Towards Large-Scale Small Object Detection: Survey and Benchmarks

Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han, *Fellow, IEEE*

***Abstract*—With the rise of deep convolutional neural networks, object detection has achieved prominent advances in past years. However, such prosperity could not camouflage the unsatisfactory situation of Small Object Detection (SOD), one of the notoriously challenging tasks in computer vision, owing to the poor visual appearance and noisy representation caused by the intrinsic structure of small targets. In addition, large-scale dataset for benchmarking small object detection methods remains a bottleneck. In this paper, we first conduct a thorough review of small object detection. Then, to catalyze the development of SOD, we construct two large-scale Small Object Detection dAtasets (SODA), SODA-D and SODA-A, which focus on the Driving and Aerial scenarios respectively. SODA-D includes 24828 high-quality traffic images and 278433 instances of nine categories. For SODA-A, we harvest 2513 high resolution aerial images and annotate 872069 instances over nine classes. The proposed datasets, as we know, are the first-ever attempt to large-scale benchmarks with a vast collection of exhaustively annotated instances tailored for multi-category SOD. Finally, we evaluate the performance of mainstream methods on SODA. We expect the released benchmarks could facilitate the development of SOD and spawn more breakthroughs in this field.**

***Index Terms*—Benchmark, convolutional neural networks, deep learning, object detection, small object detection.**

I. Introduction

OBJECT detection is an essential task which aims at categorizing and locating the objects of interest in images/videos. Thanks to the enormous volume of data and powerful learning ability of deep Convolutional Neural Networks (CNNs), object detection has scored remarkable achievements in recent years [1], [2], [3], [4], [5]. Small[1] Object Detection (SOD), as a sub-field of generic object detection, which concentrates on detecting those objects with small size, is of great theoretical and practical significance in various scenarios such as surveillance, drone scene analysis, pedestrian detection, traffic sign detection in autonomous driving, *etc*.

Albeit the substantial progresses have been made in generic object detection, the research of SOD proceeded at a relatively slow pace. To be more specific, there remains a huge performance gap in detecting small and normal sized objects even for leading detectors. Taking DyHead [9], one of the state-of-the-art detectors, as an example, the mean Average Precision (mAP) metric of small objects on COCO [6] test-dev set obtained by DyHead is only 28.3%, significantly lag behind that of objects with medium and large sizes (50.3% and 57.5% respectively). We posit such performance degradation originates the following two-fold: 1) the intrinsic difficulty of learning proper representation from limited and distorted information of small objects; 2) the scarcity of large-scale dataset for small object detection.

The low-quality feature representation of small objects can be attributed to their limited sizes and the generic feature extraction paradigm. Concretely, the current prevailing feature extractors [10], [11], [12] usually down-sample the feature maps to diminish the spatial redundancy and learn high dimensional features, which unavoidably extinguishes the representation of tiny objects. Moreover, the features of small objects are inclined to be contaminated by background and other instances after the convolution process, making the network can hardly capture the discriminative information that is pivotal for the subsequent tasks. To tackle this problem, researchers have proposed a series of work, which can be categorized into six groups: sample-oriented methods, scale-aware methods, attention-based methods, feature-imitation methods, context-modeling methods, and focus-and-detect methods. We will discuss these approaches exhaustively in the review part.

To alleviate the data scarcity, some datasets tailored for small object detection have been proposed, e.g., SOD [28] and TinyPerson [7]. However, these small-scale datasets cannot meet the needs of training supervised CNN-based algorithms, which are "hungry" for a substantial amount of labeled data. In addition, several public datasets contain a considerable number of small objects, such as WiderFace [8], SeaPerson [29] and DOTA[2] [30], *etc*. Unfortunately, these datasets are either designed for single-category detection task (face detection or pedestrian detection) which usually follows a relatively certain pattern, or among which tiny objects merely distribute in a few

Manuscript received 25 December 2022; revised 2 June 2023; accepted 25 June 2023. Date of publication 29 June 2023; date of current version 3 October 2023. This work was supported in part by the National Science Foundation of China under Grants 62136007 and U20B2068, and in part by the Natural Science Basic Research Program of Shaanxi under Grants 2021JC-16 and 2023-JC-ZD-36. Recommended for acceptance by X. Bai. *(Corresponding author: Junwei Han.)*

The authors are with the School of Automation, Northwestern Polytechnical University, Xi'an 710021, China (e-mail: gcheng@nwpu.edu.cn; shaunyuan@mail.nwpu.edu.cn; yaoxiwen@nwpu.edu.cn; 2021202443@mail.nwpu.edu.cn; zengqinghua@mail.nwpu.edu.cn; xiexing@mail.nwpu.edu.cn; junweihan2010@gmail.com).

Datasets and codes are available at: https://shaunyuan22.github.io/SODA.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2023.3290594, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2023.3290594

[1]Here by "small" we mean the size of the object is relatively limited and often determined by an area [6] or length [7], [8] threshold.

[2]The term **DOTA** in our paper represents DOTA-v2.0.

TABLE I
SUMMARY OF SEVERAL SURVEYS RELATED TO OBJECT DETECTION

Title	Publication	Descriptions
Deep Learning for Generic Object Detection: A Survey [13]	IJCV 2020	A comprehensive survey of the recent progresses driven by deep learning techniques in generic object detection
Object Detection With Deep Learning: A Review [14]	TNNLS 2019	A systematic review on deep learning-based detection frameworks for generic object detection and other subtasks
Object detection in 20 years: A survey [15]	PIEEE 2023	A survey focuses on object detection spanning over a quarter-century's time
Pedestrian detection: an evaluation of the state of the art [16]	TPAMI 2011	A detailed evaluation of pedestrian detectors in monocular images
From Handcrafted to Deep Features for Pedestrian Detection: A Survey [17]	TPAMI 2021	A through survey for pedestrian detection approaches based on handcrafted features and deep features
Text Detection and Recognition in Imagery: A Survey [18]	TPAMI 2014	A systematic survey related to automatic text detection and recognition in color images
A survey on object detection in optical remote sensing images [19]	JPRS 2016	A review of recent progress about object detection in optical remote sensing images
Object detection in optical remote sensing images: A survey and a new benchmark [20]	JPRS 2020	A thorough review of deep learning based methods for object detection in aerial images
Vision for Looking at Traffic Lights: Issues, Survey, and Perspectives [21]	TITS 2016	An overview of traffic light recognition research in relation to driver assistance systems
Object Detection Using Deep Learning Methods in Traffic Scenarios [22]	CS 2021	A survey dedicated to object detection in traffic scenarios based on deep learning methods
Imbalance Problems in Object Detection: A Review [23]	TPAMI 2020	A comprehensive review of the imbalance problems in object detection
Weakly Supervised Object Localization and Detection: A Survey [24]	TPAMI 2021	A systematic survey on weakly supervised object localization and detection
Deep learning-based detection from the perspective of small or tiny objects: A survey [25]	IVC 2022	A review of existing deep learning-based detection methods which can be utilized for small or tiny objects
A survey and performance evaluation of deep learning methods for small object detection [26]	ESWA 2021	A survey of recently developed deep learning methods for small object detection
A Survey of the Four Pillars for Small Object Detection: Multiscale Representation, Contextual Information, Super-Resolution, and Region Proposal [27]	TSMCS 2022	A review of small object detection based on four genres of techniques: multiscale representation, contextual information, super-resolution, and region-proposal

The top are the surveys focusing on the generic object detection and specific tasks, and the bottom are the existing reviews of small object detection.

categories (*small-vehicle* in DOTA dataset). In a nutshell, the currently available datasets could not support the training of deep learning-based models customized for small object detection, as well as serve as an impartial benchmark for evaluating multi-category SOD algorithms. Whilst, as a foundation for building data-driven deep CNN models, the accessibility of large-scale datasets such as PASCAL VOC [31], ImageNet [32], COCO [6], and DOTA [30] is of great significance for both the academic and industrial communities, and each of which noticeably boosts the development of object detection in related fields. This inspires us to think: can we build a large-scale dataset, where the objects of multiple categories have very limited sizes, to serve as a benchmark that can be adopted to verify the design of small object detection framework and facilitate the further research of SOD?

Taking the aforementioned problems into account, we construct two large-scale Small Object Detection dAtasets (SODA), SODA-D and SODA-A, which focus on the Driving and Aerial scenarios respectively. The proposed SODA-D is built on top of MVD [33] and our data, where the former is a dataset dedicated to pixel-level understanding of street scenes, and the latter is mainly captured by on-board cameras and mobile phones. With 24828 well-chosen and high-quality images of driving scenarios, we annotate 278433 instances of nine categories with horizontal bounding boxes. SODA-A is the benchmark specialized for SOD under aerial scenes, which has 872069 instances with oriented box annotation across nine classes. It contains 2513 high-resolution images extracted from Google Earth.

A. Problem Definition

Object detection aims to classify and locate instances. Small object detection or tiny object detection, as the term suggests, merely focuses on detecting those objects with limited sizes. In this task, the terms *tiny* and *small* are typically defined by an area threshold [6] or length threshold [7], [8]. Take COCO [6] as an example, the objects occupying an area less than and equal to 1024 pixels come to *small* category.

B. Comparisons With Previous Reviews

Quite a number of surveys about object detection have been published in recent years [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], and our review differs from the existing ones mainly in two aspects.

1. *A comprehensive and timely review dedicated to small object detection task across multiple domains:* Most of the previous reviews (as in Table I) concentrate on either generic object detection [13], [14], [15] or specific object detection task such as pedestrian detection [16], [17], text detection [18], detection in remote sensing images [19], [20], and detection under traffic scenarios [21], [22], *etc*. Furthermore, there already exist several reviews paying their attention to small object detection [25], [26], [27], however, they either fail to the comprehensiveness and in-depth analysis because only partial reviews on limited areas were conducted, or categorize considerable algorithms belonging to generic detection as small object detection methods, which is indeed not rigorous for a SOD-oriented survey. By narrowly casting our sight to small/tiny objects, we extensively review hundreds of literature related to SOD task which covers a broad spectrum of research fields, including face detection, pedestrian detection, traffic sign detection, vehicle detection, object detection in aerial images, to name a few. As a result, *we provide a systematic survey of small object detection and an understandable and highly structured taxonomy, which organizes SOD approaches into six major categories based on the techniques involved and is radically different from previous ones*.

2. *Two large-scale benchmarks customized for small object detection were proposed, on which in-depth evaluation and analysis of several representative detection algorithms were performed:* Previous reviews mainly resort to general detection datasets such as PASCAL VOC [31] and COCO [6] to conduct evaluation, which is dominated by the medium-sized and large-sized instances and thereby failing to embody the authentic performance of related methods when it comes to small objects. Instead, we present the large-scale benchmark SODA and on top of which, a thorough evaluation of several representative generic

object detection methods and newly published SOD approaches was provided.

C. Scope

Object detection in early period usually integrated hand-crafted features [34], [35], [36] and machine learning approaches [37], [38] to recognize the objects of interest. The methods following this sophisticated philosophy perform catastrophically poorly in small objects due to their limited capability of scale variation. After 2012, the powerful learning ability of deep convolutional network [39] brings a glimmer of hope to the whole detection community, especially considering that object detection had reached a plateau after 2010 [40]. The seminal work [40] broken the ice and since then, an increasing number of detection methods based on deep neural networks were proposed, whereafter, object detection entering the deep learning era [15]. Thanks to the outstanding modeling ability of deep networks for scale variation and powerful abstraction of information, small object detection obtains an unprecedented improvement. Therefore, our review focuses on the major development of deep learning-based SOD methods. To sum up, the main contributions of this paper are in three folds:

1. Reviewing the development of small object detection in the deep-learning era and providing a systematic survey of the recent progress in this field, which can be grouped into six categories: sample-oriented methods, scale-aware methods, attention-based methods, feature-imitation methods, context-modeling methods, and focus-and-detect approaches. Except for the taxonomies, in-depth analysis about the pros and cons of these methods were also provided. Meanwhile, we review dozens of datasets that span over multiple areas which relate to small object detection.

2. Releasing two large-scale benchmarks for small object detection, where the first one was dedicated to driving scenarios and the other was specialized for aerial scenes. The proposed datasets are the first-ever attempt to large-scale benchmarks tailored for SOD. We hope these two exhaustively annotated benchmarks could help the researchers to develop and verify effective frameworks for SOD and facilitate more breakthroughs in this field.

3. Investigating the performance of several representative object detection methods on our datasets, and providing in-depth analyses according to the quantitative and qualitative results, which could benefit the algorithm design of small object detection afterwards.

The remainder of this paper is organized as follows. In Section II, we conduct a comprehensive survey of small object detection. And a thorough review on several publicly available benchmarks related to small object detection is given in Section III. In Section IV, we elaborate the collection and annotation, as well as the data characteristics about the proposed benchmarks. In Section V, the results and analyses of several representative methods on our benchmarks are provided. Finally, we conclude our work and discuss the prospective research directions of small object detection.

II. Review on Small Object Detection

A. Main Challenges

In addition to some common challenges in generic object detection such as intra-class variations, inaccurate localization, occluded object detection, *etc*., typical issues exist when it comes to SOD tasks, primarily including object information loss, noisy feature representation, low tolerance for bounding box perturbation and inadequate samples.

Information Loss: Current prevailing object detectors [1], [2], [3], [4], [5], [9] usually include a backbone network and a detection head, where the latter makes decision depends on the representation output by the former. Such paradigm was proven to be effective and gives rise to the unprecedented success. However, the generic feature extractor [10], [11], [12] usually leverage sub-sampling operations to filter noisy activation [41] and reduce the spatial resolution of feature maps, thus inevitably losing the information of objects. Such information loss will scarcely impair the performance of large or medium-sized objects to a certain extent, considering that the final features still retain enough information of them. Unfortunately, this is fatal for small objects, because the detection head can hardly give accurate predictions on top of the highly structural representations, in which the weak signals of small objects were almost wiped out.

Noisy Feature Representation: Discriminative features are crucial for both the classification and localization tasks [42], [43]. Small objects often have poor-quality appearance, consequently it is intractable to learn representations with discrimination from their distorted structures. Whilst, the regional features of small objects are inclined to be contaminated by the background and other instances, introducing noise to the learned representation further. To sum up, the feature representations of small objects are apt to suffer from the noise, hindering the subsequent detection.

Low Tolerance for Bounding Box Perturbation: Localization, as one of the primary tasks of detection, is formulated as a regression problem in most detection paradigms [1], [3], [4], [44], [45], [46], [47], and generally the Intersection over Union (IoU) metric was adopted to evaluate the accuracy. Nevertheless, localizing small objects is tougher than larger ones. As shown in Fig. 1, a slight deviation (6 pixels along the diagonal direction) of predicted box for a small object causes significant drop on IoU (from 100% to 32.5%) compared to medium and large objects (56.6% and 71.8%). Meanwhile, a greater variance (say, 12 pixels) further exacerbates the situation, and the IoU drops to poorly 8.7% for small objects. That is to say, small objects have a lower tolerance for bounding box perturbation compared with larger ones, aggravating the learning of regression branch.

Inadequate Samples for Training: Selecting positive and negative samples is an indispensable step towards training a high performance detector. However, things get tougher when it comes to small objects. Concretely, small instances occupy fairly small regions and have limited overlaps to priors (anchors or points). This tremendously challenges conventional label assignment strategies [1], [3], [4], [47], [48], which collect *pos/neg* samples based on the overlaps of boxes or center regions, leading

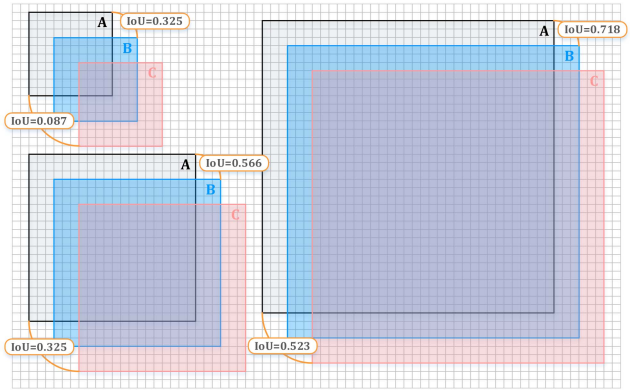

Fig. 1. Low tolerance of small objects for bounding box perturbation. Top-left, bottom-left and right represent small object (20×20 pixels, a grid denotes two pixels), medium object (40×40 pixels) and large object (70×70 pixels), respectively. A denotes the Ground Truth (GT) box, B and C are predicted boxes with slight deviations along the diagonal direction (6 pixels and 12 pixels). IoU indicates the Intersection-over-Union between the GT box and the related predicted box.

insufficient positive samples assigned for small instances during training.

B. Review of Small Object Detection Algorithms

General object detection methods based on deep learning can be categorized into two groups: two-stage and one-stage detection, where the former detects objects in a coarse-to-fine routine while the later performs the detection at one stroke. Two-stage detection methods [1], [46], [49] produce high-quality proposals with a well-designed architecture such as Region Proposal Network (RPN) [1] at first, then the detection heads take regional features as input and perform subsequent classification and localization respectively. Compared with two-stage algorithms, one-stage approaches [3], [44], [50] tile dense anchors on feature maps and predict the classification scores and coordinates directly. Benefiting from proposal-free setting, one-stage detectors enjoy high computational efficiency but often lag behind in accuracy. In addition to the above two categories, several anchor-free methods [4], [47], [48], [51] have emerged in recent years, which discard the anchor paradigm. Moreover, query-based detectors [5], [52], which formulate the detection as a set prediction task, have shown great potential. We cannot elaborate on the related frameworks in the light of space restraints. Please refer to corresponding surveys [13], [14], [15] and original papers for more details.

To address the aforementioned challenging issues, existing small object detection methods usually introduce deliberate designs to powerful paradigms working well in generic object detection. Next, we will briefly introduce these approaches and an overview of the proposed solutions is presented in Fig. 2. Moreover, the comparisons of representative approaches in each classification are exhibited in Section A.1 of Appendix, available online.

1) Sample-Oriented Methods: One of the most critical procedures of training a learning-based detector is the sampling (often coexists with assignment), which has led to significant progress in generic object detection [53], [54]. However, for SOD task, generic sampling strategies usually fail to provide adequate positive samples, thereby impairing the final performance. Such predicament originates from two aspects: the targets with limited sizes only occupy a small portion in current datasets [6], [30], [31]; current overlap-based matching schemes [1], [3], [4], [47], [48] are too rigorous to sample sufficient positive anchors or points owing to the limited overlaps between priors and the regions of small objects. In view of the two observations, a series of efforts have been made to alleviate the sample-scarcity issue and can be split into two factions: increasing the number of small objects by data augmentation or devising optimal assignment strategy to enable adequate samples for network learning.

Data-Augmentation Strategies: Kisantal et al. [55] augments small instances by copying a small object and pasting it with random transformation to different positions in the identical image. RRNet [56] introduces an adaptive augmentation strategy named AdaResampling, which follows the same philosophy as [55], the major difference lies in that a prior segmentation map was used to guide the sampling process of valid positions to be pasted, and a scale transformation for pasted objects reduces the scale discrepancy further. Zhang et al. [57] and Wang et al. [58] both employed divide-and-resize functionality-based operations to obtain more training samples of small objects. On top of object segmentation, image inpainting and image blending, DS-GAN [59] devises a novel data augmentation pipeline to generate high-quality synthetic data of small objects.

Optimized Label Assignment: Methods following this philosophy intend to alleviate the sub-optimal sampling result due to the overlap- or distance-based matching strategy and reduce the perturbation during regression. With the help of the devised scale compensation anchor matching strategy, S ^{3} FD [60] increases the matched anchors of tiny faces, thereby improving the recall rate. Zhu et al. [61] proposed Expected Max Overlapping (EMO) score, which takes anchor stride into account when computing the overlaps and enlightens better anchor setups for small faces. Xu et al. [62] employed the proposed DotD (defined as the Normalized euclidean Distance between the center points of two bounding boxes) to replace the commonly used IoU. Similarly, RFLA [63] measures the similarity between the Gaussian receptive field of each feature point and ground truth in label assignment, which boosts the performance of mainstream detectors on tiny objects.

Samples matter in object detection, especially for SOD task. Without enough positive samples, the regions of small objects are under-optimized during training and thereby hampering subsequent classification and regression. Either augmentation-based methods or devised matching strategies and appropriate prior settings intend to provide sufficient positive samples. Nevertheless, the former line of methods always suffers from inconsistent performance improvement and poor transferability. Meanwhile, current optimized label assignment schemes are prone to introduce low-quality samples and still struggle on the objects with extremely limited sizes.

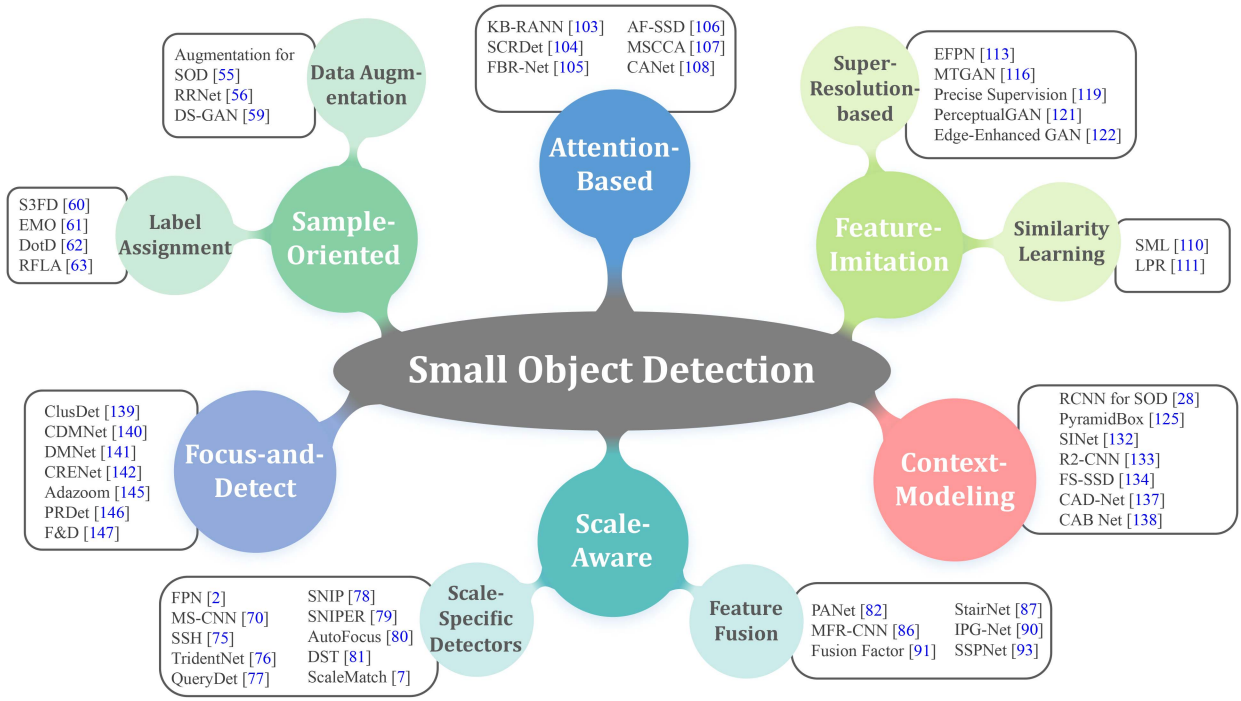

Fig. 2. Structured taxonomy of the existing deep learning-based methods for small object detection, which includes six genres. Only several representative methods of each category are demonstrated.

2) Scale-Aware Methods: Objects in an image often vary in scale and such variation could be particularly severe in traffic scenarios and remote sensing images, leading disparate detection difficulties for a single detector. Previous approaches [64], [65] usually employ image pyramid [66] with sliding window scheme to handle the scale-variance issue. However, handcrafted feature based methods, constrained by the limited representation capacity, perform catastrophically poorly on small objects. Early detection methods based on deep models still struggle in detecting tiny objects because only high-level features were used for recognition. To remedy the weakness of this paradigm and inspired by the success of reasoning across multi-level in other vision fields [67], [68], the following works mainly follow two paths. One refers to construct scale-specific detectors by devising multi-branch architecture or tailored training scheme, and the other line of efforts intends to fuse the hierarchical features for powerful representations of small objects. Both of these approaches actually minimize the information loss during feature extraction to a certain extent.

Scale-Specific Detectors: The nature behind this line is simple: the features at different depths or levels were responsible for detecting the objects of corresponding scales only. Yang et al. [69] exploited scale-dependent pooling (SDP) to select a proper feature layer for subsequent pooling operation of small objects. MS-CNN [70] generates object proposals at different intermediate layers, each of which focuses on the objects within certain scale ranges, enabling the optimal receptive field for small objects. Following this roadmap, DSFD [71] employs two-shot detector connected by the feature enhancement module to detect the faces of various scales. YOLOv3 [45] conducts multi-scale predictions by adding parallel branches where high-resolution features are responsible for small objects. Lin et al. [2] proposed Feature Pyramid Network (FPN), where the instances of various scales were assigned to different pyramid levels according to their sizes. Meanwhile, the interaction of features at different depths further guarantees the proper representation of multi-scale objects. This simple yet effective design has become an essential component in feature extractor and inspires a series of remarkable variants, e.g., NAS-FPN [72], and Recursive-FPN [73]. In addition, combining scale-wise detectors for multi-scale detection has been extensively explored. Li et al. [74] built parallel subnetworks where small-size subnetwork is learned specifically to detect small pedestrians. SSH [75] combines scale-variant face detectors, each trained for a certain scale range, to form a strong multi-scale detector to handle the faces varying extremely in scales. TridentNet [76] builds a parallel multi-branch architecture where each branch possesses optimal receptive fields for the objects of different scales. QueryDet [77] designs the cascade query strategy to avoid the redundant computation on low-level features, making it possible to detect small objects on high-resolution feature maps efficiently.

Several approaches aim to develop tailored data preparation strategies to force the detector concentrate on the instances with specific scales during training. On top of generic multi-scale training scheme, Singh et al. [78] devised a novel training paradigm, Scale Normalization for Image Pyramids (SNIP), which only takes the instances whose resolutions fall into the desired scale range for training and the remainders are simply ignored. By this setting, small instances could be tackled at the

most reasonable scales without compromising the detection performance on medium-to-large objects. Later, Sniper [79] advises to sample chips from a multi- scale image pyramid for efficient training. Najibi et al. [80] proposed a coarse-to-fine pipeline for detecting small objects. Considering that the collaboration between data preparation and model optimization is underexplored by previous methods [2], [66], [76], Chen et al. [81] designed a feedback-driven training paradigm to dynamically direct data preparation and further balance the training loss of small objects. Yu et al. [7] introduced a statistic-based match strategy for scale consistency.

Hierarchical Feature Fusion: Deep CNN architecture produces hierarchical feature maps at different spatial resolutions, in which low-level features describe finer details along with more localization cues, while high-level features capture richer semantic information [13], [43], [76], [82], [83], [84]. For SOD task, deep features may struggle with the disappeared response of small objects, and the feature maps at early stages are susceptible to variations such as illumination, deformation and object pose, making the classification task more challenging. To overcome this dilemma, extensive approaches leverage feature fusion, which integrates the features at different depths, to obtain better feature representation for small objects. Enlightened by the simple-yet-effective interaction design in FPN [2], PANet [82] enriches the feature hierarchy with bidirectional paths, enhancing deeper features with accurate localization signals. Aiming at optimizing multi-scale feature fusion with a more intuitive and principled fashion, Tan et al. [85] proposed the bi-directional feature pyramid network (BiFPN) to warrant the proper representations for small objects and better accuracy-and-efficiency trade-offs. Zhang et al. [86] concatenated the pooled features of an RoI at multiple depths with global feature to obtain more robust and discriminative representation for small traffic objects. Woo et al. [87] proposed StairNet where deconvolution was exploited to enlarge the feature map, such learning-based up-sampling function can achieve a more refined feature than naive kernel-based up-sampling and allows that the information of different pyramid levels propagates more efficiently [88]. M2Det [89] constructs parallel branches to describe features from shallow-to-deep in a cascade manner, where a Thinned U-shape Module is leveraged to capture more detailed information for small objects. Liu et al. [90] introduced IPG-Net, where a set of images at different resolutions obtained by the image pyramid [66] were input to the designed IPG transformation module to extract shallow features to complement spatial information and details. Gong et al. [91] devised a statistic-based fusion factor to control the information flow of adjacent layers. Noting that the gradient inconsistency encountered in FPN-based approaches deteriorates the representation ability of low-level features [92], SSPNet [93] highlights the features of specific scales at different layers and employs the relationship of adjacent layers in FPN to accomplish proper feature sharing.

Scale-specific architectures are committed to processing small objects at most reasonable scale, and fusion-based approaches aim to bridge the spatial and semantic gaps between lower pyramidal levels and higher ones. However, the former maps the objects of different sizes to corresponding scale levels in a heuristic manner which may confuse the detectors, because the information of a single layer is inadequate to make accurate prediction. On the other hand, in-network information flow is not always conducive to the representations of small objects. Our goal is to not only endow the low-level features with more semantics, but also prevent the original responses of small objects from overwhelmed by deeper signals. Unfortunately, you cannot have your cake and eat it, hence this dilemma needs to be addressed carefully.

3) Attention-Based Methods: Human can quickly focus and distinguish objects while ignoring those unnecessary parts by a sequence of partial glimpses at the whole scene [94], and this astonishing capacity in our perception system is generally referred as visual attention mechanism, which plays a crucial role in our visual system [95]. Not surprisingly, this powerful mechanism has been extensively investigated in the previous literature [96], [97], [98], [99], [100] and shows great potential in many vision fields [5], [9], [101], [102]. By allocating different weights to different parts of feature maps, the attention modeling indeed emphasizes the valuable regions while suppressing those dispensable ones. Naturally, one can deploy this superior scheme to highlight the small objects that are inclined to be dominated by the background and noisy patterns in an image, thereby partly minimizing the contamination in feature representation.

Enlightened by the human cognition, KB-RANN [103] exploits long-term and short-term attention neural networks to focus on the particular parts of image features, enhancing the detection of small objects. SCRDet [104] designs an oriented object detector, in which pixel attention and channel attention were trained in a supervised manner to highlight small object regions while eliminating the interference of noise. Extending the anchor-free detector FCOS [4] with the proposed level-based attention, FBR-Net [105] equilibrates the features at different pyramid levels and enhances the learning of small object under complicated situations. Lu et al. [106] designed a dual path module to highlight the key feature of small objects and suppress the non-object information. By replacing the complex convolution components with the proposed enhanced channel attention (ECA) blocks, MSCCA [107] constructs a lightweight detector with balanced channel features and less parameters. Li et al. [108] designed a cross-layer attention module to obtain stronger responses of small objects.

Drawing on the cognitive mechanism of mankind, visual attention plays an important role in nowadays vision fields, and it enables high-quality representations by screening the key parts while restraining noisy ones. Attention-series methods are highly claimed for their flexible embedding designs and can be plugged into almost all the SOD architectures, however, the performance improvement comes at the cost of heavy computation overhead owing to the correlation operations and moreover, current attention paradigms are lacking supervised signals and optimized implicitly.

4) Feature-Imitation Methods: One of the most significant challenges of SOD is the low-quality representations caused by the little information of small instances. This situation will likely get worse for those objects with extremely limited sizes [109].

Meanwhile, larger instances often embody clear visual structures and better discrimination. Hence, a straightforward way to alleviate this low-quality issue is enriching the regional features of small objects by mimicking that of larger ones [110]. To this end, several methods have been proposed and can be categorized into two genres: feature imitation by similarity learning and super-resolution-based frameworks. By mining the intrinsic relations between the objects of various scales, the methods in this classification largely ameliorate the problem of information loss and noisy feature representation.

Similarity Learning-Based Methods: The principle of this line is simple: imposing additional similarity constraints on generic detectors, thereby bridging the representation gap between small objects and large ones. Wu et al. [110] proposed Self-Mimic Learning method, where the representations of small-scale pedestrians were enforced to approach to the local average RoI features of large-scale ones. Inspired by the memory process of human visual understanding mechanism, Kim et al. [111] devised a large-scale embedding learning with the large-scale pedestrian recalling memory (LPR Memory), and the overall architecture was optimized under the recalling loss which intends to guide the small- and large-scale pedestrian features to be similar.

Super-Resolution-Based Frameworks: Methods following this roadmap aim at restoring the distorted structures of small objects instead of simply amplifying the ambiguous appearance of them. With the help of deconvolution and sub-pixel convolution [112], Zhou et al. [83] and Deng et al. [113] obtained high-resolution features specialized for small object detection. With self-supervised learning paradigm, Pan et al. [114] proposed a guided feature upsampling module to learn upscaled feature representations with detailed information. Generative Adversarial Network (GAN) [115] has remarkable capability to generate visually authentic data by following a two-player minimax game between the generator and the discriminator, which, unsurprisingly, enlightens the researchers to explore this powerful paradigm for generating high-quality representations of small objects. Deeming that directly operating the whole images incurs non-negligible computational cost at feature extraction stage [113], MTGAN [116] super-resolves the patches of RoIs with the generator network. Bai et al. [117] extended this paradigm to face detection task and Na et al. [118] applied super-resolution method to small candidate regions for better performance. Though super-resolving target patches could partly reconstruct the blurry appearance of small objects, this scheme neglects the contextual cues which play an important role for network prediction [119], [120]. To deal with this issue, Li et al. [121] devised PerceptualGAN to mine and exploit the intrinsic correlations between small-scale and large-scale objects, in which the generator learns to map the weak representations of small objects to super-resolved ones to deceive the discriminator. To go a step further, Noh et al. [119] introduced direct supervision to the super-resolution procedure. Rabbi et al. [122] and Courtrai et al. [123] both use GAN to super-resolve low-resolution remote sensing images, where the former screens the edge details to avoid high-frequency information loss during reconstructing, and the latter incorporates the cyclic GAN and residual feature aggregation to capture complex features.

By adding additional similarity loss or super-resolution architectures to prevailing detectors, feature imitation methods empower the model to mine the intrinsic correlations between small-scale objects and large-scale ones, thereby enhancing the semantic representation of small objects. Nevertheless, either similarity learning-based methods or super-resolution-based approaches have to avoid the collapse problem and sustain the feature diversity. Moreover, GAN-based methods are inclined to fabricate spurious textures and artifacts, imposing negative impacts on detection. Worse still, the existence of super-resolution architecture complicates the end-to-end optimization.

5) Context-Modeling Methods: We human can effectively utilize the relationship between the environment and the objects or the relation of objects to facilitate the recognition of objects and scenes [124], [125]. Such prior knowledge that captures the sematic or spatial associations is known as context, which conveys the evidence or cues beyond the object regions. The contextual information is of critical importance not only in visual systems of human [120], [124], but also in scene understanding tasks such as object recognition [126], semantic segmentation [127] and instance segmentation [128], *etc*. Interestingly, informative context sometimes can provide more decision support than the object itself, especially when it comes to recognizing the objects with poor viewing quality [124]. To this end, several methods exploit the contextual cues to boost the detection of small objects, thereby overcoming the loss issue in decision making.

ION et [129] computes global contextual features by two four-directional IRNN structures [130] for better detection of small and heavily occluded objects. Chen et al. [28] employed the representations of context regions which encompass the proposal patches for subsequent recognition. Hu et al. [131] investigated how to effectively encode the regions beyond the object extent and model the local context information in a scale-invariant manner to detect tiny faces. PyramidBox [125] makes full use of contextual cues to find small and blur faces that are indistinguishable from background. Assuming that the original RoI pooling operation would break up the structures of small objects, SINet [132] introduces a context-aware RoI pooling layer to maintain the contextual information. $\mathcal{R}^2$-CNN [133] employs a global attention block to suppress false alarms and efficiently detect small objects in large-scale remote sensing images. The intrinsic correlations of objects in an image can be regarded as context likewise. FS-SSD [134] exploits the implicit spatial context information, the distances between intra-class and inter-class instances, to redetect the objects with low confidences. Similarly, a context reasoning module was introduced by Fu et al. [135] to capture the intrinsic relationships and propagate the semantic- and spatial-relatedness between different regions. Pato et al. [136] leveraged the contextual information from predictions to restore the confidences and improve the final precision. Zhang et al. [137] captured the correlations between small objects and global scene (global context), as well as their neighboring instances (local context) to improve the performance. Cui et al. [138] devised a context-aware block

to integrate multi-scale context cues with pyramidal dilated convolutions, endowing the high-resolution features with strong semantics that are conducive to small instances.

From the information theory perspective, the more types of features are considered, the more likely higher detection accuracy can be obtained [86]. Inspired by the consensus, context priming has been extensively studied to generate more discriminative features, especially for small objects who have inadequate cues, enabling precise recognition. Unfortunately, both holistic context modeling or local context priming confuse about which regions should be encoded as context. In other words, current context modeling mechanisms determine the contextual regions in a heuristic and empirical fashion, which cannot guarantee the constructed representations are interpretable enough for detection.

6) Focus-and-Detect Methods: Small objects in high-resolution images tend to distribute non-uniformly [139], and the general divide-and-detect scheme consumes too much computation on those empty patches, leading the inefficiency during inference. Can we filter out those regions with no object thereby reducing the useless operations to boost the detection? The answer is YES! Efforts in this area break the chain of generic pipeline for processing high-resolution images. They first abstract the regions contain targets, on which the detection performs subsequently. Such paradigm guarantees that small objects can be processed at higher resolutions, thereby easing the information loss and improving the representation quality.

Yang et al. [139] proposed a Clustered Detection network (ClusDet) that fully exploits the semantic and spatial information between objects to generate cluster chips and then performs the detection. Following this paradigm, Duan et al. [140] and Li et al. [141] both exploited pixel-wise supervision to density estimation, achieving more accurate density maps which characterize the distribution of objects well. CRENet [142] designs a clustering algorithm to adaptively search cluster regions.

Deeming that the fixed-size input processing pipeline incurs missing detection of small objects, [143] exploits tilling method to detect pedestrians and vehicles in high-resolution aerial images in real time. Sharing similar philosophy, Deng et al. [144] and Xu et al. [145] devised a super-resolution network and a reinforcement learning framework to increase the spatial resolutions of local patches for finer detection and adaptively zoom the focus regions, respectively. Except the conventional region-mining procedure, Leng et al. [146] employed a region-specific context learning module to enhance the perception of size-limited instances in challenging areas. F & D [147] introduces a Focus & Detect framework, where Focusing Network detects candidate regions which then were cropped and resized to higher resolution, enabling the accurate detection of small objects.

Compared to generic sliding window mechanism, focus-and-detect methods empower adaptive crops and flexible zoom-in operation, i.e., smaller objects can be processed at higher resolutions while larger ones can be detected in a relatively lower resolution, which significantly saves memory footprint at inference and reduces the interference of background. Methods following this roadmap have to answer the key question: *where to focus*? Current approaches resort to either manually additional annotations or auxiliary architectures like segmentation network or Gaussian Mixture Model, yet the former requires laborious labeling while the latter complicates the end-to-end optimization.

III. Review of Datasets for Small Object Detection

A. Datasets for Small Object Detection

Datasets are the cornerstone of learning-based object detection methods, especially for data-driven deep learning approaches. In the past decades, various research institutions have launched plenty of high-quality datasets [6], [30], [31], [32], and these publicly available benchmarks significantly boost the development of related fields. Unfortunately, very few benchmarks are designed for small object detection. For the sake of integrity, we still retrospect a dozen datasets which contain considerable number of small objects, and expect to provide a comprehensive review of datasets. Instead of restricting our scope to specific tasks, we investigate the related datasets which span over a wide range of research areas, including face detection [8], pedestrian detection [7], [148], [149], object detection in aerial images [20], [30], [160], [162], to name a few. The statistics of these benchmarks are given in Table II, and here only the most representative among them were introduced below in detail due to the space restriction. More details please refer to Section A.2 of Appendix, available online.

COCO: Pioneering works [31], [32], though push forward the development of vision recognition tasks, have been criticized for their ideal condition, where objects usually have large sizes and center on the images, bearing little resemblance to the real-world scenarios. To bridge this gap and foster fine level image understanding, COCO [6] was launched in 2014, its trainval set annotates 886 K objects distributed in 123 K images with instance-level mask, covering 80 common categories under complex everyday scenes. Comparing to previous datasets for object detection, COCO contains more small objects (about 30% instances in COCO trainset have an area less than 1024 pixels) and more densely packed instances, both of which challenge the detectors. Moreover, the fully segmented annotation and the reasonable evaluation metric encourage more accurate localization. All these features help COCO be the de facto standard for validating the effectiveness of object detection methods in past years.

WiderFace: WiderFace [8] is a large-scale benchmark towards accurate face detection, in which faces vary significantly in scale, pose, occlusion, expression, appearance and illumination. It contains 32203 images with a total of 393703 instances. Except common bounding box annotations, attributes including occlusion, pose and event categories were also provided, which allows thorough investigation for existing approaches. The faces in WiderFace are divided into three subsets, namely small (between 10-50 pixels), medium (between 50-300 pixels) and large (larger than 300 pixels), where small subset accounts for half of all instances.

TinyPerson: TinyPerson [7] focuses on the seaside pedestrian detection. TinyPerson annotates 72561 persons in 1610 images

TABLE II
STATISTICS OF SOME BENCHMARKS AVAILABLE FOR SMALL OBJECT DETECTION

Dataset name	Task field	Publication	#Images	#Instances	Descriptions and Characteristics
COCO [6]	ODNI	ECCV 2014	123K	886K	One of the most popular datasets for generic object detection
SOD [28]	ODNI	ACCV 2016	4925	8393	A small-scale dataset for small object detection
WiderFace [8]	Face detection	CVPR 2016	32K	393K	A large-scale benchmark with rich annotations for face detection
EuroCity Persons [148]	Pedestrian detection	TPAMI 2019	47K	219K	The largest dataset for pedestrian detection captured from dozens of Europe cities
WiderPerson [149]	Pedestrian detection	TMM 2020	13K	39K	Pedestrian detection benchmark in traffic scenarios
TinyPerson [7]	Pedestrian detection	WACV 2020	1610	72K	The first dataset dedicated to tiny-scale pedestrian detection
STS Dataset [150]	Traffic sign detection	SCIA 2011	20000	3488	The first publicly available traffic sign dataset for detection
LISA [151]	Traffic sign detection	TITS 2012	6610	7855	A traffic sign dataset allowing for detection and tracking
GTSDB [152]	Traffic sign detection	IJCNN 2013	900	1206	A benchmark for traffic sign detection collected under different scenarios
TT100K [153]	Traffic sign detection	CVPR 2016	100K	30K	A realistic and large-scale benchmark for traffic sign detection
BSTLD [154]	Traffic light detection	ICRA 2017	13427	24000	A large dataset for detecting traffic lights whose sizes down to 1 pixel in width
UCAS-AOD [155]	ODAI	ICIP 2015	910	6029	A aerial dataset collected from Google Earth for detection
VEDAI [156]	ODAI	JVC 2016	1268	2950	A database dedicated to small vehicle detection in aerial images
xView [157]	ODAI	arXiv 2018	1128	1M	One of the largest and most diverse available dataset of overhead imagery
DIOR [20]	ODAI	JPRS 2020	23K	192K	One of the most frequently used benchmarks for object detection in aerial images
UAVDT [158]	ODAI	IJCV 2020	80K	841K	A dataset collected by Unmanned Aerial Vehicles for object detection and tracking
VisDrone [159]	ODAI	TPAMI 2021	189K	2.5M	A large-scale drone-captured benchmark for detection and tracking
DOTA [30]	ODAI	TPAMI 2022	11K	1.79M	The largest remote sensing detection dataset including considerable small objects
AI-TOD [160]	ODAI	JPRS 2022	28K	700K	A tiny object detection dataset based on previous available datasets
NWPU-Crowd [161]	Crowd counting	TPAMI 2021	5109	2.13M	The largest dataset for crowd counting and localization to date

ODNI stands for object detection in natural images and ODAI denotes object detection in aerial images. (1K= 1000, 1M = 1000K).

which are categorized into two subsets: tiny and small, according to their lengths. Due to the extremely tiny size, an ignore label was assigned to those regions that cannot be certainly recognized. As the first dataset dedicated to tiny-scale pedestrian detection, TinyPerson is a concrete step towards for tiny object detection. However, its limited number of instances and single pattern restrict its capacity to serve as a benchmark for SOD.

TT100 K: TT100K [153] is a dataset for realistic traffic sign detection which includes 30000 traffic sign instances in 100000 images, covering 45 common Chinese traffic-sign classes. Each sign in TT100 K is annotated with precise bounding box and instance-level mask. The images in TT100 K are captured from Tencent Street Views, holding a high variability in weather conditions and illumination. Moreover, TT100 K contains considerable small instances (80% of instances occupy less than 0.1% in the whole image area) and the entire dataset follows a long-tail distribution.

VisDrone: VisDrone [159] is a large-scale drone-captured dataset which is collected over various urban/suburban areas of 14 different cities across China. Concentrating on two essential tasks in computer vision, VisDrone supports four tracks: image object detection, video object detection, single object tracking and multi-object tracking. For image object detection track, there are 10209 images with a resolution of 2000×1500 pixels and 542 K instances covering 10 common object categories in traffic scenarios. The images in VisDrone are captured with drones from various urban scenes, thereby containing a mass of small objects due to viewpoint variations and heavy occlusions.

DOTA: DOTA [30] is proposed to facilitate the object detection in Earth Vision. It contains 18 common categories and 1793658 instances in 11268 images. Each object has been annotated with horizontal/oriented bounding box. Owing to the high diversity of orientations in overhead view images and large-scale variations among instances, DOTA dataset has numerous small objects, but they only distribute in a few categories (*small-vehicle*).

B. Evaluation Metrics

Before diving into the evaluation criteria of small object detection, we first introduce related preliminary concepts. Given a ground-truth bounding box b_{g} and a predicted box b_{p} output by the detector, if the IoU between b_{g} and b_{p} is greater than the predefined threshold, and the predicted label is in accordance with the ground-truth, the current detected box will be identified as a potential prediction to this object, also known as True Positive (TP), otherwise it will be regarded as a False Positive (FP). Once we obtain the number of TP, FP and False Negative (FN, also known as missed positives), the Average Precision (AP) can be computed to evaluate the performance of detectors.

Average Precision: Average Precision (AP) is originally introduced in VOC2007 Challenge [31] and usually adopted in a category-wise manner. Concretely, given a confidence threshold and an IoU threshold β, the Recall (R) and Precision (P) can be calculated afterwards. By varying the confidence threshold α, one can obtain different pairs (P, R) and ultimately, AP can be determined by averaging the precision scores under different recalls. This fixed IoU based AP metric once dominated the community for years.

A new evaluation metric was introduced with the launch of COCO dataset after 2014, which averages AP across multiple IoU thresholds between 0.5 and 0.95 (with an interval of 0.05). Apart from merely considering fixed IoU threshold, this criterion also takes the higher IoU thresholds into account, encouraging more accurate localization. This reasonable metric has been

TABLE III
AREA SUBSETS AND CORRESPONDING AREA RANGES OF OBJECTS IN SODA BENCHMARK

Area Subset	*Small*			*Normal*
	extremely Small	*relatively Small*	*generally Small*	
Area Range	$(0, 144]$	$(144, 400]$	$(400, 1024]$	$(1024, 2000]$

used as the "gold standard" in detection community and widely adopted by the following works [153], [163]. Noting that the overall AP is computed by averaging the APs of all categories in practice.

IV. BENCHMARKS

In this section, we briefly introduce the data acquisition and annotation process for building SODA-D and SODA-A. Then, we shed light on the characteristics of our benchmarks and the main differences between our datasets and related existing ones. Moreover, other details such as scene selection, image collection, data cleaning, license declaration, annotation principles and quality assurance will be discussed in the Section B of Appendix, available online.

A. Data Acquisition and Annotation

Our aim is to build datasets tailored for small object detection, hence the point is *how to define a valuable object*.

Definition About a Valuable Object: Generally, a bounding box B can be represented as (x, y, w, h, θ), where (x, y) denotes the center location and (w, h) indicates the width and height of the box respectively, the parameter θ stands for the orientation angle and is unused for horizontal annotation. Moreover, we use $S = w \times h$ to denote the pixel area of an object. In line with the definition of small or tiny objects in previous works [6], [7], [160], we adopt the absolute area criterion and regard an instance who has an area smaller than 1024 pixels, i.e., $S \leq 1024$, as a *Small* object. Meanwhile, an object whose area between 1024 and 2000 pixels will be annotated as a *Normal* object. Otherwise, the object comes to the *ignore* category and will not influence the final evaluation results. Considering the detection difficulty increases sharply when the object size gets smaller, we further divide the *Small* objects into three subsets: *extremely Small* (*eS*), *relatively Small* (*rS*) and *generally Small* (*gS*), as demonstrated in Table III.

Data Source: The images in SODA-D are mainly from MVD [33], self-shooting and the Internet. MVD is a large-scale dataset for semantic understanding of street scenes, of which 25000 high-quality images are captured from road views, highways, rural areas and off-road. Thanks to the high-quality and high-resolution property with MVD, we can obtain a large set of valuable instances with clear visual structure. For self-shooting part, we use on-board cameras and mobile phones to collect images of typical driving scenes in several Chinese cities, including Beijing, Shenzhen, Shanghai, Xi'an, Qingdao, Guangzhou, *etc*. In addition, we also crawl images by searching keywords on the image search engines (Google, Bing, Baidu, *etc*.). Finally, we obtained 24828 images of traffic scene.

TABLE IV
NUMBERS OF INSTANCES OF EACH CATEGORY AND THREE SPLITS OF SODA-D (LEFT) AND SODA-A (RIGHT)

Category	#Instances	Category	#Instances
people	35928	airplane	31622
rider	4636	helicopter	1395
bicycle	2560	small-vehicle	526047
motor	3896	large-vehicle	17006
vehicle	69197	ship	65690
traffic-sign	85905	container	138242
traffic-light	62729	storage-tank	35331
traffic-camera	7636	swimming-pool	29735
warning-cone	5946	windmill	27001
Train	134301	Train	344228
Validation	56050	Validation	231439
Test	88082	Test	296402
Total	278433	Total	872069

Enlightened by the pioneering works [20], [30], Google Earth[3] was leveraged to collect images for SODA-A, we extract 2513 images from hundreds of cities around the world suggested by the experts. It is noting that numerous images with cluttered background and high density which are closer to realistic challenges are captured. In addition, the images in SODA-A have a relatively high resolution and most of them enjoy a resolution larger than 4700×2700, enabling the finer details and adequate context that are of great significance to small object detection [124], [125].

Dataset Split: Following the pioneering works [6], [33], we split the full image-set into three subsets: train-set, validation-set and test-set, and each subset occupies approximately $50\% : 20\% : 30\%$ for SODA-D and $40\% : 25\% : 35\%$ for SODA-A.

Category Selection: Take the realistic value for applications and the intrinsic size into consideration, we select nine valuable categories for SODA-D: *people*, *rider*, *bicycle*, *motor*, *vehicle*, *traffic-sign*, *traffic-light*, *traffic-camera*, and *warning-cone*. For SODA-A, we also annotate nine object classes: *airplane*, *helicopter*, *small-vehicle*, *large-vehicle*, *ship*, *container*, *storage-tank*, *swimming-pool*, and *windmill*.

Instance-Level Annotation: The general principle to annotate SODA resembles that of general detection benchmarks [6], [20], [30], [31], [32], and the only difference lies in the *ignore* regions. Enlightened by the previous works [7], [158], [159], we assign *ignore* label to the two datasets when: 1) the instances belonging to the preset categories but with an area greater than 2000; 2) the objects that are excessively small and heavily occluded thus cannot be distinguished. In addition, we merge the *ignore* regions as possible while avoiding surround valuable foreground instances.

B. Statistical Analysis

We annotate 278433 instances for SODA-D and 872069 objects for SODA-A, and the number of instances for each category and that for three subsets are shown in Table IV. Also the example instances of each category are shown in Fig. 3.

[3] https://earth.google.com/

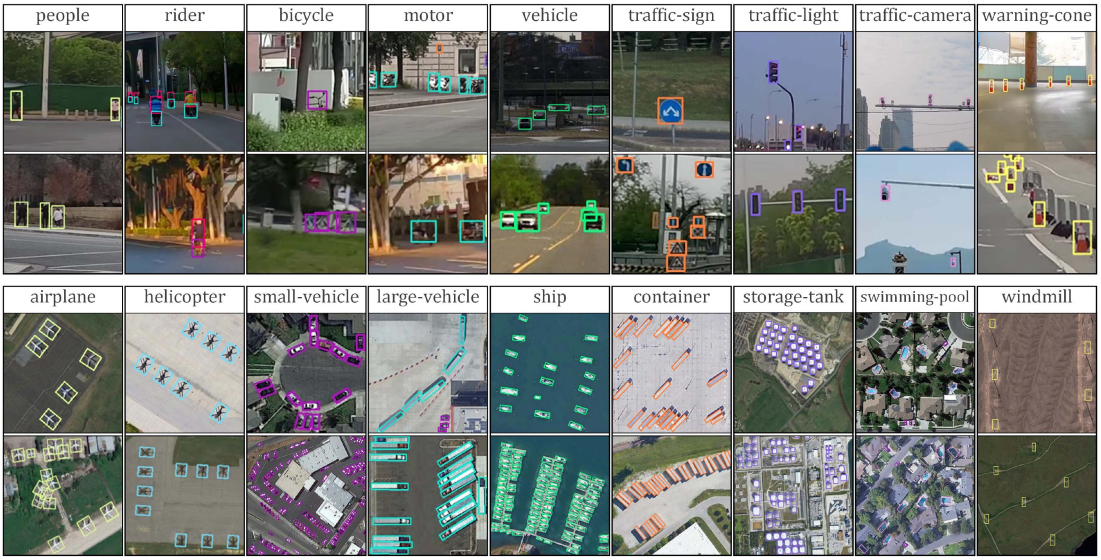

Fig. 3. Example instances of each category in SODA-D (Top) and SODA-A (Bottom).

TABLE V
COMPARISONS BETWEEN SODA-D AND SEVERAL RELATED DETECTION DATASETS UNDER DRIVING SCENE (TOP), LIKEWISE FOR SODA-A AND SOME DETECTION DATASETS UNDER AERIAL SCENARIO (BOTTOM)

Dataset	#Images	#Categories	#Instances			Split	Avg. Res. ($W \times H$)	Year
			eS	*rS*	*gS*			
TT100K [153]	8876	45	71	2800	6430	train/test	2048×2048	2016
EuroCity Persons [148]	32605	18	5318	28048	59190	train/val	1920×1024	2019
TJU-DHD Traffic [163]	50266	5	82	1189	20366	train/val	1624×1200	2021
SODA-10M [164]	10000	6	33	3061	10056	train/val	1920×1080	2021
SODA-D	24828	9	25834	71064	102066	train/val/test	3407×2470	2022

Dataset	Annotation	#Images	#Categories	#Instances			Split	Avg. Res. ($W \times H$)	Year
				eS	*rS*	*gS*			
CARPK [165]	HBB	1448	1	220	1716	1378	train/test	1280×720	2017
VisDrone [159]	HBB	8629	10	78999	97251	108793	train/val/test-dev	1490×957	2021
AI-TOD [160]	HBB	28036	8	193200	135566	17200	train/val	800×800	2021
DOTA [30]	OBB	2423	18	114045	94867	69934	train/val	2217×2074	2022
DIOR-R [162]	OBB	23463	20	30938	37471	39697	train/val/test	800×800	2022
SODA-A	OBB	2513	9	304900	363738	168874	train/val/test	4761×2777	2022

Note that eS, rS and gS stand for extremely Small, relatively Small and generally Small according to our definition (see Table III). And for each dataset, we only count the subsets whose annotations are available, see Split column. Avg. Res. denotes the average image resolution of the dataset. HBB/OBB denotes horizontal /oriented bounding box.

Next we highlight the most prominent feature of our dataset: *small size*. From Table V, SODA-D and SODA-A both far exceed the existing mainstream object detection datasets under traffic and aerial scenarios on the amount of *Small* objects, especially for *extremely Small* ones. Moreover, we also show the category-wise area distribution and overall scale distribution of instances in SODA-D and SODA-A in Fig. 4. As can be seen from (a) and (b), the area of objects in our benchmarks falls into a relatively tight range (especially for *traffic-camera* in SODA-D and *small-vehicle* and *ship* in SODA-A). Moreover, from (c) and (d) in Fig. 4, the size range of objects in SODA-D mainly comes to $[10, 30]$ and for SODA-A, it is strikingly $[5, 15]$. If we shed our light on the *Small* objects, the average absolute size of SODA-D and SODA-A is 20.31 pixels and 14.75 pixels, respectively.

Except the small size and large volume, our SODA-D and SODA-A also exhibit several unique characters.

1) Data Properties of SODA-D

Rich diversity: Our SODA-D dataset inherits one of the most preeminent virtues of MVD: the rich diversity in terms of locations, weathers, period, shooting views and scenarios. Fig. 5 shows some examples of our dataset covering various weather, view and illumination conditions. We believe that our diverse data could empower the model with the ability to generalize to different situations.

High Spatial Resolution: The images in SODA-D enjoy very high resolution and high quality, which is entailed for small or tiny object detection. In Fig. 6, we demonstrate the distribution of image resolution in SODA-D, and the average resolution at 3407×2470 shows a clear predominance in comparison with previous datasets who focus on object detection under traffic scenes, as illustrated in Table V.

Ignore Regions: Our benchmark contains a mass of *ignore* annotations (especially for SODA-D which has 153976 well-annotated *ignore* regions), which is one of the most highlighted features. The *ignore* definitions of *Instance -level annotation* part in Section IV-A could maintain the stability of training and evaluation. Concretely, we deem that the prevailing detectors [1], [3], [4], [5], [9], [45], [47], [48], [51], [52], [166]

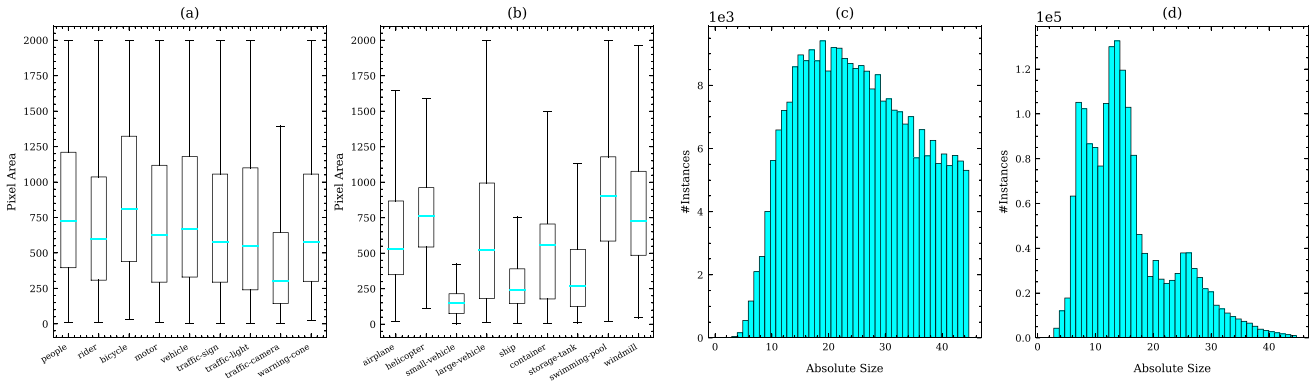

Fig. 4. Category-wise area distribution of instances in (a) SODA-D and (b) SODA-A, and overall scale distribution of instances in (c) SODA-D and (d) SODA-A.

Fig. 5. Example images under diversified conditions in our SODA-D dataset, where masked bounding boxes represent *ignore* regions. Best viewed in zoom-in windows.

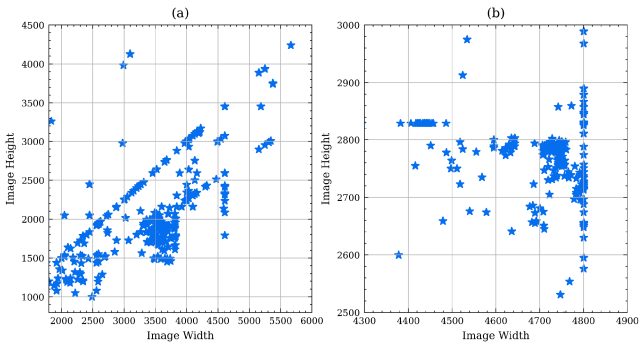

Fig. 6. Distribution of image resolution in (a) SODA-D and (b) SODA-A. Note that we randomly sample 2000 images to obtain the size profile for clear illustration.

can handle the first situation well, hence it is not our concern. For the latter condition, our well-trained annotators are called for cautiously labeling the regions as ignore, when they cannot make confident judgment even at highest zoom-in level. And it will only bring error and instability if we insist on annotating these regions as foreground objects. To put it in another way, *can we expect current algorithms to outperform human's eyes*? Therefore, categorizing these regions into *ignore* will not impose negative impact during evaluation process, and can guarantee the models concentrate on the authentic and valuable small objects.

2)Data Properties of SODA-A

We show an example image of SODA-A in Fig. 7 and the local zoom-in windows exhibit the details of annotated instances.

Large Density Variation: As in Fig. 8, the number of instances per image in SODA-A varies significantly from 1 to 11134, implying that our benchmark not only contains sparse condition but also includes numerous images where the objects positioned in extremely close proximity. Moreover, the average number of instances per image in SODA-A is 347.02, which is more than twice the number of DOTA (159.18). This literally calls for a robust model with the capacity of handling excessively clustered situation.

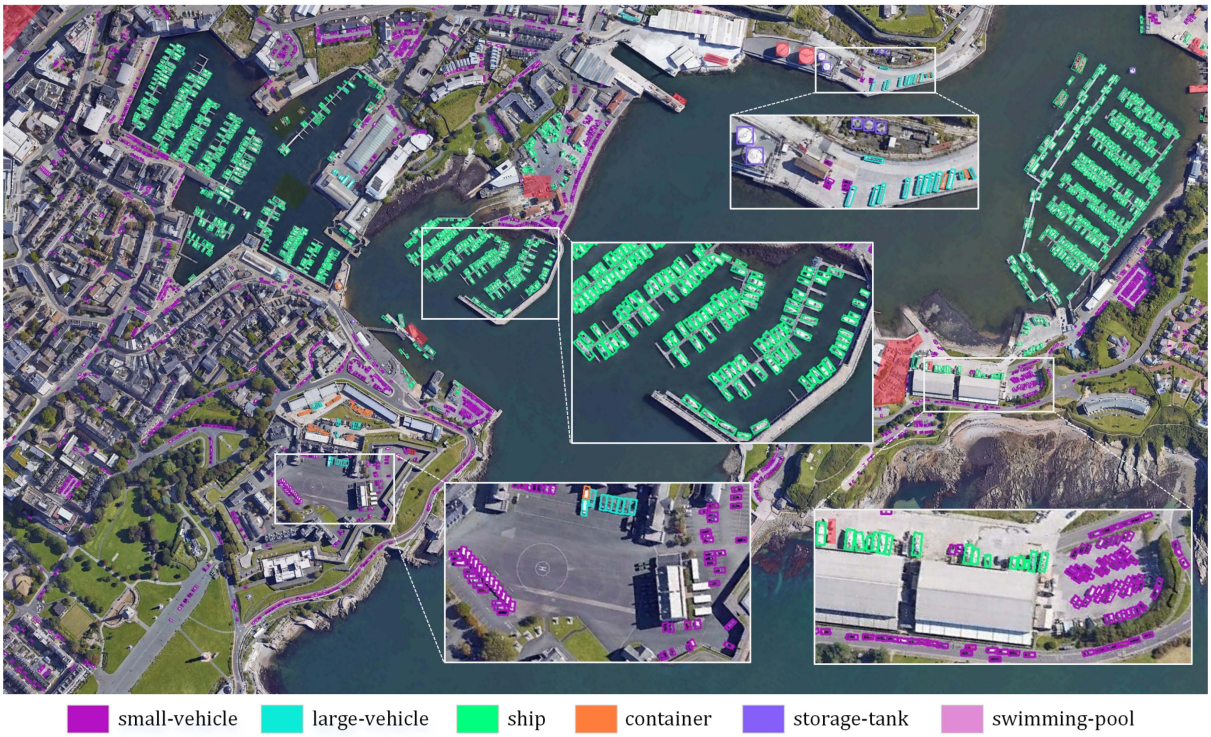

Fig. 7. Example image in SODA-A. The instances of different categories are best viewed in color and zoom-in windows, where masked areas denote the *ignore* regions.

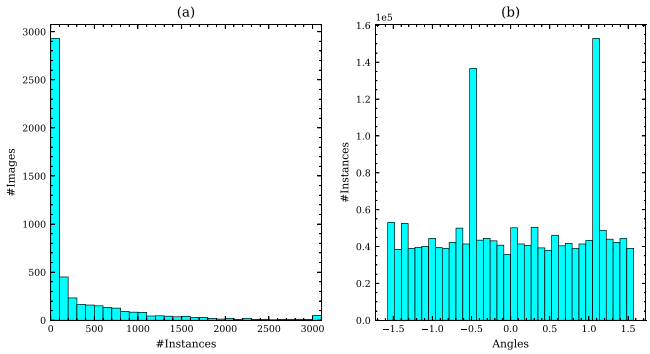

Fig. 8. Density distribution per image (a) and the orientation profile (b) of instances in SODA-A. Note that the number of images with more than 3000 instances were accumulated for clear demonstration of (a).

Various Orientations: The instances in SODA-A can appear in an arbitrary-rotated fashion. We indicate the orientation distribution of SODA-A in Fig. 8, and the tilt angle of annotated instances distributes from $-\pi/2$ to $\pi/2$. Note that we do not follow the orientation definition in DOTA, because most objects with tiny size cannot convey sufficient visual cues to determine their head or tail.

Diverse Locations: The images in SODA-A are collected from hundreds of cities around the world, which substantially enhances the data diversity in fact (e.g., the appearance of *airplane* objects in our SODA-A can vary considerably). Furthermore, the concomitant intra-class variation and complicated background bring more challenges.

C. Comparisons With Previous Benchmarks

Although there have been tremendous datasets for object detection, few of them dedicated to SOD task. Even so, we compare several related benchmarks with SODA to highlight its uniqueness.

1) SODA-D

MVD: Despite the SODA-D dataset is constructed on top of MVD, our intention is completely different from MVD. To be more specific, MVD concentrates on the pixel-level understanding of street scenes, while the proposed SODA-D highlights the detection of those objects with extremely small size under complicated driving scenarios.

2) SODA-A

AI-TOD: AI-TOD is built on several publicly available datasets, including DIOR [20], DOTA [30], VisDrone [159], xView [157], and Airbus-Ship[4]. However, the above datasets were not designed for SOD task, hence more than 88% instances of AI-TOD come from the category *vehicle*, leading to a non-negligible imbalance issue as shown in Fig. 9. Meanwhile, each category in our SODA-A contains adequate instances, except *helicopter* class, and this advantage becomes more pronounced when considering the data volume (our SODA-A contains 837512 instances belonging to *Small* object subset). In addition, the images in AI-TOD are cropped from existing datasets and the image resolution is fixed to 800×800. More importantly, AI-TOD only provides horizontal annotations, which severely limits its capacity to approach objects accurately and to handle

[4]https://www.kaggle.com/c/airbus-ship-detection

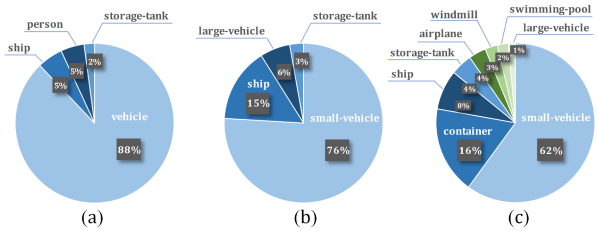

Fig. 9. Class distribution of *Small* instances in (a) AI-TOD, (b) DOTA, and (c) SODA-A. Those categories with instances less than 2000 are not included.

the densely-packed situation that is common and challenging for SOD in aerial images. In contrast, from Table V and Fig. 6, our SODA-A possesses an average image resolution of 4761×2777, and the well-annotated oriented boxes allow for large density cases and encourage more accurate localization.

DOTA: DOTA is the largest dataset for object detection in aerial images to date. Compared to DOTA, who puts emphasis on scale variation issue, we mainly focus on the small-scale objects which confuse current detectors. Moreover, though DOTA contains substantial amounts of small objects (more than 110 K instances come to the *extremely Small* and more than 270 K instances possess an area within 1024 pixels, as exhibited in Table V), most of them centralized at *small-vehicle*, as in Fig. 9.

V. Experiments

A. Evaluation Protocol

Following the evaluation protocols in COCO [6], we use the Average Precision (AP) to evaluate the performance of detectors. Concretely, as the paramount metric, the overall AP is obtained by averaging the AP over 10 IoU thresholds between 0.5 and 0.95 on *Small* objects. AP_{50} and AP_{75} are computed at the single IoU thresholds of 0.5 and 0.75, respectively. Moreover, to highlight our concern for size-limited objects, the AP of four area subsets also are demonstrated, namely, AP_{eS}, AP_{rS}, AP_{gS} and AP_N.

B. Implementation Details

To conduct fair comparisons of several benchmarking baselines, all the experiments on SODA-D and SODA-A are implemented on top of the toolbox mmdetection[5][171] and mmrotate[6][172], respectively. Directly feeding the high-resolution images in SODA to deep model is infeasible due to the GPU memory limitation, hence we crop original images into a series of 800×800 patches with a stride of 650. These patches will be resized to 1200×1200 during training and testing, which could partly alleviate the information loss caused in the feature extraction stage. Noting the patch-wise detection results will be first mapped to the original images, on which Non Maximum Suppression (NMS) was performed to prune out redundant predictions. We use 4 NVIDIA GeForce RTX 3090 GPUs to train the models, and the batch size is set to 8 for the experiments of SODA-D and 4 for that of SODA-A, where the angle ranges is $[-\pi/2, \pi/2)$. Only random flip was used for augmentation during training, and more details and hyperparamter settings please refer to Sections C.1 and D.1 in Appendix, available online.

[5]https://github.com/open-mmlab/mmdetection
[6]https://github.com/open-mmlab/mmrotate

C. Results Analysis on SODA-D

In this section, we perform a rigorous evaluation of several representative methods on our SODA-D dataset, and provide in-depth analyses on top of the results. Moreover, we conduct several experiments to investigate the effect of label assignment and loss designs to SOD. More details can be found in Sections C.2 and C.3 in Appendix, available online.

1) Benchmarking Results: Table VI reports the results of 12 representative methods on SODA-D test-set. From the table, we can find that Faster RCNN [1] scores 28.9% on AP, and benefiting from the cascade structure, Cascade RCNN [166] attains the best performance with an AP of 31.2% and an impressive AP_{75} of 27.8%, which steadily outperform other detectors. On top of Faster RCNN, RFLA [63] achieves 29.7% AP, though meanwhile, the AP_{eS} actually drops 0.7 points, showing that the devised assignment might not be suitable for those instances with excessively limited sizes. One-stage detector RetinaNet [3] scores 28.2% AP which is close to Faster RCNN, but there exists a huge gap (11.9% *v.s.* 13.9%) when comes to the AP_{eS}, and when the object size gets larger, such difference becomes smaller, which reveals that the misalignment issue imposes a significantly severe impact on tiny objects. Similarly, though RepPoints [168] can obtain an overall AP of 28.0%, but the AP_{eS} metric (10.1%) is largely behind Faster RCNN and RetinaNet. This phenomenon indicates that point representation, in comparison to its box counterpart, may not be a good choice for small objects, but shows great potential for large ones. For anchor-free detectors, ATSS [169] can achieve 26.8% AP on our SODA-D test-set, which is superior to FCOS [4] (23.9%), and the latter behaves badly on *extremely Small* objects (6.9%). This may partly originates from the occlusion challenge of our dataset, also known as the ambiguous sample problem. CenterNet [47] and CornerNet [51] only obtain an AP of 21.5% and 24.6%, respectively. It can be noticed that even with more training epochs, the performances of CenterNet and CornerNet are remarkably inferior to that of anchor-based methods, and the disparity becomes more staggering for *extremely Small* and *relatively Small* objects. YOLOX [167] can obtain competitive results (26.7% AP and 13.6% AP_{eS}) when compared to other anchor-free counterparts though meanwhile struggles on the objects of large areas. For the query-based detector, Sparse RCNN [170] achieves 24.2% AP which is comparable to FCOS. Though exploiting multi-scale deformable attention to reduce high computation in encoder and enabling the access of high-resolution features, Deformable DETR [52] only delivers 19.2% AP, lagging noticeably behind other competitors even with more training epochs. This performance gap may reveal that the sparse query paradigm could not cover small objects adequately.

2) Category-Wise Results: We also list the category-wise results on Table VII, in which the AP of *rider*, *bicycle*, *motor* and *traffic-camera* are clearly inferior to other categories, we deem

TABLE VI
BASELINE RESULTS ON SODA-D TEST-SET

Method	Publication	Schedule	AP	AP_{50}	AP_{75}	AP_{eS}	AP_{rS}	AP_{gS}	AP_N	#Param.	FLOPs
Faster RCNN [1]	TPAMI 2017	1×	28.9	59.7	24.2	13.9	25.6	34.3	43.2	41.16M	292.28G
Cascade RCNN [166]	TPAMI 2021	1×	31.2	59.9	27.8	14.1	27.5	37.1	46.9	68.95M	320.07G
RetianNet [3]	TPAMI 2020	1×	28.2	57.6	23.7	11.9	25.2	34.1	44.2	35.68M	299.50G
CornerNet [51]	ECCV 2018	2×	24.6	49.5	21.7	6.5	20.5	32.2	43.8	200.96M	1104.06G
CenterNet [47]	ArXiv 2019	70e	21.5	48.8	15.6	5.1	16.2	29.6	42.4	70.75M	137.21G
FCOS [4]	TPAMI 2022	1×	23.9	49.5	19.9	6.9	19.4	30.9	40.9	31.86M	284.53G
RepPoints [168]	ICCV 2019	1×	28.0	55.6	24.7	10.1	23.8	35.1	45.3	36.60M	273.96G
ATSS [169]	CVPR 2020	1×	26.8	55.6	22.1	11.7	23.9	32.2	41.3	31.32M	290.79G
Deformable-DETR [52]	ICLR 2021	50e	19.2	44.8	13.7	6.3	15.4	24.9	34.2	35.17M	739.11G
Sparse RCNN [170]	CVPR 2021	1×	24.2	50.3	20.3	8.8	20.4	30.2	39.4	105.96M	213.00G
YOLOX [167]	ArXiv 2021	70e	26.7	53.4	23.0	13.6	25.1	30.9	30.4	8.94M	48.11G
RFLA [63]	ECCV 2022	1×	29.7	60.2	25.2	13.2	26.9	35.4	44.6	41.16M	292.06G

All the models are trained with a ResNet-50 [10] as the backbone except YOLOX (CSP-darknet) [167] and cornernet (HourglassNet-104) [51]. Schedule denotes the epoch setting during training, where '1×' and '50e' represent 12 and 50 epochs.

TABLE VII
CATEGORY-WISE AP OF BASELINE DETECTORS ON SODA-D TEST-SET

Method	people	rider	bicycle	motor	vehicle	t-sign	t-light	t-camera	w-cone	AP
Faster RCNN [1]	35.8	16.5	12.5	23.1	44.1	45.8	37.8	14.3	30.5	28.9
Cascade RCNN [166]	39.2	18.0	14.5	24.2	47.4	48.1	39.8	15.2	33.4	31.2
RetianNet [3]	34.0	16.9	11.1	22.5	44.3	45.6	36.3	14.2	29.0	28.2
CornerNet [51]	30.5	15.7	11.3	22.8	37.3	40.3	31.8	8.0	24.3	24.6
CenterNet [47]	25.6	12.8	9.5	19.9	32.9	35.8	27.6	9.3	20.4	21.5
FCOS [4]	29.7	13.9	10.4	19.5	40.2	38.0	31.6	8.9	23.0	23.9
RepPoints [168]	36.0	15.9	10.8	21.6	44.8	45.6	37.3	12.7	27.7	28.0
ATSS [169]	33.3	16.0	10.6	21.3	42.7	43.3	34.8	11.6	27.8	26.8
Deformable-DETR [52]	23.0	10.2	7.6	16.4	30.2	33.8	24.9	7.5	19.6	19.2
Sparse RCNN [170]	31.4	12.4	8.5	19.0	39.9	42.1	33.1	8.5	23.0	24.2
YOLOX [167]	33.8	14.7	8.0	20.1	43.6	43.3	35.9	11.5	29.5	26.7
RFLA [63]	37.1	19.1	13.4	24.2	45.0	45.8	37.5	14.7	30.5	29.7

The training settings are consistent with table VI and the full names of class abbreviation are as follows: T-Sign (traffic-sign), T-Light (traffic-light), T-Camera (traffic-camera) and W-Cone (warning-cone).

TABLE VIII
THE AP PERFORMANCE OF BASELINE DETECTORS WITH DIFFERENT BACKBONE NETWORKS

Method	ResNet-50	ResNet-101	Swin-T	ConvNext-T
Faster RCNN [1]	28.9	28.7	30.3	31.9
Cascade RCNN [166]	31.2	30.6	32.8	34.3
RetianNet [3]	28.2	27.8	28.8	29.7
FCOS [4]	23.9	24.3	29.2	29.1
RepPoints [168]	28.0	28.2	28.4	30.2
ATSS [169]	26.8	26.7	27.2	27.9
Sparse RCNN [170]	24.2	25.3	24.9	25.2

All the models were trained for '1×' schedule.

that the root cause of this phenomenon comes from two-fold. 1) Class-imbalance issue. These categories contain less samples compared to other classes, e.g., only 2560 samples included in *bicycle* category. 2) The limited area. For instance, nearly half of the *traffic-camera* objects possess an area less than 256 pixels, as demonstrated in Fig. 4. In other words, this phenomenon corroborates previous findings, i.e., the detection difficulty increases sharply when the object size gets smaller.

3) Baseline Detectors With Different Backbones: Table VIII shows the performance of baseline detectors with different backbone networks. Compared to ResNet-50, ResNet-101 only brings a slight improvement even degrades the performance (see Cascade RCNN and RetinaNet). This phenomenon substantiates previous hypothesis that deeper models might not be better for the size-limited objects and moreover, the highly structural representations in deeper layers which hardly contain small object cues are suboptimal for detection. Swin-T [173] yields substantial improvements for all detectors, especially for FCOS (+5.3 points). This impressive performance reveals the powerful representation ability of shifted-window scheme for small objects, and could shed more light on the subsequent feature extractor design of SOD. Not surprisingly, most detectors with ConvNext-T [174] as the backbone achieve the best performance, exhibiting good robustness and potential in capturing the finer representations of small objects.

4) Qualitative Results: Fig. 10 demonstrates the visualization results of Cascade RCNN on SODA-D test-set. The first pair shows the challenge under complicated background and heavy occlusion, where the detector can hardly learn discriminative representation from small instances and is inclined to lose instances resembling the background. In addition, identifying those partly occluded objects is even more challenging. The second pair represents the detections of low illumination, in which the detector fails to recognize those instances under the shadow, still less predicts accurate bounding boxes.

D. Results Analysis on SODA-A

Based on SODA-A, we investigate the performance of several leading methods of oriented object detection. Also, considering

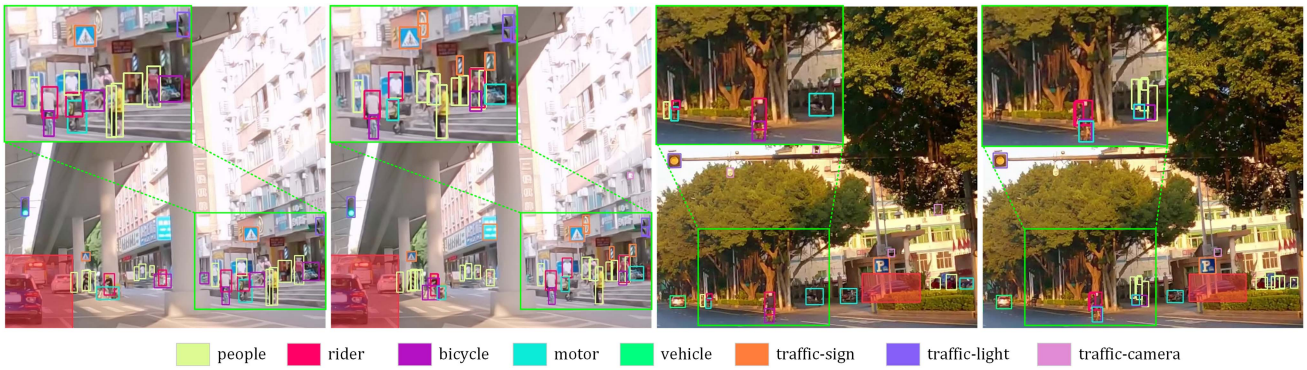

Fig. 10. Qualitative results of Cascade RCNN [166] on SODA-D test-set. Columns 1 and 3 denote the ground-truth annotations and columns 2 and 4 stand for the predictions. Best viewed in color and zoom-in windows, where masked bounding boxes represent *ignore* regions. Only predictions with confidence scores larger than 0.3 are demonstrated.

TABLE IX
BASELINE RESULTS ON SODA-A TEST-SET

Method	Publication	Schedule	AP	AP_{50}	AP_{75}	AP_{eS}	AP_{rS}	AP_{gS}	AP_N	#Param.	FLOPs
Rotated Faster RCNN [1]	TPAMI 2017	1×	32.5	70.1	24.3	11.9	27.3	42.2	34.4	41.14M	292.25G
Rotated RetinaNet [3]	TPAMI 2020	1×	26.8	63.4	16.2	9.1	22.0	35.4	28.2	36.16M	800.21G
RoI Transformer [175]	CVPR 2019	1×	36.0	73.0	30.1	13.5	30.3	46.1	39.5	55.08M	306.20G
Gliding Vertex [176]	TPAMI 2021	1×	31.7	70.8	22.6	11.7	27.0	41.1	33.8	41.14M	292.25G
Oriented RCNN [177]	ICCV 2021	1×	34.4	70.7	28.6	12.5	28.6	44.5	36.7	41.13M	292.44G
S^2A-Net [178]	TGRS 2022	1×	28.3	69.6	13.1	10.2	22.8	35.8	29.5	38.64M	732.74G
DODet [179]	TGRS 2022	1×	31.6	68.1	23.4	11.3	26.3	41.0	33.5	69.34M	555.49G
Oriented RepPoints [180]	CVPR 2022	1×	26.3	58.8	19.0	9.4	22.6	32.4	28.5	55.66M	827.21G
DHRec [181]	TPAMI 2022	1×	30.1	68.8	19.8	10.6	24.6	40.3	34.6	31.99M	792.76G

All the models are trained with a ResNet-50 as the backbone. Schedule denotes the epoch setting during training, where '1×' refers to 12 epochs.

our SODA-A contains densely packed issue, we explore the impact of proposal number for the final performance, please refer to Section D.2 of Appendix, available online.

1) Benchmarking Results: Table IX shows the results of nine representative methods on SODA-A test-set. RoI Transformer [175] achieves top performance with 36.0% AP. This remarkable success can be attributed to its powerful proposal generator, in which rotated proposals produced by the RRoI Learner can guarantee the high recall of small objects. By revising vanilla Faster RCNN to output an additional angle prediction, Rotated Faster RCNN [1] scores 32.5% on AP, which validates the robustness of this prevailing method again. Oriented RCNN [177] obtains a relatively high performance both at overall AP (34.4%). Thanks to its efficient oriented RPN, Oriented RCNN can generate high-quality proposals with negligible parameter grow. From the results of RoI Transformer and Oriented RCNN, we can see that high-quality proposals are of great significance to small object detection, particularly for the densely packed objects. Gliding Vertex [176] and DODet [179] both resort to novel representations for oriented objects, the former learns four gliding offsets to corresponding sides while the latter utilizes aspect ratio and area to denote an object. Gliding Vertex achieves 31.7% AP which is comparable to DODet (31.6%). For one-stage detectors, Rotated RetinaNet [3] achieves 26.8% AP and lags largely behind two-stage ones. This is because SODA-A contains considerable excessively small objects that one-stage paradigm cannot handle well, as discussed in Section V-C1. S^2A-Net [178] designs feature alignment module to alleviate the misalignment problem, and finally achieves an AP with 28.3%. Though it can substantially increase the score of AP_{50}, the concomitant performance decline on the AP_{75} metric can be non-negligible (-3.3 points) when compared to Rotated RetinaNet, which indicates that the performance gain of S^2A-Net is likely to come at the cost of subsequent regression accuracy. Oriented RepPoints [180] achieves 26.3% points on AP metric which is slightly inferior to Rotated RetinaNet, exhibiting such point set representation is unamiable for small objects in aerial scenario, especially for those with large aspect ratios which will be discussed in next section. By exploiting two horizontal rectangles to encode the multi-oriented object, DHRec [181] disposes the discontinuity problem subtly and achieves 30.1% AP which is significantly superior to its one-stage counterparts with least parameters.

2) Category-Wise Results: Category-wise results of baseline algorithms on SODA-A test-set are shown in Table X. The AP of *helicopter* category is observably below that of other classes due to limited instance numbers. The objects of *large-vehicle* and *container* with elongated structure challenge the regression branch especially for Oriented RepPoints, and moreover, Gliding Vertex and DODet have comparable results yet perform variably on different categories, which can be attributed to the different representation about oriented objects.

3) Baseline Detectors With Different Backbones: Table XI shows the performance of baseline detectors with different backbone networks. Similar to the results on SODA-D, we can see that ResNet-101 only brings slight performance improvement

TABLE X
CATEGORY-WISE AP OF BASELINE DETECTORS ON SODA-A TEST-SET

Method	airplane	helicopter	s-vehicle	l-vehicle	ship	container	s-tank	s-pool	windmill	AP
Rotated Faster RCNN [1]	49.4	18.1	33.4	19.6	43.5	29.8	42.8	34.1	21.9	32.5
Rotated RetinaNet [3]	42.0	16.8	29.9	10.0	35.1	23.7	35.1	30.7	18.1	26.8
RoI Transformer [175]	53.2	21.4	36.1	25.9	46.4	35.7	44.6	36.9	23.5	36.0
Gliding Vertex [176]	46.7	12.8	33.3	21.9	43.4	29.8	43.3	31.2	22.7	31.7
Oriented RCNN [177]	52.2	20.2	34.4	24.4	45.2	32.1	43.1	36.3	22.2	34.4
S^2A-Net [178]	41.5	20.4	31.2	14.0	36.7	26.1	29.6	33.8	21.6	28.3
DODet [179]	49.4	19.8	32.1	17.3	41.3	26.0	42.2	34.7	21.3	31.6
Oriented RepPoints [180]	51.7	8.5	30.3	2.6	28.0	19.6	40.3	33.2	21.9	26.3
DHRec [181]	45.5	17.2	31.0	15.6	38.5	28.5	38.8	34.5	20.9	30.1

The training settings are consistent with table IX and the full names of class abbreviation are as follows: S-Vehicle (small-vehicle), L-Vehicle (large-vehicle), S-Tank (storage-tank) and S-Pool (swimming-pool).

Fig. 11. Qualitative results of Oriented RCNN [177] on SODA-A test-set. Columns 1 and 3 represent the ground-truth annotations and columns 2 and 4 denote the predictions. Best viewed in color. Only predictions with confidence scores larger than 0.3 are demonstrated.

TABLE XI
AP PERFORMANCE OF BASELINE DETECTORS ON SODA-A TEST-SET WITH DIFFERENT BACKBONE NETWORKS

Method	ResNet-50	ResNet-101	Swin-T	ConvNext-T
Rotated Faster RCNN [1]	32.5	32.7	33.6	34.3
Rotated RetinaNet [3]	26.8	26.8	23.3	21.7
RoI Transformer [175]	36.0	35.8	36.1	37.5
Gliding Vertex [176]	31.7	32.0	32.9	34.0
Oriented RCNN [177]	34.4	34.4	35.1	35.9
S^2A-Net [178]	28.3	28.3	26.0	/
Oriented RepPoints [180]	26.3	26.7	26.2	25.7

All the models were trained for '1×' schedule.

even decline. However, when Swin-T backbone was employed to extract the features, two fundamentally distinct phenomena occur simultaneously. For RPN-based detectors, Swin-T can yield varying levels of performance gain (from 0.1 points to 1.2 points), but for RPN-free detectors, Swin-T causes substantial performance decline (-3.5 points for Rotated RetinaNet and -2.3 points for S^2A-Net), which is completely different from the results on SODA-D. We conjecture this disparity lies in the limited ability of Swin-T to cope with dense distribution when the detector suffers from misalignment issue, particularly for those objects with extremely close proximity. When taking ConvNext-T as the backbone network the general trend is similar to Swin-T, those RPN-free detectors suffer from more severe misalignment issue because there exists a huge gap between the object regions and horizontal priors.

4) Qualitative Results: We visualize the detection results of Oriented RCNN on SODA-D test-set in Fig. 11. The first pair shows the results of tiny instances and only very few of them were detected, demonstrating that detecting tiny objects is a massive challenge for current detectors, even with top performance. The second pair exhibits the detections of low contrast, of which *airplane* instances possess similar visual feature with background and the model confuses them with *helicopter*. Moreover, because the detailed information which is conducive for identification is hardly retained, the model is likely to utilize visual appearance for recognition instead, which unavoidably results in false positives and incorrect predictions (see the *container* predictions). More qualitative results are exhibited in the Supplementary material, available online.

VI. CONCLUSION AND OUTLOOK

We presented a systematic study on small object detection. Concretely, we exhaustively reviewed hundreds of literature for SOD from the perspective of algorithms and datasets. Moreover, to catalyze the progress of SOD, we constructed two large-scale benchmarks under driving scenario and aerial scene, dubbed SODA-D and SODA-A. SODA-D comprises 278433 instances annotated with horizontal boxes, while SODA-A includes 872069 objects with oriented boxes. The well-annotated datasets, to the best of our knowledge, are the first attempt to large-scale benchmarks tailored for small object detection, and could serve as an impartial platform for benchmarking various SOD methods. On top of SODA, we performed a thorough evaluation and comparison of several representative algorithms. Based on the results, we discuss several potential solutions and directions for future development of SOD task.

Effective Feature Extractor for Small Objects: As alluded to in the results, deeper backbone networks might not be conducive to extract high-quality feature representations for small objects. Designing an effective backbone, which enjoys powerful feature extraction capability while avoiding high computational cost and information loss, is of paramount importance.

High-Quality Hierarchical Representation: FPN is an indispensable part in small object detection. Nevertheless, current feature pyramid architecture is suboptimal for SOD, owing to the heuristic pyramid level assignment strategy, few samples were assigned to higher levels (actually only P_2 feature is responsible to the detection during our benchmark experiments). Consequently, the high-level layers are optimized in an implicit and indirect manner which may hamper the fusion quality. Moreover, detecting on low-level feature maps brings heavy computational burden. Thus, an efficient hierarchical feature architecture tailored for SOD task is in high demand.

Optimized Label Assignment Strategy: As we discussed in Sections II-B1 and C.2 of Appendix, available online, albeit the current label assignment schemes perform well on generic object detection and large objects, they still struggle on the instances of extremely small sizes, neither the overlap-based strategies nor the distribution-based ones. Therefore, designing an optimized strategy to assign sufficient positive samples for size-limited instances can substantially stabilize the training procedure and boost the performance further.

Proper Evaluation Metric for SOD: The multiple IoU thresholds-based evaluation process has been the de facto standard for validating the effectiveness of methods in generic object detection. However, such ubiquitous metric is too stringent for those instances with extremely sizes. In other words, the top priority of small object detection under some specific scenarios is to recognize the objects and obtain their rough locations instead of obsessing how accurate they are. Hence, it is impractical to pursue precise detections of small objects when the model cannot find them. Consequently, borrowing the experience of other fields such as crowd counting and devising a proper metric to guide the training and inference of SOD architectures under some specific scenes plays a significant role in future development.

Acknowledgments

The authors would to thank Peter Kontschieder for the constructive discussions and feedback, as well as their high-quality MVD.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

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

[3] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár,, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[4] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1922–1933, Apr. 2022.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[6] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[7] X. Yu, Y. Gong, N. Jiang, Q. Ye, and Z. Han, "Scale match for tiny person detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1257–1265.

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

[9] X. Dai et al., "Dynamic head: Unifying object detection heads with attentions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7373–7382.

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

[11] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500.

[12] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[13] L. Liu et al., "Deep learning for generic object detection: A survey,," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020.

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

[15] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023.

[16] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2011.

[17] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, "From handcrafted to deep features for pedestrian detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4913–4934, Sep. 2022.

[18] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1480–1500, Jul. 2015.

[19] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 117, pp. 11–28, 2016.

[20] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.

[21] M. Jensen, M. P. Philipsen, A. Mφgelmose, T. B. Moeslund, and M. M. Trivedi, "Vision for looking at traffic lights: Issues, survey, and perspectives," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1800–1815, Jul. 2016.

[22] A. Boukerche and Z. Hou, "Object detection using deep learning methods in traffic scenarios,," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–35, 2021.

[23] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "Imbalance problems in object detection: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3388–3415, Oct. 2021.

[24] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5866–5885, Sep. 2022.

[25] K. Tong and Y. Wu, "Deep learning-based detection from the perspective of small or tiny objects: A survey," *Image Vis Comput*, vol. 123, 2022, Art. no. 104471.

[26] Y. Liu, P. Sun, N. Wergeles, and Y. Shang, "A survey and performance evaluation of deep learning methods for small object detection," *Expert Syst. Appl.*, vol. 172, 2021, Art. no. 114602.

[27] G. Chen et al., "A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 52, no. 2, pp. 936–953, Feb. 2022.

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

[29] X. Yu et al., "Object localization under single coarse point supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4858–4867.

[30] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022.

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

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

[33] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5000–5009.

[34] D. G. Lowe, "Distinctive image features from scale-invariant keypoints,," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

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

[36] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF),," *Comput. Vis. Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[37] C. Cortes and V. Vapnik, "Support-vector networks,," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[38] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, 1995, pp. 278–282.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[40] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[41] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[42] L. Chen, H. Zheng, Z. Yan, and Y. Li, "Discriminative region mining for object detection," *IEEE Trans. Multimedia*, vol. 23, pp. 4297–4310, 2021.

[43] Z. Qin et al., "ThunderNet: Towards real-time generic object detection on mobile devices," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6717–6726.

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

[45] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

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

[47] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.

[48] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6569–6578.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

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

[51] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.

[52] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–16.

[53] Y. Ma, S. Liu, Z. Li, and J. Sun, "IQDET: Instance-wise quality distribution sampling for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1717–1725.

[54] K. Kim and H. S. Lee, "Probabilistic anchor assignment with iou prediction for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 355–371.

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

[56] C. Chen et al., "RRNet: A hybrid detector for object detection in drone-captured images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, 2019, pp. 100–108.

[57] X. Zhang, E. Izquierdo, and K. Chandramouli, "Dense and small object detection in UAV vision based on cascade network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, 2019, pp. 118–126.

[58] X. Wang, D. Zhu, and Y. Yan, "Towards efficient detection for small objects via attention-guided detection network and data augmentation," *Sensors*, vol. 22, no. 19, 2022, Art. no. 7663.

[59] B. Bosquet et al., "A full data augmentation pipeline for small object detection based on generative adversarial networks," *Pattern Recognit.*, vol. 133, 2023, Art. no. 108998.

[60] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S^3FD: Single shot scale-invariant face detector," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 192–201.

[61] C. Zhu, R. Tao, K. Luu, and M. Savvides, "Seeing small faces from robust anchor's perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5127–5136.

[62] C. Xu, J. Wang, W. Yang, and L. Yu, "Dot distance for tiny object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 1192–1201.

[63] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "RFLA: Gaussian receptive based label assignment for tiny object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 526–543.

[64] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[65] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[66] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing,," *RCA Eng.*, vol. 29, no. 6, pp. 33–41, 1984.

[67] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, "Object instance segmentation and fine-grained localization using hypercolumns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 627–639, Apr. 2017.

[68] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[69] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2129–2137.

[70] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 354–370.

[71] J. Li et al., "Dsfd: Dual shot face detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5055–5064.

[72] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7029–7038.

[73] S. Qiao, L.-C. Chen, and A. Yuille, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10208–10219.

[74] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.

[75] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "SSH: Single stage headless face detector," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4885–4894.

[76] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, "Scale-aware trident networks for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6053–6062.

[77] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13668–13677.

[78] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection-snip," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3578–3587.

[79] B. Singh, M. Najibi, and L. S. Davis, "SNIPER: Efficient multi-scale training," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 9310–9320.

[80] M. Najibi, B. Singh, and L. Davis, "AutoFocus: Efficient multi-scale inference," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9745–9755.

[81] Y. Chen et al., "Dynamic scale training for object detection," 2020, *arXiv:2004.12432*.

[82] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.

[83] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu, "Scale-transferrable object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 528–537.

[84] J. Wang, Y. Yuan, and G. Yu, "Face attention network: An effective face detector for the occluded faces," 2017, *arXiv:1711.07246*.

[85] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10778–10787.

[86] H. Zhang, K. Wang, Y. Tian, C. Gou, and F.-Y. Wang, "MFR-CNN: Incorporating multi-scale features and global information for traffic object detection," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8019–8030, Sep. 2018.

[87] S. Woo, S. Hwang, and I. S. Kweon, "StairNet: Top-down semantic aggregation for accurate one shot detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1093–1102.

[88] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.

[89] Q. Zhao et al., "M2Det: A single-shot object detector based on multi-level feature pyramid network," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 9259–9266.

[90] Z. Liu, G. Gao, L. Sun, and L. Fang, "IPG-Net: Image pyramid guidance network for small object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 4422–4430.

[91] Y. Gong, X. Yu, Y. Ding, X. Peng, J. Zhao, and Z. Han, "Effective fusion factor in FPN for tiny object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1159–1167.

[92] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, *arXiv:1911.09516*.

[93] M. Hong, S. Li, Y. Yang, F. Zhu, Q. Zhao, and L. Lu, "SSPNet: Scale selection pyramid network for tiny person detection from UAV images," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, 2021, Art. no. 8018505.

[94] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[95] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain,," *Nat. Rev. Neurosci.*, vol. 3, no. 3, pp. 201–215, 2002.

[96] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[97] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[98] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[99] M. Jaderberg et al., "Spatial transformer networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[100] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[101] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned one-stage object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 3490–3499.

[102] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

[103] K. Yi, Z. Jian, S. Chen, and N. Zheng, "Feature selective small object detection via knowledge-based recurrent attentive neural network," 2018, *arXiv:1803.05263*.

[104] X. Yang et al., "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8231–8240.

[105] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1331–1344, Feb. 2021.

[106] X. Lu, J. Ji, Z. Xing, and Q. Miao, "Attention and feature fusion SSD for remote sensing object detection," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art no. 5501309.

[107] Q. Ran, Q. Wang, B. Zhao, Y. Wu, S. Pu, and Z. Li, "Lightweight oriented object detection using multiscale context and enhanced channel attention in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ.*, vol. 14, pp. 5786–5795, 2021.

[108] Y. Li, Q. Huang, X. Pei, Y. Chen, L. Jiao, and R. Shang, "Cross-layer attention network for small object detection in remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ.*, vol. 14, pp. 2148–2161, 2021.

[109] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "CARAFE: Content-aware reassembly of features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3007–3016.

[110] J. Wu, C. Zhou, Q. Zhang, M. Yang, and J. Yuan, "Self-mimic learning for small-scale pedestrian detection," in *Proc. ACM Multimedia*, 2020, pp. 2012–2020.

[111] J. U. Kim, S. Park, and Y. M. Ro, "Robust small-scale pedestrian detection with cued recall via memory learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 3030–3039.

[112] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.

[113] C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, "Extended feature pyramid network for small object detection," *IEEE Trans. Multimedia*, vol. 24, pp. 1968–1979, 2021.

[114] X. Pan et al., "Self-supervised feature augmentation for large image object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 6745–6758, 2020.

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

[116] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 210–226.

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

[118] B. Na and G. C. Fox, "Object detection by a super-resolution method and a convolutional neural networks," in *Proc. BigData*, 2018, pp. 2263–2269.

[119] J. Noh, W. Bae, W. Lee, J. Seo, and G. Kim, "Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9724–9733.

[120] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1271–1278.

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

[122] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1432.

[123] S. M. A. Bashir and Y. Wang, "Small object detection in remote sensing images with residual feature aggregation-based super-resolution and object detector network," *Remote Sens.*, vol. 13, no. 9, 2021, Art. no. 1854.

[124] A. Torralba, "Contextual priming for object detection,," *Int. J. Comput. Vis.*, vol. 53, no. 2, pp. 169–191, 2003.

[125] X. Tang, D. K. Du, Z. He, and J. Liu, "PyramidBox: A context-assisted single shot face detector," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 812–828.

[126] D. Parikh, C. L. Zitnick, and T. Chen, "Exploring tiny images: The roles of appearance and contextual information for machine and human object recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1978–1991, Oct. 2011.

[127] H. Zhang et al., "Context encoding for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7151–7160.

[128] K. Chen et al., "Hybrid task cascade for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4969–4978.

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

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

[131] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 951–959.

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

[133] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "$\mathcal{R}^2$-CNN: Fast tiny object detection in large-scale remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5512–5524, Aug. 2019.

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

[135] K. Fu, J. Li, L. Ma, K. Mu, and Y. Tian, "Intrinsic relationship reasoning for small object detection," 2020, *arXiv:2009.00833*.
[136] L. V. Pato, R. Negrinho, and P. M. Aguiar, "Seeing without looking: Contextual rescoring of object detections for ap maximization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14610–14618.
[137] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Dec. 2019.
[138] L. Cui et al., "Context-aware block net for small object detection," *IEEE Trans. Cybern.*, vol. 52, no. 4, pp. 2300–2313, Apr. 2022.
[139] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8311–8320.
[140] C. Duan, Z. Wei, C. Zhang, S. Qu, and H. Wang, "Coarse-grained density map guided object detection in aerial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 2789–2798.
[141] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 737–746.
[142] Y. Wang, Y. Yang, and X. Zhao, "Object detection using clustering algorithm adaptive searching regions in aerial images," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 651–664.
[143] F. Ozge Unel, B. O. Ozkalayci, and C. Cigla, "The power of tiling for small object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 582–591.
[144] S. Deng et al., "A global-local self-adaptive network for drone-view object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1556–1569, 2020.
[145] J. Xu, Y. Li, and S. Wang, "Adazoom: Adaptive zoom network for multi-scale object detection in large scenes," 2021, *arXiv:2106.10409*.
[146] J. Leng, M. Mo, Y. Zhou, C. Gao, W. Li, and X. Gao, "Pareto refocusing for drone-view object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1320–1334, Mar. 2023.
[147] O. C. Koyun, R. K. Keser, İ. B. Akkaya, and B. U. Töreyin, "Focus-and-detect: A small object detection framework for aerial images," *Signal Process. Image Commun.*, vol. 104, 2022, Art. no. 116675.
[148] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "Eurocity persons: A novel benchmark for person detection in traffic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1844–1861, Aug. 2019.
[149] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, "WiderPerson: A diverse dataset for dense pedestrian detection in the wild," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 380–393, Feb. 2020.
[150] F. Larsson and M. Felsberg, "Using fourier descriptors and spatial models for traffic sign recognition," in *Proc. Scand. Conf. Image Anal.*, 2011, pp. 238–249.
[151] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1484–1497, Dec. 2012.
[152] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark," in *Proc. Int. Joint Conf. Neural Netw.*, 2013, pp. 1–8.
[153] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2110–2118.
[154] K. Behrendt, L. Novak, and R. Botros, "A deep learning approach to traffic lights: Detection, tracking, and classification," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 1370–1377.
[155] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 3735–3739.
[156] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun.*, vol. 34, pp. 187–203, 2016.
[157] D. Lam et al., "xview: Objects in context in overhead imagery," 2018, *arXiv:1802.07856*.
[158] H. Yu et al., "The unmanned aerial vehicle benchmark: Object detection, tracking and baseline,," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1141–1159, 2020.
[159] P. Zhu et al., "Detection and tracking meet drones challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7380–7399, Nov. 2022.
[160] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 190, pp. 79–93, 2022.
[161] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-crowd: A large-scale benchmark for crowd counting and localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2141–2149, Jun. 2021.
[162] G. Cheng et al., "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625411.
[163] Y. Pang, J. Cao, Y. Li, J. Xie, H. Sun, and J. Gong, "TJU-DHD: A diverse high-resolution dataset for object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 207–219, 2021.
[164] J. Han et al., "SODA10M: A large-scale 2D self/semi-supervised object detection dataset for autonomous driving," 2021, *arXiv:2106.11118*.
[165] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4165–4173.
[166] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2021.
[167] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.
[168] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9656–9665.
[169] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9759–9768.
[170] P. Sun et al., "Sparse R-CNN: End-to-end object detection with learnable proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14449–14458.
[171] K. Chen et al., "MMDetection: OpenMMlab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
[172] Y. Zhou et al., "MMRotate: A rotated object detection benchmark using pytorch," 2022, *arXiv:2204.13317*.
[173] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.
[174] Z. Liu et al., "A convnet for the 2020s," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11976–11986.
[175] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning ROI transformer for oriented object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2844–2853.
[176] Y. Xu et al., "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.
[177] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 3520–3529.
[178] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602511.
[179] G. Cheng et al., "Dual-aligned oriented detector," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618111.
[180] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1829–1838.
[181] G. Nie and H. Huang, "Multi-oriented object detection in aerial images with double horizontal rectangles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4932–4944, Apr. 2023.

Gong Cheng received the BS degree from Xidian University, Xi'an, China, in 2007, and the MS and PhD degrees from Northwestern Polytechnical University, Xi'an, China, in 2010 and 2013, respectively. He is a professor with Northwestern Polytechnical University. His main research interests include computer vision, pattern recognition, and remote sensing image understanding.

Xiang Yuan received the BS degree from Chang'an University, Xi'an, China, in 2017, and the MS degree from Northwestern Polytechnical University, Xi'an, China, in 2021. He is currently working toward the PhD degree with the School of Automation, Northwestern Polytechnical University. His main research interests include computer vision and object detection.

Xiwen Yao received the BS and PhD degrees from Northwestern Polytechnical University, Xi'an, China, in 2010 and 2016, respectively. He is an associate professor with Northwestern Polytechnical University. His research interests include computer vision and remote sensing image processing, especially on fine-grained image classification and object detection.

Kebing Yan received the BS degree from Northwestern Polytechnical University, Xi'an, China, in 2021. She is currently working toward the MS degree with Northwestern Polytechnical University. Her main research interests include pattern recognition and object detection.

Qinghua Zeng received his BS degree from Northwestern Polytechnical University, Xi'an, China, in 2021, where he is currently pursuing the MS degree.

His main research interests include object detection.

Xingxing Xie received the BS degree from Inner Mongolia University, Huhhot, China, in 2015, and the MS degree from Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently working toward the PhD degree with Northwestern Polytechnical University. His main research interests include computer vision and pattern recognition.

Junwei Han (Fellow, IEEE) received the BS, MS, and PhD degrees in pattern recognition and intelligent systems from Northwestern Polytechnical University, Xi'an, China, in 1999, 2001, and 2003, respectively. He was a research fellow with Nanyang Technological University, Singapore, The Chinese University of Hong Kong, Hong Kong, The Dublin City University, Dublin, Ireland, and The University of Dundee, Dundee, U.K., from 2003 to 2010. He is a professor with Northwestern Polytechnical University. His research interests include computer vision and brain imaging analysis.