) [159] which selectively back-propagates the gradients of object instances of different sizes. In training stage, they merely select ground truth boxes and proposals which fall in a specified size range at a particular resolution. Similarly, during testing, they produce proposals using RPN for each resolution and classify them independently at each resolution. Meanwhile, they only pick detections which fall in a specified range at each resolution. Later, Singh et al. present another approach for efficient multi-scale training, namely SNIPER [160]. It uses patches as training data instead of regular images. SNIPER crops selected regions around the ground-truth instances as positive chips and samples background as negative chips. It samples low resolution regions from a multi-scale image pyramid to accelerate multi-scale training. SNIP and SNIPER can effectively improve the detection performance of small objects. Furthermore, Kim et al. develop a scale-aware network called SAN [161], and introduce a new learning scheme which considers purely the relationship between channels without the spatial information. To make CNN-based detectors more robust to the scale variation, SAN maps the convolutional features gained from the different scales onto a scale-invariant subspace. It first extracts convolutional features from scale normalized patches. Then SAN and detection network are trained simultaneously via using these extracted features. Different from the above three approaches, Zhou et al. [162] propose montage pre-training, a general pre-training paradigm for object detection. Compared to the widely used ImageNet pre-training, montage pre-training needs only the target detection dataset while taking only one quarter of computational resources. Specifically, they reduce the potential redundancy via carefully extracting useful samples from the original images, assembling samples in a montage manner as input, and utilizing an ERF-adaptive dense classification scheme for model pre-training. These designs largely consider the network utilization and improve the learning efficiency and final performance. In addition to multi-scale training and montage pre-training strategies, Chen et al. [163] introduce a collage fashion of down-scaled images, and propose a dynamic scale training (DST) strategy to mitigate scale variation challenge in object detection. They utilize feedback information from the optimization process to dynamically guide the data preparation. In each training iteration, they fetch the loss proportion owing to small objects as feedback. It could be calculated after each forward propagation

Table 7
Summary of strengths and weaknesses of anchor-based and anchor-free mechanism

Method	Superiority	Drawback
ISOD [72] SCRDet [134] Cascade R-CNN [135] CasMaskGF [136] Eggert et al. [137] AMS-Net [138] FaceBoxes [139] S3FD [140] LSFHI [141] MetaAnchor [142] Guided anchoring [143] CPT-Matching [144]	A well-devised region proposal strategy can take advantage of limited anchor size and anchor amount, reduce computational cost in producing interested region, and efficiently detect small objects. Hence, these well-designed methods adopt predefined anchors to enumerate possible locations, scales and aspect ratios for the search of the objects, which are conducive to the detection of small objects to a certain extent.	The usage of anchor introduces a large number of hyper-parameters, which makes the network hard to train. Besides, improper usage of anchor could cause imbalance between the positive and negative samples of small objects, which makes the model pay more attention to large objects. This is not beneficial for the detection improvement of small objects. The existing anchor design is difficult to substantially balance the contradiction between the detection accuracy and the calculation cost for small objects.
CornerNet [145] Grid R-CNN [146] CornerNet-Lite [147] Centernet-KT [148] ExtremeNet [149] CenterNet [150] M-CenterNet [48] RepPoints [151] FCOS [152] FoveaBox [153]	These approaches are anchor box free, as well as proposal free. By eliminating the predefined set of anchor boxes, these methods completely avoid the complicated computation related to anchor boxes such as calculating overlapping during training. Also, they avoid all hyper-parameters about anchor boxes. Thus, these schemes are relatively easy to train. Moreover, these works are more efficient and effective for the final detection performance.	In object detection, methods based on key points, center points, corner points and their improvements usually encounter the drawback of a large number of incorrect object bounding boxes, arguably due to the lack of an additional evaluation inside cropped regions.

[2] https://github.com/PaddlePaddle/PaddleDetection

during model training. The proposed strategy is simple yet obtains significant benefits, outperforming previous schemes. Table 8 shows the advantages and disadvantages of the above training strategies.

3.2.6. Data augmentation

Data are at the core of any deep learning model. Insufficient training samples are every so often answerable for deprived performances in deep learning solutions to problems. Thus, using a large amount of data for training is vital for good performance of any deep learning model [164]. Data augmentation is a technique that can be utilized to extend the size of dataset required by deep learning model for training through artificially producing variations of existing actual images in the dataset. It can be broadly divided into four categories: geometric transformations (e.g. scaling, rotating, flipping, cropping, padding etc.), color transformations (such as changing the contrast, brightness, hue, saturation, noise in an image), random occlusion (for instance random cutout, erase, hide and seek, grid mask, cutmix [165] and mosaic augmentation etc.) and schemes based on deep learning. Gao et al. [155] augment training data by random horizontal flip with the probability of 0.5. Moreover, they also adopt random cropping and expanding. Similarly, Zhou et al. [162] crops foreground patches to construct jigsaw assembly for upstream classification. YOLOv4 [94] uses the mosaic augmentation for small object detection in images for the first time. They join four images into one single image. Consequently, the objects in the joined image appear at a smaller scale than the original image. This kind of augmentation is conductive to ameliorating the detection of small objects in images. Geometric transformations, color transformations, and random occlusion these three kind of augmentations do not always capture all the disparities in the environments. Also there are chances of lost information or features of the original dataset because these techniques attempt to change the geometry or lighting conditions of the images [164].

Nowadays data augmentation approaches based on deep learning are providing more convincing proofs. Compared with classification task, it is more difficult to design plausible augmentation schemes for object detection. Zoph et al. [166] try to adopt AutoAugment [167] to devise better data augmentation strategies for object detection. They also employ this strategy to evaluate the value of data augmentation in object detection and compare it against the value of model architectures. They achieve state-of-the-art without changing any network architecture. Customized augmentations like [166,167] plausibly relieve the variation problem to a certain degree. These approaches involve thousands of GPU days for optimizing the policy controller before actual re-training. Furthermore, the searched strategy is also fixed during re-training without adapting the optimization. To tackle the following two problems: 1) only a few images contain small objects 2) the lack of diversity in the locations of small objects, Kisantal et al. [168] propose two schemes respectively: 1) they solve the first issue through oversampling those images containing small objects 2) they address the second problem via copy-pasting small objects multiple times in each image containing small objects. These two approaches have significantly improved the detection of small objects. However, we can't simply paste the cropped object randomly in the drone captured image. There is an obvious position prior in the drones captured image. For example, car flies in the sky is impossible. Thus, Chen et al. [169] introduce a novel adaptive data augmentation strategy called adaptive resampling (AdaResampling) to logically augment the data.

Focusing on face detection, Tang et al. [98] exploit data-anchor-sampling (DAS) scheme to augment the training samples across different scales, which increases the diversity of training data for smaller faces. In short, data-anchor-sampling resizes train images through reshaping a random face in this image to a random smaller anchor size. To further boost the performance of face detector, Li et al. [99] combine the original SSD-sampling and data-anchor-sampling scheme, where color distort, random crop and horizontal flip are done on the photo with a specified probability value. Thus, they introduce a balanced-data-anchor-sampling (BDAS) strategy. It picks the anchor size with equal probability, and then the selected size will be obtained in the interval nearby the anchor size with equal probability too. In the implementation, they use BDAS with probability of 0.8 and SSD-sampling with probability of 0.2, severally.

Unlike the above methods, Chen et al. [170] put forward a novel feedback-driven data provider called Stitcher. In the Stitcher, images are resized into smaller components and then stitched into the same size to regular images. Stitched images contain inevitable smaller objects, which would be beneficial to guide next-iteration update by utilizing the loss statistics as feedback. In addition, there are some approaches [50,158,171] dedicated to the detection of tiny person. Yu et al. [50] put forward an effective scale match (SM) strategy. According to different object sizes, it crops the images to narrow the gap between objects of different sizes, in order to avoid the situation that small objects' information is easy to be lost in conventional scaling operations. Feng et al. [158] develop a suit of strategies to further improve the detectors performance including data augmentation based on SM that aligns the object scales between the existing large-scale dataset and TinyPerson dataset. This strategy can obtain the favorable tiny-object representation. Subsequently, Jiang et al. [171] comprehensively analyze the scale information of TinyPerson, and propose a novel refined scale match scheme, namely SM+, for tiny person detection. Unlike the SM that only considers the whole image, SM+ focuses on every instance. This method effectively promotes the similarity between pre-training and target dataset, which improves the detection performance over the state-of-the-art detectors with a large margin. The strengths and weaknesses of data augmentation approaches are presented in Table 9.

Table 8
Summarization of strengths and weaknesses of training strategies

Method	Merit	Shortcoming
HR [96] SFA [154] Gao et al. [155] Feng et al. [158] SNIP [159] SNIPER [160] SAN [161] Montage pre-training [162] DST [163]	It is acknowledged that models trained with multi-scale training could further enhance the detection performance with matching multi-scale testing. Besides, the montage pre-training scheme and the dynamic scale training strategy can also help for small or tiny object detection.	While features learned in the way of multi-scale training and testing are more robust to scale variation, it also affects inference speed. For example, SNIP and SNIPER rely on multi-scale testing that suffers from inference burden. Moreover, they operate in a static manner during training, rendering them unable to provide scale-sensitive data that the network desires, which overlooks the dynamic merits.

Table 9
Summarization of strengths and weaknesses of data augmentation methods

Method	Advantage	Weakness
Gao et al. [155] Zhou et al. [162] YOLOv4 [94] Zoph et al. [166] AutoAugment [167] DAS [98] BDAS [99] Augmentation [168] AdaResampling [169] Stitcher [170] SM [50] Feng et al. [158] SM+ [171]	It is a very cumbersome to capture a large number of new images for any domain. From this perspective, data augmentation methods save time and cost. Besides, changing the model architecture comes at the cost of adding more complexity to inference, making models slower. However, data augmentation strategies do not add any inference complexity in this regard.	It is more difficult to devise plausible augmentation strategies for object detection than for classification. Compared to advances in network architectures, data augmentation schemes attract less research attention perhaps because it is believed to add less value in detection performance and to transfer poorly.

3.2.7. Schemes based on loss function

Optimizing the loss function is an effective strategy to improve the detection performance of small objects. The loss contribution usually utilized on deep CNN mainly comes from samples that are easy to be detected. The cross entropy loss (CEL) function is not beneficial for detecting some hard samples, such as small-scale pedestrians. Thus, Han et al. [172] devise a novel loss function based on CEL to increase the loss contribution from hard-to-detect examples. Besides, Wu et al. [173] design a mimic loss function to force the feature representations of small-scale pedestrians to approach those of large-scale pedestrians. The experimental results show that the detector trained with the mimic loss is significantly effective for small-scale pedestrian detection.

Chen et al. [174] propose a novel framework to substitute the classification task in one-stage detectors with a ranking task, and using the average-precision loss (AP loss) for the ranking problem. Moreover, they put forward a new error-driven learning algorithm to effectively optimize the AP based objective function. However, the AP loss focuses on the original pairs and is non-differentiable. A specific method has to be designed to minimize the AP-loss. To this end, Qian et al. [175] develop a novel distributional ranking loss (DR loss) to deal with this challenge. The DR loss ranks the expectations of distributions in lieu of original pairs. Furthermore, DR loss is differentiable and can be optimized with stochastic gradient descent (SGD) in the standard training pipeline. The above two loss function based schemes balance the foreground and background examples by reweighting examples of imbalanced categories in the loss function. Unlike these two methods, Lin et al. [38] attempt to handle the class imbalance via reshaping the standard CEL such that it down-weights the loss assigned to well-classified examples. The proposed new focal loss focuses training on a sparse set of hard examples and prevents the vast number of easy negatives from overwhelming the detector during training. To evaluate the effectiveness of focal loss, they devise and train a simple dense detector, namely RetinaNet. Subsequently, Zhang et al. [176] apply the cascade idea to the RetinaNet and propose a new object detector named Cas-RetinaNet, which further improve the accuracy of small object detection. To further alleviate the negative impact for detection performance brought by the issue of sample imbalance, Ji et al. [57] propose a novel ω-focal loss function, which significantly improves the detection accuracy of the categories with few objects. By incorporating multi-scale receptive field attention and ω-focal loss into an end-to-end architecture, they develop a one-stage framework called CommodityNet for small and dense commodity detection. Inspired by focal loss, Wang et al. [177] devise a focal-area loss function which is learned via focusing the area change of small objects such as stones on road. This loss function can heighten the significance of the small target in the classification loss, so as to promote the accurate detection of small targets on the road.

To generalize focal loss from its discrete form to the continuous version for successful optimization, Li et al. [178] propose generalized focal loss (GFL). It can be divided into quality focal loss (QFL) and distribution focal loss (DFL), for optimizing the improved two representations severally. More specifically, QFL focuses on a sparse set of hard examples and simultaneously generates their continuous 0~1 quality estimations on the corresponding class. DFL makes the network to rapidly focus on learning the probabilities of values around the continuous locations of object bounding boxes, under an arbitrary and flexible distribution. The bounding box distributions are introduced as "general distribution" in GFL, which describes the uncertainty of the predicted bounding boxes well. Different from GFL, Li et al. [179] explore a novel perspective to perform localization quality estimation (LQE) based on the learned distributions of the four parameters of the bounding box. Via using the close correlation between distribution statistics and the real localization quality, they introduce a lightweight distribution-guided quality predictor (DGQP) for reliable LQE based on GFL, thus building generalized focal loss v2 (GFLv2).

Schemes based on loss function will also be conductive to object localization and accelerate detection. Yu et al. [180] propose a novel intersection over union loss (IoU loss) function for bounding box prediction, which regresses the four bounds of a predicted box as a whole unit. Through taking the advantages of IoU loss and deep fully convolutional networks, the UnitBox is introduced, which performs accurate and efficient localization, reveals robust to objects of varied shapes and scales. Afterwards, Tychsen-Smith et al. [181] derive bounded IoU loss, a new bounding box regression loss based on a set of IoU upper bounds that better matches the goal of IoU maximization while still maintaining good convergence properties. Nevertheless, IoU has a plateau making it infeasible to optimize in the case of non-overlapping bounding boxes. To this end, Rezatofighi et al. [182] address the weaknesses of IoU via developing a generalized version as both a new loss and metric. By incorporating this generalized IoU as a loss (GIoU loss) into the existing detector, they show a consistent performance improvement on detection benchmarks using both the IoU based and GIoU based. Although IoU loss and GIoU loss have been proposed, they still suffer from the problems of slow convergence and inaccurate regression. Zheng et al. [183] propose a distance-IoU loss (DIoU loss) through incorporating the normalized distance between the predicted box and the target box, which converges much faster in training than IoU and GIoU losses.

In addition to the above schemes, there are also some methods based on loss function. He et al. [184] introduce a novel bounding box regression loss, namely KL loss, for learning bounding box transformation and localization variance together. This kind of loss greatly improves the localization accuracies of various architectures without additional computation nearly. The learned localization variance can merge neighboring bounding boxes during non-maximum suppression, which further heightens the localization performance of small objects. The intrinsic relation between the localization and classification subnets in CNN framework is not exploited explicitly for object detection. To this end, Xu et al. [185] put forward a novel association loss, namely the proxy squared error loss (PSE loss), to entangle the two subnets, thus employ the dependency between the classification and localization scores gained from these two subnets to enhance the detection performance. Via the analysis of loss distribution over different scales in the iterations, there is a remarkable gap between the loss provided by the small objects and large objects. To balance the loss distribution and

Table 10
Summary of strengths and weaknesses of schemes based on loss function

Method	Superiority	Weakness
Improved CEL [172]	These two schemes are effective for small-scale pedestrian detection.	Mimic learning will cause the localization problem of objects.
Mimic loss [173]		
AP loss [174]	These approaches can alleviate the category imbalance of small objects by designing or optimizing loss function, which improve the small object detection performance, such as small instances in common life.	The complexity and stability of the ranking schemes used in AP loss and DR loss need to be further discussed. In addition, some other methods have been improved through modifying the focal loss. Although the accuracy of these methods has been enhanced to a certain extent, their operation steps have become more complicated.
DR loss [175]		
Focal loss [38]		
Cas-RetinaNet [176]		
ω-focal loss [57]		
Focal-area loss [177]		
GFL [178]		
GFLv2 [179]	These methods based on loss function are conductive to object localization and accelerate detection. To some extent, small object detection can be improved.	
IoU loss [180]		If the target box and the prediction box do not intersect, then the loss is zero. Without gradients of backpropagation, the training cannot be performed. When the target box completely surrounds the prediction box, GIoU degenerates to IoU. Besides, DIoU does not fully consider the impact of the aspect ratio of the detection box on the loss.
Bounded IoU loss [181]		
GIoU loss [182]		
DIoU loss [183]		
KL loss [184]		
PSE loss [185]		
Feedback-driven loss [186]	It can supervise small objects more effectively.	

Table 11
An overview of some representative methods on different datasets according to our taxonomy of small/tiny object detection

Type	Method	Dataset	Type	Method	Dataset
Super-resolution techniques	Perceptual GAN [76]	TT100K	Featurized image pyramids	IPG RCNN [113]	MS-COCO
	PKG [77]	TT100K		HRDNet [85]	MS-COCO
	JSC-Net [78]	KITTI	Pyramidal feature hierarchy (e.g. SSD, and some SSD-based approaches)	SSD [45]	AI-TOD
	FaceGAN [81]	WIDER FACE		FFSSD [115]	PASCAL VOC
	TFPGAN [82]	WIDER FACE		MDSSD [116]	PASCAL VOC/MS-COCO
	MTGAN [83]	WIDERFACE/MS-COCO		RFBNet [118]	TT100K
	Feature SR [84]	TT100K/MS-COCO		TridentNet [119]	AI-TOD
	ISOD [72]	SOD		FPN [122]	UAVDT/DIOR/TinyPerson/STDdb
	HRDNet [85]	MS-COCO		DFPN [127]	TT100K
	GDL [87]	MS-COCO		Libra R-CNN[128]	TinyPerson
	SimultaneousSR [88]	PASCAL VOC	Feature pyramid network (In addition to FPN, some variants of FPN are includEd.)	RHFNet [130]	MS-COCO
	QueryDet [89]	MS-COCO		IMFRE [129]	PASCAL VOC/MS-COCO
	STDnet [53]	UAVDT/STDdb		SFRF [131]	TinyPerson
	STDnet-ST [91]	UAVDT/STDdb		S-α [132]	TinyPerson
	Pyramid-Box [98]	WIDER FACE		BSCF [125]	PASCAL VOC
	Pyramid-Box++[99]	WIDER FACE		CANet [54]	DIOR
Context-based information	HR [96]	WIDER FACE	Schemes based on loss function	SSPNet [112]	TinyPerson
	ACNet [97]	WIDER FACE		AP loss [174]	MS-COCO
	BC-ESS [101]	WIDER FACE		DR loss [175]	MS-COCO
	IENet [102]	WIDERFACE/MS-COCO		Focal loss [38]	AI-TOD/TinyPerson/MS-COCO
	IR RCNN [103]	MS-COCO		Cas-RetinaNet[176]	MS-COCO
	FA-SSD [104]	PASCAL VOC		GFL [178]	MS-COCO
	LocalNet [105]	MS-COCO		GFLv2 [179]	MS-COCO
	RCN [73]	UAVDT/STDdb		Bounded IoU loss [181]	MS-COCO
	SCAN [110]	KITTI		KL loss [184]	MS-COCO
	CABNet [111]	TT100K		PSE loss [185]	MS-COCO
	SSPNet [112]	TinyPerson		Feedback-drivenloss [186]	DOTA/PASCAL VOC/MS-COCO
Training strategy	HR [96]	WIDER FACE	Data augmentation	DAS [98]	WIDER FACE
	SFA [154]	WIDER FACE		BDAS [99]	WIDER FACE
	SNIP [159]	MS-COCO		Augmentation[168]	MS-COCO
	SNIPER [160]	MS-COCO		Stitcher [170]	PASCAL VOC/MS-COCO
	SAN [161]	MS-COCO		SM [50]	TinyPerson
	DST [163]	PASCALVOC/MS-COCO		SM+ [171]	TinyPerson
	ISOD [72]	SOD		CornerNet [145]	MS-COCO
	SCRDet [134]	DOTA		Grid R-CNN [146]	AI-TOD/TinyPerson
Anchor-based mechanism	CascadeR-CNN [135]	AI-TOD/ STDdb	Anchor-free mechanism	Centernet-KT [148]	MS-COCO
	CasMaskGF [136]	TT100K		ExtremeNet [149]	MS-COCO
	AMS-Net [138]	KITTI		CenterNet [150]	AI-TOD
	S3FD [140]	WIDER FACE		M-CenterNet [48]	AI-TOD
	LSFHI [141]	WIDER FACE		RepPoints [151]	AI-TOD
	CPT-Matching [144]	PASCALVOC/MS-COCO		FCOS [152]	AI-TOD/TinyPerson

alleviate the insufficient supervisory on small objects, Liu et al. [186] propose a novel feedback-driven loss function. Compared with the original loss function, the feedback-driven loss function can supervise small objects more effectively. It utilizes the loss distribution information as the feedback signal, and trains the model in a more balanced way. Table 10 exhibits the merits and drawbacks of schemes based on loss function.

4. Performance analysis and discussion

In this section, we evaluate some representative small or tiny object detection techniques on 12 popular datasets in chronological order, including DOTA, UAVDT, AI-TOD, DIOR, KITTI, TinyPerson, TT100K, WIDER FACE, PASCAL-VOC, MS-COCO, SOD and USC-GRAD-STDdb.

DOTA, UAVDT, AI-TOD and DIOR datasets are for remote sensing small target detection. KITTI, TinyPerson, TT100K and WIDER FACE datasets are for pedestrian detection, tinyperson detection, traffic signs detection and face detection respectively. PASCAL-VOC, MS-COCO and SOD datasets are for small object detection in daily life. USC-GRAD-STDdb database is for small object detection in videos.

We have comprehensively summarized the methods of detecting small/tiny objects from seven aspects. They are super-resolution techniques, context-based information, multi-scale representation learning,

Table 12
Category-wise object detection results on DOTA dataset. The abbreviations of the 15 categories are as follows: Plane-PL, Baseball Diamond-BD, Bridge-BR, Ground Field Track-GFT, Small Vehicle-SV, Large Vehicle-LV, Ship-SH, Tennis Court-TC, Basketball Court-BC, Storage Tank-ST, Soccer-ball Field-SF, Roundabout-RO, Harbor-HA, Swimming Pool-SP, Helicopter-HE. Bold fonts indicate the best result. (in %)

Year	Method	PL	BD	BR	GFT	SV	LV	SH	TC	BC	ST	SF	RO	HA	SP	HE	mAP
2021	Feedback-driven loss [186]	79.0	38.2	28.7	36.9	44.2	40.9	57.8	65.4	54.6	36.3	34.1	39.6	42.9	45.1	16.3	43.2
2019	SCRDet [134]	**90.2**	**81.9**	**55.3**	**73.3**	**72.1**	**77.7**	**78.1**	**90.9**	**82.4**	**86.4**	**64.5**	**63.5**	**75.8**	**78.2**	**60.1**	**75.4**

Table 13
Detection performance on the very small and small subsets of the UAVDT dataset. Average Precision when IoU is at least 0.5 ($AP_{0.5}$) and Average Precision when the IoU goes from 0.5 to 0.95 in 5% steps (AP_{all}). The best result for each metric is highlighted in bold. (in %)

Year	Method	Very small subset		Small subset	
		AP_{all}	$AP_{0.5}$	AP_{all}	$AP_{0.5}$
2021	STDnet-ST [91]	**13.3**	**36.4**	--	--
2020	STDnet [53]	12.6	35.4	--	--
2018	RCN [73]	12.5	35.1	**27.0**	**53.7**
2017	FPN [122]	11.8	29.7	24.8	49.2

anchor mechanism, training strategy, data augmentation and schemes based on loss function. According to our taxonomy of small or tiny object detection, we show an overview of some representative approaches on the above datasets in Table 11. Subsequently, we discuss the possible future research directions based on performance analysis.

4.1. Performance analysis

We first report the detection results of some representative methods on the DOTA and UAVDT dataset in Table 12 and Table 13 respectively. Obviously, the detection performance of SCRDet [134] is significantly better than that of Feedback-driven loss [186] on the 15 categories of the DOTA dataset. In Table 13, we find that STDnet-ST [91] and RCN [73] obtain the best detection results on the very small subsets of UAVDT and small subsets of UAVDT dataset, respectively.

The detection results of different small or tiny object detection algorithms on the DIOR dataset are shown in Table 14. From Table 14, we can observe that CANet [54] obtains 74.3% mAP on DIOR dataset, which outperforms other five approaches. Particularly, it gains better results in most categories, including airplane, bridge, chimney, expressway toll station, ground track field, overpass, ship, stadium, storage tank, tennis court and vehicle. Furthermore, Cheng et al. [126] achieve better performance in the other nine categories: airport, baseball field, basketball court, dam, expressway service area, golf field, harbor, train station and windmill.

The performance of different detectors on AI-TOD dataset is reported in Table 15. M-CenterNet [48] obtains the best performance in five metrics, including AP, $AP_{0.5}$, AP_{vt}, AP_t, and oLRP. Especially, in the AP_{vt} and AP_t metrics, M-CenterNet far surpasses other detectors. Moreover, as shown in Table 16, M-CenterNet achieves the best performance on five categories, such as bridge, storage-tank, vehicle, person and windmill.

Table 17 and Table 18 show the performance of different methods on TinyPerson benchmark in terms of MR and AP metrics. We find that Faster R-CNN with SFRF [131] achieves the best result in all approaches for tiny1, tiny2 and tiny3 objects. Besides, these detectors, i.e., RetinaNet, Faster R-CNN, and Cascade R-CNN, gain further improvements with using SSPNet [112]. Especially, Cascade R-CNN-SSPNet [112] and Faster R-CNN-SSPNet [112] gain the best detection performance for small objects and tiny objects, respectively. In Table 18, Faster R-CNN-SSPNet [112] exceeds other methods with a large margin in terms of MR metric. It is not difficult to find that SSPNet adopting multi-scale representation learning and context-based information could well detect all kinds of size objects, such as tiny, tiny1 ~ 3 and small objects.

We compare the detection results of AMS-Net [138], JCS-Net [78], SqueezeDet [25] and SCAN [110] on the KITTI challenge. As shown in Table 19, JCS-Net outperforms the AMS-Net in terms of moderate-level and hard-level. The column "mAP" of SCAN exhibits substantial gains (1.6 points) over SqueezeDet. Specifically, the column "Cyclists" achieves much better performance than the SqueezeDet, where the column "Cyclists-Moderate" obtains 15.4 points higher than SqueezeDet.

Similarly, we compare Feature SR [84], DFPN [127], CasMaskGF [136] and Perceptual GAN [76] these four methods on TT100K dataset. From Table 20, we can observe that Feature SR offers a substantial improvement in terms of Overall-F1 among these four approaches. More concretely, the columns "Small-F1", "Medium-F1" and "Large-F1" reveal substantial gains (2.2, 2.6, and 5.5 points severally) over Perceptual-GAN. The column "Medium-F1" of CasMaskGF achieves the highest F1 of 96.9% among four methods. In addition, among three object sizes, the detection performance of small objects is the lowest. We also show the AP of most commonly used traffic signs in Table 21. DFPN achieves excellent performance in most categories than other four algorithms. Some categories reach 98% AP, such as "il60", "p27", "pl100", "pl70" and "pm20". In particular, category "pl120" obtains the highest AP of 99% among all categories.

We exhibit the detection results of different detectors on WIDER FACE dataset. As shown in Table 22, Pyramid-Box++ [99] using context-based information and data augmentation outperforms other schemes and achieves the highest AP among easy-level, medium-level and hard-level. Similarly, Pyramid-Box [98] also outperforms other

Table 14
Overall performance evaluation of different methods on DIOR dataset. * indicates new implementation by Li et al. [54]. The best result is highlighted in bold. (in %)

Year	2017	2021	2021	2021	2021	2022
Method	FPN [122]	CSFF [187]	PANet* [54]	Libra R-CNN* [54]	CANet [54]	Cheng et al. [126]
Backbone	ResNet101	ResNet101	ResNet101	ResNet101	ResNet101	ResNet101
Airplane	54.0	57.2	62.8	60.8	**70.3**	62.8
Airport	74.5	79.6	79.1	79.8	82.4	**86.5**
Baseball field	63.3	70.1	71.5	71.7	72.0	**74.8**
Basketball court	80.7	87.4	88.3	87.8	87.8	**89.2**
Bridge	44.8	46.1	50.7	49.5	**55.7**	49.2
Chimney	72.5	76.6	79.3	79.8	**79.9**	76.6
Dam	60.0	62.7	66.4	64.8	67.7	**72.5**
Expressway service area	75.6	82.6	82.1	82.0	83.5	**85.7**
Expressway toll station	62.3	73.2	70.8	74.8	**77.2**	75.1
Golf field	76.0	78.2	78.0	79.7	77.3	**81.3**
Ground track field	76.8	81.6	82.8	82.5	**83.6**	83.3
Harbor	46.4	50.7	49.4	42.9	56.0	**60.2**
Overpass	57.2	59.5	62.7	63.0	**63.6**	62.7
Ship	71.8	73.3	72.1	72.0	**81.0**	72.7
Stadium	68.3	63.4	67.1	75.4	**79.8**	77.3
Storage tank	53.8	58.5	62.8	62.7	**70.8**	61.9
Tennis court	81.1	85.9	81.2	81.3	**88.2**	88.0
Train station	59.5	61.9	61.6	64.3	67.6	**69.9**
Vehicle	43.1	42.9	50.8	49.9	**51.2**	47.0
Windmill	81.2	86.9	88.2	88.5	89.6	**89.7**
mAP	65.1	68.0	70.4	70.7	**74.3**	73.3

Table 15
Performance of different detectors on AI-TOD testset. AP stands for average precision. oLRP denotes the Optimal Localization Recall Precision. AP_{vt}, AP_t, AP_s, AP_m, are average precision for very tiny, tiny, small, medium scales, respectively. Note that AP is a higher-is-better measure, while oLRP is an error metric, and thus it is a lower-is-better measure. Bold and underline fonts indicate the best and suboptimal results for each metric, respectively. (in %)

Year	Method	Backbone	AP	$AP_{0.5}$	$AP_{0.75}$	AP_{vt}	APt	APs	AP_m	oLRP
2020	M-CenterNet [48]	DLA34	**14.5**	**40.7**	6.4	**6.1**	**15.0**	19.4	20.4	**85.8**
2019	TridentNet [119]	ResNet50	7.5	20.9	3.6	1.0	5.8	12.6	14.0	92.7
2019	FCOS [152]	ResNet50-FPN	9.8	24.1	5.9	1.4	8.0	15.1	17.4	90.8
2019	RepPoints [151]	ResNet50-FPN	9.2	23.6	5.3	2.5	9.2	12.9	14.4	91.5
2019	Grid R-CNN [146]	ResNet50-FPN	12.2	27.7	9.0	0.2	10.3	22.6	23.3	88.6
2019	CenterNet [150]	DLA34	13.4	39.2	5.0	3.8	12.1	17.7	18.9	87.1
2018	Cascade R-CNN [135]	ResNet50-FPN	13.8	30.8	**10.5**	0.0	10.6	**25.5**	**26.6**	87.6
2017	RetinaNet [38]	ResNet50-FPN	4.7	13.6	2.1	2.0	5.4	6.3	7.6	94.7
2016	SSD512 [45]	VGG16	7.0	21.7	2.8	1.0	4.7	11.5	13.5	92.8

Table 16
Class-wise object detection results on AI-TOD testset. oLRP denotes the Optimal Localization Recall Precision. AP stands for average precision. The best and suboptimal results for each object category are highlighted in bold and underline, severally. (in %)

Year	Method	Airplane	Bridge	Storage-tank	Ship	Swimming pool	Vehicle	Person	Wind mill
		AP/oLRP	AP/oLRP	AP/oLRP	AP/oLRP	AP/oLRP	AP/oLRP	AP/oLRP	AP/oLRP
2020	M-CenterNet [48]	18.59/83.00	**10.58/89.23**	**27.55/74.50**	22.27/79.47	7.53/92.06	**18.60/81.19**	**9.17/90.49**	**2.03/96.73**
2019	TridentNet [119]	9.67/89.84	0.77/98.56	12.28/88.00	17.11/85.00	3.20/97.00	11.87/88.66	3.98/95.80	0.94/98.38
2019	FCOS [152]	14.30/86.46	4.75/94.83	19.77/82.89	22.24/80.97	0.65/98.29	12.51/88.10	3.98/95.62	0.17/99.57
2019	RepPoints [151]	2.92/96.18	2.34/97.32	21.37/80.92	**26.40/77.23**	0.00/100.00	15.16/85.90	5.39/94.53	0.00/100.00
2019	Grid R-CNN [146]	22.55/78.59	8.59/91.46	18.93/82.74	21.99/81.21	7.28/92.72	12.94/87.68	4.81/94.99	0.35/99.28
2019	CenterNet [150]	17.43/84.27	9.46/90.61	25.93/75.46	21.86/80.97	6.21/93.42	16.54/82.32	8.12/91.82	1.94/97.73
2018	Cascade R-CNN [135]	**25.57/77.62**	7.47/92.87	23.33/79.07	23.55/79.69	**10.81/89.75**	14.09/86.80	5.34/94.55	0.00/100.00
2017	RetinaNet [38]	0.01/99.88	6.62/93.51	1.84/96.34	20.87/79.40	0.06/99.82	5.67/92.41	1.75/97.04	0.53/99.17
2016	SSD512 [45]	14.52/86.49	3.13/96.24	10.89/89.40	13.05/87.95	1.92/96.67	7.84/91.22	3.12/96.53	1.48/97.61

Table 17
APs of different methods on TinyPerson benchmark. AP represents average precision. Tiny, tiny1, tiny2, tiny3, and small reflect the object size in range [2,20], [2,8], [8,12], [12,20] and [20,32] respectively. Bold and underline fonts indicate the best and suboptimal results for each metric, respectively. (in %)

Year	Method	AP_{50}					AP_{25}	AP_{75}
		Tiny	Tiny1	Tiny2	Tiny3	Small	Tiny	Tiny
2021	Faster RCNN-FPN-MSM + [171]	52.61	34.20	57.60	63.61	67.37	72.54	6.72
2021	RetinaNet with S-α [132]	48.34	28.61	54.59	59.38	61.73	71.18	5.34
2021	Faster RCNN-FPN with S-α [132]	48.39	31.68	52.20	60.01	65.15	69.32	5.78
2021	RetinaNet + SM with S-α [132]	52.56	33.90	58.00	63.72	65.69	73.09	6.64
2021	RetinaNet + MSM with S-α [132]	51.60	33.21	56.88	62.86	64.39	72.60	6.43
2021	Faster RCNN-FPN + SM with S-α [132]	51.76	34.58	55.93	62.31	66.81	72.19	6.81
2021	Faster RCNN-FPN + MSM with S-α [132]	51.41	34.64	55.73	61.95	65.97	72.25	6.69
2021	RetinaNet-SSPNet [112]	54.66	42.72	60.16	61.52	65.24	77.03	6.31
2021	Cascade R-CNN-SSPNet [112]	58.59	45.75	62.03	65.83	**71.80**	78.72	8.24
2021	Faster R-CNN-SSPNet [112]	**59.13**	47.56	62.36	66.15	71.17	**79.47**	**8.62**
2021	Faster R-CNN with SFRF [131]	57.24	**51.49**	**64.51**	**67.78**	65.33	78.65	6.42
2020	RetinaNet-SM [50]	48.48	29.01	54.28	59.95	63.01	69.41	5.83
2020	RetinaNet-MSM [50]	49.59	31.63	56.01	60.78	63.38	71.24	6.16
2020	Faster R-CNN-FPN-SM [50]	51.33	33.91	55.16	62.58	66.96	71.55	6.46
2020	Faster R-CNN-FPN-MSM [50]	50.89	33.79	55.55	61.29	65.76	71.28	6.66
2019	Grid RCNN [146]	47.14	30.65	52.21	57.21	62.48	68.89	6.38
2019	Libra RCNN [128] (Balanced FPN)	44.68	27.08	49.27	55.21	62.65	64.77	6.26
2019	FCOS [152]	17.90	2.88	12.95	31.15	40.54	41.95	1.50
2017	RetinaNet [38]	33.53	12.24	38.79	47.38	48.26	61.51	2.28
2017	Faster R-CNN-FPN [122]	47.35	30.25	51.58	58.95	63.18	63.18	5.83

detectors (except Pyramid-Box++). Actually, IENet and LSFHI obtain good performance at the easy-level and hard-level, respectively.

We calculate AP of each of the five categories in the PASCAL-VOC2007 testset. Comparative results of some approaches on PASCAL-VOC2007 testset are exhibited in Table 23. It is noting that Feedback-driven loss [186] achieves 80.7% mAP, which is better than other methods. Especially, it obtains 86.4% AP, 78.0% AP, 65.9% AP and 88.4% AP for category "boat", "bottle", "potted plant" and "sheep" respectively. Moreover, DST [163] and MDSSD512 [116] achieve the best detection results in "car" category with AP of 89.3%.

We report the detection results of different methods on MS-COCO test-dev dataset, as shown in Table 24. IENet [102] achieves the best performance in five metrics, including AP, AP_{50}, AP_{75}, AP_S, and AP_L. Especially, in the AP_S metric, IENet surpasses other detectors with a large margin. For the AP_M metric, GFLv2 [179] gains 54.3% AP on the test-dev set, which outperforms other algorithms. Besides, other algorithms such as QueryDet [89], GFLv2 [179], SNIP [159] and SNIPER [160], obtain more than 29% AP on small object detection. Schemes based on loss function, such as Feedback-driven loss [186] and AP loss [174], both gain over 25% AP on small objects.

Table 18
MRs of different methods on TinyPerson benchmark. MR denotes miss rate. The lower the MR value, the better the performance of the detector. The best result in each MR is marked in bold. (in %)

Year	Method	MR_{50}					MR_{25}	MR_{75}
		Tiny	Tiny1	Tiny2	Tiny3	Small	Tiny	Tiny
2021	RetinaNet with S-α [132]	87.73	89.51	81.11	79.49	72.82	74.85	98.57
2021	Faster RCNN-FPN with S-α [132]	87.29	87.69	81.76	78.57	70.75	76.58	98.42
2021	RetinaNet + SM with S-α [132]	87.00	87.62	79.47	77.39	69.25	74.72	98.41
2021	RetinaNet + MSM with S-α [132]	87.07	88.34	79.76	77.76	70.35	75.38	98.41
2021	Faster R-CNN-FPN + SM with S-α [132]	85.96	86.57	79.14	77.22	69.35	73.92	98.30
2021	Faster R-CNN-FPN + MSM with S-α [132]	86.18	86.51	79.05	77.08	69.28	73.90	98.24
2021	RetinaNet-SSPNet [112]	85.30	82.87	76.73	77.20	72.37	69.25	98.63
2021	Cascade R-CNN-SSPNet [112]	83.47	82.80	75.02	73.52	62.06	68.93	98.27
2021	Faster R-CNN-SSPNet [112]	**82.79**	**81.88**	**73.93**	**72.43**	**61.26**	**66.80**	**98.06**
2020	RetinaNet-SM [50]	88.87	89.83	81.19	80.89	71.82	77.88	98.57
2020	RetinaNet-MSM [50]	88.39	87.80	79.23	79.77	72.18	76.25	98.57
2020	Faster R-CNN-FPN-SM [50]	86.22	87.14	79.60	76.14	68.59	74.16	98.28
2020	Faster R-CNN-FPN-MSM [50]	85.86	86.54	79.20	76.86	68.76	74.33	98.23
2019	Grid R-CNN [146]	87.96	88.31	82.79	79.55	73.16	78.27	98.21
2019	Libra R-CNN [128] (Balanced FPN)	89.22	90.93	84.64	81.62	74.86	82.44	98.39
2019	FCOS [152]	96.28	99.23	96.56	91.67	84.16	90.34	99.56
2017	RetinaNet [38]	88.31	89.65	81.03	81.08	74.05	76.33	98.76
2017	Faster R-CNN-FPN [122]	87.57	87.86	82.02	78.78	72.56	76.59	98.39

Table 19
Summary of detection results of different methods on KITTI dataset. (in %)

Year	Method		Easy			Moderate			Hard		
2020	AMS-Net [138]		85.63			70.52			69.31		
2019	JCS-Net [78]		85.62			74.99			69.65		
Year	Method	mAP	Cars			Pedestrians			Cyclists		
			Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
2018	SCAN [110]	82.0	96.7	80.7	70.4	87.2	78.2	70.3	87.9	86.7	79.7
2017	SqueezeDet [25]	80.4	90.4	87.1	78.9	87.6	80.3	78.1	81.4	71.3	68.5

Table 20
Performance comparison with different methods on TT100K dataset. F1 represents F1-measure, which computed by recall and accuracy. (in %)

Year	Method	Small			Medium			Large			Overall		
		Rec.	Acc.	F1	Rec.	Acc.	F1	Rec.	Acc.	F1	Rec.	Acc.	F1
2019	Feature SR [84]	92.6	84.9	88.6	97.5	94.5	96.0	97.5	93.3	95.4	95.7	90.6	93.1
2018	DFPN [127]	93.0	84.0	88.3	97.0	95.0	95.9	92.0	96.0	93.9	--	--	--
2018	CasMaskGF [136]	--	--	88.2	--	--	96.9	--	--	94.8	--	--	92.8
2017	Perceptual GAN [76]	89.0	84.0	86.4	96.0	91.0	93.4	89.0	91.0	89.9	--	--	90.4

Table 21
Comparisons of AP for each class on the TT100K testset. PKG shows the results of categories that contain instances over 100. Bold fonts indicate the best results for each object category. (in %)

Year	Method	i2	i4	i5	il100	il60	il80	io	ip	p10	p11	p12	p19	p23	p26	p27
2020	CABNet [111]	76	88	89	81	90	85	81	78	69	78	74	**88**	87	81	81
2019	PKG [77]	68	79	85	76	--	58	--	75	70	79	78	84	77	77	66
2018	DFPN [127]	**90**	**92**	**94**	93	**98**	**94**	**86**	**90**	**89**	**90**	**94**	75	**93**	**89**	**98**
2018	RFBNet [118]	76	79	88	87	90	88	77	79	66	67	71	73	83	75	80
2017	Perceptual GAN [76]	85	**92**	**94**	**97**	95	83	79	**90**	84	85	88	84	92	83	**98**
Year	Method	p3	p5	p6	pg	ph4	ph4.5	ph5	pl100	pl120	pl20	pl30	pl40	pl5	pl50	pl60
2020	CABNet [111]	75	85	83	88	72	64	**79**	88	88	69	73	75	79	75	76
2019	PKG [77]	--	88	62	69	66	73	61	88	80	68	78	89	70	82	84
2018	DFPN [127]	81	**91**	**90**	**93**	94	**80**	78	**98**	**99**	90	**92**	**91**	**92**	**90**	**95**
2018	RFBNet [118]	69	78	69	89	68	63	76	89	85	67	72	72	75	63	70
2017	Perceptual GAN [76]	**92**	90	83	**93**	**97**	68	69	97	98	**92**	91	90	86	87	92
Year	Method	pl70	pl80	pm20	pm30	pm55	pn	pne	po	pr40	w13	w32	w55	w57	w59	wo
2020	CABNet [111]	73	79	74	67	81	85	90	64	89	71	67	84	79	68	47
2019	PKG [77]	78	89	70	--	81	--	87	--	88	--	50	--	--	--	--
2018	DFPN [127]	**98**	**92**	**98**	**97**	86	**90**	**97**	**81**	**97**	**90**	**95**	**95**	90	68	50
2018	RFBNet [118]	65	72	74	54	**87**	78	88	60	85	65	70	72	82	69	44
2017	Perceptual GAN [76]	97	86	90	77	81	89	93	78	92	66	83	88	**93**	**71**	**54**

Table 22
Detection results of different detectors on WIDER FACE dataset. The best and suboptimal results are marked in bold and underline, respectively. (in %)

Year	Method	AP		
		Easy	Medium	Hard
2021	IENet [102]	96.1	94.7	89.6
2020	TFPGAN [82]	93.7	93.4	87.4
2020	LSFHI [141]	95.7	94.9	89.7
2020	BC-ESS [101]	93.0	91.2	80.9
2020	MTGAN [83]	95.9	94.8	89.3
2019	ACNet [97]	91.1	89.8	80.0
2019	SFA [154]	94.9	93.6	86.6
2019	Pyramid-Box++ [99]	**96.5**	**95.9**	**91.2**
2018	Pyramid-Box [98]	96.1	95.0	88.9
2018	FaceGAN [81]	94.4	93.3	87.3
2017	HR [96]	92.5	91.0	80.6
2017	S3FD [140]	93.7	92.4	85.2

Table 23
AP results of small objects on PASCAL-VOC2007 testset. Bold and underline fonts indicate the best and suboptimal results, respectively. AP and mAP refer to average precision, and the average AP of these five categories, respectively. (in %)

Year	Method	mAP	potted plant	sheep	boat	bottle	car
2021	BSCF [125]	72.2	57.3	79.3	74.1	62.7	87.4
2021	DST [163]	76.3	59.8	86.7	73.0	72.7	**89.3**
2021	Feedback-driven loss [186]	**80.7**	**65.9**	**88.4**	**86.4**	**78.0**	84.8
2020	CADNet320 (CPT-Matching) [144]	70.6	54.8	78.5	73.1	58.4	88.0
2020	CADNet512 (CPT-Matching) [144]	73.2	58.8	80.2	74.8	63.6	88.8
2020	Simultaneous SR [88]	58.7	43.4	70.8	54.3	47.4	77.5
2020	Stitcher [170]	74.9	58.7	85.1	70.7	71.5	88.7
2020	MDSSD300 [116]	68.7	52.3	78.6	70.6	55.0	87.0
2020	MDSSD512 [116]	72.4	56.7	84.0	73.7	58.3	**89.3**
2017	FFSSD [115]	69.2	53.9	80.6	71.7	52.9	86.9

To further perform a fair comparison on MS-COCO dataset, we unify the backbone including VGG16, ResNet50, ResNet101, ResNet101-FPN and ResNeXt101. As shown in Table 25, we also introduce an item named the degree of reduction in AP (DoR-AP) to illustrate the large gap of detection performance among different scale objects. Obviously, the detection performance of large objects is the best among large, medium and small objects, while that of small objects is the worst. From Table 25, we observe that the accuracy gap between small objects and large objects is much larger than that between small objects and medium objects. This once again shows the difficulty of small object detection.

A comparison on MS-COCO test-dev set between single-scale testing and multi-scale testing is also exhibited in Table 26. From the Table 26, we find that multi-scale testing is very effective for deep learning-based detectors. It can further improve detection performance on large, medium and small objects. In particular, through using multi-scale testing, IMFRE512 [129] improves the detection performance of small objects by 6%.

In Table 27 and Table 28, we compare some detectors from the SOD and USC-GRAD-STDdb dataset, respectively. ISOD [72] using RPN and super-resolution obtains excellent detection precision on SOD dataset. It is more accurate than another method [52] with RPN and upscaling for six categories, including mouse, telephone, toilet paper, faucet, plate and jar. Besides, STDnet-ST achieves better detection performance on USC-GRAD-STDdb dataset than other four detectors in Table 28.

4.2. Discussion

We can obtain the following remarks based on the above comparative analysis from Tables 12 to 27.

(1) Super-resolution techniques, context-based information, multi-scale representation learning, anchor mechanism, training strategy, data augmentation and schemes based on loss function can achieve optimistic detection results of small objects on 12 datasets, including DOTA, UAVDT, AI-TOD, DIOR, KITTI, TinyPerson, TT100K, WIDER FACE, PASCAL-VOC, MS-COCO, SOD and USC-GRAD-STDdb. Moreover, combining above multiple approaches can also definitely improve the final detection performance of small or tiny objects, such as HR [96] using context-based information and training strategy; both SSPNet [112] and FA-SSD [104] adopting multi-scale representation learning and context-based information at the same time; both Pyramid-Box [98] and Pyramid-Box++[99] exploiting simultaneously context-based information and data augmentation; Libra R-CNN [128] taking into account data augmentation, loss function and multi-scale representation learning concurrently.
(2) Utilizing different input resolutions could influence the performance of small objects. Compared with a lower input resolution, a higher input resolution can gain more improvements, e.g. LocalNet300 [105] and LocalNet512 [105], RHFNet320 [130] and RHFNet416 [130], IMFRE300 [129] and IMFRE512 [129], etc. Besides, the detection performance of small objects is also affected by different backbone networks, such as QueryDet [89] employing ResNet101 and ResNeXt101, respectively; GFLv2 [179] using ResNet101 and Res2Net101-DCN, severally. Namely, more powerful backbone and higher input resolution can enhance the small object detection performance further.
(3) To obtain more accurate precision of small objects, multi-scale testing is very effective for deep learning-based detectors.

The results on these datasets and the above remarks show that small or tiny object detection has achieved the promising progress. However, there is still a huge gap between the state-of-the-art small object detection methods and that of the general-size object. Much work remains to be done, which we discuss at the following aspects:

(1) Datasets and evaluation criteria for small or tiny object detection

Datasets are vital to object detection. Nevertheless, there are few generally accepted small object datasets. Most researchers use small object datasets built by themselves to evaluate their algorithms. So establishing a large-scale and generic small object dataset is absolutely necessary. It can also provide a universal performance evaluation. In addition, the IoU based evaluation metric (AP and mAP) may not be suitable for small or tiny objects because even a small shift of bounding box in the image would cause a large difference in IoU value. To alleviate the problem that IoU itself and its extensions are very sensitive to the location deviation of the tiny objects, Wang et al. [189] first model the bounding boxes as 2D Gaussian distributions and then propose a novel evaluation metric called Normalized Wasserstein Distance (NWD) to compute the similarity between them by their corresponding Gaussian distributions. The proposed NWD metric can be easily embedded into anchor-based detectors for tiny object detection. Thus, a new evaluation criteria tailored for small object detection is also extremely significant.

(2) Weakly supervised small or tiny object detection

Most deep learning-based detectors use the models learned from well-annotated images with bounding boxes to detect different objects. However, it is hard to train a common model for small or tiny object detection by utilizing a fully-supervised learning method. Meanwhile, the annotation process of fully-supervised learning is very time-consuming.

Table 24
Detection results of different methods on MS-COCO test-dev dataset. "++" denotes multi-scales during inference. Feature SR: the results are presented on MS-COCO 2017 test-dev. The best and suboptimal results are marked in bold and underline. (in %)

Year	Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
2021	LocalNet300 [105]	LN-ResNet	32.3	50.6	34.4	15.1	38.7	46.7
2021	LocalNet512 [105]	LN-ResNet	34.4	53.2	36.9	20.3	42.3	47.1
2021	DST [163]	ResNet101-FPN	40.8	--	--	25.8	44.1	51.9
2021	Feedback-driven loss [186] (Cascade RCNN)	ResNet101-FPN	43.9	63.2	47.0	25.1	46.8	57.7
2021	IMFRE300 [129]	ResNet101	35.9	56.4	38.6	17.0	41.6	52.2
2021	IMFRE512 [129]	ResNet101	37.3	58.2	40.3	22.9	44.0	49.2
2021	QueryDet [89]	ResNet101	43.8	64.3	46.9	27.5	46.4	53.0
2021	QueryDet [89]	ResNeXt101	44.7	65.8	47.4	29.1	47.5	53.1
2021	HRDNet++ [85]	ResNet101 + 152	47.4	66.9	51.8	32.1	50.5	55.8
2021	IENet [102]	ResNet101	**51.2**	**69.3**	**56.1**	**34.5**	53.8	**63.6**
2021	GFLv2 [179]	ResNet101	46.2	64.3	50.5	27.8	49.9	57.0
2021	GFLv2 [179]	Res2Net101-DCN	50.6	69.0	55.3	31.3	**54.3**	63.5
2020	CADNet320 (CPT-Matching) [144]	VGG16	27.8	47.1	29.0	8.5	30.1	43.8
2020	CADNet512 (CPT-Matching) [144]	VGG16	30.5	50.8	32.1	11.4	35.0	44.8
2020	GDL [87]	ResNet50	34.8	55.2	37.6	23.5	44.2	35.7
2020	GDL [87]	ResNet101-FPN	39.2	59.7	43.0	28.8	46.8	38.8
2020	RHFNet320 [130]	DarkNet53	32.3	53.4	34.8	13.8	36.5	48.9
2020	RHFNet416 [130]	DarkNet53	35.2	57.5	37.6	15.9	37.5	48.9
2020	RHFNet512 [130]	ResNet101	37.7	59.8	40.1	19.9	42.9	51.5
2020	Stitcher [170] (Faster R-CNN)	ResNeXt101	43.1	65.6	47.4	28.0	46.7	54.2
2020	IR RCNN [103]	ResNet50	37.6	60.0	40.6	21.9	39.7	47.0
2020	IR RCNN [103]	ResNet101	39.7	62.0	43.2	22.9	42.4	50.2
2020	IPG RCNN [113]	IPG-Net101	45.7	64.3	49.9	26.6	48.6	58.3
2020	MTGAN [83]	ResNet101	41.4	63.2	45.4	24.7	44.2	52.6
2020	MDSSD300 [116]	VGG16	26.8	46.0	27.7	10.8	--	--
2020	MDSSD512 [116]	VGG16	30.1	50.5	31.4	13.9	--	--
2020	DR loss [175]	ResNet101-FPN	41.7	60.9	44.8	23.5	44.9	53.1
2020	GFL [178]	ResNet101-DCN	47.3	66.3	51.4	28.0	51.1	59.2
2020	PSE loss [185]	ResNet50	35.4	53.7	37.8	20.2	39.8	47.1
2019	Feature SR [84] (Faster R-CNN)	ResNet101	34.2	57.2	36.1	16.2	35.7	48.1
2019	Augmentation [168]	ResNet50	30.4	--	--	17.9	32.9	38.6
2019	AP loss [174]	ResNet101	42.1	63.5	46.4	25.6	45.0	53.9
2019	Cas-RetinaNet [176]	ResNet101	39.3	59.0	42.8	22.4	42.6	50.0
2019	KL loss [184] (Faster R-CNN)	ResNet50-FPN	38.5	57.8	41.2	20.9	41.2	51.5
2018	Bounded IoU loss [181]	DeNet101	41.8	60.9	44.9	21.5	45.0	57.5
2018	SNIP [159]	DPN98 [188]	45.7	67.3	51.1	29.3	48.8	57.1
2018	SNIPER [160]	ResNet101	46.1	67.0	51.6	29.6	48.9	58.1
2018	SAN [161]	R-FCN	36.3	59.6	--	16.7	40.5	55.5
2017	Focal loss [38]	ResNet101-FPN	39.1	59.1	42.3	21.8	42.7	50.2

Table 25
Detection results of different methods on MS-COCO test-dev dataset. DoR-AP-SM represents the gap of detection performance between small objects and medium objects. Similarly, DoR-AP-SL denotes the gap of detection performance between small objects and large objects. (in %)

Year	Method	Backbone	DoR-AP-SM	DoR-AP-SL	AP_S	AP_M	AP_L
2021	LocalNet512 [105]	VGG16	17.8	28.6	17.0	34.8	45.6
2021	IMFRE512 [129]		19.5	28.7	17.7	37.2	46.4
2020	CADNet512 (CPT-Matching) [144]		23.6	33.4	11.4	35.0	44.8
2020	IR RCNN [103]	ResNet50	17.8	25.1	21.9	39.7	47.0
2020	PSE loss [185]		19.6	26.9	20.2	39.8	47.1
2019	Augmentation [168]		15.0	20.7	17.9	32.9	38.6
2021	IMFRE512 [129]	ResNet101	21.1	26.3	22.9	44.0	49.2
2021	QueryDet [89]		18.9	25.5	27.5	46.4	53.0
2021	IENet [102]		19.3	29.1	34.5	53.8	63.6
2021	GFLv2 [179]		22.1	29.2	27.8	49.9	57.0
2020	RHFNet512 [130]		23.0	31.6	19.9	42.9	51.5
2020	IR RCNN [103]		19.5	27.3	22.9	42.4	50.2
2020	MTGAN [83]		19.5	27.9	24.7	44.2	52.6
2019	AP loss [174]		19.4	28.3	25.6	45.0	53.9
2019	Cas-RetinaNet [176]		20.2	27.6	22.4	42.6	50.0
2018	SNIPER [160]		19.3	28.5	29.6	48.9	58.1
2021	DST [163]	ResNet101-FPN	18.3	26.1	25.8	44.1	51.9
2021	Feedback-driven loss [186] (Cascade RCNN)		21.7	32.6	25.1	46.8	57.7
2020	DR loss [175]		21.4	29.6	23.5	44.9	53.1
2017	Focal loss [38]		20.9	28.4	21.8	42.7	50.2
2021	QueryDet [89]	ResNeXt101	18.4	24.0	29.1	47.5	53.1
2020	Stitcher [170] (Faster R-CNN)		18.7	26.2	28.0	46.7	54.2

Table 26
Comparative results of different methods on MS-COCO test-dev in terms of single-scale and multi-scale testing. SS and MS denote single-scale and multi-scale testing, severally. We also report the performance gain of small, medium and large objects. (in %)

Year	Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
2021	GFLv2 (SS) [179]	Res2Net101-DCN	50.6	69.0	55.3	31.3	54.3	63.5
2021	GFLv2 (MS) [179]	Res2Net101-DCN	53.3	70.9	59.2	35.7 (+4.4)	56.1 (+1.8)	65.6 (+2.1)
2021	QueryDet (SS) [89]	ResNeXt101	44.7	65.8	47.4	29.1	47.5	53.1
2021	QueryDet (MS) [89]	ResNeXt101	46.6	67.2	50.2	31.1 (+2)	49.1 (+1.6)	55.0 (+1.9)
2021	IMFRE512 (SS) [129]	ResNet101	37.3	58.2	40.3	22.9	44.0	49.2
2021	IMFRE512 (MS) [129]	ResNet101	41.1	62.3	44.3	28.9 (+6)	47.0 (+3)	53.0 (+3.8)
2020	DR loss (SS) [175]	ResNet101-FPN	41.7	60.9	44.8	23.5	44.9	53.1
2020	DR loss (MS) [175]	ResNet101-FPN	43.4	62.1	47.0	26.7 (+3.2)	46.1 (+1.2)	55.0 (+1.9)
2019	CenterNet-KT511 (SS) [148]	Hourglass104	44.9	62.4	48.1	25.6	47.4	57.4
2019	CenterNet-KT511 (MS) [148]	Hourglass104	47.0	64.5	50.7	28.9 (+3.3)	49.9 (+2.5)	58.9 (+1.5)
2019	ExtremeNet511 (SS) [149]	Hourglass104	40.2	55.5	43.2	20.4	43.2	53.1
2019	ExtremeNet511 (MS) [149]	Hourglass104	43.7	60.5	47.0	24.1 (+3.7)	46.9 (+3.7)	57.6 (+4.5)
2018	CornerNet511 (SS) [145]	Hourglass104	40.5	56.5	43.1	19.4	42.7	53.9
2018	CornerNet511 (MS) [145]	Hourglass104	42.1	57.8	45.3	20.8 (+1.4)	44.8 (+2.1)	56.7 (+2.8)

Table 27
Comparative results on the SOD [52] dataset. The last column is the weighted average precision. The best result is highlighted in bold. (in %)

Year	Method	mouse	telephone	switch	outlet	clock	toilet paper	tissue box	faucet	plate	jar	Average
2017	RPN and super-resolution [72]	**60.1**	**16.9**	16.2	23.5	30.3	**34.1**	12.8	**38.0**	**15.2**	**4.7**	**25.2**
2017	Faster RCNN with anchors [72]	57.7	14.3	15.4	22.1	26.0	31.7	8.1	35.1	11.9	3.1	22.6
2016	RPN and upscaling [52]	56.8	16.4	**31.1**	**29.4**	**31.9**	29.4	**23.4**	31.3	9.3	4.2	24.8

Table 28
Detection performance on the very small subset of USC-GRAD-STDdb database. Average Precision when IoU is at least 0.5 ($AP_{0.5}$) and Average Precision when the IoU goes from 0.5 to 0.95 in 5% steps (AP_{all}). The best result is highlighted in bold. (in %)

Year	Method	AP_{all}	$AP_{0.5}$
2021	STDnet-ST [91]	**21.4**	**63.4**
2020	STDnet [53]	18.9	59.1
2018	RCN [73]	18.3	57.8
2018	Cascade-FPN [135]	17.4	55.9
2017	FPN [122]	17.3	54.5

Also, it is not possible to give every possible labelled example related to a problem domain in the real world scenario. Therefore, Yao et al. [190] only adopt image-level annotations to learn the object detectors through a dynamic curriculum learning scheme. As weakly supervised information, image-level annotations (category information) are help for the classification of objects. Besides, only using bounding boxes to train the object detection model maybe enhance the localization of objects. The clever usage of these two kinds of weakly-supervised details may improve the detection performance of small or tiny objects better.

(3) High accuracy or real-time small/tiny object detection

How to make a trade-off between the accuracy and inference speed is of importance in small object detection. Actually, it is up to application scenarios. Different application scenarios have different preferences. For example, high accuracy may be the first choice when tumors are observed by detecting small masses successful in medical images. Real-time detection would be the key point when techniques for small or tiny object detection are applied to autonomous driving because it is necessary to avoid obstacles in time. Furthermore, in the field of national defense and military, it simultaneously needs to have high precision and high detection speed to detect any targets. Thus, how to make a balance between the high accuracy and real-time detection is also the focus for this domain.

5. Conclusion

This review comprehensively discusses small or tiny object datasets, the definitions of small or tiny objects, techniques for small or tiny object detection, small or tiny object detection performance analysis, and promising directions of small object detection. This work brings an up-to-date and thorough survey to the small object detection community. Also, we hope this review could provide researchers guidance for further research on small object detection systems.

Declaration of Competing Interest

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] M. Najibi, P. Samangouei, R. Chellappa, L.S. Davis, "SSH: Single Stage Headless Face Detector," presented at the IEEE International Conference on Computer Vision, Venice, Italy 2017.
[2] P. Samangouei, R. Chellappa, M. Najibi, L.S. Davis, "Face-MagNet: Magnifying Feature Maps to Detect Small Faces," presented at the IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 2018.
[3] Z. Liu, D. Li, S.S. Ge, F. Tian, Small Traffic Sign Detection from Large Image, Appl. Intell. 50 (1) (2020) 1–13.
[4] L. Liu, et al., Deep-learning and depth-map based approach for detection and 3-D localization of small traffic signs, IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 13 (2020) 2096–2111.
[5] T. Song, L. Sun, D. Xie, H. Sun, S. Pu, "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation," presented at the European Conference on Computer Vision, Munich, Germany, 2018.
[6] A. Wolpert, M. Teutsch, M.S. Sarfraz, R. Stiefelhagen, "Anchor-free small-scale multispectral pedestrian detection," presented at the British Machine Vision Conference, Virtual Event, UK, 2020.
[7] W. Zhang, S. Wang, S. Thachan, J. Chen, Y. Qian, "Deconv R-CNN for small object detection on remote sensing images," presented at the IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 2018.
[8] X. Zhang, E. Izquierdo, K. Chandramouli, "Dense and small object detection in UAV Vision based on cascade network," presented at the IEEE International Conference on Computer Vision Workshops, Seoul, South Korea, 2019.

[9] W. Sun, D. Yan, J. Huang, C. Sun, Small-scale moving target detection in aerial image by deep inverse reinforcement learning, Soft. Comput. 24 (8) (2020) 5897–5908.
[10] G. Cheng, et al., Dual-aligned oriented detector, IEEE Trans. Geosci. Remote Sens. 60 (2022) 5618111.
[11] X.Q. Zhou, Y.X. Zou, Y. Wang, "Accurate small object detection via density map aided saliency estimation," presented at the IEEE International Conference on Image Processing, Beijing, China, 2017.
[12] M. Menikdiwela, C.V. Nguyen, H. Li, M. Shaw, "CNN-based Small Object Detection and Visualization with Feature Activation Mapping," presented at the International Conference on Image and Vision Computing New Zealand, Christchurch, New Zealand, 2017.
[13] C. Eggert, S. Brehm, A. Winschel, D. Zecha, R. Lienhart, "A Closer Look: Small Object Detection in Faster R-CNN," presented at the International Conference on Multimedia and Expo, Hong Kong, China, 2017.
[14] P. Pham, D. Nguyen, T. Do, T.D. Ngo, D.-D. Le, "Evaluation of Deep Models for Real-Time Small Object Detection," presented at the International Conference on Neural Information Processing, Guangzhou, China, 2017.
[15] Z. Yang, W. Xu, Z. Wang, X. He, F. Yang, Z. Yin, "Combining Yolov3-tiny Model with Dropblock for Tiny-face Detection," presented at the IEEE International Conference on Communication Technology, Xi'an, China, 2019.
[16] L. Fang, X. Zhao, S. Zhang, Small-objectness sensitive detection based on shifted single shot detector, Multimed. Tools Appl. 78 (10) (2019) 13227–13245.
[17] B.A. Mudassar, S. Mukhopadhyay, "Rethinking Convolutional Feature Extraction for Small Object Detection," presented at the British Machine Vision Conference, Cardiff, UK, 2019.
[18] F.O. Unel, B.O. Özkalayci, C. Çigla, "The Power of Tiling for Small Object Detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 2019.
[19] C. Cao, et al., An improved faster R-CNN for small object detection, IEEE Access 7 (2019) 106838–106846.
[20] Y. Zhang, M. Wang, Z. Li, "An Efficient Object Detection Framework with Modified Dense Connections for Small Objects Optimizations," presented at the ACM International Conference on Computing Frontiers, Catania, Sicily, Italy, 2020.
[21] G. Lee, S. Hong, D. Cho, Self-supervised feature enhancement networks for small object detection in noisy images, IEEE Signal Process. Lett. 28 (2021) 1026–1030.
[22] X. Yu, et al., "The 1st Tiny Object Detection Challenge: Methods and Results," presented at the European Conference on Computer Vision Workshops, Glasgow, UK, 2020.
[23] Y. Li, S. Li, H. Du, L. Chen, D. Zhang, Y. Li, YOLO-ACN: focusing on small target and occluded object detection, IEEE Access 8 (2020) 227288–227303.
[24] C. Zhang, D. He, Z. Li, Z. Wang, "Parallel Connecting Deep and Shallow CNNs for Simultaneous Detection of Big and Small Objects," presented at the Chinese Conference on Pattern Recognition and Computer Vision, Guangzhou, China, 2018.
[25] B. Wu, F.N. Iandola, P.H. Jin, K. Keutzer, "SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving," presented at the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 2017.
[26] H. Levi and S. Ullman, "Efficient Coarse-to-Fine Non-Local Module for the Detection of Small Objects," presented at the British Machine Vision Conference, Cardiff, UK, 2019.
[27] Z.Q. Zhao, P. Zheng, S.T. Xu, X. Wu, Object detection with deep learning: a review, IEEE Trans. Neural Networks Learn. Syst. 30 (11) (2019) 3212–3232.
[28] Z. Zou, Z. Shi, Y. Guo, J. Ye, "Object Detection in 20 Years: A Survey," *arXiv: 1905.05055v2*, 2019 1–39.
[29] L. Liu, et al., Deep learning for generic object detection: a survey, Int. J. Comput. Vis. 128 (2) (2020) 261–318.
[30] X. Wu, D. Sahoo, S.C.H. Hoi, Recent advances in deep learning for object detection, Neurocomputing 396 (2020) 39–64.
[31] K. Oksuz, B. Cam, S. Kalkan, E. Akbas, Imbalance problems in object detection: a review, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2021) 3388–3415.
[32] D.-N. Nguyen, T. Do, T.D. Ngo, D.-D. Le, An evaluation of deep learning methods for small object detection, J. Electrical Comput. Eng. 2020 (27) (2020) 3189691–3189709.
[33] K. Tong, Y. Wu, F. Zhou, Recent advances in small object detection based on deep learning: a review, Image Vis. Comput. 97 (2020) Art. no. 103910.
[34] G. Chen, et al., A survey of the four pillars for small object detection: multiscale representation, contextual information, super-resolution, and region proposal, IEEE Trans. Syst. Man Cybernet. Syst. (2020) 1–18.
[35] Y. Liu, P. Sun, N. Wergeles, YiShang, A survey and performance evaluation of deep learning methods for small object detection, Expert Syst. Appl. 172 (6) (2021) 1–14 Art. no. 114602.
[36] R. Girshick, "Fast R-CNN," presented at the IEEE International Conference on Computer Vision Santiago, Chile 2015.
[37] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," presented at the Advances in Neural Information Processing Systems, Montreal, Quebec, Canada, 2015.
[38] T.-Y. Lin, P. Goyal, R.B. Girshick, K. He, P. Dollár, "Focal Loss for Dense Object Detection," presented at the IEEE International Conference on Computer Vision, Venice, Italy, 2017.
[39] J. Redmon, A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv: 1804.02767v1*, 2018.
[40] T.Y. Lin, et al., "Microsoft COCO: Common Objects in Context," European conference on computer vision, 2014 740–755.
[41] 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.
[42] C. Wojek, P. Dollar, B. Schiele, P. Perona, Pedestrian detection: an evaluation of the state of the art, IEEE Trans. Pattern Analysis Machine Intelligence 34 (4) (2012) 743–761.
[43] A. Geiger, P. Lenz, R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," presented at the Computer Vision and Pattern Recognition, 2012.
[44] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, S. Hu, "Traffic-sign detection and classification in the wild," presented at the Computer Vision and Pattern Recognition, 2016.
[45] W. Liu, et al., "SSD: Single Shot MultiBox Detector," presented at the European Conference on Computer Vision, Amsterdam, The Netherlands, 2016.
[46] G.-S. Xia, et al., "DOTA: A Large-Scale Dataset for Object Detection in Aerial Images," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.
[47] D. Du, et al., "The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking," presented at the European Conference on Computer Vision, Munich, Germany, 2018.
[48] J. Wang, W. Yang, H. Guo, R. Zhang, G.-S. Xia, "Tiny Object Detection in Aerial Images," presented at the International Conference on Pattern Recognition Milan, Italy, 2020.
[49] K. Li, G. Wan, G. Cheng, L. Meng, J. Han, Object detection in optical remote sensing images: a survey and a new benchmark, ISPRS J. Photogramm. Remote Sens. 159 (2020) 296–307.
[50] X. Yu, Y. Gong, N. Jiang, Q. Ye, Z. Han, "Scale Match for Tiny Person Detection," presented at the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 2020.
[51] S. Yang, P. Luo, C.C. Loy, X. Tang, "WIDER FACE: A Face Detection Benchmark," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016.
[52] C. Chen, M.-Y. Liu, O. Tuzel, J. Xiao, "R-CNN for Small Object Detection," presented at the Asian Conference on Computer Vision, Taipei, Taiwan, 2016.
[53] B. Bosquet, M. Mucientes, V.M. Brea, STDnet: exploiting high resolution feature maps for small object detection, Eng. Appl. Artif. Intell. 91 (2020) Art. no. 103615.
[54] Y. Li, Q. Huang, X. Pei, Y. Chen, L. Jiao, R. Shang, Cross-layer attention network for small object detection in remote sensing imagery, IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 14 (2021) 2148–2161.
[55] Z.-Z. Wang, K. Xie, X.-Y. Zhang, H.-Q. Chen, C. Wen, J.-B. He, Small-object detection based on YOLO and dense block via image super-resolution, IEEE Access 9 (2021) 56416–56429.
[56] S. Zhang, R. Benenson, B. Schiele, "CityPersons: A Diverse Dataset for Pedestrian Detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017.
[57] Z. Ji, Q. Kong, H. Wang, Y. Pang, "Small and Dense Commodity Object Detection with Multi-Scale Receptive Field Attention," presented at the ACM International Conference on Multimedia, Nice, France, 2019.
[58] P. Zhu, L. Wen, D. Du, X. Bian, Q. Hu, H. Ling, "Vision Meets Drones: Past, Present and Future," *arXiv:* 2001.*06303v2*, 2020.
[59] D. Du, et al., "VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results," presented at the IEEE International Conference on Computer Vision Workshops, Seoul, South Korea, 2019.
[60] M. Braun, S. Krebs, F. Flohr, D.M. Gavrila, "The EuroCity Persons Dataset: A Novel Benchmark for Object Detection," *arXiv:*1805.*07193v2*, 2018.
[61] L. Tuggener, I. Elezi, J. Schmidhuber, M. Pelillo, T. Stadelmann, "DeepScores-A Dataset for Segmentation, Detection and Classification of Tiny Objects," presented at the International Conference on Pattern Recognition, Beijing, China, 2018.
[62] P. Fang, Y. Shi, "Small Object Detection Using Context Information Fusion in Faster R-CNN," presented at the IEEE International Conference on Computer and Communications, Chengdu, China, 2018.
[63] M. Cordts, et al., "The Cityscapes Dataset for Semantic Urban Scene Understanding," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016.
[64] K. Behrendt, L. Novak, R. Botros, "A Deep Learning Approach to Traffic Lights: Detection, Tracking, and Classification," presented at the IEEE International Conference on Robotics and Automation, Singapore, 2017.
[65] 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.
[66] P. Pinggera, S. Ramos, S. Gehrig, U. Franke, C. Rother, R. Mester, "Lost and Found: Detecting Small Road Hazards for Self-driving Vehicles," presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems, Daejeon, South Korea, 2016.
[67] A. Robicquet, A. Sadeghian, A. Alahi, S. Savarese, "Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes," presented at the European Conference on Computer Vision, Amsterdam, The Netherlands, 2016.
[68] K. Liu, G. Máttyus, Fast multiclass vehicle detection on aerial images, IEEE Geosci. Remote Sens. Lett. 12 (9) (2015) 1938–1942.
[69] J. Yan, X. Zhang, Z. Lei, S.Z. Li, Face detection by structural models, Image Vis. Comput. 32 (10) (2014) 790–799.
[70] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, C. Igel, "Detection of Traffic Signs in Real-world Images: The German Traffic Sign Detection Benchmark," presented at the International Joint Conference on Neural Networks, Dallas, TX, USA, 2013.
[71] A. Møgelmose, 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.
[72] H. Krishna, C.V. Jawahar, "Improving Small Object Detection," presented at the Asian Conference on Pattern Recognition, Nanjing, China, 2017.

[73] B. Bosquet, M. Mucientes, V.M. Brea, "STDnet: A ConvNet for Small Target Detection," presented at the British Machine Vision Conference, Newcastle, UK, 2018.
[74] I.J. Goodfellow, et al., "Generative Adversarial Nets," presented at the Neural Information Processing Systems, Montreal, Quebec, Canada, 2014.
[75] C. Ledig, et al., "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017.
[76] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, S. Yan, "Perceptual Generative Adversarial Networks for Small Object Detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017.
[77] Z. Yang, et al., "Prior Knowledge Guided Small Object Detection on High-Resolution Images," presented at the IEEE International Conference on Image Processing, Taipei, Taiwan, 2019.
[78] Y. Pang, J. Cao, J. Wang, J. Han, JCS-Net: joint classification and super-resolution network for small-scale pedestrian detection in surveillance images, IEEE Trans. Informat Forensics Security 14 (12) (2019) 3322–3331.
[79] P. Dollár, Z. Tu, P. Perona, S.J. Belongie, "Integral Channel Features," presented at the British Machine Vision Conference, London, UK, 2009.
[80] J. Cao, Y. Pang, X. Li, Learning multilayer channel features for pedestrian detection, IEEE Trans. Image Process. 26 (7) (2017) 3210–3220.
[81] Y. Bai, Y. Zhang, M. Ding, B. Ghanem, "Finding Tiny Faces in the Wild With Generative Adversarial Network," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.
[82] D. Liu, Z.-Q. Zhao, W. Tian, "TFPGAN: Tiny Face Detection with Prior Information and GAN," presented at the International Conference on Intelligent Computing, Bari, Italy, 2020.
[83] Y. Zhang, Y. Bai, M. Ding, B. Ghanem, Multi-task generative adversarial network for detecting small objects in the wild, Int. J. Comput. Vis. 128 (6) (2020) 1810–1828.
[84] J. Noh, W. Bae, W. Lee, J. Seo, G. Kim, "Better to Follow, Follow to Be Better: Towards Precise Supervision of Feature Super-Resolution for Small Object Detection," presented at the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019.
[85] Z. Liu, G. Gao, L. Sun, Z. Fang, "HRDNet: High-resolution Detection Network for Small Objects," presented at the IEEE International Conference on Multimedia and Expo, Shenzhen, China, 2021.
[86] J. Pang, C. Li, J. Shi, Z. Xu, H. Feng, R2-CNN: fast tiny object detection in large-scale remote sensing images, IEEE Trans. Geosci. Remote Sens. 57 (8) (2019) 5512–5524.
[87] Y. Gu, J. Li, C. Wu, W. Jia, J. Chen, "Small Object Detection by Generative and Discriminative Learning," presented at the International Conference on Pattern Recognition, Milan, Italy, 2020.
[88] H. Ji, Z. Gao, X. Liu, Y. Zhang, T. Mei, "Small Object Detection Leveraging on Simultaneous Super-resolution," presented at the International Conference on Pattern Recognition, Milan, Italy, 2020.
[89] C. Yang, Z. Huang, N. Wang, "QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection," *arXiv:2103.09136v1*, 2021.
[90] B. Bosquet, M. Mucientes, V.M. Brea, "Correlation-based ConvNet for Small Object Detection in Videos," presented at the International Conference on Pattern Recognition, Milan, Italy, 2020.
[91] B. Bosquet, M. Mucientes, V.M. Brea, STDnet-ST: spatio-temporal ConvNet for small object detection, Pattern Recogn. 116 (2021) Art. no. 107929.
[92] 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.
[93] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, "Path Aggregation Network for Instance Segmentation," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.
[94] A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, YOLOv4: optimal speed and accuracy of object detection, 2020 Art. no. arXiv:2004.10934.
[95] A. Oliva, A. Torralba, The Role of Context in Object Recognition, Trends Cogn. Sci. 11 (12) (2007) 520–527.
[96] P. Hu, D. Ramanan, "Finding Tiny Faces," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017.
[97] C. Zhang, T. Li, S. Guo, N. Li, Y. Gao, K. Wang, "Aggregation Connection Network For Tiny Face Detection," presented at the IEEE International Joint Conference on Neural Network, Budapest, Hungary, 2019.
[98] X. Tang, D.K. Du, Z. He, J. Liu, "PyramidBox: A Context-Assisted Single Shot Face Detector," presented at the European Conference on Computer Vision, Munich, Germany, 2018.
[99] Z. Li, X. Tang, J. Han, J. Liu, R. He, "PyramidBox++: High Performance Detector for Finding Tiny Face," *arXiv:1904.00386v1*, 2019.
[100] Y. Xi, J. Zheng, X. He, W. Jia, H. Li, "Beyond Context: Exploring Semantic Similarity for Tiny Face Detection," presented at the IEEE International Conference on Image Processing, Athens, Greece, 2018.
[101] Y. Xi, et al., Beyond Context: Exploring Semantic Similarity for Small Object Detection in Crowded Scenes, Pattern Recogn. Lett. 137 (2020) 53–60.
[102] J. Leng, Y. Ren, W. Jiang, X. Sun, Y. Wang, Realize your surroundings: exploiting context information for small object detection, Neurocomputing 433 (2021) 287–299.
[103] K. Fu, J. Li, L. Ma, K. Mu, Y. Tian, "Intrinsic Relationship Reasoning for Small Object Detection," *arXiv:2009.00833v1*, 2020.
[104] J.-S. Lim, M. Astrid, H.-J. Yoon, S.-I. Lee, "Small Object Detection using Context and Attention," presented at the International Conference on Artificial Intelligence in Information and Communication, Jeju Island, South Korea, 2021.
[105] Z. Yan, H. Zheng, Y. Li, L. Chen, Detection-oriented backbone trained from near scratch and local feature refinement for small object detection, Neural. Process. Lett. 53 (3) (2021) 1921–1943.
[106] J. Liu, et al., Multi-component fusion network for small object detection in remote sensing images, IEEE Access 7 (2019) 128339–128352.
[107] S. Yang, et al., "Inception Parallel Attention Network for Small Object Detection in Remote Sensing Images," presented at the Chinese Conference on Pattern Recognition and Computer Vision, Nanjing, China, 2020.
[108] X. Liang, J. Zhang, L. Zhuo, Y. Li, 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. 30 (6) (2020) 1758–1770.
[109] G. Cheng, C. Lang, M. Wu, X. Xie, X. Yao, J. Han, Feature enhancement network for object detection in optical remote sensing images, J. Remote Sens. 2021 (2021) Art. no. 9805389.
[110] L. Guan, Y. Wu, J. Zhao, SCAN: semantic context aware network for accurate small object detection, Int. J. Comput. Intelligence Syst. 11 (1) (2018) 936–950.
[111] L. Cui, et al., Context-aware block net for small object detection, IEEE Trans. Cybernet. (2020) 1–14.
[112] M. Hong, S. Li, Y. Yang, F. Zhu, Q. Zhao, L. Lu, "SSPNet: Scale Selection Pyramid Network for Tiny Person Detection from UAV Images," *arXiv:2107.01548v1*, 2021.
[113] Z. Liu, G. Gao, L. Sun, L. Fang, "IPG-Net: Image Pyramid Guidance Network for Small Object Detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 2020.
[114] J. Dai, Y. Li, K. He, J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks," presented at the Advances in Neural Information Processing Systems, Barcelona, Spain 2016.
[115] G. Cao, X. Xie, W. Yang, Q. Liao, G. Shi, J. Wu, "Feature-Fused SSD: Fast Detection for Small Objects," presented at the 9th International Conference on Graphic and Image Processing, Qindao, China, 2017.
[116] M. Xu, et al., MDSSD: multi-scale deconvolutional single shot detector for small objects, SCIENCE CHINA Inf. Sci. 63 (2) (2020) 120113.
[117] C. Sun, Y. Ai, S. Wang, W. Zhang, Mask-guided SSD for small-object detection, Appl. Intell. 51 (6) (2021) 3311–3322.
[118] S. Liu, D. Huang, Y. Wang, "Receptive Field Block Net for Accurate and Fast Object Detection," presented at the European Conference on Computer Vision, Munich, Germany, 2018.
[119] Y. Li, Y. Chen, N. Wang, Z.-X. Zhang, "Scale-Aware Trident Networks for Object Detection," presented at the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019.
[120] M. Razaak, H. Kerdegari, V. Argyriou, P. Remagnino, "Multi-scale Feature Fused Single Shot Detector for Small Object Detection in UAV Images," presented at the International Conference on Computer Vision Systems, Thessaloniki, Greece, 2019.
[121] S. Han, J. Kwon, S. Kwon, "Real-time Small Object Detection Model in the Bird-view UAV Imagery," presented at the International Conference on Vision, Image and Signal Processing, Vancouver, BC, Canada, 2019.
[122] T.-Y. Lin, P. Dollár, R.B. Girshick, K. He, B. Hariharan, S.J. Belongie, "Feature Pyramid Networks for Object Detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017.
[123] J. Qu, C. Su, Z. Zhang, A. Razi, Dilated convolution and feature fusion SSD network for small object detection in remote sensing images, IEEE Access 8 (2020) 82832–82843.
[124] Y. Liu, F. Yang, P. Hu, Small-object detection in UAV-captured images via multi-branch parallel feature pyramid networks, IEEE Access 8 (2020) 145740–145750.
[125] Q. Zheng, Y. Chen, Feature pyramid of bi-directional stepped concatenation for small object detection, Multimed. Tools Appl. 80 (13) (2021) 20283–20305.
[126] G. Cheng, M. He, H. Hong, X. Yao, X. Qian, L. Guo, Guiding clean features for object detection in remote sensing images, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5.
[127] Z. Liang, J. Shao, D. Zhang, L. Gao, "Small Object Detection Using Deep Feature Pyramid Networks," presented at the Pacific-Rim Conference on Multimedia, Hefei, China, 2018.
[128] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, D. Lin, "Libra R-CNN: Towards balanced learning for object detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[129] Q. Zheng, Y. Chen, Interactive multi-scale feature representation enhancement for small object detection, Image Vis. Comput. 108 (2021) Art. no. 104128.
[130] P.-Y. Chen, J.-W. Hsieh, C.-Y. Wang, H.-Y.M. Liao, "Recursive Hybrid Fusion Pyramid Network for Real-Time Small Object Detection on Embedded Devices," presented at the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 2020.
[131] J. Liu, Y. Gu, S. Han, Z. Zhang, J. Guo, X. Cheng, Feature rescaling and fusion for tiny object detection, IEEE Access 9 (2021) 62946–62955.
[132] Y. Gong, X. Yu, Y. Ding, X. Peng, J. Zhao, Z. Han, "Effective Fusion Factor in FPN for Tiny Object Detection," presented at the IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2021.
[133] C. Zhu, R. Tao, K. Luu, M. Savvides, "Seeing Small Faces From Robust Anchor's Perspective," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.
[134] X. Yang, et al., "SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects," presented at the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019.
[135] Z. Cai, N. Vasconcelos, "Cascade R-CNN: Delving Into High Quality Object Detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.

[136] G. Wang, Z. Xiong, D. Liu, C. Luo, "Cascade Mask Generation Framework for Fast Small Object Detection," presented at the IEEE International Conference on Multimedia and Expo, San Diego, CA, USA, 2018.
[137] C. Eggert, D. Zecha, S. Brehm, R. Lienhart, "Improving small object proposals for company logo detection," presented at the International Conference on Multimedia Retrieval, Bucharest, Romania, 2017.
[138] S. Zhang, X. Yang, Y. Liu, C. Xu, Asymmetric multi-stage CNNs for small-scale pedestrian detection, Neurocomputing 409 (2020) 12–26.
[139] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, S.Z. Li, "FaceBoxes: A CPU Real-time Face Detector with High Accuracy," presented at the IEEE International Joint Conference on Biometrics, Denver, CO, USA, 2017.
[140] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, S.Z. Li, "S3FD: Single Shot Scale-Invariant Face Detector," presented at the IEEE International Conference on Computer Vision, Venice, Italy, 2017.
[141] Z. Zhang, W. Shen, S. Qiao, Y. Wang, B. Wang, A.L. Yuille, "Robust Face Detection via Learning Small Faces on Hard Images," presented at the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 2020.
[142] T. Yang, X. Zhang, Z. Li, W. Zhang, J. Sun, "MetaAnchor: Learning to Detect Objects with Customized Anchors," presented at the Advances in Neural Information Processing Systems, Montréal, Canada, 2018.
[143] J. Wang, K. Chen, S. Yang, C.C. Loy, D. Lin, "Region Proposal by Guided Anchoring," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019.
[144] K. Duan, D. Du, H. Qi, Q. Huang, Detecting small objects using a channel-aware deconvolutional network, IEEE Trans. Circuits Syst. Video Technol. 30 (6) (2020) 1639–1652.
[145] H. Law, J. Deng, "CornerNet: Detecting Objects as Paired Keypoints," presented at the European Conference on Computer Vision, Munich, Germany, 2018.
[146] X. Lu, B. Li, Y. Yue, Q. Li, J. Yan, "Grid R-CNN," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019.
[147] H. Law, Y. Teng, O. Russakovsky, J. Deng, "CornerNet-Lite: Efficient Keypoint based Object Detection," presented at the British Machine Vision Conference, Virtual Event, UK, 2020.
[148] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, "CenterNet: Keypoint Triplets for Object Detection," presented at the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019.
[149] X. Zhou, J. Zhuo, P. Krähenbühl, "Bottom-Up Object Detection by Grouping Extreme and Center Points," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019.
[150] X. Zhou, D. Wang, P. Krähenbühl, "Objects as Points," *arXiv:*1904.*07850v2*, 2019.
[151] Z. Yang, S. Liu, H. Hu, L. Wang, S. Lin, "RepPoints: Point Set Representation for Object Detection," presented at the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019.
[152] Z. Tian, C. Shen, H. Chen, T. He, "FCOS: Fully Convolutional One-Stage Object Detection," presented at the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019.
[153] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, J. Shi, FoveaBox: beyound anchor-based object detection, IEEE Trans. Image Process. 29 (2020) 7389–7398.
[154] S. Luo, X. Li, R. Zhu, X. Zhang, SFA: small faces attention face detector, IEEE Access 7 (2019) 171609–171620.
[155] C. Gao, W. Tang, L. Jin, Y. Jun, "Exploring Effective Methods to Improve the Performance of Tiny Object Detection," presented at the European Conference on Computer Vision Workshops, Glasgow, UK, 2020.
[156] K. Chen, et al., MMDetection: Open MMLab detection toolbox and benchmark, Comput. Res. Reposit. (2019)https://doi.org/10.48550/arXiv.1906.07155 Technical report of MMDetection.
[157] N. Bodla, B. Singh, R. Chellappa, L.S. Davis, "Soft-NMS - Improving Object Detection with One Line of Code," presented at the IEEE International Conference on Computer Vision, Venice, Italy, 2017.
[158] Y. Feng, et al., "Effective Feature Enhancement and Model Ensemble Strategies in Tiny Object Detection," presented at the European Conference on Computer Vision Workshops, Glasgow, UK, 2020.
[159] B. Singh, L.S. Davis, "An Analysis of Scale Invariance in Object Detection SNIP," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.
[160] B. Singh, M. Najibi, L.S. Davis, "SNIPER: Efficient Multi-Scale Training," presented at the Advances in Neural Information Processing Systems, Montréal, Canada, 2018.
[161] Y. Kim, B.-N. Kang, D. Kim, "SAN: Learning Relationship Between Convolutional Features for Multi-scale Object Detection," presented at the European Conference on Computer Vision, Munich, Germany, 2018.
[162] D. Zhou, X. Zhou, H. Zhang, S. Yi, W. Ouyang, "Cheaper Pre-training Lunch: An Efficient Paradigm for Object Detection," presented at the European Conference on Computer Vision, Glasgow, UK, 2020.
[163] Y. Chen, et al., "Dynamic Scale Training for Object Detection," *arXiv:*2004.*12432v2*, 2021.
[164] P. Kaur, B.S. Khehra, B.S. Mavi, "Data Augmentation for Object Detection: A Review," presented at the IEEE International Midwest Symposium on Circuits and Systems, Lansing, MI, USA, 2021.
[165] S. Yun, D. Han, S. Chun, S.J. Oh, Y. Yoo, J. Choe, "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features," presented at the IEEE International Conference on Computer Vision, Seoul, South Korea, 2019.
[166] B. Zoph, E.D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, Q.V. Le, "Learning Data Augmentation Strategies for Object Detection," presented at the European Conference on Computer Vision, Glasgow, UK, 2020.
[167] E.D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, Q.V. Le, "AutoAugment: Learning Augmentation Policies from Data," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019.
[168] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, K. Cho, "Augmentation for Small Object Detection," presented at the The 9th International Conference on Advances in Computing and Information Technology, Sydney, Australia, 2019.
[169] C. Chen, et al., "RRNet: A Hybrid Detector for Object Detection in Drone-Captured Images," presented at the IEEE International Conference on Computer Vision Workshops, Seoul, South Korea, 2019.
[170] Y. Chen, et al., Stitcher: feedback-driven data provider for object detection, Comput. Res. Reposit. (2020) arXiv:2004.12432v1.
[171] N. Jiang, X. Yu, X. Peng, Y. Gong, Z. Han, "SM+: Refined Scale Match for Tiny Person Detection," presented at the IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 2021.
[172] B. Han, Y. Wang, Z. Yang, X. Gao, Small-scale pedestrian detection based on deep neural Network, IEEE Trans. Intell. Transp. Syst. 21 (7) (2020) 3046–3055.
[173] J. Wu, C. Zhou, Q. Zhang, M. Yang, J. Yuan, "Self-Mimic Learning for Small-scale Pedestrian Detection," presented at the ACM International Conference on Multimedia, Seattle, WA, USA, 2020.
[174] K. Chen, et al., "Towards Accurate One-Stage Object Detection With AP-Loss," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019.
[175] Q. Qian, L. Chen, H. Li, R. Jin, "DR Loss: Improving Object Detection by Distributional Ranking," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2020.
[176] H. Zhang, H. Chang, B. Ma, S. Shan, X. Chen, "Cascade RetinaNet: Maintaining Consistency for Single-Stage Object Detection," presented at the British Machine Vision Conference, Cardiff, UK, 2019.
[177] Z. Wang, J. Fang, J. Dou, J. Xue, "Small Object Detection on Road by Embedding Focal-Area Loss," presented at the 10th International Conference on Image and Graphics, Beijing, China, 2019.
[178] X. Li, et al., "Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection," presented at the Advances in Neural Information Processing Systems virtual, 2020.
[179] X. Li, W. Wang, X. Hu, J. Li, J. Tang, J. Yang, "Generalized Focal Loss V2: Learning Reliable Localization Quality Estimation for Dense Object Detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition, 2021.
[180] J. Yu, Y. Jiang, Z. Wang, Z. Cao, T.S. Huang, "UnitBox: An Advanced Object Detection Network," presented at the ACM Conference on Multimedia Conference, Amsterdam, The Netherlands, 2016.
[181] L. Tychsen-Smith, L. Petersson, "Improving Object Localization With Fitness NMS and Bounded IoU Loss," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.
[182] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I.D. Reid, S. Savarese, "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019.
[183] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," presented at the AAAI Conference on Artificial Intelligence, New York, NY, USA, 2020.
[184] Y. He, C. Zhu, J. Wang, M. Savvides, X. Zhang, "Bounding Box Regression With Uncertainty for Accurate Object Detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019.
[185] D. Xu, J. Guan, P. Feng, W. Wang, Association loss for visual object detection, IEEE Signal Process. Lett. 27 (2020) 1435–1439.
[186] G. Liu, J. Han, W. Rong, Feedback-driven loss function for small object detection, Image Vis. Comput. 111 (2021) Art. no. 104197.
[187] G. Cheng, Y. Si, H. Hong, X. Yao, L. Guo, Cross-scale feature fusion for object detection in optical remote sensing images, IEEE Geosci. Remote Sens. Lett. 18 (3) (2021) 431–435.
[188] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, J. Feng, "Dual Path Networks," presented at the Neural Information Processing Systems, 2017.
[189] J. Wang, C. Xu, W. Yang, L. Yu, "A Normalized Gaussian Wasserstein Distance for Tiny Object Detection," *arXiv:2110.13389v1*, 2021.
[190] X. Yao, X. Feng, J. Han, G. Cheng, L. Guo, Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning, IEEE Trans. Geosci. Remote Sens. 59 (1) (2021) 675–685.