Mathematical Biosciences and Engineering

MBE, 20(4): 6551–6590.
DOI: 10.3934/mbe.2023282
Received: 18 October 2022
Revised: 21 December 2022
Accepted: 26 December 2022
Published: 02 February 2023

http://www.aimspress.com/journal/MBE

Survey

Deep learning-based small object detection: A survey

Qihan Feng[1], Xinzheng Xu[1] and Zhixiao Wang[1,2,*]

[1] College of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

[2] Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou 221116, China

* **Correspondence:** Email: zhxwang@cumt.edu.cn.

Abstract: Small object detection (SOD) is significant for many real-world applications, including criminal investigation, autonomous driving and remote sensing images. SOD has been one of the most challenging tasks in computer vision due to its low resolution and noise representation. With the development of deep learning, it has been introduced to boost the performance of SOD. In this paper, focusing on the difficulties of SOD, we analyze the deep learning-based SOD research papers from four perspectives, including boosting the resolution of input features, scale-aware training, incorporating contextual information and data augmentation. We also review the literature on crucial SOD tasks, including small face detection, small pedestrian detection and aerial image object detection. In addition, we conduct a thorough performance evaluation of generic SOD algorithms and methods for crucial SOD tasks on four well-known small object datasets. Our experimental results show that network configuring to boost the resolution of input features can enable significant performance gains on WIDER FACE and Tiny Person. Finally, several potential directions for future research in the area of SOD are provided.

Keywords: small object detection; deep learning; computer vision; neural network; benchmark

1. Introduction

Object detection (OD) [1–8] is an essential task that forms the basis of many other computer vision tasks, such as object tracking [9,10], instance segmentation [11–13], action recognition [14–18], environment surveillance [19–21], video checking in sports [22,23], scene understanding [24–28], etc.

Thanks to the powerful feature-learning ability of deep convolutional neural networks (CNNs) [29–33], object detection research has been experiencing rapid growth over the last decade. Deep learning-based object detection techniques can be divided into two categories: one-stage models and two-stage models. The successful two-stage models include the R-CNN [2], Fast R-CNN [3], Faster R-CNN [4], SPP-Net [34] and feature pyramid network (FPN) models [35]. These models generate a region of interest (ROI) in the first stage; they then fine-tune the ROI to classify the objects and localize the bounding box in the second stage. YOLO [6], SSD [36] and several anchor-free models, including feature selection anchor-free module (FSAF) [37], CornerNet [38], FCOS [39] and CenterNet [40], are one-stage models that directly classify and localize the object from the feature map without completing the ROI stage.

Small object detection (SOD) [41] is an emerging research area within object detection. SOD has been widely used in medical image analysis, maritime rescue, face recognition in surveillance video, drone scene analysis and others. Many promising deep learning-based SOD works have been published in recent years. Small objects can be defined in two major ways. One definition method is relative size [42], where the ratio of the object's bounding box's width and height to the image's width and height is less than 0.1, or the ratio of the object's bounding box's area to the image's area is less than 0.03; the other definition method is absolute size, where the COCO [43] dataset indicates that an object is small if its size is less than 32 × 32 pixels. Examples are shown in Figure 1. These definitions mean that the visual feature of a small object is limited.

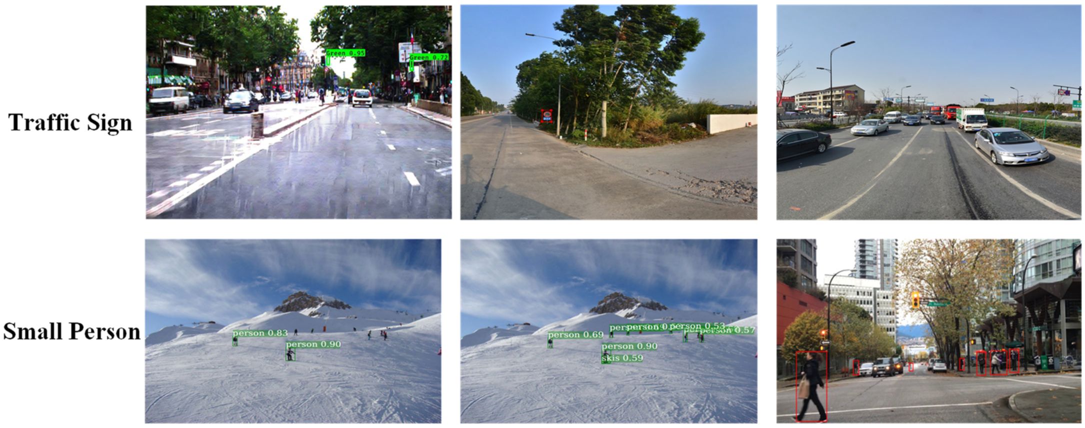

Figure 1. Small object examples.

Though large-scale datasets, such as Microsoft Common Objects in Context (MS COCO) [43], ImageNet [44], and PASCAL VOC [45] and, have contributed to the growth of object detection methods, these methods fail to accurately detect small objects. Taking Co-DETR [46], i.e., one of the state-of-art methods, as an example, the mean average precision (mAP) metric of small objects on COCO obtained by Co-DETR was only 48.4%, which significantly lags behind that of objects with medium and large sizes (67.1 and 77.3% respectively). The main reason for the poor performance for SOD is that small objects have lower resolution and occupy fewer pixels than larger objects; the spatial position information loss by performing down-sampling and a pooling operation in the convolutional networks makes it more challenging for the detection head to locate the small objects. The large scarcity of small object datasets is another obstacle to the advancement of SOD. Existing small object datasets mainly concentrate on specific scenarios; see [47] for human faces, [48–51] for pedestrians and [52–56] for traffic scenes; networks trained on them are unsuitable for general SOD. To overcome these challenges, researchers have developed a series of strategies to improve the performance of SOD. We summarize these techniques from the perspectives of boosting the resolution of input features, scale-aware training, incorporating contextual information and data augmentation. An in-depth comparative performance evaluation of these methods on well-known datasets is also used to draw meaningful conclusions.

As shown in Table 1, recent object detection reviews aim to present progress in the area of deep learning-based object detection. Zhao et al. [57] reviewed the deep learning-based object detection frameworks and mentioned the difficulties of SOD. A large-scale, freely accessible benchmark (DIOR) for object detection in optical remote sensing images has been proposed by Li et al. [58]. A thorough analysis of the imbalance issues in object detection has been provided by Oksuz et al. [59]. To improve the effectiveness of object detection-purposed deep learning-based approaches, researchers working at the intersection of deep learning and computer vision are developing multidisciplinary solutions. Current approaches suggested to deal with the issue of class-incremental object detection have been examined by Menezes et al. [60]. However, these reviews mainly focus on the object detection of regular-sized rather than small-sized objects.

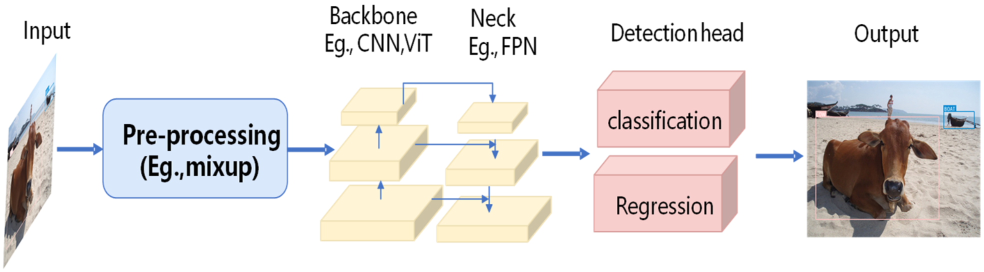

Figure 2. General framework of object detection models.

There are also recent surveys on SOD. In the review by Chen et al. [63], the four pillars of SOD are discussed. However, they did not connect the basic module design of the detector to the challenges in SOD; rather, they only reviewed studies on SOD from the viewpoint of the model framework (Figure 2), such as MMDetection [64], which divides the framework of the detector into a backbone, neck and head. The current SOD methods based on deep learning have been reviewed by Tong et al. [65] from five perspectives; they analyzed the evaluation results for two general datasets. Tong et al. limited their work to generic SOD and did not consider a model developed for SOD tasks. In addition to summarizing and contrasting current deep learning approaches for SOD, Liu et al. [66] also provided a brief overview of related methods, such as traditional object detection, face detection, picture segmentation and remote sensing images. However, they only evaluated the performance of a few networks: Faster R-CNN, SSD, YOLO and SSD. Partial performance evaluation cannot illustrate the broad picture of SOD. Tong and Wu refined and differentiated between small and tiny objects [67]. To examine this expanding field, Rekavandi et al. [68] presented a thorough analysis of more than 160 research publications released between 2017 and 2022. Other significant works include that by Cheng et al. [69], who constructed two large-scale SOD datasets, SODA-D and SODA-A, focusing on the driving and aerial scenarios, respectively. In contrast to these earlier object detection surveys, we focus on the difficulties related to SOD, investigate recent deep learning-based SOD algorithms and thus present a taxonomy to illustrate the novel strategies developed to improve SOD performance. In addition to providing an in-depth description of deep learning-based SOD algorithms developed in three areas, our study also offers meaningful comparisons of the associated experimental results.

Table 1. Summary of related reviews.

Title	Publication	Strengths	Limitations
Object detection with deep learning: A review [57]	TNNLS 2019	It reviews the deep learning-based object detection models and the difficulties of SOD.	These reviews offer a thorough summary of object detection. However, they concentrate on regular-sized object detection rather than small objects.
Object detection in optical remote sensing images: A survey and a new benchmark [58]	ISPRS 2020	It constructs DIOR, a large dataset of remote sensing.	
Imbalance problems in object detection: A review [59]	arXiv 2020	It reviews the imbalance problem of object detection.	
Continual object detection: A review of definitions, strategies, and challenges [60]	arXiv 2022	This survey investigates continual object detection.	
New generation deep learning for video object detection: A survey [61]	TNNLS 2022	It systematizes the latest video object detection models and analyzes the performance of these models on two datasets.	
A survey of deep learning-based object detection [62]	IEEE Access 2022	It reviews detection methods, general datasets and typical applications.	
A survey of the four pillars for small object detection: Multi-scale representation, contextual information, super-resolution, and region proposal [64]	TSMC 2020	It discusses the four pillars of SOD and reports on the performance of SOD on three datasets.	These studies do not contain a complete assessment of the most recent SOD approaches.
Recent advances in small object detection based on deep learning: A review [65]	IVC 2020	It reviews the SOD from five perspectives and analyzes the evaluation results for two general datasets.	
A survey and performance evaluation of deep learning methods for small object detection [66]	ESWA 2021	The solutions are summarized for the four challenges of SOD and some experiment analyses are provided.	It only analyzes the performance of three classical object detection algorithms (Faster R-CNN, SSD, YOLO).
Deep learning-based detection from the perspective of small or tiny objects: A survey [67]	IVC 2022	Aims to discuss small- or tiny-object datasets, detection techniques and the performance of these techniques.	These surveys systematically reviewed the development of SOD. Nevertheless, they all lack a comprehensive review of techniques deliberately designed for critical SOD tasks.
A guide to image and video based small object detection using deep learning: Case study of maritime surveillance [68]	arXiv 2022	Reviews the SOD methods and investigates the performance of SOD in maritime environments.	
Towards large-scale small object detection: Survey and benchmarks [69]	arXiv 2022	It presents a detailed study of SOD and yields two large-scale benchmarks for a driving scenario and aerial scene.	

Continued on next page

Title	Publication	Strengths	Limitations
Deep learning based small object detection: A survey	**Ours**	We comprehensively discuss the definition of small objects, the challenges encountered in detecting small objects, the strengths and weaknesses of generic SOD algorithms and three crucial SOD tasks. We also analyze the performance of SOD on three datasets and summarize meaningful conclusions.	

In summary, our contributions are as follows:

1) Systematic overview of deep learning-based SOD algorithms. We analyze state-of-the-art deep learning-based SOD algorithms in accordance with the challenges in SOD, and we provide a taxonomy that summarizes the strategies for improving SOD performance from the perspective of boosting the resolution of input features, scale-aware training, incorporating contextual information and data augmentation. Additionally, we provide a thorough review of the methods of crucial SOD tasks, including small face detection, small pedestrian recognition and aerial image detection.

2) Performance evaluation of SOTA deep learning-based SOD algorithms. We not only analyze the performance of generic SOD methods with the general large-scale dataset, but we also evaluate the performance of state-of-the-art SOD methods on three crucial SOD tasks, including small face detection, small pedestrian detection and aerial image detection.

3) Finally, according to the taxonomy methods and performance analysis of SOD, we discuss potential directions for future research, including suitable metrics for SOD optimization, weakly supervised SOD methods, multi-task joint optimization, and open world or few-shot SOD.

The remainder of the paper is organized as follows. The generic SOD algorithms are discussed in Section 2. Section 3 summarizes the methods proposed for three SOD tasks. We provide datasets and evaluation metrics in Section 4 and evaluate generic SOD methods, small face, small pedestrian, and aerial image SOD methods. Future directions are discussed in Section 5. Finally, the conclusion of this paper is presented in Section 6.

2. Generic SOD algorithms

In this section, we will extensively review the methods of generic SOD. To deal with the challenges of SOD, existing SOD methods typically have complex designs added to the current pipeline that excels at generic object detection. We will describe these methods from four perspectives, including boosting the resolution of input features, scale-aware training, incorporating contextual information and data augmentation. The advantages and disadvantages of each perspective, as shown in Tables 2–6, are then discussed in detail.

2.1. Boosting resolution of input features

The difficulty in precisely locating small objects is mainly due to the down-sampling operation of the CNN, which causes the features of small objects to disappear, and the low spatial resolution of the high-level feature map seriously loses the spatial position information of small objects. A fairly rational solution to that is to use high-resolution feature maps or high-resolution images. However, employing high-quality images or increasing the feature map resolution will result in higher computing costs. Numerous scholars have constructed feature pyramids by reusing multi-scale feature maps produced by network forward propagation, followed by the use of low-level high-resolution feature maps with more minute spatial details to detect small objects. Additionally, some models have learned the mapping function from low-resolution features to high-resolution features to achieve the same detection effect as large objects. Both approaches substantially increase the resolution of the predictive feature layer. Several typical models that boost the resolution of input features are shown in Figure 3.

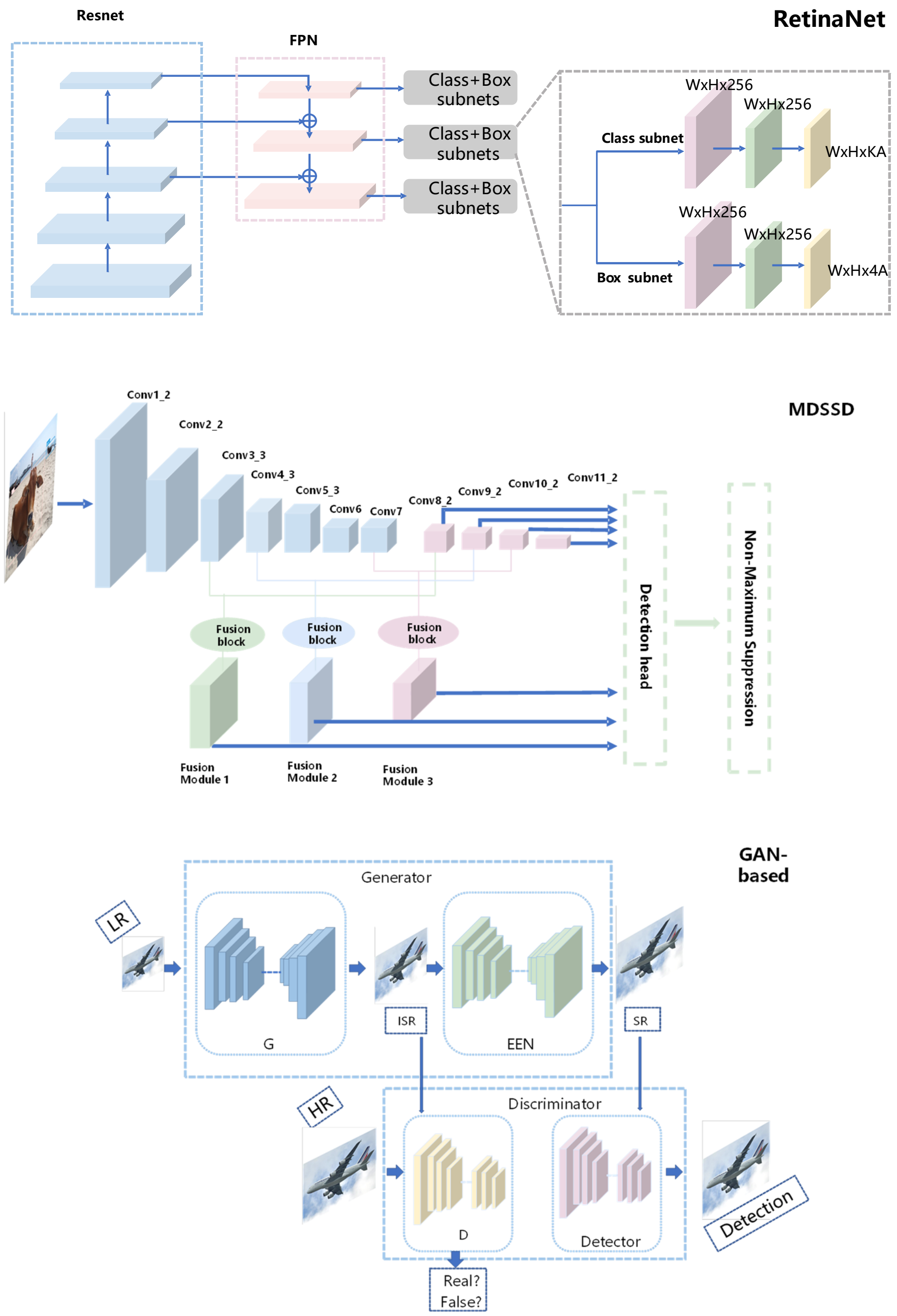

Figure 3. Structures of RetinaNet [74], MDSSD [75] and GAN-based SOD [87].

SSD [36] is a multi-scale object detection technique that detects objects by placing reference windows of different scales in different layers of the networks. The detection accuracy of small objects has not greatly improved. The primary explanation is that low-level feature maps have a limited receptive field and a significantly poorer ability to represent features than deep feature maps. Therefore, Lin et al. proposed FPNs [35]. The core idea behind FPNs is to use forward propagation of the network to create four feature maps of different scales, merge the high-level feature maps with the lower-level feature maps through layer-by-layer up-sampling, fuse the features from different network depths to achieve feature enhancement and then make predictions by using the fused feature map that each layer needs to only predict one scale of objects. The results of the experiments show that the FPN significantly increases SOD accuracy and can guarantee a detection speed of 6 FPS. Since the FPN was proposed, numerous enhanced variants have been developed, including the PANet [70], BiFPN [71], ASFF [72], NAS-FPN [73], etc. The object-proposal-based detection technique has long had a modestly better detection accuracy, despite the integrated convolutional network-based detection model having a significantly faster detection speed. After investigating the reasons behind this, Lin et al. presented RetinaNet [74]. The one-stage network initially outperformed the two-stage network. Lin et al. argued that the foreground-background class imbalance mostly accounts for the integrated convolutional network's inferior detection performance. So, focal loss was proposed to improve cross-entropy loss. Focal loss is given by Eq (1):

$$FL(p) = \begin{cases} -\alpha(1-p)^{\gamma} log(p) if y = 1 \\ -(1-\alpha)p^{\gamma} log(1-p) otherwise \end{cases} \tag{1}$$

α is the balancing variant; $p \in [0,1]$ stands for the probability when y = 1 (positive sample). The rate at which simple examples are down-weighted is adjusted by the focusing parameter $\gamma(\gamma \geq 0)$. RetinaNet can achieve the "focus" of hard samples and the redistribution of network learning ability by reducing the learning weight of simple background samples during the network training process. The lightweight feature fusion module proposed for FSSD [75] uses down-sampling to create a new feature pyramid. MDSSD [76] involves applying deconvolution to a high-level feature map with both powerful semantic information and then fusing it with low-level feature maps by using the fusion module to preserve rich spatial details and high feature representation capabilities for small objects. The architectures of RetinaNet and MDSSD are shown in Figure 3.

At the last layer of a backbone, small object features have almost disappeared. The top-down path makes it nearly impossible for FPNs to fuse the features of small objects. Additionally, as the network gets deeper, the deep feature map gains more semantic information but loses out on spatial information. This causes an offset between anchors and convolutional features, meaning that, after several convolutions, the position of the anchor on the deep feature map differs from the position on the original map. Additionally, the deep features and shallow features cannot be effectively aligned by the FPN fusion. Gong et al. [77] proposed a fusion factor for describing the coupling degree of adjacent layers in FPNs which can be calculated by using the dataset statistical data or learned through implicit learning. Adjusting the fusion factor of adjacent layers in an FPN can adaptively drive the shallow layers to focus on learning tiny objects, thus improving the detection of tiny objects. The high-resolution detection network (HRDNet) [78] accepts multiple resolution inputs via multi-depth backbones. To cut down on computational costs, the multi-depth image pyramid network (MD-IPN) uses a multi-depth backbone to output multi-scale, multi-level feature maps, which means that high-resolution input will be fed into a shallow network to reserve more positional information, and that low-resolution data will be fed into a deep network to extract more semantics. Multi-scale FPNs align and fuse multi-scale feature groups produced by an MD-IPN to decrease the information mismatch between these multi-scale, multi-level features. Liu et al. [79] proposed the IPG-Net to mitigate the disappearance of small object features following serial down-sampling and the dislocation between spatial information and semantic information; it includes an IPG transformation and IPG fusion module. IPG-Net receives an image pyramid as the input; the IPG transformation module extracts shallow features from image pyramids of various resolutions that include rich spatial information and detailed information; the IPG fusion module fuses the shallow features extracted by the IPG transformation module and the deep features of the backbone. RHF-Net [80] applies top-down and bottom-up feature fusion. It contains a recursive execution of the hybrid fusion module that enables RHF-Net to both connect high-level semantic features to the low-level features (top-down direction) and reshape the rich spatial features of low-level feature maps to the deeper layer (bottom-up direction), thus improving the contextual features of objects of all scales.

The spatial distribution of small objects on the high-resolution feature map of the feature pyramid is very sparse, accounting for only a small part of the high-resolution feature map. QueryDet [81] uses the query technique to accelerate the reasoning speed of the object detector based on the feature pyramid by preventing the detection head from doing resource-intensive calculations on the entire high-resolution feature map. It includes a query head in parallel with classification and regression to predict the locations (query keys) of a possible small object in the features of the previous layer. The current layer uses these locations to generate a sparse value feature map (query value). Then, it predicts the query keys of this layer to be given to the following layer.

Super-resolution is another effective method that directly enriches the information of small objects by increasing the resolution of the input image. EFPNs [82] add a super-resolution layer to an FPN, as it uses the feature texture transfer module to super-resolve features by extracting regional texture features from the reference features. This adds convincing details to the EFPN and improves the accuracy of SOD. To eliminate the representational disparity between large and small objects and allow a small object to attain the same detection accuracy as large objects, Li et al. [83] used a GAN to enhance the small object's feature representation to a super-resolved representation. But, the super-resolved feature might not be convincing, as the large object image and the small object image are not from the same image. The SOD-MTGAN [84] learns the mapping between low-resolution image patches and high-resolution image patches, which reduces the computational cost. Noh et al. [85] used high-resolution features for direct supervision. And, under the guidance of a super-resolution discriminator, low-resolution features are transferred to the super-resolution feature generator to generate high-resolution features. MARE [86] uses a network to obtain attention weights, which are considered as weights for each level of feature maps, to generate the final attention feature maps; it then performs feature fusion to further enhance the information that is useful for small targets. The EESRGAN [87] adds edge-enhanced sub-networks (EENs) [88] to the ESRGAN [89]. EENs perform edge enhancement on the intermediate super-resolution (ISR) images generated by the generator to produce the final super-resolution image. Together, the discriminator and detector perform the role of the discriminator, and the discriminator trains the generator by using relativistic loss [90]. The following Eqs (2) and (3) show the relativistic loss of the discriminator and the adversarial loss [91] of the generator. Where D_{Ra} is the probability that a real image (I_{HR}) is relatively more realistic than a generated intermedia image (I_{ISR}), $E_{I_{SR}}$ is the operation that calculates the average of all generated intermediate images in a mini-batch, and $E_{I_{HR}}$ is the operation that calculates the average of all real images in a mini-batch. Additionally, the EESRGAN employs end-to-end training to backpropagate the detector loss to the generator. Thus, the generator receives gradients from both the detector and the discriminator to enhance the quality of super-resolution images. Cao et al. proposed the MHN [92], which splits the network into three distinct branches (branch-l, branch-m, branch-s), where each branch produced equivalent high-level semantic feature maps with a variety of resolutions, allowing it to better match objects of various scales.

$$L_G^{Ra} = -E_{I_{HR}}\left[log\left(1 - D_{Ra}(I_{HR}, I_{ISR})\right)\right] - E_{I_{SR}}\left[log\left(D_{Ra}(I_{ISR}, I_{HR})\right)\right] \tag{2}$$

$$L_D^{Ra} = -E_{I_{HR}}\left[log\left(D_{Ra}(I_{HR}, I_{ISR})\right)\right] - E_{I_{SR}}\left[log\left(1 - D_{Ra}(I_{ISR}, I_{HR})\right)\right] \tag{3}$$

2.2. Scale-aware training

The largest object in the COCO dataset is 20 times larger than the smallest, and the scale invariance of CNNs is not robust against such large-scale variances. Scale-aware training strategies can make the detector more robust against scale variance. A common process of the scale-aware training model is shown in Figure 4.

Previously proposed approaches use image pyramids [93,94] to improve the accuracy of object detection at various scales, which have larger memory requirements. Scale normalization for image pyramids (SNIP) [95] is a training strategy that uses the image pyramid training model and only backpropagates the loss of object size within the predetermined range. To go further, SNIPER [96] chooses chips with a fixed resolution of 512 × 512 pixels from each layer of the pyramid to act as the training unit, unlike SNIP, which analyzes every pixel in an image. Because of the smaller chip resolution, it can train with a larger batch size, which improves both training efficiency and detection accuracy. Kim et al. proposed a scale-aware network (SAN) [97] that maps the convolutional features from the different scales onto a scale-invariant subspace to make CNN-based detection methods more robust against the scale variation, and also to construct a unique learning method that purely considers the relationship between channels without the spatial information for the efficient learning of the SAN. This method essentially improves the quality of convolutional features in the scale space and can be generally applied to many CNN-based detection methods to enhance the detection accuracy with a slight increase in the computing time.

Trident [98] is a multi-branch parallel network, where each branch adopts an appropriate dilated ratio to provide the receptive field size that can align with the object size. Moreover, a scale-sensitive training approach is applied to enhance each branch's capacity for scale perception and prevent the training of objects of extreme scale on branches with unmatched receptive fields. Each branch's effective range, *l*, is given by Eq (4):

$$l_i \leq \sqrt{wh} \leq u_i \tag{4}$$

Peng et al. [99] show that the local and dense continuous scales which are hard to optimize are not necessary, and that, through a collaboration of well-learned global scales on layers, a network could be granted the scale-awareness. Therefore, they designed a global scale learning module to replace the normal convolutional module and learn the appropriate global scale for different layers.

Table 2. Summary of strengths and weaknesses resulting from boosting the resolution of input feature methods.

Method	Publication	Techniques	Strengths	Weaknesses
SSD [36]	ECCV16	Pyramidal feature hierarchy without fusing features.	SSD can detect objects of various sizes.	The low-level prediction feature map has no strong semantics.
FPN [35]	CVPR17	Feature pyramid network (including feature fusion and multi-scaled fusion modules, etc.).	FPN dramatically improves the detection accuracy of small objects.	The feature representation capability will be diminished by the semantic gap between feature layers of various scales.
RetinaNet [74]	ICCV17		RetinaNet alleviates the foreground-background class imbalance problem.	
FSSD [75]	arXiv17		Lightweight feature fusion module.	
MDSSD [76]	arXiv18		MDSSD incorporates contextual information that is more conducive to SOD.	Lower detection speed than SSD.
[77]	CVPR21	FPN with a learnable fusion factor.	The fusion factor further improves FPN performance for small objects.	
HRDNet [78]	arXiv20	FPN with image pyramid.	HRDNet acquires more details for small objects with high resolution.	Large numbers of parameters.
IPG-Net [79]	CVPR20		IPG-Net alleviates the vanishment of the small object features.	This method is inefficient.
RHF-Net [80]	CVPR20	Recursive hybrid fusion pyramid network.	Low computational cost and high accuracy.	
QueryDet [81]	CVPR22	Query mechanism.	Accelerating inference with sparse query.	
EFPN [82]	TMM22	Super-resolution (include super-resolution layer and enhancing representations of small objects to be similar to large ones).	EFPN adds a high-resolution layer to FPN to increase the accuracy of the SOD.	Super-resolution feature extraction leads to more computational costs.
[83]	CVPR17		The GAN-based approaches effectively enhance the level of detail of information on small objects.	
MTGAN [84]	ECCV18			
TPS [85]	ICCV19			
MRAE [86]	CVPR22	FPN with attention weight.	It provides a practical solution for multi-resolution feature extraction without using a GAN, and it is time-efficient.	

Table 3. Summary of strengths and weaknesses of scale-aware training methods.

Method	Publication	Techniques	Strengths	Weaknesses
SNIP [95] SNIPER [96]	CVPR18	Scale normalization training strategy for image pyramids.	SNIP and SNIPER can effectively improve the detection performance of small objects.	It requires an input image pyramid that brings a high computational cost.
SAN [97]	CVPR18	Scale-aware training.	SAN makes the network more robust against scale invariance.	
Trident [98]	ICCV19	Multi-branch architecture and scale-aware training.	Multi-branch technique makes the receptive field size align with the object size.	It may bring about the over-fitting problem in each branch, as caused by too few effective samples.
POD [99]	ICCV19	Global scale learning.	This method makes the network more sensitive to scale invariance.	

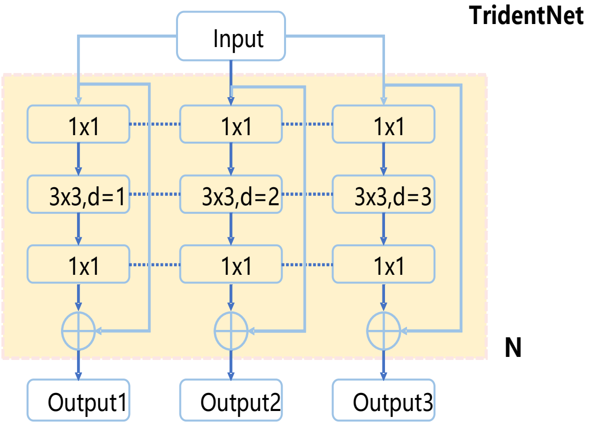

Figure 4. Architecture of TridentNet [98]; *d* is the dilation rate.

2.3. Incorporating contextual information

Visual objects frequently coexist with other relevant objects in a certain setting, which provides rich contextual associations to be exploited. Researchers [100] have shown that utilizing the context as extra information can help to detect small objects with obscure features. Two typical models of incorporating contextual information are shown in Figure 5.

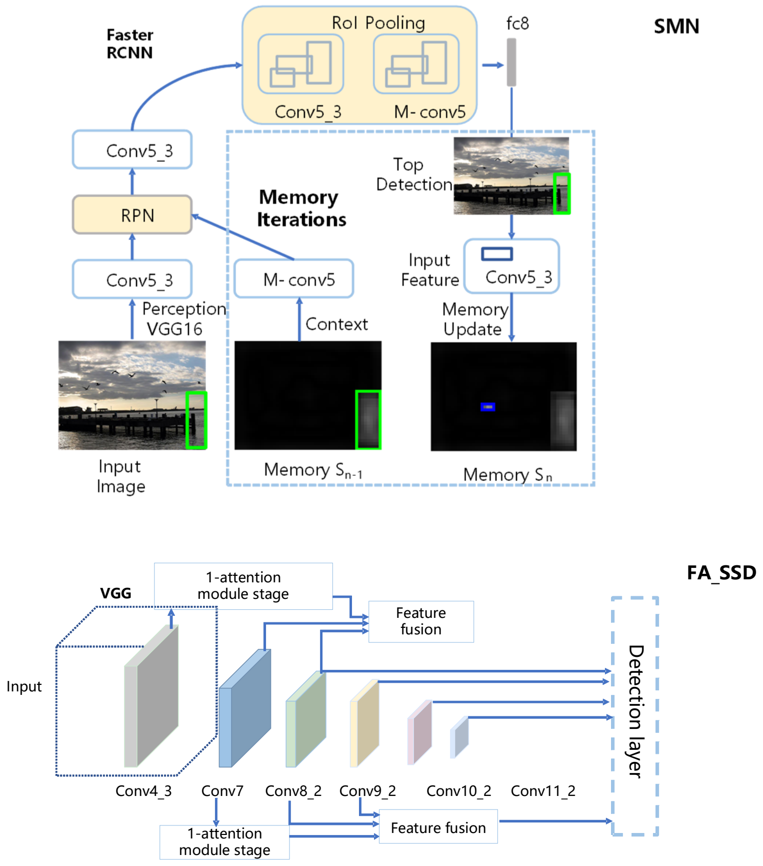

Figure 5. Architectures of SMN [86] and FA_SSD [88].

Chen et al. [42] extended the R-CNN model by using ContextNet and a small region proposal generator to improve SOD. Regarding the region proposal network (RPN), Chen et al. used smaller RPN anchor sizes (16^2, 40^2, 100^2 vs. 128^2, 256^2, 512^2). ContextNet integrates contextual information to calculate the final classification score. Bell et al. [101] proposed ION, which utilizes information inside and outside of the ROI to improve detection performance. Regarding the inside part, ION extracts the features of the ROI at several levels at different scales by using skip pooling to enhance the ability to detect small objects. Regarding the outside part, ION extracts the contextual information outside of the ROI by using a spatial recurrent neural network to enhance the feature information and promote the subsequent classification and regression performance. The DSSD [102] fuses deep semantic information as context with shallow semantic information. The CSSD [103] is a context-aware framework that incorporates context by integrating deconvolutional or dilated convolutional layers into SSD. In object detection, there are two common contexts. Image-level context refers to modeling the contextual information of each pixel in the whole image, which is implicitly incorporated into the deep convolutional network, while the instance-level context, which models object-object relationships, is an important clue for object detection and reasoning. A spatial memory network (SMN) [104] was proposed to get the instance-level context. The network detects an object, remembers it and then uses it as a priori knowledge to help detect the previously missed target in the next iteration. Fu et al. [105] introduced a unique contextual reasoning method for SOD that models and infers the relationships between objects' inherent semantic and spatial layouts. The learnable semantic association functions are defined by the semantic module from the standpoint that proposals belonging to the same category share semantic co-occurrence information. The formula is given by Eq (5):

$$s'_{ij} = \sigma(i,j).\, f\left(p_i^o, p_j^o\right) = \sigma(i,j).\, \phi(\mathrm{p_i^o})\phi(\mathrm{p_j^o})^{\mathrm{T}} \tag{5}$$

where $\sigma(i,j)$ denotes an indicator function and ϕ maps the initial region features p_i^o to latent representations. The spatial layout module disregards semantic similarity and builds relationships based on spatial similarity and spatial distance in the internal spatial layout, allowing small objects that have a high degree of spatial similarity and appear in clusters to communicate contextual information about the spatial layout to one another. FA-SSD [106] is a combination of F-SSD and A-SSD. F-SSD uses higher-level feature maps as context to concatenate with low-level feature maps. A-SSD uses an attention mechanism to minimize unnecessary shallow features in the background. Both image-level context and instance-level context are commonly used by SOD.

Table 4. Summary of strengths and weaknesses associated with incorporating contextual information methods.

Method	Publication	Techniques	Strengths	Weaknesses
ION [101]	CVPR16	Integrate contextual information.	ION exploits context and multi-scale representations to improve SOD.	Underutilization of early feature layers.
DSSD [102] CSSD [103]	arXiv17 WACV18		Fusing contextual information in different ways to improve SOD performance.	Slower detection speed than SSD.
SMN [104]	ICCV17	Spatial memory for contextual reasoning.	SMN models the instance-level context to improve the performance of SOD.	The gradient will vanish as the reasoning signal and perceptual signal cancel each other out.
IRR [105]	arXiv20	Contextual reasoning integrates intrinsic relations.	IRR updates the initial regional features to boost SOD.	Small objects are associated with difficulty in extracting semantic features.
FA-SSD [106]	ICAIIC21	Context with attention.	FA-SSD is more accurate than SSD.	It has lower accurate than DSSD.

2.4. Data augmentation

High-quality large-scale datasets can greatly improve the performance of deep learning SOD. However, the amount of labeled data is still far from sufficient due to the high cost of annotation. Data augmentation is a common method to enrich the diversity of the dataset, thus improving the generality and robustness of the model to some extent. This can also help to mitigate the degradation of object detection accuracy due to the uneven distribution of different scale objects in the dataset.

A lot of data augmentation techniques have been developed, such as affine transformation, Mosaic [107], MixUp [108] and CutMix [109], but these methods have better performance on medium- or large-sized objects than small objects. Kisantal et al. [110] thoroughly investigated the MS

COCO dataset and discovered a sample imbalance problem: images with small objects in the dataset are only a small fraction; particularly, the number of small objects in each image is less and the site of occurrence lacks diversity. Kisantal et al. proposed oversampling images with small objects to increase the number of small objects during training. Chen et al. [111] found that random copying and pasting led to background mismatch and object size mismatch. To solve that, they employed adaptive data augmentation, which uses a semantic segmentation network to obtain an a priori roadmap and samples an effective position to place the object enhanced by the roadmap. Ünel et al. [112] proposed a tiling-based technique where the input images are deliberately split into overlapping tiles to increase the relative pixel area of small objects.

To address scale variance, DST [113] receives the loss proportion caused by small objects as feedback. If the loss proportion is smaller than the predetermined threshold, the training images are enlarged and spliced in the following iteration to compensate for the missing small objects. Zoph et al. [114] used AutoAugment [115] to find the optimal data augmentation method for object detection by applying an augmentation strategy search to the training set. An RNN controller and a reinforcement learning methodology are included in the search strategy. Chen et al. [116] proposed scale-aware automatic data augmentation, which includes a scale-aware search space with augmentations at the image and box levels, as well as a search metric called the Pareto scale balance. The metric is realized by recording accumulated loss and accuracy over various scales.

Table 5. Summary of strengths and weaknesses of data augmentation methods.

Method	Publication	Techniques	Strengths	Weaknesses
Kisantal et al. [110]	arXiv19	Oversampling and random copy-pasting.	This approach achieves better object detection accuracy for small objects.	Random copying and pasting may cause background mismatch.
RRNet [111]	ICCV19	Adaptive resampling augmentation strategy.	Free-anchor and adaptive resampling result in excellent performance for very small objects.	
Ünel et al. [112]	CVPR19	Tiling-based augmentation.	The method provides a good trade-off between accuracy and time cost.	
DST [113]	arXiv20	Uses the feedback information to guide data preparation.	Feedback-driven and dynamic data preparation paradigms mitigate the scale-invariant issue.	
Zoph et al. [114]	ECCV20	Automatic data augmentation	This approach has no additional inference cost and minimal training cost.	The strategy is intricate.
Chen et al. [116]	CVPR21		It can be transferred to other datasets and tasks and is scale-sensitive.	The high time cost of auto-augmentation approaches for searching.

2.5. Other strategies

Samet et al. [117] proposed a new labeling technique in which the predictions derived from individual features are aggregated into one prediction to reduce the labeling noise of the anchor-free detector. Duan et al. proposed CenterNet++ [118], which uses a triplet of a center key point and a pair of corners to represent an object. The corners can locate objects with any geometry. Wang et al. [119] evaluated the sensitivity of the Intersection over Union (IoU) to position variations of small objects, and they suggest replacing the IoU with a new measuring technique that models each box as a Gaussian distribution and uses the normal Wasserstein distance (NWD) to determine the similarity of the two distributions to one another. Xu et al. [120] presented receptive field distance to quantify the similarity between the Gaussian receptive field and ground truth directly, rather than assigning samples with IoU sampling strategies. C3Det [121], an interactive, multi-class, tiny-object annotation framework that Lee et al. suggested, allays concerns about the demands and expense of annotation in the actual world. SAHI [122] entails dividing the input images into overlapping slices to yield a higher percentage of small objects in the image of the input network.

Table 6. Summary of strengths and weaknesses of other methods.

Method	Publication	Techniques	Strengths	Weaknesses
PPDet [117]	BMVC20	Anchor-free with a new label strategy.	It reduces the contributions of non-discriminatory features during training.	
CenterNet++ [118]	CVPR21	An anchor-free detector that uses triplet key points to represent objects.	This model with multi-resolution performs better.	
NWD [119]	arXiv21	A new metric to replace IoU.	These two metrics are more effective than the IoU metric for small object detection.	
RFLA [120]	ECCV22			
C3Det [121]	CVPR22	Annotation framework for tiny objects.	It alleviates the expense of tiny-object annotation.	
SAHI [122]	arXiv22	Slicing-aided inference.	This scheme is plug-and-play, does not require pre-training and improves the accuracy of detecting small objects.	Larger feature maps require more memory and computing cost.

3. Crucial SOD tasks

In this section, we present a systematic review of SOD in terms of small face detection, small pedestrian detection and aerial image detection tasks. We first thoroughly describe the current approach to each task. Then, a comprehensive summary of the strengths and weaknesses of each method is presented.

3.1. Small face detection

Multi-scale modeling [123] was proposed following a thorough investigation of image resolution, object scale variation and contextual information. This algorithm uses SSD as the foundation, and the sparse discrete image pyramid is fused to handle the scale shift of objects. Rich contextual information is necessary for SOD, but low-level feature maps are used because SOD lacks semantic information; however, deep feature maps contain rich contextual and semantic information. As a result, multi-layer feature fusion is incorporated into SOD, which enhances the performance of small face detection. S^{3}FD [124] incorporates a scale-equitable face detection network to adapt face detection at various scales. Additionally, the effective receptive field and equal-proportion interval principles are used to define the scales of the anchors, ensuring that different scales of anchors are distributed uniformly across the image, and that anchors at different layers match their corresponding effective receptive fields. Then, by using a scale compensation anchor-matching approach, the recall rate of small faces is increased. Lastly, the false positive rate of small faces is decreased by predicting the number of background anchors for each matched small anchor. [125] uses generative adversarial network to generate high-resolution face. Face-MagNet [126] employs ConvTranspose (kernel = 8, stride = 4) layers that pass the features of small faces from the lower feature layer to the prediction layer inside an RPN and classifier to magnify the feature maps for the better detection of small faces.

Table 7. Summary of strengths and weaknesses of small face detection methods.

Method	Publication	Techniques	Strengths	Weaknesses
Hu and Ramanan [123]	CVPR17	Super-resolution by GAN.	The joint super-resolution and refinement model is effective.	No fusion of contextual information.
S^{3}FD [124]	ICCV17	Scale-invariant strategy.	The three-trick, scale-equitable framework, max-out, and scale compensation anchor-matching achieve superior performance.	
MagNet [126]	WACV18	Feature fusion approach to integrating contextual information.	ConvTranspose is more helpful than skip connections or context pooling.	The improvement is not obvious.
Zhu et al. [127]	CVPR18	EMO metric to get a high IoU.	The EMO score inspired several effective strategies for a new anchor design to obtain a higher facial IoU score.	
TinaFace [128]	arXiv21	Geometric transformations and multi-scale representation.	Simple improvements of RetinaNet achieve better performance.	
Zhang et al. [132]	WACV20	Hard example mining and super-resolution.	It handles the imbalance between images.	

Zhu et al. [127] pointed out that the anchor-based face detector does not process small faces well because the anchor and small faces cannot overlap perfectly, so it is difficult to adjust the anchor to be close to ground truth. Therefore, Zhu et al. proposed an expected max overlapping (EMO) score, which improves the ability of the anchor and face to obtain a high IoU. And, by increasing the number of small-scale anchors, it enhances the likelihood of matching a face. Additionally, to get a high IoU for these faces with the anchor, the algorithm randomly moves the face positions during training. Finally, a compensation strategy of anchor matching was also proposed to improve the chance of detecting hard faces. TinaFace [128] involves modifications to RetinaNet, and it achieved a 92.4% average precision (AP). First, a DCN [129] is introduced as the backbone to learn complex geometric transformations; then, Inception is used to improve the multi-scale representation. And, the loss of bounding box regression is changed from smooth L1 to DIoU [130] due to DIoU being more accommodating for small objects. Finally, an IoU-aware branch is included to address the mismatch between localization accuracy and classification scores. Hard example mining techniques like OHEM [131] identify hard positive and hard negative examples and focus more effort on training on those hard instances to improve detector performance. Zhang et al. [132] increased the effectiveness of OHEM by combining OHEM with hard image-level mining to train the face detector; it automatically alters training weights on images according to their difficulty. Additionally, they used a detector that only produces a single high-resolution feature map with small anchors to specifically learn small faces and train it by using the hard image mining strategy. The strengths and disadvantages of small face detection methods are shown in Table 7.

3.2. Small pedestrian detection

Song et al. [133] proposed a topological line localization (TLL) network, i.e., a topological line detection network based on the pedestrian torso, which was designed to reduce the effects of small-scale pedestrian boundary blur, appearance blur and the annotation method of the bounding box that brings too much of a noisy background to small objects. And, combining TLL and ConvLSTM into a single time-aware architecture to aggregate the features of consecutive frames in the video thus enhances the performance of small pedestrian detection. Furthermore, a Markov random field, as a post-processing strategy, is employed to deal with crowd occlusion. Das et al. [134] constructed the ISI pedestrian dataset, which includes 13,129 annotated video frames with 82.3 thousands labeled pedestrians. Additionally, Das et al. provided a three-phase detection algorithm. First, the prospective regions in each frame are identified using a zone classifier, which uses an improved Inception network to lower the error. The frames per second is then significantly improved by solely using the possible regions to locate the pedestrian's position. Finally, using non-maximum suppression (NMS) is applied to remove the redundant bounding box of the same pedestrian.

CNNs can not only learn low-level features, but it also has a strong ability to learn high-level semantic features. Therefore, CSP [135] simplifies pedestrian detection into pedestrian scale prediction and center tasks through a convolutional operation. The detection head applies a convolutional operation to the feature map generated by the feature extractor and adds two parallel 1 × 1 convolutions to generate, respectively, a centroid heat map and a scale size prediction map. Cross-entropy loss is employed in center point prediction and L1 loss is employed in scale prediction. Yu et al. [136] constructed the TinyPerson dataset, which focuses on persons on, at and around the seaside for maritime quick rescue. Pedestrians in TinyPerson are much smaller than those in other datasets, with most having pixel ranges of under 20 pixels and a wide variance in the person's aspect ratio. And, to solve the problem that the distribution of the pre-training dataset differs greatly from the distribution of the dataset for the specified task, this algorithm proposes a scale match to make the feature distribution consistent between the pre-trained dataset E and the task-specific dataset D, as shown in Eq (6), where $P_{(s,D)}$ is defined as the probability density function of objects of size s in the dataset D, and T is the scale change function.

$$P_{(s,T(E))\approx P_{(s,D)}} \tag{6}$$

The FSAF [37] allows each instance to freely choose the best layer to optimize the network, instead of using the traditional pyramid, which puts several anchors of a fixed size at each level. The best feature layer for each instance is dynamically selected throughout the training phase based on the content of the instance, rather than just its size; the selection function is given by Eq (7):

$$l' = \lfloor l_0 + log_2\left(\frac{\sqrt{wh}}{224}\right)\rfloor \tag{7}$$

where 224 is the ImageNet pre-training size and l_0 is the initial feature layer. The strengths and disadvantages of small pedestrian methods are shown in Table 8.

Table 8. Summary of strengths and weaknesses of small pedestrian detection methods.

Method	Publication	Techniques	Strengths	Weaknesses
Song et al. [133]	ECCV18	A topological line detection.	TLL can automatically adapt to small-scale pedestrians.	No mitigation of information loss for small pedestrians.
SaYwF [134]	arXiv19	A three-phase detection model.	Achieves a trade-off between detection accuracy and detection speed.	
CSP [135]	CVPR19	Pedestrian detection is converted to high-level semantic feature prediction.	No additional post-processing is required for CSP.	Objects with a large variance in aspect ratio need to be examined.
FSAF [37]	CVPR19	Feature-selective anchor-free module.	Dynamically assigning each instance to the most suitable feature level is more robust.	Separate anchor-free branches do not have many advantages over anchor-based branches.
Yu et al. [136]	WACV20	Scale match of the pre-trained dataset to the task-specified dataset.	Scale match can better utilize the existing annotated data.	It has poor performance on TinyPerson.

3.3. SOD in aerial images

Object detection in aerial images is crucial in many real-world applications, including urban planning, emergency rescue [137], traffic detection [138,139], etc. Since aerial images are usually taken from high altitudes looking down, the rotation of objects varies greatly and is displayed in arbitrary directions. In addition, aerial images contain highly dense scenes and many small objects, making SOD a complex problem for aerial remote sensing images. Innovative detection algorithms have emerged to address these issues.

S^{2}A-Net [140] contains a feature alignment module and an oriented detection module to keep consistency between the classification score and localization accuracy. SCRDet [141] designed a supervised multidimensional attention to highlight small object regions and reduce the effect of background noise. Oriented R-CNN [142] and MRDet [143] both proposed a lightweight regional proposal network to generate oriented proposals. [144] proposed a novel model which contains four parts. To extract feature maps from the input photos, the first component serves as the backbone. The backbone incorporates a ResNet50 network with deformable convolutional layers because a regular convolution cannot adjust to variations in viewpoint in images taken by drones. The second part seeks to use an FPN to exploit and improve the feature maps obtained from ResNet50. The RPN, which can be used to extract prospective proposals of objects in the image, is the third component. The last section is a task head for certain goals. Bounding box and mask prediction are assigned by using an interleaved cascade architecture by the component. Yi et al. [145] extended the center key-point object detector for oriented object detection. A U-shaped network [146] is the foundation of the model. In the process of up-sampling, skip connections are used to combine feature maps. Four maps make up the output of the architecture: the heat map, offset map, box parameter map and orientation map. The heat map and offset map are used to deduce the locations of the center points. After the center points are detected, the box boundary-aware vectors (BBAVectors) are regressed to capture the oriented bounding boxes.

Table 9. Summary of strengths and weaknesses of detection methods on aerial images.

Method	Publication	Techniques	Strengths	Weaknesses
Zhang et al. [144]	ICCV19	Model fusion, cascade network, deformable convolution and data augmentation.	The joint optimization of four strategies makes the model perform well on VisDrone.	The efficiency and detection speed of the network is poor.
Yi et al. [145]	WACV21	Oriented anchor-free object detector.	Extended BBAVector technique on CenterNet is simple and effective.	
ReDet [147]	arXiv21	Rotation-invariant feature representation.	Smaller models and better results for small- and medium-sized objects.	
DarkNet-RI [149]	TGRS21		The multi-scale and rotation-invariant feature representation is robust against scale-variance.	Need to enhance the overlapping and occluded object detection.
Li et al. [151]	CVPR22	Adaptive points learning approach.	This model can classify and localize objects with arbitrary orientation.	It requires large computing cost.
DotD [152]	CVPRW21	A new metric DotD.	It's valid for defining positive and negative anchors in training.	

According to Han et al. [147], CNNs lack rotation invariance, which means that, after an image is rotated, the features it extracted will also change. ReCNN was therefore proposed, allowing CNNs to have rotation invariance. They incorporate rotation-equivariant networks into the backbone to extract rotation-equivariant features, which allows for precise prediction of the orientation. Then, the rotation-invariant RoI Align module was developed based on RROI Align [148] to align both the channel dimension and the spatial dimension to obtain the rotation invariance features. DarkNet-RI [149] uses DarkNet53 [7] as a backbone that contains a rotation-invariant layer to extract rotation-invariant multi-scale features and use classification solutions to directly predict the location of objects. After that, a box refinement module is utilized to carry out additional NMS to eliminate overlapping and redundant bounding boxes. RepPoints [150] develops adaptive point sets and can capture the geometric structure of airborne objects with abrupt changes in direction in a chaotic environment. Three oriented conversion functions were presented by Li et al. [151] to transform adaptive points into oriented bounding boxes for various oriented objects. They apply MinAeraRect in the post-processing to provide the usually rotated rectangle prediction, and the NearestGTCorner and MinAeraRect functions are applied to enhance adaptive point learning during training. Xu et al. [152] proposed Dot Distance (DotD), i.e., a normalized Euclidean distance between the centroids of two bounding boxes, to solve the problem of IoU being sensitive to minute offsets between bounding boxes when detecting tiny objects. S^{2}ANET-SR [153] uses super-resolution to enhance the feature extraction of small objects in remote sensing images and incorporates perceptual loss and texture matching loss to train S^{2}ANET-SR jointly with the detection loss. The authors of [154] developed a cross-layer attention module to extract non-local features from small objects to enhance their features. The authors of [155] utilized a Gaussian mixture model to generate focal regions, as well as an incomplete box suppression method to mitigate the truncated box problem, which improved the performance of SOD. The strengths and weaknesses of aerial image methods are shown in Table 9.

4. Evaluation of SOD

This section provides an overview of the SOD datasets that are currently available. The performance of SOTA SOD approaches is also evaluated by using three large-scale datasets. We chose well-known image datasets: MS COCO for the general SOD evaluation, WiderFace for SOD tasks with small faces, TinyPersons for SOD tasks with small pedestrians and DOTA for SOD tasks for aerial images.

4.1. Dataset

A high-quality dataset is important for developing advanced object detection algorithms. In recent years, many well-known datasets for object detection have been published, such as MS COCO [43] and VOC [45]. VOC is a dataset for the Pascal VOC challenge object detection subtask, which has two versions: VOC2007 and VOC2012. More than 27 thousands object instance bounding boxes are labeled in 33,043 images in VOC2012. MS COCO is a sizable multi-task dataset, as it has 91 object categories in all (80 object categories are used for object detection tasks) and 2500 thousands labeled instances in 328 thousands images. The tasks on the COCO dataset are more challenging because, in contrast to VOC, COCO contains more small objects and more complicated backgrounds in the images. COCO also has a more balanced object distribution. Less than 20% of the images in the COCO dataset have only one category and an average of 3.5 categories and 7.7 instance objects of each image. Over 70% of the images in the VOC dataset have only one category; on average, there are only 1.4 categories and 2.3 instance objects per image. These benchmarks boost the development of detecting regular-sized objects. Unfortunately, the detection of small objects is still insufficient. It is caused by both the characteristics of small objects themselves, as well as the fewer benchmarks designed for SOD. To provide a comprehensive review of a dataset, we investigated datasets containing a large number of small objects that span various SOD tasks, such as face detection, pedestrian detection, traffic sign/light detection and aerial image object detection, as shown in Tables 10–13.

Table 10. Overview of face detection datasets.

Dataset	Year	Description
WIDER FACE [47]	2016	WIDER FACE is a large-scale dataset of face images. Images are selected from the publicly available WIDER dataset.
IJB [156]	2015	IJB-A/B/C is a dataset for face detection and recognition. IJB-A contains 1845 objects, 11,754 images, 55,026 video frames, 7011 videos and 10,044 non-facial images.
DarkFace [157]	2019	The DarkFace dataset offers 6000 nighttime low-light photos from real-world locations, all labeled with bounding boxes of human faces. Additionally, this dataset has 9000 unlabeled low-light images taken in the same environment.
UFDD [158]	2018	UFDD, an unconstrained face detection dataset, consists of more than 6000 images and 11,000 faces, and it contains seven scenes: rain, snow, haze, blur, illumination, lens impediments and distractors.
WildestFaces [159]	2018	The WildestFaces dataset includes 67,889 pictures. Along with annotations for face detection and recognition, it also includes tags for blur severity, scale and occlusion.

Table 11. Overview of pedestrian detection datasets.

Dataset	Year	Description
TinyPerson [136]	2020	TinyPerson is a challenging benchmark for tiny object detection in a complex context and at a long distance. A total of 72,651 labeled very small objects are included in the dataset.
WiderPerson [160]	2020	The WiderPerson dataset, which contains 32203 images with a total of 393703 instances.
EuroCity [161]	2018	The EuroCity person dataset was collected in several European countries by in-vehicle cameras; it includes about 47,300 images with more than 238,200 annotated instances of people.
Citypersons [162]	2017	The Citypersons dataset is a subset of a cityscape; it offers 5,000 images from 27 cities with 30 fine-grained, pixel-level annotations.
Caltech [163]	2009	Caltech is a challenging dataset that contains low-resolution, frequently obstructed objects. There are 192,000 and 155,000 pedestrian instances in the training and testing sets, respectively.

4.2. Evaluation metrics

Frames per second refers to the speed of object detection, and it indicates the number of images that can be processed within each second. A higher value implies that the method is faster and can potentially be applied to real-time SOD.

IoU measures the similarity between the areas of the prediction bounding box ($bbox_{pred}$) and the ground truth bounding box, $bbox_{GT}$. The IoU function is given by Eq (8).

$$IoU = \frac{Area(bbox_{pred} \cap bbox\ \ )}{Area(bbox_{pred} \cup bbox\ \ )} \tag{8}$$

AP is a common metric for object detection tasks, and the following definitions are used in the AP calculation.

1) Positive sample: a sample that contains a detection object, and the prediction bbox confidence score is larger than the set threshold.

2) Negative samples: samples that do not contain detection objects, and the prediction bbox confidence score is larger than the set threshold.

3) True Positive (TP): positive samples that are predicted correctly.

4) True Negative (TN): negative samples that are predicted to be correct.

5) False Positive (FP): positive samples that are predicted to be wrong.

6) False Negative (FN): negative samples that are predicted to be wrong.

Table 12. Overview of aerial image object detection datasets.

Dataset	Year	Description
DIOR [58]	2020	DIOR is made up of 20 common object categories, 23,463 optimum remote sensing images and 192,472 hand-annotated object instances with axis-aligned bounding boxes.
VisDrone [164]	2022	VisDrone was collected by the AISKYEYE team at Tianjin University in China while utilizing several UAVs; it includes pedestrians, automobiles, bicycles and other categories.
UAVDT [165]	2018	UAVDT is a sizable UAV-based video dataset with 80,000 total frames that are intended for vehicle detection and tracking.
DOTA [166]	2018	DOTA has three versions so far; DOTA-v1.0 includes 188,282 instances of 2806 aerial images in 15 main categories.
NWPU VHR-10 [167]	2016	The NWPU VHR-10 dataset contains a total of 800 very high-resolution optical remote sensing images, which were acquired from Google Earth and Vaihingen.
UCAS-AOD [168]	2015	The UCAS-AOD datasets include many small objects with intricate backgrounds with a total of 2420 images and 14,596 instances.

In the VOC dataset, the IoU threshold is typically set to 0.5. Positive samples with IoU values higher than 0.5 are labeled as TP, and positive samples with IoU values lower than 0.5 are labeled as FP. FN indicates the number of objects in ground truth that are not found. Then, the precision rate and recall rate are given in Eqs (9) and (10). AP is calculated across different recalls. Specifically, for a given recall value *r*, the precision value takes the maximum of all recall values that are greater than or equal to *r*. Then, the area under the precision-recall (P-R) curve is referred to as the AP value. The mAP is the mean AP value across all categories. AP and mAP are given in Eqs (11) and (12).

$$precision = \frac{TP}{TP+FP} = \frac{TP}{allpredictedbox} \tag{9}$$

$$recall = \frac{TP}{TP+FN} = \frac{TP}{allgroundtruth} \tag{10}$$

$$AP = \int_0^1 P(R)dR \tag{11}$$

$$mAP = \frac{\sum AP}{N} \tag{12}$$

The stricter COCO evaluation metric is more widely used than the PASCAL VOC evaluation metric. The IoU thresholds of it typically range from 0.5 to 0.95, with a 0.05 step size. A special AP is also calculated separately for small (the square of the area < 32^2), medium (32^2 < area < 96^2), and large (area > 96^2) objects in MS COCO.

Table 13. Overview of traffic scene and other detection datasets.

Dataset	Year	Scenario	Description
SOD [42]	2021	Generic	SOD is a subset of the SUN [171] and MS COCO datasets. Ten types of objects that appear extremely small in the images were manually chosen by the authors.
TT100K [56]	2016	Traffic Sign	TT100K has 100,000 images and 30,000 traffic sign instances across 128 classes.
DeepScores [169]	2018	Stradivarius	DeepScores includes high-quality images of sheet music, with around 100 million small objects.
KITTI [170]	2012	Traffic Scene	KITTI has up to 15 vehicles and 30 pedestrians in each image captured in Karlsruhe, Germany.

4.3. Performance on generic SOD

Table 14 shows the performance evaluation results for generic SOD algorithms applied to the COCO dataset; note that AP has the same meaning as mAP. AP50 and AP75 denote the AP when the IoU is set to 0.5 or 0.75, respectively, while AP_s, AP_m and AP_l denote the average accuracy for small, medium and large objects, respectively. As shown, IENet [179 achieves the best AP (51.2). In general, the detection performance for large objects is much higher than that for other sizes. HRDNet [78] achieves a value of 32.1 for small objects, whereas MRCenterNet [118] achieves a value of 27.8 for small objects. These results show that increasing the resolution of the input feature with multi-scale training can yield better performance on small objects. All experiments were conducted on a Linux system with NVIDIA GeForce RTX 2080Ti, CUDA 11.7.

Table 14. Detection result on MS COCO test-dev dataset for typical SOD algorithms.

Model	Year	Backbone	AP	AP_{50}	AP_{75}	APs	AP_m	AP_l	FPS
Faster R-CNN [4]	2015	R101-FPN	36.5	58.3	39.3	18.4	40.6	50.6	6
Mask R-CNN [5]	2017	R101-FPN	38.2	60.3	41.7	20.1	41.1	50.2	8.6
YOLOv7-tiny [8]	2022		38.7						286
YOLOv7-E6E [8]	2022		56.8	74.4	62.1				36
FPN [35]	2017	R101-FPN	36.2	59.1	39.0	18.2	39.0	48.2	6
SSD [36]	2015	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8	28
CornerNet* [38]	2018	Hourglass	40.5	56.5	43.1	19.4	42.7	53.9	4.1
FCOS [39]	2019	R101-FPN	41.8	60.3	45.3	25.6	47.7	56.1	7
Efficientdet [71]	2020	Efficientdet	33.8	52.2	35.8	12.0	38.3	51.2	98
RetinaNet [74]	2017	R101-FPN	39.1	59.1	42.3	21.8	42.7	50.2	13.6
FSSD [75]	2017	VGG16	31.8	52.8	33.5	14.2	35.1	45.0	65

Continued on next page

Model	Year	Backbone	AP	AP_{50}	AP_{75}	APs	AP_m	AP_l	FPS
HRDNet* [78]	2021	R101+152	**47.4**	**66.9**	**51.8**	**32.1**	**50.5**	**55.8**	2.8
RHF-Net [79]	2020	ResNet-101	37.7	59.8	40.1	19.9	42.9	51.5	29.1
QueryDet [81]	2021	R50-FPN	38.2	58.6	40.9	23.7	42.0	49.5	13.6
SNIP [95]	2018	DPN [174]	45.7	67.3	51.1	29.3	48.8	57.1	5
SNIPER [96]	2018	ResNet101	46.1	67.0	51.6	29.6	48.9	58.1	5
FR-FDWT [99]	2019	ResNet-101	42.1	63.4	45.7	21.8	45.1	57.1	7
ION [101]	2016	VGG16	24.6	46.3	23.3	7.4	26.2	38.8	1.3
DSSD [102]	2017	ResNet101	33.2	53.3	35.2	13.0	35.4	51.1	6.4
FRCNN-DST [113]	2021	R101-FPN	40.1	59.3	43.2	25.6	43.9	50.9	9
Retina-DST [113]	2021	R101-FPN	41.3	59.9	43.8	25.4	45.1	54.0	13.6
FCOS-DST [113]	2021	R101-FPN	41.6	60.0	44.6	26.5	45.4	53.1	7
PPDet [117]	2020	R101-FPN	39.6	58.0	43.4	23.9	44.1	51.0	7.5
CenterNet++ [118]	2022	ResNet-101	**47.7**	**65.1**	**51.9**	**27.8**	**50.5**	**60.6**	104
DCN* [125]	2017	AlignedIncR	37.5	58.0	40.8	19.4	40.1	52.5	7
RefineDet [172]	2018	ResNet-101	36.4	57.5	39.5	16.6	39.9	51.4	24
D2Det [173]	2020	R101-FPN	45.4	64.0	49.5	25.8	48.7	58.1	4
CoupleNet [175]	2017	ResNet101	33.1	53.5	35.4	11.6	36.3	50.1	8.2
Regionlets [176]	2018	ResNet-101	39.3	59.8	–	21.7	43.7	50.9	–
FitnessNMS [177]	2018	ResNet-101	41.8	60.9	44.9	21.5	45.0	57.5	–
PPYOLOE [178]	2022	CSPRepRes	43.1	60.5	46.6	23.2	45.2	56.9	208
IENet [180]	2021	ResNet-101	**51.2**	**69.3**	**56.1**	**34.5**	**53.8**	**63.6**	3

Table 15. Performance evaluation on the WIDERFACE dataset.

Method	Year	Backbone	AP		
			Easy	Medium	Hard
Faster R-CNN [4]	2015	ResNet50	84.0	72.4	34.7
RetinaNet [74]	2017	ResNet50	94.8	93.8	89.6
S^3FD [124]	2017	VGG16	93.4	92.7	85.4
TFD with GAN [125]	2018	VGG16	93.2	92.2	85.8
Face-MagNet [126]	2018	ResNet101	92.5	91.4	83.1
TinaFace [128]	2020	**ResNet50**	**96.3**	**95.7**	**92.1**
Small Hard Face [132]	2020	VGG16	95.0	93.8	88.5
IENet [180]	2021	ResNet50	96.1	94.7	89.6
RetinaNet	2019	Mobilenet [181]	87.9	80.7	40.3
PyramidBox [182]	2018	ResNet50	95.5	94.6	88.8
RetinaFace [183]	2019	ResNet50	88.6	87.0	80.1

4.4. Performance for small face detection

In Table 15, we evaluate small face detection methods on WIDERFACE [47]. WIDERFACE defines three levels of difficulty: 'easy', 'medium' and 'hard' based on the detection rate of

EdgeBox [180]. As shown, TinaFace [128] achieves the best AP; the AP values for the easy, medium and hard test sets are 96.3, 95.7 and 92.1 respectively. IENet [180] achieves relatively better results, the AP values for the easy, medium and hard test sets are 96.1, 94.7 and 89.6, respectively. TinaFace and IENet both increase the resolution of the prediction feature map, which fully utilizes the fused feature map. IENET also fully incorporates the contextual information. It shows that boosting the resolution of the prediction feature map and incorporating contextual information may be the key to enhancing face detection.

Table 16. Performance evaluation on the TinyPerson dataset.

Method	Year	MR_{50}^{tiny}	MR_{50}^{small}	MR_{25}^{tiny}	MR_{75}^{tiny}	AP_{50}^{tiny}	AP_{50}^{small}	AP_{25}^{tiny}	AP_{75}^{tiny}
Faster R-CNN [4]	2015	87.78	71.31	77.35	98.4	43.55	56.69	64.07	5.35
FPN [36]	2017	87.57	72.56	76.59	98.39	**47.35**	**63.18**	68.43	5.83
FCOS [39]	2019	**96.12**	**84.14**	**89.56**	**99.56**	17.9	35.75	40.49	1.45
RetinaNet [74]	2017	92.66	82.84	81.95	99.13	33.53	48.26	61.51	2.28
Grid R-CNN [185]	2018	87.96	73.16	78.27	98.21	47.14	62.48	**68.89**	**6.38**
DSFD [186]	2019	93.47	78.72	78.02	99.48	31.15	51.64	59.58	1.99
FreeAnchor [187]	2022	88.97	73.67	77.62	98.7	41.36	53.36	63.73	4.00
Li-RCNN [188]	2019	89.22	74.86	82.44	98.78	44.68	62.65	64.77	6.26

Table 17. Performance evaluation on the DOTA-v1.0 dataset; '-O' indicates the detection results with an oriented bounding box.

Method	Year	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Faster R-CNN-O [4]	2015	88.4	73.1	44.9	59.1	73.3	71.5	77.1	90.8	78.9	83.9	48.6	63.0	62.2	64.9	56.2	69.1
Mask R-CNN [5]	2020	76.8	73.5	49.9	57.8	51.3	71.3	79.7	90.4	75.1	67.3	48.5	**70.6**	64.8	64.5	55.9	63.4
CenterNet-O [40]	2019	81.0	64.0	22.6	56.6	38.6	64.0	64.9	90.8	78.0	72.5	44.0	41.1	55.5	55.0	57.4	59.1
RetinaNet-O [74]	2017	88.6	77.6	42.1	58.1	74.5	71.6	79.1	90.8	82.1	74.3	54.7	60.6	62.5	69.5	60.4	68.2
S²A-Net [140]	2019	89.1	82.8	48.3	71.1	78.1	78.3	87.2	90.8	84.9	85.6	60.3	62.6	65.2	69.1	57.9	74.1
SCRDet [141]	2019	**89.9**	80.7	52.1	68.4	68.4	60.3	72.4	**90.9**	87.9	**86.9**	65.0	66.7	66.3	68.2	65.2	72.6
Oriented R-CNN [142]	2019	88.9	83.5	55.3	**76.9**	74.3	82.1	87.5	**90.9**	85.6	85.3	**65.5**	66.8	74.4	70.2	57.3	**76.3**
MRDet [143]	2019	89.5	**84.0**	**55.4**	66.7	76.3	82.1	87.9	90.8	86.9	85.0	52.3	66.0	**76.2**	**76.8**	**67.5**	76.2
BBAVectors [145]	2021	88.4	80.0	50.7	62.2	78.4	79.0	87.9	**90.9**	83.6	84.4	54.1	60.2	65.2	64.3	55.7	72.3
ReDet [147]	2021	88.8	82.6	54.0	74.0	78.1	**84.1**	**88.0**	**90.9**	**87.8**	85.8	61.8	60.4	76.0	68.1	63.6	**76.3**
ROI-Trans [148]	2019	88.6	78.5	43.4	75.9	68.8	73.7	83.6	90.7	77.3	81.5	58.4	53.5	62.8	58.9	47.7	69.6
RepPoints-O [151]	2021	87.0	83.2	54.1	71.2	**80.2**	78.4	87.3	**90.9**	86.0	86.3	59.9	70.5	73.5	72.3	59.0	76.0
CAD-Net [189]	2019	87.8	82.4	49.4	73.5	71.1	63.5	76.6	**90.9**	79.2	73.3	48.4	60.9	62.0	67.0	62.2	69.9

4.5. Performance on small pedestrian detection

Table 16 shows the typical small pedestrian SOD methods on the TinyPerson [136] dataset. MR [184] denotes the miss rate. The size divides are indicated by the superscripts MR and AP, where tiny denotes the size range (2, 20) and small denotes the size range (20, 32). The IoU thresholds utilized for the evaluation are indicated by the subscripts of MR and AP. Among these algorithms, FCOS [39] achieves the best results for all MR evaluations. With an IoU of 0.5, the FPN produced the best AP for small and tiny objects, whereas the Grid R-CNN [185] did so with IoU values of 0.25 and 0.75, respectively.

4.6. Performance on aerial images

In Table 17, we compare the performance of state-of-the-art aerial image object detection algorithms on DOTA-v1.0 [166], which consists of 15 categories: plane (PL), baseball diamond (BD), bridge (BR), ground field track (GTF), small vehicle (SV), large vehicle (LV), tennis court (TC), basketball court (BC), storage tank (SC), soccer-ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP) and helicopter (HC). ReDet and Oriented R-CNN achieve the best mAP value of 76.3. The best AP in each category is marked in bold.

4.7. Further discussion

Based on the experimental results, we further discuss some limitations of existing SOD methods as follows.

1) The framework of SOD is generally modified by popular models like Faster R-CNN, SSD and YOLO; these architectures may not be suitable for small objects, leading to poor performance.

2) Using super-resolution to enhance the resolution of a small object can improve the precision of SOD, but the detection speed will be significantly lower and unable to fulfill the demands of real-world scenarios like real-time monitoring.

3) Transformers have been widely applied in the computer vision field, like DETR [190] in object detection. However, there has not been much research on using transformers for SOD.

4) CNNs are not sensitive to scale changes. There is a need to design feature extractors that are more suitable for scale-aware.

5) MS COCO may not be an ideal benchmark for small objects because small objects account for a relatively small percentage of the dataset.

5. Challenges and future directions

5.1. Challenges of SOD

In addition to the common challenges in object detection, such as continual object detection, imbalance problems, etc. There are typical challenges when it comes to SOD, including feature representation with noise, small object information loss, the effect of the receptive field, location variation sensitivity and the scarcity of small object datasets.

Feature representation with noise. The features of small objects are often contaminated by noise in the background after CNN implementation, making it difficult for the network to capture the discriminative information that is pivotal for the localization and classification tasks. Besides, small objects are often occluded and clustered, so it is particularly difficult to distinguish small objects from noisy clutter and precisely locate their boundaries.

Small object information loss. The features of a small object are virtually eliminated after the down-sampling operations in deeper neural networks due to the small number of pixels occupying each small object. The weak information wipeout of small objects is fatal to SOD because it is hard for the detection head to give accurate predictions in the presence of highly structural representations.

Effect of the receptive field. Large receptive fields are typically chosen by deep neural networks to prevent the loss of information. However, the receptive field for the prediction low-resolution feature map may not match the size of small objects. If the receptive field is larger than the small object, it will cause the object to be detected to become the background, and no features will be extracted by backbone networks, resulting in poor SOD performance.

Location variation sensitivity. Small location deviation of the bounding box in the IoU-based metric produces a more significant disturbance for small objects than for larger objects, which makes it difficult to find a suitable IoU threshold and deliver high-quality positive and negative samples to train the networks.

Scarcity of small object datasets. There are still not enough large-scale general small object datasets to match the cost of annotating small objects. Although MS COCO has a reasonably large amount of small objects (31.62%), each image has too many instances, which leads to the uneven distribution of small objects.

5.2. Future directions

According to the challenges of SOD and the analysis of performance results, we discuss several potential directions for future research in SOD:

1) Weakly supervised, unsupervised and self-supervised SOD. Existing deep learning-based SOD techniques use a fully supervised model. For model training, a sizable number of images with bounding-box annotations (fully supervised information) are required. However, the annotation work is labor-intensive and time-consuming. Weakly supervised object detection can use image-level labels (such as image categories) as supervised signals to train object localization models without the need for pixel-level annotation, which lessens the workload associated with the annotation. Unsupervised salient object detection [191] and self-supervised learning tasks [192] based on contrastive learning have become hot research topics in the past 2 years. Therefore, it is crucial to continue researching the development of weakly supervised learning-based SOD algorithms.

2) Suitable metric for SOD. IoU-based metrics, including the original IoU and its extensions (DIoU, GIou, etc.), are extremely sensitive to the position deviation of small objects and significantly reduce the detection performance when utilized in anchor-based detectors. The authors of [119] use a new Wasserstein distance-based SOD metric, which outperformed the standard fine-tuning baseline by an AP value of 6.7 AP, as well as the state-of-the-art SOTA model by an AP value of 6.0. Therefore, designing a suitable metric for small objects will be crucial to further research.

3) Multi-task joint optimization. Even though techniques like scale-aware training strategies, incorporating contextual information, data augmentation and increasing the resolution of input features help to improve SOD performance, they are still far from adequate, and the combined use of these methods may be able to further improve SOD performance.

4) Open world or few-shot SOD. Few-shot object detection [193] has produced prominent achievements, and SOD in the few-shot scenario is also in urgent need of solutions. Open world SOD seeks to overcome the SOD conundrum while enabling incremental learning in the model, and this type of issue will be a significant research topic in the future.

6. Conclusions

An in-depth review of state-of-the-art deep learning-based SOD algorithms is provided in this paper. We focus on SOD optimization approaches that aim to address the challenges of SOD, including scale-aware training, contextual information incorporation, data augmentation and boosting the resolution of input features. We have summarized the strengths and limitations of these approaches. We have also reviewed methods for crucial SOD tasks, including tiny face detection, tiny pedestrian detection and aerial image object detection. Additionally, detailed experiments were carried out to evaluate the performance of generic SOD algorithms, as well as methods for crucial SOD tasks; we found that boosting the resolution of input features is the most efficient way to improve SOD performance. Finally, we have presented four potential future directions for SOD.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61876186) and Xuzhou Science and Technology Project (KC21300).

Conflict of interest

The authors declare no conflict of interest.

References

1. S. Agarwal, J. O. D. Terrail, F. Jurie, Recent advances in object detection in the age of deep convolutional neural networks, preprint, arXiv:1809.03193.
2. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, (2014), 580–587. https://doi.org/10.1109/CVPR.2014.81
3. R. Girshick, Fast R-CNN, in 2*015 IEEE International Conference on Computer Vision (ICCV)*, (2015), 1440–1448. https://doi.org/10.1109/ICCV.2015.169
4. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2016), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
5. K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, *IEEE Trans. Pattern Anal. Mach. Intell.*, **42** (2020), 386–397. https://doi.org/10.1109/TPAMI.2018.2844175
6. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, real-time object detection, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 779–88. https://doi.org/10.1109/CVPR.2016.91
7. J. Redmon, A. Farhadi, YOLOv3: An incremental improvement, preprint, arXiv:1804.02767.
8. J. C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, preprint, arXiv:2207.02696.
9. K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, et al., T-CNN: tubelets with convolutional neural networks for object detection from videos, *IEEE Trans. Circuits Syst. Video Technol.*, (2017), 2896–2907. https://doi.org/10.1109/TCSVT.2017.2736553

10. T. Yin, X. Zhou, P. Krahenbuhl, Center-based 3d object detection and tracking, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 11784–11793. https://doi.org/10.1109/CVPR46437.2021.01161
11. J. Dai, K. He, J. Sun, Instance-aware semantic segmentation via multi-task network cascades, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 3150–3158. https://doi.org/10.1109/CVPR.2016.343
12. B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 447–456. https://doi.org/10.1109/CVPR.2015.7298642
13. B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Simultaneous detection and segmentation, in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, (2014), 297–312. https://doi.org/10.1007/978-3-319-10584-0_20
14. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., Going deeper with convolutions, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 1–9. https://doi.org/10.1109/CVPR.2015.7298594
15. H. Wang, F. He, Z. Peng, T. Shao, Y. L. Yang, K. Zhou, et al., Understanding the robustness of skeleton-based action recognition under adversarial attack, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* (2021), 14656–14665. https://doi.org/10.1109/CVPR46437.2021.01442
16. L. Wang, Z. Tong, B. Ji, G. Wu, TDN: Temporal difference networks for efficient action recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 1895–1904. https://doi.org/10.48550/arXiv.2012.10071
17. D. Li, Z. Qiu, Y. Pan, T. Yao, H. Li, T. Mei, Representing videos as discriminative sub-graphs for action recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 3310–3319. https://doi.org/10.48550/arXiv.2201.04027
18. C. F. R. Chen, R. Panda, K. Ramakrishnan, R. Feris, J. Cohn, A. Oliva, et al., Deep analysis of cnn-based spatio-temporal representations for action recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 6165–6175. https://doi.org/10.1109/CVPR46437.2021.00610
19. S. Jha, C. Seo, E. Yang, G. P. Joshi, Real time object detection and trackingsystem for video surveillance system, *Multimed. Tools Appl.*, **80** (2021), 3981–3996. https://doi.org/10.1007/s11042-020-09749-x
20. M. A. Farooq, A. A. Khan, A. Ahmad, R. H. Raza, Effectiveness of state-of-the-art super resolution algorithms in surveillance environment, in *Conference on Multimedia, Interaction, Design and Innovation*, **1376** (2021), 79–88. https://doi.org/10.48550/arXiv.2107.04133
21. X. Zheng, X. Li, K. Xu, X. Jiang, T. Sun, Gait identification under surveillance environment based on human skeleton, preprint, arXiv:2111.11720.
22. F. Wu, Q. Wang, J. Bian, H. Xiong, N. Ding, F. Lu, et al., A survey on video action recognition in sports: datasets, methods and applications, preprint, arXiv:2206.01038.
23. C. J. Roros, A. C. Kak, maskGRU: Tracking small objects in the presence of large background motions, preprint, arXiv:2201.00467.
24. Y. B. Can, A. Liniger, D. P. Paudel, L. Van Gool, Structured bird's-eye-view traffic scene understanding from onboard images, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 15641–15650. https://doi.org/10.1109/ICCV48922.2021.01537

25. S. Hampali, S. Stekovic, S. D. Sarkar, C. S. Kumar, F. Fraundorfer, V. Lepetit, Monte carlo scene search for 3d scene understanding, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 13804–13813. https://doi.org/10.1109/CVPR46437.2021.01359
26. J. Hou, B. Graham, M. Niessner, S. Xie, Exploring data-efficient 3d scene understanding with contrastive scene contexts, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 15587–15597. https://doi.org/10.1109/CVPR46437.2021.01533
27. Y. Liu, R. Wang, S. Shan, X. Chen, Structure inference net: object detection using scene-level context and instance-level relationships, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 6985–6994. https://doi.org/10.1109/CVPR.2018.00730
28. M. Schön, M. Buchholz, K. Dietmayer, MGNet: monocular geometric scene understanding for autonomous driving, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 15784–15795. https://doi.org/10.1109/ICCV48922.2021.01551
29. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
30. S. H. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, M. H. Yang, P. Torr, Res2Net: a new multi-scale backbone architecture, in *IEEE Trans. Pattern Anal. Mach. Intell.*, **43** (2021), 652–662. https://doi.org/10.1109/TPAMI.2019.2938758
31. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.
32. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., MobileNets: efficient convolutional neural networks for mobile vision applications, preprint, arXiv:1704.04861.
33. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. C. Chen, MobileNetV2: inverted residuals and linear bottlenecks, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 4510–4520. https://doi.org/10.48550/arXiv.1801.04381
34. 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** (2015), 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824
35. T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 936–944. https://doi.org/10.1109/CVPR.2017.106
36. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, et al., SSD: single shot multibox detector, in *European Conference on Computer Vision*, (2016), 21–37. https://doi.org/10.1007/978-3-319- 46448-0_2
37. C. Zhu, Y. He, M. Savvides, Feature selective anchor-free module for single-shot object detection, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 840–849.
38. H. Law, J. Deng, CornerNet: Detecting objects as paired keypoints, in *European Conference on Computer Vision*, (2018), 765–781. https://doi.org/10.1007/978-3-030-01264-9_45
39. Z. Tian, C. Shen, H. Chen, T. He, FCOS: fully convolutional one-stage object detection, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 9626–9635. https://doi.org/10.1109/ICCV.2019.00972

40. X. Zhou, D. Wang, P. Krähenbühl, Objects as points, preprint, arXiv:1904.07850.
41. C. Eggert, S. Brehm, A. Winschel, D. Zecha, R. Lienhart, A closer look: small object detection in faster R-CNN, in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, (2017), 421–426. https://doi.org/10.1109/ICME.2017.8019550
42. C. Chen, M. Y. Liu, O. Tuzel, J. Xiao, R-CNN for small object detection, in *Asian Conference on Computer Vision*, **10115** (2017), 214–230. https://doi.org/10.1007/978-3-319-54193-8_14
43. T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft COCO: common objects in context, in *European Conference on Computer Vision*, (2014), 740–755. https://doi.org/10.48550/arXiv.1405.0312
44. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, F. Li, ImageNet: a large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), 248–255. https://doi.org/10.1109/CVPR.2009.5206848
45. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, *Int. J. Comput. Vis.*, **88** (2010), 303–338. https://doi.org/10.1007/s11263-009-0275-4
46. Z. Zong, G. Song, Y. Liu, DETRs with collaborative hybrid assignments training, preprint, arXiv:2211.12860.
47. S. Yang, P. Luo, C. C. Loy, X. Tang, WIDER FACE: a face detection benchmark, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 5525–5533. https://doi.org/10.1109/CVPR.2016.596
48. A. B. Chan, Z. S. J. Liang, N. Vasconcelos, Privacy preserving crowd monitoring: counting people without people models or tracking, in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, (2008), 1–7. https://doi.org/10.1109/CVPR.2008.4587569
49. L. Wang, J. Shi, G. Song, Object detection combining recognition and segmentation, in *Asian Conference on Computer Vision*, **4843** (2007), 189.
50. E. Bondi, R. Jain, P. Aggrawal, S. Anand, R. Hannaford, A. Kapoor, et al., BIRDSAI: a dataset for detection and tracking in aerial thermal infrared videos, in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, (2020), 1736–1745. https://doi.org/10.1109/WACV45572.2020.9093284
51. L. Neumann, M. Karg, S. Zhang, C. Scharfenberger, E. Piegert, S. Mistr, et al., NightOwls: a pedestrians at night dataset, in *Asian Conference on Computer Vision*, (2019), 691–705. https://doi.org/10.1007/978-3-030-20887-5_43
52. K. Behrendt, L. Novak, R. Botros, A deep learning approach to traffic lights: Detection, tracking, and classification, in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, (2017), 1370–1377. https://doi.org/10.1109/ICRA.2017.7989163
53. C. Ertler, J. Mislej, T. Ollmann, L. Porzi, G. Neuhold, Y. Kuang, The Mapillary Traffic sign dataset for detection and classification on a global scale, in *European Conference on Computer Vision*, (2020), 68–84. https://doi.org/10.48550/arXiv.1909.04422
54. J. Zhang, M. Huang, X. Jin, X. Li, A real-time chinese traffic sign detection algorithm based on modified yolov2, *Algorithms*, **10** (2017), 127. https://doi.org/10.3390/a10040127
55. D. Tabernik, D. Skočaj, Deep learning for large-scale traffic-sign detection and recognition, preprint, arXiv:1904.00649.

56. Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, S. Hu, Traffic-sign detection and classification in the wild, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 2110–2118. https://doi.org/10.1109/CVPR.2016.232
57. Z. Zhao, P. Zheng, S. T. Xu, X. Wu, Object detection with deep learning: a review, *IEEE Trans. Neural Networks Learn. Syst.*, **30** (2019), 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865
58. 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. https://doi.org/10.1016/j.isprsjprs.2019.11.023
59. K. Oksuz, B. C. Cam, S. Kalkan, E. Akbas, Imbalance problems in object detection: a review, preprint, arXiv:1909.00169.
60. A. G. Menezes, G. de Moura, C. Alves, A. C. P. L. F. de Carvalho, Continual object detection: a review of definitions, strategies, and challenges, preprint, arXiv:2205.15445.
61. L. Jiao, R. Zhang, F. Liu, S. Yang, B. Hou, L. Li, et al., New generation deep learning for video object detection: a survey, *IEEE Trans. Neural Networks Learn. Syst.*, **33** (2022), 3195–3215. https://doi.org/10.1109/TNNLS.2021.3053249
62. L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, et al., A survey of deep learning-based object detection, *IEEE Access*, **7** (2019), 128837–128868. https://doi.org/10.1109/ACCESS.2019.2939201
63. G. Chen, H. Wang, K. Chen, Z. Li, Z. Song, Y. Liu, 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 Cybern, Syst.,* **52** (2022), 936–953. https://doi.org/10.1109/TSMC.2020.3005231
64. K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, et al., MMDetection: open mmlab detection toolbox and benchmark, preprint, arXiv:1906.07155.
65. K. Tong, Y. Wu, F. Zhou, Recent advances in small object detection based on deep learning: A review, *Image Vis. Comput.*, **97** (2020), 103910. https://doi.org/10.1016/j.imavis.2020.103910
66. Y. Liu, P. Sun, N. Wergeles, Y. Shang, A survey and performance evaluation of deep learning methods for small object detection, *Expert Syst. Appl.*, **172** (2021), 114602. https://doi.org/10.1016/j.eswa.2021.114602
67. K. Tong, Y. Wu, Deep learning-based detection from the perspective of small or tiny objects: A survey, *Image Vis. Comput.*, **123** (2022), 104471. https://doi.org/10.1016/j.imavis.2022.104471
68. A. M. Rekavandi, L. Xu, F. Boussaid, A. K. Seghouane, S. Hoefs, M. Bennamoun, A guide to image and video based small object detection using deep learning: case study of maritime surveillance, preprint, arXiv:2207.12926.
69. G. Cheng, X. Yuan, X. Yao, K. Yan, Q. Zeng, J. Han, Towards large-scale small object detection: survey and benchmarks, preprint, arXiv:2207.14096.
70. S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 8759–8768. https://doi.org/10.1109/CVPR.2018.00913
71. M. Tan, R. Pang, Q. V. Le, EfficientDet: scalable and efficient object detection, preprint, arXiv:1911.09070.
72. S. Liu, D. Huang, Y. Wang, Learning spatial fusion for single-shot object detection, preprint, arXiv:1911.09516.

73. G. Ghiasi, T. Y. Lin, R. Pang, Q. V. Le, NAS-FPN: learning scalable feature pyramid architecture for object detection, preprint, arXiv:1904.07392.

74. T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 2999–3007. https://doi.org/10.1109/ICCV.2017.324

75. Z. Li, F. Zhou, FSSD: feature fusion single shot multibox detector, preprint, arXiv:1712.00960.

76. L. Cui, R. Ma, P. Lv, X. Jiang, Z. Gao, B. Zhou, et al., MDSSD: multi-scale deconvolutional single shot detector for small objects, preprint, arXiv:1805.07009.

77. Y. Gong, X. Yu, Y. Ding, X. Peng, J. Zhao, Z. Han, Effective fusion factor in fpn for tiny object detection, preprint, arXiv:2011.02298.

78. Z. Liu, G. Gao, L. Sun, Z. Fang, HRDNet: High-resolution detection network for small objects, preprint, arXiv:2006.07607.

79. Z. Liu, G. Gao, L. Sun, L. Fang, IPG-Net: image pyramid guidance network for small object detection, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2020), 4422–4430. https://doi.org/10.1109/CVPRW50498.2020.00521

80. 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, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2020), 1612–1621. https://doi.org/10.1109/CVPRW50498.2020.00209

81. C. Yang, Z. Huang, N. Wang, QueryDet: cascaded sparse query for accelerating high-resolution small object detection, preprint, arXiv:2103.09136.

82. C. Deng, M. Wang, L. Liu, Y. Liu, Y. Jiang, Extended feature pyramid network for small object detection, *IEEE Trans. Multimedia*, **24** (2022), 1968–1979. https://doi.org/10.1109/TMM.2021.3074273

83. J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, S. Yan, Perceptual generative adversarial networks for small object detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1951–1959. https://doi.org/10.1109/CVPR.2017.211

84. Y. Bai, Y. Zhang, M. Ding, B. Ghanem, SOD-MTGAN: small object detection via multi-task generative adversarial network, in *European Conference on Computer Vision*, **11217** (2018), 210–226. https://doi.org/10.1007/978-3-030-01261-8_13

85. 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, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 9724–9733. https://doi.org/10.1109/ICCV.2019.00982

86. F. Zhang, L. Jiao, L. Li, F. Liu, X. Liu, MultiResolution attention extractor for small object detection, preprint, arXiv:2006.05941.

87. J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, D. Chao, Small-object detection in remote sensing images with end-to-end edge-enhanced gan and object detector network, preprint, arXiv:2003.09085.

88. K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, J. Jiang, Edge-enhanced GAN for remote sensing image super-resolution, *IEEE Trans. Geosci. Remote Sens.*, **57** (2019), 5799–5812. https://doi.org/10.1109/TGRS.2019.2902431

89. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, et al., ESRGAN: enhanced super-resolution generative adversarial networks, in *Proceedings of the European conference on computer vision (ECCV)*, (2018). https://doi.org/10.1007/978-3-030-11021-5_5

90. A. Jolicoeur-Martineau, The relativistic discriminator: a key element missing from standard gan, preprint, arXiv:1807.00734.

91. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial nets, *Adv. Neural Inf. Process Syst.*, **27** (2014). https://doi.org/10.48550/arXiv.1406.2661

92. J. Cao, Y. Pang, S. Zhao, X. Li, High-level semantic networks for multi-scale object detection, *IEEE Trans. Circuits Syst. Video Technol.*, **30** (2020), 3372–3386. https://doi.org/10.1109/TCSVT.2019.2950526

93. K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, *IEEE Signal Process. Lett.*, **23** (2016), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342

94. Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, X. Hu, Scale-aware face detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1913–1922. https://doi.org/10.1109/CVPR.2017.207

95. B. Singh, L. S. Davis, An analysis of scale invariance in object detection - snip, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 3578–3587. https://doi.org/10.1109/CVPR.2018.00377

96. B. Singh, M. Najibi, L. S. Davis, SNIPER: efficient multi-scale training, *Adv. Neural Inf. Process Syst.*, **31** (2018). https://doi.org/10.48550/arXiv.1805.09300

97. Y. Kim, B. N. Kang, D. Kim, SAN: learning relationship between convolutional features for multi-scale object detection, in *European Conference on Computer Vision*, **11209** (2018), 328–343. https://doi.org/10.1007/978-3-030-01228-1_20

98. Y. Li, Y. Chen, N. Wang, Z. Zhang, Scale-aware trident networks for object detection, preprint, arXiv:1901.01892.

99. J. Peng, M. Sun, Z. X. Zhang, T. Tan, J. Yan, POD: practical object detection with scale-sensitive network, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 9606–9615. https://doi.org/10.1109/ICCV.2019.00970

100. A. Oliva, A. Torralba, The role of context in object recognition, *Trends Cogn. Sci.*, **11** (2007), 520–527. https://doi.org/10.1016/j.tics.2007.09.009

101. S. Bell, C. L. Zitnick, K. Bala, R. Girshick, Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 2874–2883. https://doi.org/10.1109/CVPR.2016.314

102. C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, DSSD: deconvolutional single shot detector, preprint, arXiv:1701.06659.

103. W. Xiang, D. Q. Zhang, H. Yu, V. Athitsos, Context-aware single-shot detector, in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, (2018), 1784–1793. https://doi.org/10.1109/WACV.2018.00198

104. X. Chen, A. Gupta, Spatial memory for context reasoning in object detection, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 4106–4116. https://doi.org/10.1109/ICCV.2017.440

105. K. Fu, J. Li, L. Ma, K. Mu, Y. Tian, Intrinsic relationship reasoning for small object detection, preprint, arXiv:2009.00833.

106. J. S. Lim, M. Astrid, H. J. Yoon, S. I. Lee, Small object detection using context and attention, in *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, (2021), 181–186. https://doi.org/10.1109/ICAIIC51459.2021.9415217

107. A. Bochkovskiy, C. Y. Wang, H. Y. M. Liao, YOLOv4: optimal speed and accuracy of object detection, preprint, arXiv:2004.10934.

108. H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, Mixup: beyond empirical risk minimization, preprint, arXiv:1710.09412.

109. S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, CutMix: regularization strategy to train strong classifiers with localizable features, in *Proceedings of the IEEE International Conference on Computer Vision*, (2019), 6023–6032. https://doi.org/10.1109/ICCV.2019.00612

110. M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, K. Cho, Augmentation for small object detection, preprint, arXiv:1902.07296.

111. C. Chen, Y. Zhang, Q. Lv, S. Wei, X. Wang, X. Sun, et al., RRNet: a hybrid detector for object detection in drone-captured images, in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, (2019), 100–108. https://doi.org/10.1109/ICCVW.2019.00018

112. F. O. Unel, B. O. Ozkalayci, C. Cigla, The power of tiling for small object detection, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2019), 582–591. https://doi.org/10.1109/CVPRW.2019.00084

113. Y. Chen, P. Zhang, Z. Li, Y. Li, X. Zhang, L. Qi, et al., Dynamic scale training for object detection, preprint, arXiv:2004.12432.

114. B. Zoph, E. D. Cubuk, G. Ghiasi, T. Y. Lin, J. Shlens, Q. V. Le, Learning data augmentation strategies for object detection, in *European Conference on Computer Vision*, (2020), 566–583. https://doi.org/10.1007/978-3-030-58583-9_34

115. E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q. V. Le, AutoAugment: learning augmentation policies from data, preprint, arXiv:1805.09501.

116. Y. Chen, Y. Li, T. Kong, L. Qi, R. Chu, L. Li, et al., Scale-aware automatic augmentation for object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 9563–9572. https://doi.org/10.1109/CVPR46437.2021.00944

117. N. Samet, S. Hicsonmez, E. Akbas, Reducing label noise in anchor-free object detection, preprint, arXiv:2008.01167.

118. K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, CenterNet++ for object detection, preprint, arXiv:2204.08394.

119. J. Wang, C. Xu, W. Yang, L. Yu, A normalized gaussian wasserstein distance for tiny object detection, preprint, arXiv:2110.13389.

120. C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, G. Xia, RFLA: Gaussian receptive field based label assignment for tiny object detection, in *Proceedings of the European conference on computer vision (ECCV)*, (2022). https://doi.org/10.1007/978-3-031-20077-9_31

121. C. Lee, S. Park, H. Song, J. Ryu, S. Kim, H. Kim, et al., Interactive multi-class tiny-object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 14136–14145. https://doi.org/10.1109/CVPR52688.2022.01374

122. F. C. Akyon, S. Altinuc, A. Temizel, Slicing aided hyper inference and fine-tuning for small object detection, preprint, arXiv:2202.06934.

123. P. Hu, D. Ramanan, Finding tiny faces, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1522–1530. https://doi.org/10.1109/CVPR.2017.166

124. S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, S. Z. Li, S^3FD: single shot scale-invariant face detector, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 192–201. https://doi.org/10.1109/ICCV.2017.30

125. Y. Bai, Y. Zhang, M. Ding, B. Ghanem, Finding tiny faces in the wild with generative adversarial network, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 21–30. https://doi.org/10.1109/CVPR.2018.00010

126. P. Samangouei, M. Najibi, L. Davis, R. Chellappa, Face-magnet: magnifying feature maps to detect small faces, preprint, arXiv:1803.05258.

127. C. Zhu, R. Tao, K. Luu, M. Savvides, Seeing small faces from robust anchor's perspective, preprint, arXiv:1802.09058.

128. Y. Zhu, H. Cai, S. Zhang, C. Wang, Y. Xiong, TinaFace: strong but simple baseline for face detection, preprint, arXiv:2011.13183.

129. J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, et al., Deformable convolutional networks, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 764–773. https://doi.org/10.1109/ICCV.2017.89

130. Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, Distance-IoU loss: faster and better learning for bounding box regression, in *Proceedings of the AAAI conference on artificial intelligence*, **34** (2019), 12993–13000. https://doi.org/10.1609/aaai.v34i07.6999

131. A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), 761–769. https://doi.org/10.1109/CVPR.2016.89

132. Z. Zhang, W. Shen, S. Qiao, Y. Wang, B. Wang, A. Yuille, Robust face detection via learning small faces on hard images, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, (2020), 1361–1370. https://doi.org/10.48550/arXiv.1811.11662

133. T. Song, L. Sun, D. Xie, H. Sun, S. Pu, Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation, preprint, arXiv:1807.01438.

134. S. Das, P. S. Mukherjee, U. Bhattacharya, Seek and you will find: a new optimized framework for efficient detection of pedestrian, preprint, arXiv:1912.10241.

135. W. Liu, S. Liao, W. Ren, W. Hu, Y. Yu, High-level semantic feature detection: a new perspective for pedestrian detection, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 5182–5191. https://doi.org/10.1109/CVPR.2019.00533

136. X. Yu, Y. Gong, N. Jiang, Q. Ye, Z. Han, Scale match for tiny person detection, in 2020 *IEEE Winter Conference on Applications of Computer Vision (WACV)*, (2020), 1246–1254. https://doi.org/10.1109/WACV45572.2020.9093394

137. D. Božić-Štulić, Ž. Marušić, S. Gotovac, Deep learning approach in aerial imagery for supporting land search and rescue missions, *Int. J. Comput Vis.*, **127** (2019), 1256–1278. https://doi.org/10.1007/s11263-019-01177-1

138. G. Adaimi, S. Kreiss, A. Alahi, Perceiving traffic from aerial images, preprint, arXiv:2009.07611.

139. C. Gheorghe, N. Filip, Road traffic analysis using unmanned aerial vehicle and image processing algorithms, in *2022 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)*, (2022), 1–5. https://doi.org/10.1109/AQTR55203.2022.9802058

140. J. Han, J. Ding, J. Li, G. S. Xia, Align deep features for oriented object detection, *IEEE Trans. Geosci. Remote Sens.*, **60** (2022), 5602511. https://doi.org/10.1109/TGRS.2021.3062048
141. X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, et al., SCRDet: towards more robust detection for small, cluttered and rotated objects, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 8231–8240. https://doi.org/10.1109/ICCV.2019.00832
142. X. Xie, G. Cheng, J. Wang, X. Yao, J. Han, Oriented r-cnn for object detection, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 3500–3509. https://doi.org/10.1109/ICCV48922.2021.00350
143. R. Qin, Q. Liu, G. Gao, D. Huang, Y. Wang, MRDet: a multi-head network for accurate oriented object detection in aerial images, preprint, arXiv:2012.13135.
144. X. Zhang, E. Izquierdo, K. Chandramouli, Dense and small object detection in uav vision based on cascade network, in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, (2019), 118–126. https://doi.org/10.1109/ICCVW.2019.00020
145. J. Yi, P. Wu, B. Liu, Q. Huang, H. Qu, D. Metaxas, Oriented object detection in aerial images with box boundary-aware vectors, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, (2021), 2150–2159. https://doi.org/10.1109/WACV48630.2021.00220
146. O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention*, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
147. J. Han, J. Ding, N. Xue, G. S. Xia, ReDet: a rotation-equivariant detector for aerial object detection, preprint, arXiv:2103.07733.
148. J. Ding, N. Xue, Y. Long, G. S. Xia, Q. Lu, Learning ROI transformer for oriented object detection in aerial images, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 2849–2858. https://doi.org/10.1109/CVPR.2019.00296
149. M. Zand, A. Etemad, M. Greenspan, Oriented bounding boxes for small and freely rotated objects, *IEEE Trans. Geosci. Remote Sensing*, **60** (2022), 1–15. https://doi.org/10.1109/TGRS.2021.3076050
150. Z. Yang, S. Liu, H. Hu, L. Wang, S. Lin, RepPoints: point set representation for object detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 9657–9666. https://doi.org/10.1109/ICCV.2019.00975
151. W. Li, Y. Chen, K. Hu, J. Zhu, Oriented reppoints for aerial object detection, preprint, arXiv:2105.11111.
152. C. Xu, J. Wang, W. Yang, L. Yu, Dot distance for tiny object detection in aerial images, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2021), 1192–1201, https://doi.org/10.1109/CVPRW53098.2021.00130
153. X. Fang, F. Hu, M. Yang, T. Zhu, R. Bi, Z. Zhang, Z. Gao, Small object detection in remote sensing images based on super-resolution, *Pattern Recognit. Lett.*, **153** (2022), 107–112. https://doi.org/10.1016/j.patrec.2021.11.027.5
154. 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. Sel. Top Appl. Earth Obs. Remote Sens.*, **14** (2021), 2148–2161. https://doi.org/10.1109/JSTARS.2020.3046482
155. O. C. Koyun, R. K. Keser, İ. B. Akkaya, B. U. Töreyin, Focus-and-detect:a small object detection framework for aerial images, *Signal Process. Image Commun.*, **104** (2022), 116675. https://doi.org/10.1016/j.image.2022.116675

156. B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, et al., Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 1931–1939. https://doi.org/10.1109/CVPR.2015.7298803

157. Y. Yuan, W. Yang, W. Ren, J. Liu, W. J. Scheirer, Z. Wang, UG^{2+}: a collective benchmark effort for evaluating and advancing image understanding in poor visibility environments, preprint, arXiv:1904.04474.

158. H. Nada, V. A. Sindagi, H. Zhang, V. M. Patel, Pushing the limits of unconstrained face detection: a challenge ataset and baseline results, in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, (2018), 1–10. https://doi.org/10.1109/BTAS.2018.8698561

159. M. K. Yucel, Y. C. Bilge, O. Oguz, N. Ikizler-Cinbis, P. Duygulu, R. G. Cinbis, Wildest faces: face detection and recognition in violent settings, preprint, arXiv:1805.07566.

160. S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, G. Guo, WiderPerson: A diverse dataset for dense pedestrian detection in the wild, *IEEE Trans. Multimedia*, **22** (2020), 380–393. https://doi.org/10.1109/TMM.2019.2929005

161. M. Braun, S. Krebs, F. Flohr, D. M. Gavrila, The eurocity persons dataset: a novel benchmark for object detection, preprint, arXiv:1805.07193.

162. S. Zhang, R. Benenson, B. Schiele, CityPersons: a diverse dataset for pedestrian detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 4457–4465. https://doi.org/10.1109/CVPR.2017.474

163. P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: a benchmark, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), 304–311. https://doi.org/10.1109/CVPR.2009.5206631

164. P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, et al., Detection and tracking meet drones challenge, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 7380–7399. https://doi.org/10.1109/TPAMI.2021.3119563

165. D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, et al., The unmanned aerial vehicle benchmark: object detection and tracking, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 370–386. https://doi.org/10.1007/s11263-019-01266-1

166. G. S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, et al., DOTA: a large-scale dataset for object detection in aerial images, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 3974–3983. https://doi.org/10.1109/CVPR.2018.00418

167. G. Cheng, J. Han, P. Zhou, L. Guo, Multi-class geospatial object detection and geographic image classification based on collection of part detectors, *ISPRS J. Photogramm. Remote Sens.*, **98** (2014), 119–132. https://doi.org/10.1016/j.isprsjprs.2014.10.002

168. H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, J. Jiao, Orientation robust object detection in aerial images using deep convolutional neural network, in *2015 IEEE International Conference on Image Processing (ICIP)*, (2015), 3735–3739. https://doi.org/10.1109/ICIP.2015.7351502

169. L. Tuggener, I. Elezi, J. Schmidhuber, M. Pelillo, T. Stadelmann, DeepScores-a dataset for segmentation, detection and classification of tiny objects, in *2018 24th International Conference on Pattern Recognition (ICPR)*, (2018), 3704–3709. https://doi.org/10.1109/ICPR.2018.8545307

170. A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, (2012), 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074
171. S. Song, S. P. Lichtenberg, J. Xiao, SUN RGB-D: a rgb-d scene understanding benchmark suite, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 567–576. https://doi.org/10.1109/CVPR.2015.7298655
172. S. Zhang, L. Wen, X. Bian, Z. Lei, S. Z. Li, Single-shot refinement neural network for object detection, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 4203–4212. https://doi.org/10.1109/CVPR.2018.00442
173. J. Cao, H. Cholakkal, R. M. Anwer, F. S. Khan, Y. Pang, L. Shao, D2Det: towards high quality object detection and instance segmentation, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 11482–11491.
174. Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, J. Feng, Dual path networks, *Adv. Neural Inf. Process Syst.*, **30** (2017). https://doi.org/10.48550/arXiv.1707.01629
175. Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, H. Lu, CoupleNet: coupling global structure with local parts for object detection, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 4146–4154. https://doi.org/10.1109/ICCV.2017.444
176. H. Hu, J. Gu, Z. Zhang, J. Dai, Y. Wei, Relation networks for object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 3588–3597. https://doi.org/ 10.1109/CVPR.2018.00378
177. L. Tychsen-Smith, L. Petersson, Improving object localization with fitness nms and bounded iou loss, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 6877–6885. https://doi.org/10.1109/CVPR.2018.00719
178. S. Xu, X. Wang, W. Lv, Q. Chang, C. Cui, K. Deng, et al., PP-YOLOE: an evolved version of YOLO, preprint, arXiv:2203.16250.
179. J. Leng, Y. Ren, W. Jiang, X. Sun, Y. Wang, Realize your surroundings: exploiting context information for small object detection, *Neurocomputing*, **433** (2021). https://doi.org/10.1016/j.neucom.2020.12.093
180. C. L. Zitnick, P. Dollár, Edge Boxes: locating object proposals from edges, in *European Conference on Computer Vision*, (2014), 391–405. https://doi.org/10.1007/978-3-319-10602-1_26
181. A. Howard, M. Sandler, G. Chu, L. C. Chen, B. Chen, M. Tan, et al., Searching for MobileNetV3, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 1314–1324. https://doi.org/10.1109/ICCV.2019.00140
182. X. Tang, D. K. Du, Z. He, J. Liu, PyramidBox: a context-assisted single shot face detector, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 797–813. https://doi.org/10.1007/978-3-030-01240-3_49
183. J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, S. Zafeiriou, RetinaFace: single-stage dense face localisation in the wild, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 5203–5212. https://doi.org/10.1109/CVPR42600.2020.00525
184. Z. Liu, J. Du, F. Tian, J. Wen, MR-CNN: a multi-scale region-based convolutional neural network for small traffic sign recognition, *IEEE Access*, **7** (2019), 57120–57128. https://doi.org/10.1109/ACCESS.2019.2913882

185. X. Lu, B. Li, Y. Yue, Q. Li, J. Yan, Grid R-CNN, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 7355–7364, https://doi.org/10.1109/CVPR.2019.00754. (2018).

186. J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, et al., DSFD: dual shot face detector, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 5060–5069. https://doi.org/10.1109/CVPR.2019.00520

187. X. Zhang, F. Wan, C. Liu, R. Ji, Q. Ye, FreeAnchor: learning to match anchors for visual object detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 3096–3109. https://doi.org/10.48550/arXiv.1909.02466

188. J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, D. Lin, Libra R-CNN: towards balanced learning for object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 821–830. https://doi.org/10.1109/CVPR.2019.00091

189. G. Zhang, S. Lu, W. Zhang, CAD-Net: a context-aware detection network for objects in remote sensing imagery, *IEEE Trans. Geosci. Remote Sens.*, **57** (2019), 10015–10024. https://doi.org/10.1109/TGRS.2019.2930982

190. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in *European Conference on Computer Vision*, **12346** (2020), 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

191. S. Li, F. Liu, L. Jiao, X. Liu, P. Chen, Learning salient feature for salient object detection without labels, *IEEE Trans. Cybern.*,**53** (2022), 1012–1025. https://doi.org/10.1109/TCYB.2022.3209978

192. F. Liu, X. Qian, L. Jiao, X. Zhang, L. Li, Y. Cui, Contrastive learning-based dual dynamic gcn for sar image scene classification, *IEEE Trans. Neural Networks Learn Syst.*, (2022), 1–15. https://doi.org/10.1109/TNNLS.2022.3174873

193. Y. Du, F. Liu, L. Jiao, Z. Hao, S. Li, X. Liu, et al., Augmentative contrastive learning for one-shot object detection, *Neurocomputing*, **513** (2022), 13–24. https://doi.org/10.1016/j.neucom.2022.09.125