A Review on Small Object Detection: The Synergistic Evolution of CNNs and Transformers

Ke Wang, *Member, IEEE,* Yang Chen, Sheng Li, Jianbo Lu, *Fellow, IEEE,*

***Abstract*—Small object detection is a significant challenge due to small target size, difficult feature extraction, susceptibility to occlusion, and dataset imbalance. Traditional convolutional neural networks (CNNs) have limitations in handling long-range dependencies and sequential data due to restricted local receptive fields. In contrast, the Transformer architecture, through its self-attention mechanism, effectively optimizes long-range dependency processing and excels in various domains. Recently, hybrid architectures combining the local feature extraction of CNNs with the global information processing of Transformers have emerged, offering new solutions for small object detection. However, few systematic reviews exist on the application of CNNs and Transformers in this area. Researchers often struggle to understand the strengths, weaknesses, and appropriate scenarios for these methods.Therefore, this paper provides a comprehensive review and comparison of the application of CNNs, Transformers, and their hybrid architectures in small object detection, delving into the key technologies and optimization strategies of each. Additionally, we review existing classical and domain-specific small object detection datasets and introduce a custom traffic road dataset for small object detection experiments in intelligent vehicles. Finally, we summarize potential research directions to support future development in small object detection. We hope this review inspires further research and advances progress in addressing this critical issue.To facilitate future research, we create a repository that includes links to relevant reviews and methodological papers for learning at https://github.com/chenyang447/A-Review-on-Small-Object-Detection.**

***Index Terms*—Small Object Detection, Convolutional Neural Networks, Transformer Architecture, Intelligent Vehicles.**

I. Introduction

Small object detection is a pivotal research direction in the field of computer vision, focused on identifying and locating objects of smaller size in images or videos. With continuous technological advancement, the importance of small object detection in numerous practical applications has become increasingly prominent, including autonomous driving, security surveillance, medical diagnostics, and remote sensing. In these scenarios, accurately detecting small objects is not only valuable but also high-risk, with its precision being crucial to the overall system performance. The development of deep learning techniques and the enhancement of computational resources have led to breakthrough progress in small object detection in recent years. These advancements are primarily reflected in improved convolutional neural network architectures (such as FPN [1] and RetinaNet [2]), the application of Transformer models (such as DETR [3]), multi-scale feature fusion, data augmentation techniques, optimized loss functions, and training strategies. These technological innovations have significantly enhanced the performance of small object detection, making it more accurate and reliable in complex environments. Nevertheless, small object detection still faces challenges such as object diversity, high background interference, and data scarcity. Future research will continue to explore model architectures, dataset richness, and training strategies to further improve the effectiveness and application scope of small object detection technology. This paper reviews various algorithms applicable to small object detection and provides an in-depth analysis of key technologies and optimization strategies, offering valuable guidance to researchers in the field.

The definition of small objects varies with specific applications and contexts, usually determined by the ratio of the object size to the image resolution. For example, in the COCO dataset [4], small objects refer to those with an area less than 32×32 pixels, while in datasets such as PASCAL VOC [5], TT100K [6], or ImageNet [7], researchers generally define small objects based on their proportion in the image.

As illustrated in Fig 1, small object detection faces a unique set of challenges within the realm of traditional object detection:

1)Small size and low resolution: Small objects occupy a limited number of pixels in an image, making their features potentially weak at the pixel level and difficult for detection algorithms to capture accurately. This scarcity is particularly pronounced against complex backgrounds, where other elements may create visual interference, blending the features of small objects with their surroundings and reducing the recognition capability of detection algorithms.

2)Background noise impact: Due to their low-resolution nature, the distinguishability of small objects' features is further weakened. In image processing, the pixel representation of small objects may be insufficient to convey their full features, making them hard to differentiate from background noise.

3)Obstruction and overlap: Due to their smaller size, small objects are more prone to occlusion by other objects in the image, especially in complex scenes. Additionally, small objects are more sensitive to environmental factors such as

K. Wang was with the State Key Laboratory of Mechanical Transmission for Advanced Equipment and the College of Mechanical and Vehicle Engineering, Chongqing University, China, 400044 (e-mail: kewang@cqu.edu.cn).

Y. Chen was with the College of Mechanical and Vehicle Engineering, Chongqing University, China, 400044,Chongqing University, China, 400044 (e-mail: chenyang@stu.cqu.edu.cn).

S. Li was with the College of Mechanical and Vehicle Engineering, Chongqing University, China, 400044,Chongqing University, China, 400044 (e-mail: 2321326148@qq.com).

J. Lu is with Vehicle Engineering, Nikola Motor Company, Phoenix, AZ 85040, USA (e-mail: jianbo.lu@ieee.org)

Manuscript received August 17, 2005 (Corresponding author: Ke wang)

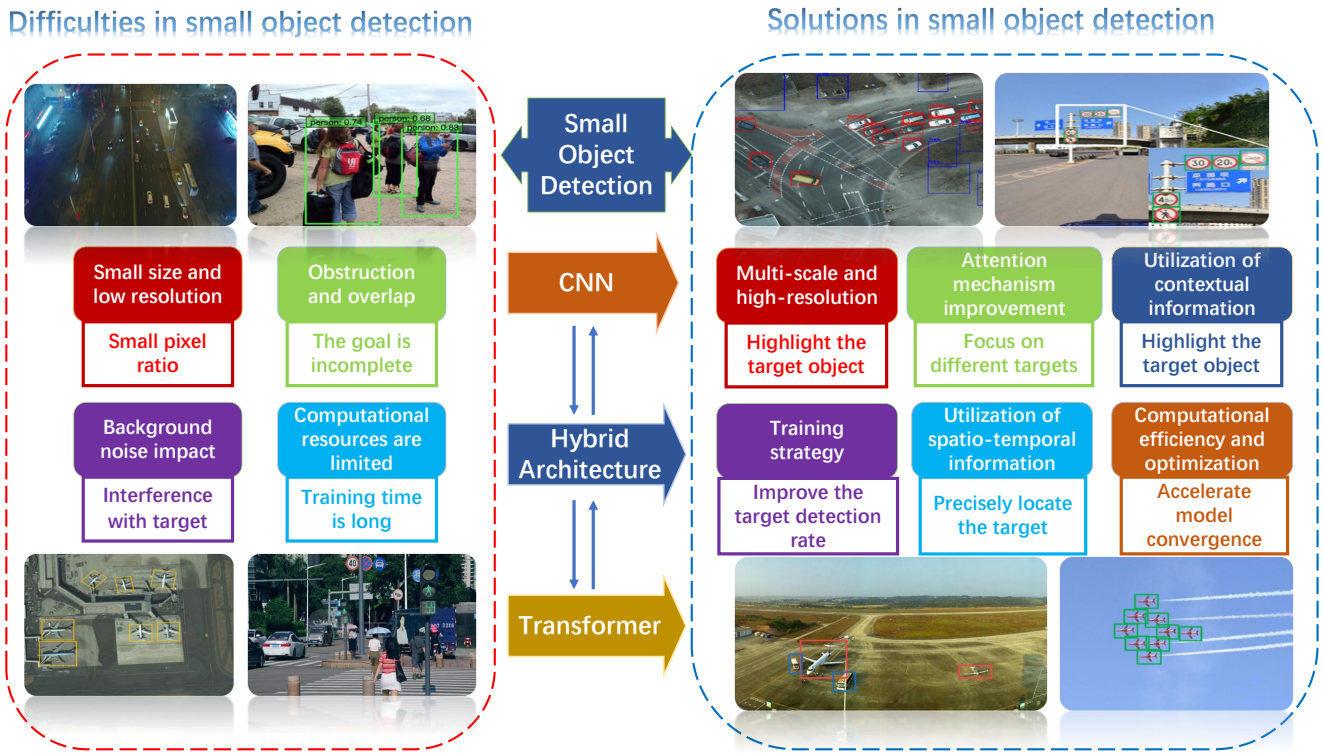


Fig. 1. Difficulties and Solutions in small object detection.

lighting conditions and weather. For instance, under nighttime or adverse weather conditions like rain, fog, or snow, the visibility and feature recognizability of small objects significantly decrease, posing a considerable challenge for detection algorithms.

4)Computational resources are limited: Small object detection requires processing more details and complex scenes, necessitating algorithms that support high-resolution image processing and intricate feature extraction. Consequently, these algorithms typically demand more powerful computational capabilities to ensure real-time and accurate detection of small objects.

In response to these challenges in the field of small object detection, researchers have developed a series of innovative methods to enhance detection performance. These methods are optimized for the characteristics of small objects and consider the needs and constraints of practical applications. We will now delve into these methods and analyze how they effectively address the key issues in small object detection.

1)Multi-scale and high-resolution: To overcome the problem of scarce feature information for small objects, multi-scale feature extraction techniques are widely applied in small object detection. By extracting features at different resolution levels, models can capture a range of details from coarse to fine, thereby enhancing the recognition of small objects. Moreover, high-resolution image inputs provide more pixel details, enabling models to better distinguish small objects from background noise.

2)Attention mechanism improvement: Attention mechanisms help models focus on important parts of the image, thus improving the detection performance of small objects. By incorporating attention modules, models can learn to ignore irrelevant information in complex backgrounds and concentrate on the key features of small objects, helping to reduce false detections and increase accuracy.

3)Utilization of contextual information: Small object detection can be improved by leveraging contextual information. Contextual cues, such as objects surrounding the target and the overall layout of the scene, can provide additional clues for recognizing small objects. Integrating this information allows models to better understand the role and significance of small objects within the entire scene.

4)Training strategy improvement: Effective training strategies are crucial for enhancing the performance of small object detection. This includes using appropriate loss functions to handle imbalanced class distributions and employing techniques like hard negative mining to focus on difficult cases in model learning. Additionally, data augmentation techniques such as random cropping and scaling can increase the model's robustness to small objects.

5)Utilization of spatiotemporal information: In the case of videos or consecutive frames, the use of spatiotemporal information is particularly important for small object detection. By analyzing the motion and changes of targets over time, models can better track and identify small objects, even when they are temporarily invisible due to occlusion or environmental factors.

6)Computational efficiency and optimization: To meet the requirements of real-time performance, small object detection algorithms need to optimize computational efficiency while maintaining high accuracy. This can be achieved by designing lightweight network architectures, reducing computation-

ally intensive operations, and adopting model compression techniques. Furthermore, efficient algorithms can reduce the reliance on high-performance computing resources, enabling small object detection on resource-constrained devices.

As shown in Table 1, numerous scholars have previously proposed various technical approaches based on CNNs to address the challenges of small object detection. These include multi-scale feature learning, data augmentation, training strategies, context-based detection, detection enhanced by Generative Adversarial Networks (GANs) [8], and super-resolution techniques. However, with the widespread adoption of Transformer architectures, many past reviews based on CNNs are no longer able to deeply adapt to the development of current technological solutions. In light of this, this paper focuses on introducing the latest research based on Transformer architectures, covering both recent and classic papers on Transformer and its hybrid architectures, and delves into the impact of Transformer architectures on the performance of small object detection.

This article comprehensively summarizes the recent research progress on CNNs, detectors based on Transformer architectures, and their hybrid structures. It delves into the respective strengths of CNNs and Transformers, discusses the deep-seated reasons why Transformers challenge CNNs, and highlights the advantages of Transformer architectures in handling long-range dependencies, deeply understanding context, and enhancing the connections between global features. Additionally, the paper outlines the technical solutions for small object detection from multiple key dimensions, including multi-scale and resolution methods, the application of attention mechanisms, effective utilization of contextual information, improved training strategies, architectural optimization, the use of spatiotemporal information (especially in video images or dynamic scenes), and the enhancement of computational efficiency. The structure of the article is as follows: Section 2 provides an overview of CNN architectures, Transformer architectures, and their hybrid structures, illustrating the strengths and weaknesses of different architectures and the functions of each component, with a detailed analysis of the pioneering works of Transformer architecture such as ViT [9], DETR [3], Deformable DETR [10], and the hybrid structure ViT-FRCNN [11]. Section 3 conducts a detailed study of small object detection techniques based on CNNs and Transformers. Section 4 summarizes the datasets for small object detection and presents them categorized by domain. Section 5 discusses the development prospects of small object detection and potential future technical directions. Finally, Section 6 concludes the entire review.

Specifically, our main contributions include:

1)We deeply dissect the current methods of small object detection from six dimensions: multi-Scale and Super-Resolution Techniques, application of attention mechanisms, utilization of contextual information, improvement of training strategies, Exploitation of Spatio-Temporal Information, and Efficiency and Optimization in Computing.We evaluate various algorithms from aspects such as real-time performance, accuracy, robustness, and cost-effectiveness.

2)We provide a comprehensive review of existing classic and domain-specific small object detection datasets and introduce our custom traffic road dataset, specifically designed for small object detection experiments in the field of intelligent vehicles.

3)We point out the current challenges and potential future research directions, which will facilitate the further development of small object detection methods in various fields. Small objects typically refer to objects that occupy a relatively small pixel area in an image. Despite their small proportion, they are indispensable in fields such as remote sensing image analysis, urban traffic monitoring, intelligent vehicles, and video processing.

II. BACKGROUND OF CNN AND TRANSFORMER

This section will review the development background of small object detection technology, with a special focus on the evolution from CNNs to the emergence of the Transformer architecture, and the impact of the latter on small object detection technology. As illustrated in Fig 2, the three architectures are presented.By combing through the evolution of these key technologies, we aim to clarify the central position of small object detection in the field of computer vision and its practical value in different application scenarios.

A. Object Detection Models Based on CNNs

In the field of computer vision, mainstream object detection models based on CNNs are primarily divided into two categories: single-stage object detectors and two-stage object detectors based on region proposals. Single-stage object detectors, such as the YOLO series [18], [19], [20], [21], [22], [23], [24], SSD [25], RetinaNet [2], and EfficientDet [26], use a convolutional neural network to directly predict the categories and locations of different objects. Single-stage detectors avoid the use of candidate boxes and transform the object localization problem into a regression problem, directly generating the category probabilities and location coordinates of objects, allowing for the final results to be obtained through one detection pass. These models are fast in detection speed but may be slightly lacking in accuracy. To understand in depth, we analyze YOLOv4 [19] as an example. The network structure of YOLOv4 mainly includes three parts: Backbone (CSPDarknet53), Neck (SPP and PAN), and Head. CSPDarknet53 optimizes Darknet53 by introducing the CSP module, which itself contains fifty-three convolutional layers for extracting deep features of images. The CSP module divides the feature map into two parts, one part directly convolves, and the other part connects across stages, reducing the amount of computation while retaining more original information. The feature pyramid structure of CSPDarknet53 enables it to extract and fuse features at different scales, thus effectively handling targets of different sizes and proportions. The Neck part uses Spatial Pyramid Pooling (SPP) and Path Aggregation Network (PAN) to merge multi-scale feature information to enhance detection performance. The SPP module performs feature fusion through different sizes of max pooling layers and channel-wise concatenation, while the PAN structure optimizes based on FPN, achieving more effective

TABLE I
SUMMARY SEVERAL REVIEWS RELATED TO SMALL OBJECT DETECTION

Title	Ref	Year	Content	Advantages	Disadvantages
Transformers in Small Object Detection: A Benchmark and Survey of State-of-the-Art	[12]	2023	This paper provides a detailed exploration of the application of transformers in Small Object Detection (SOD).	Detailed introduction	It is not comprehensive enough and lacks a review of CNN and hybrid architectures.
Toward Detection of Small Objects Using Deep Learning Methods: A Review	[13]	2022	This paper provides an analysis of contemporary general object detectors and investigates several techniques to improve object detection performance.	Targeted	Not comprehensive enough.
Deep learning-based detection from the perspective of small or tiny objects: A survey	[14]	2022	This survey comprehensively discusses small or tiny object datasets, techniques for small or tiny object detection.	Comprehensive	It lacks an in-depth analysis of existing methods.
A survey and performance evaluation of deep learning methods for small object detection	[15]	2021	According to the four challenges of small object detection, the corresponding solutions are summarized and the experimental analysis is carried out.	Comprehensive	It lacks of in-depth experimentation.
A Survey of the Four Pillars for Small Object Detection: Multiscale Representation, Contextual Information, Super-Resolution, and Region Proposal	[16]	2020	This paper reviews the four pillars of small object detection based on deep learning.	Systemic	The summary of the methods is not comprehensive enough, and the analysis is not sufficiently in-depth.
Recent advances in small object detection based on deep learning: A review	[17]	2020	Reviews the existing deep learning-based small object detection methods from five aspects, analyses experimental results on five datasets, and points out five promising directions.	Cutting-edge	The comparison and evaluation of different methods are not sufficiently in-depth.

cross-scale information transfer. The Head part consists of multiple convolutional layers responsible for predicting and outputting the location and category information of objects.

On the other hand, two-stage object detectors based on region proposals, such as the R-CNN series (including R-CNN [26], Fast R-CNN [27], Faster R-CNN [28], and Mask R-CNN [29]), first use a region proposal network to generate candidate areas, then classify and regress these areas, obtaining the final results through two stages. Although these methods have higher accuracy, their two-stage detection approach sacrifices some detection speed. Taking R-CNN as an example, its introduction in 2013 marked the application of deep learning in the field of object detection. R-CNN uses methods such as selective search to generate candidate areas, extracts features through a convolutional neural network, and then performs classification and bounding box regression on the fully connected layers. Despite R-CNN's excellent accuracy, its detection speed is slow and computationally intensive.

B. Detectors based on the Transformer architecture

Vaswani and others introduced the Transformer architecture [30] in 2017, which is entirely based on attention mechanisms. The Transformer architecture consists of two main parts: the encoder and the decoder. The encoder module first divides the input image's feature map into image patches, encodes them into a sequence through convolutional layers, and combines them with positional embeddings to preserve spatial location information. Then, all feature vectors enter the self-attention mechanism layer, where each position's feature vector calculates attention with features from other positions, capturing global image information and dynamically adjusting weights. Subsequently, the feature vectors undergo a nonlinear transformation through a feed-forward neural network containing multi-layer perceptrons (MLP). The decoder layer receives feature representations from the encoder, aiding in understanding the relationship between input and output. The self-attention layer in the decoder is similar to that in the encoder but also includes a cross-attention layer, which performs interactive calculations between the encoder's output and the decoder's input, enabling the decoder to generate the correct output sequence. These different attention mechanisms play various roles in the Transformer decoder: self-attention captures dependencies within the decoder's input sequence, while cross-attention interacts the information between the decoder's current position and the encoder's output sequence. This architecture is optimized for sequence-to-sequence tasks, simplifying the design of traditional recurrent neural networks (RNNs) [31] and CNNs, while also enhancing parallel performance. Its exceptional modeling capabilities and generalization performance have completely surpassed traditional models based on RNNs and CNNs, marking significant progress in the field of deep learning. The Transformer architecture was first applied to computer vision tasks in 2020. ViT [9] (Vision Transformer) and DETR [3] (Detection Transformer) respectively applied it to image classification and object detection tasks, achieving image preprocessing by segmenting images into patches as sequential input. The specific implementation of ViT includes segmenting the input image into fixed-size patches, linearly transforming each patch into a low-dimensional vector representation, and adding positional encoding to retain spatial location information in the sequence. Additionally, a learnable class token is embedded for image classification tasks. The Transformer encoder in ViT consists of alternating layers of multi-head self-attention and multi-layer perceptrons, with layer normalization and residual connections applied. Finally, the class token is input into an MLP Head for classification prediction. The introduction of ViT brought new ideas to the field of computer vision, demonstrating the potential of Transformers in image classification tasks and promoting the development of subsequent research. The launch of DETR marked an important milestone. A notable feature of the DETR framework is that it discards prior knowledge and constraints such as Non-Maximum Suppression (NMS) and anchors, adopting an end-to-end design that greatly simplifies the object

detection process. DETR showed excellent performance on equivalent datasets. However, due to the generally large image sizes in object detection, direct input to the Transformer would lead to an excessive computational load of parameters. To address this, DETR uses a CNN at the front end of its architecture for feature extraction and dimensionality reduction of feature maps. One of the core elements of DETR is its set-based global loss function, which helps reduce reliance on CNN-based techniques in post-processing. This loss function eliminates duplicate predictions through a bipartite matching mechanism, ensuring that each prediction matches its corresponding ground truth bounding box. Although DETR excels in some aspects, it requires a longer time to converge during training and performs less well in detecting small objects. To overcome these limitations, Deformable DETR [10], an improved version of DETR, was proposed. Unlike DETR, Deformable DETR employs deformable attention modules, allowing the model to focus only on key sampling points around reference points. By introducing offsets, it dynamically adjusts the sampling positions of attention, enabling the model to adaptively focus on different areas of the target and capture more detailed features. This method not only reduces the model's parameter count but also accelerates the model's convergence speed. To improve the detection accuracy of small objects, Deformable DETR constructs a multi-scale feature pyramid on the input features. This allows the model to perform object detection at different scales, with the multi-scale feature pyramid providing more detailed information, thereby enhancing the perception and localization of small targets. With these improvements, Deformable DETR not only enhances the accuracy of small object detection but also optimizes the application of Transformers in object detection tasks, showcasing the tremendous potential of utilizing the Transformer architecture in complex computer vision tasks.

C. Hybrid architectures combining CNNs and Transformers

The analysis of DETR reveals that many Transformer-based studies rely on Convolutional Neural Networks (CNNs) as the backbone to extract image features and adjust image sizes to fit the requirements of Transformers. In such cases, both Transformers and CNNs play crucial roles within the hybrid architecture. For instance, VIT-FRCNN is a typical example of this kind of hybrid structure. In VIT-FRCNN, the Vision Transformer (ViT) contains only the encoder part of the Transformer but effectively completes the image classification task through the class token, effectively replacing the backbone in FRCNN. In the architecture of VIT-FRCNN, the sequence formed by ViT processing image patches contains a wealth of local information, which is reorganized into feature maps at the junction of ViT and FRCNN for subsequent classification and localization tasks.

In terms of enhancing the performance of small object detection, CNNs and Transformers each have their distinct advantages. A key issue is that maintaining a high spatial resolution within the Transformer architecture is crucial, but this often comes at the cost of a quadratic increase in the computational load of the detector. To overcome these limitations and challenges, further research and exploration have been undertaken. These studies are dedicated to finding ways to reduce computational complexity while maintaining high spatial resolution or developing new model architectures to use the capabilities of Transformers more efficiently in small object detection tasks.

For example, some research focuses on developing more effective feature dimensionality reduction and feature fusion strategies to reduce the computational burden while retaining sufficient spatial information. Other studies introduce new modules or optimize existing Transformer structures, such as using local attention mechanisms or adding specific network layers to improve the efficiency and accuracy of the model when dealing with small objects. These innovative research efforts and technological advancements will be discussed in detail in the following sections. Through these explorations, we hope to make greater progress in the challenging field of small object detection and further enhance the overall performance of computer vision systems based on deep learning technology.

III. FROM CNN TO TRANSFORMER FOR SMALL OBJECT DETECTION

In this Section, we will delve into algorithms related to small object detection. As shown in Fig 3, small object detection methods can be categorized into several types. We primarily classify these methods based on the following criteria: multi-Scale and Super-Resolution Techniques, application of attention mechanisms, utilization of contextual information, improvement of training strategies, Exploitation of Spatio-Temporal Information, and Efficiency and Optimization in Computing. In the following sections, we will discuss each category in detail.

A. Multi-Scale and Super-Resolution Techniques

In the field of computer vision and image processing, multiscale and super-resolution techniques have significantly enhanced image analysis by integrating information across different scales and resolutions. In object detection algorithms, shallow networks capture rich details with their high resolution and small receptive fields, while deep networks understand images profoundly with their large receptive fields and strong semantic information processing capabilities. Combining the strengths of both networks, multiscale feature learning has achieved more precise object detection. Super-resolution technology further enhances this process, reconstructing high-resolution details from low-resolution features. The combined use of multiscale and super-resolution techniques has become the norm in the practice of small object detection,several typical algorithms are shown in Fig 4.

To fully leverage shallow and deep features, traditional methods primarily employ techniques such as dilated convolution and deconvolution. For instance, N. Zeng proposed a multi-scale feature fusion method, ABFPN [32]. This method uses dilated convolutions with different expansion rates to capture contextual information and achieve efficient feature fusion through skip connections. MR-CNN [33] upsamples deep features using multi-scale deconvolution operations and

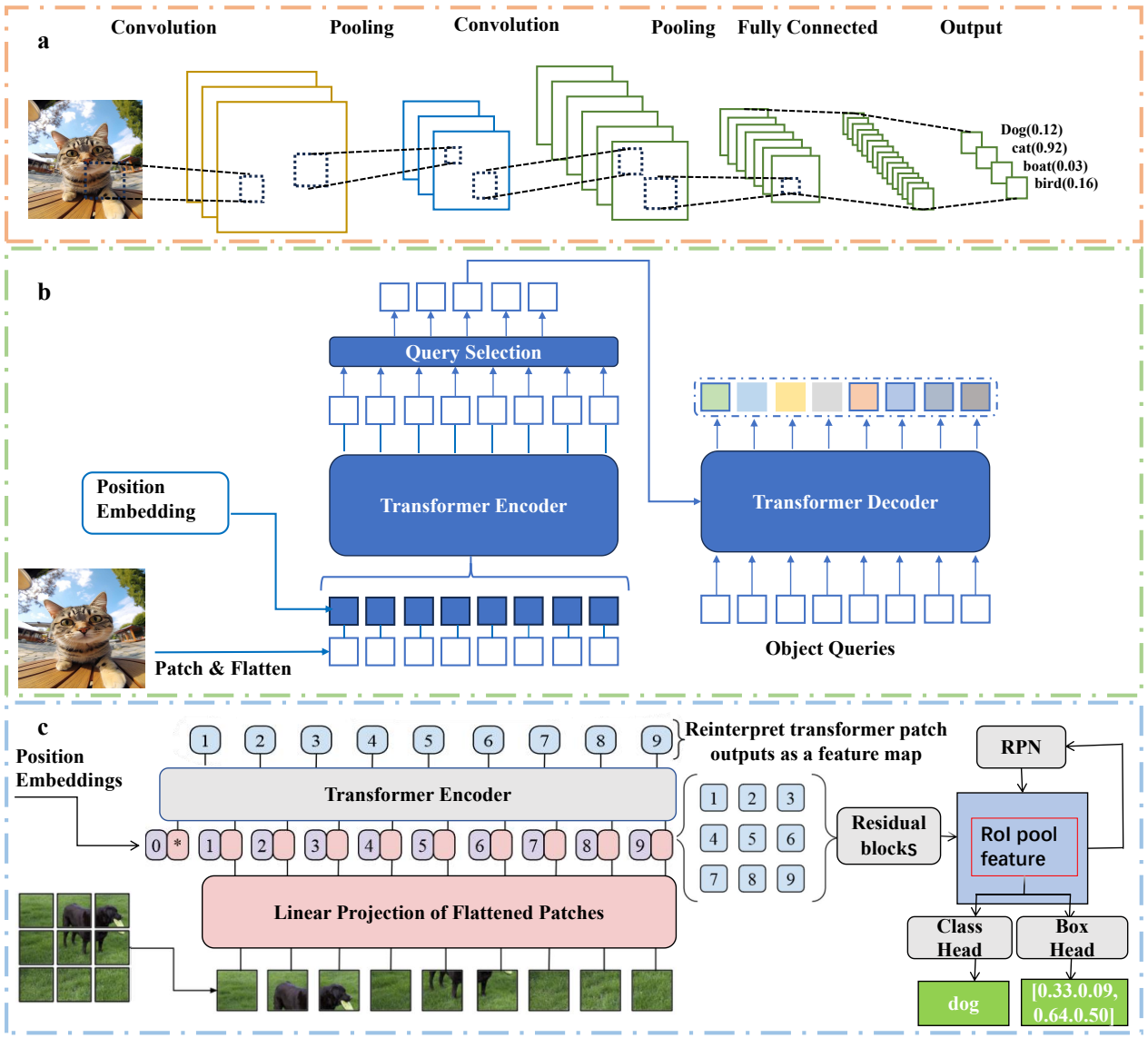


Fig. 2. [a] CNN Model Architecture. [b] Transformer Model Architecture. [c] Vit-Frcnn Model Architecture.

combines them with shallow features to form comprehensive feature maps. Z. Liu's DR-CNN [34] integrates features at different levels by incorporating deconvolution and normalization layers into the output of convolutional layers. These methods improve multi-scale feature fusion to enhance object detection and semantic segmentation performance.Unlike the traditional FPN [1] and its variants, Deformable DETR [10] directly generates multi-layer feature maps using convolution and achieves natural feature fusion through an attention mechanism, eliminating the need for a feature pyramid structure. Building on this, Dynamic DETR [16] employs deformable convolutions for scale fusion and utilizes the SE attention module [35] for channel fusion, enhancing multi-scale and multi-channel feature processing efficiency. CF-DETR [36] introduces the Transformer FPN (TEF) module to enhance multi-scale information and designs a novel CF decoder layer to fully exploit multi-scale information. To reduce computational costs, O2DETR [37] uses different scale feature maps from ResNet [38] and performs spatial and channel-level information fusion through depthwise separable convolutions, significantly reducing memory and computational demands when using multi-scale features. These Transformer-based models improve the feature fusion process in unique ways, thereby increasing object detection accuracy and efficiency while reducing computational resource consumption.

In terms of detecting low-pixel and noise-affected small objects, multi-scale feature fusion methods also demonstrate advantages. N. Zeng [32] improved Faster R-CNN [28] by fusing multi-level feature maps and optimizing anchor sizes, significantly enhancing feature map resolution and detection performance. [39] adopted a multi-scale feature detection method, extracting features from different convolutional layers to enhance detection accuracy. W. Liu [40] designed a deep network for precise vehicle detection, maintaining the spatial layout of ROI features for accurate representation and detection of small vehicles. D. W. Ma [41] proposed a multi-scale

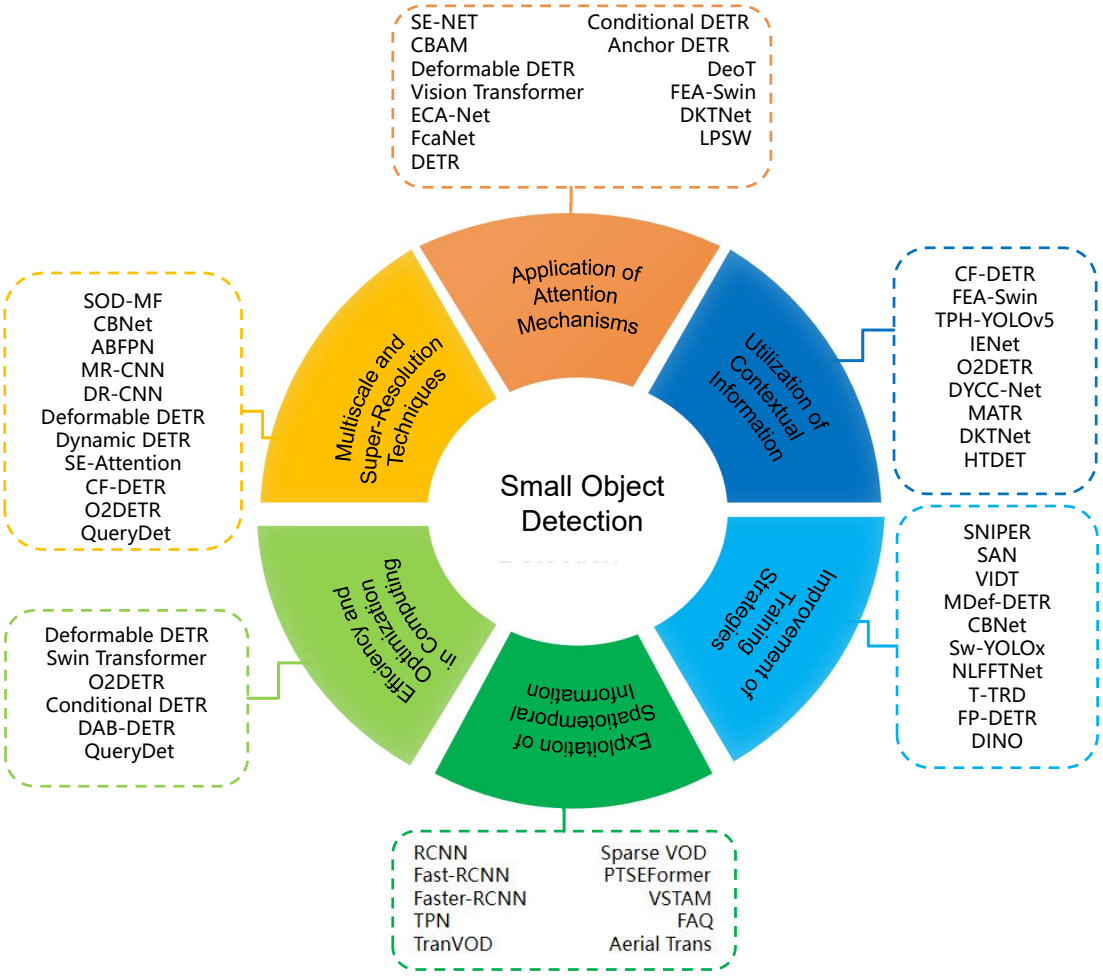


Fig. 3. Classification of small object detection using CNN and Transformer architectures and popular object detection methods in each category

multi-task region proposal method for effectively locating and identifying small objects. Q. Meng [42] built an image pyramid by inputting divided original images into a VGG-16 network [43] and adding feature map fusion post-CNN to match the original object sizes. Z. Liang [44] developed a two-stage detector that enhances the semantic features of small objects using a feature pyramid architecture during the region proposal stage and employs a dense convolutional network for better feature transmission and reuse during the classification stage, achieving more accurate classification. P. Du [45] proposed a novel network architecture and feature fusion mechanism based on YOLOv3, optimizing feature fusion weight selection through multi-scale convolution kernels and differential receptive field application. The CBNet [46] architecture connects multiple identical backbone networks in a composite manner, fusing high and low-level features and gradually expanding the receptive field for more effective object detection. T-TRD [47] introduced an end-to-end transformer-based RSI object detection framework capable of aggregating features at multiple scales and simulating interactions between paired instances, enhancing detection accuracy and efficiency. QueryDet [48] reduces the computational cost of feature pyramid-based object detectors by coarsely locating small objects on high-level feature maps and then querying details on lower-level feature maps.

By incorporating multi-scale feature fusion and super-resolution methods, researchers have made significant progress in detecting low-pixel and noise-affected small objects. Current strategies primarily enhance small object detection accuracy through multi-level feature map fusion, anchor size optimization, feature extraction from different convolutional layers, specialized deep networks, and novel region proposal methods. Future research can further explore these methods' potential, optimizing multi-scale feature fusion techniques to achieve more efficient and accurate small object detection.

B. Application of Attention Mechanisms

The attention mechanism plays a crucial role in small object detection by guiding the model to focus on key areas within an image, significantly enhancing detection performance. This mechanism is primarily implemented through spatial attention, channel attention, or a combination of both. The spatial attention mechanism mimics the human visual system's focus shifts, enabling the model to concentrate on critical parts of the image while ignoring irrelevant regions. This is typically achieved by generating a two-dimensional spatial attention

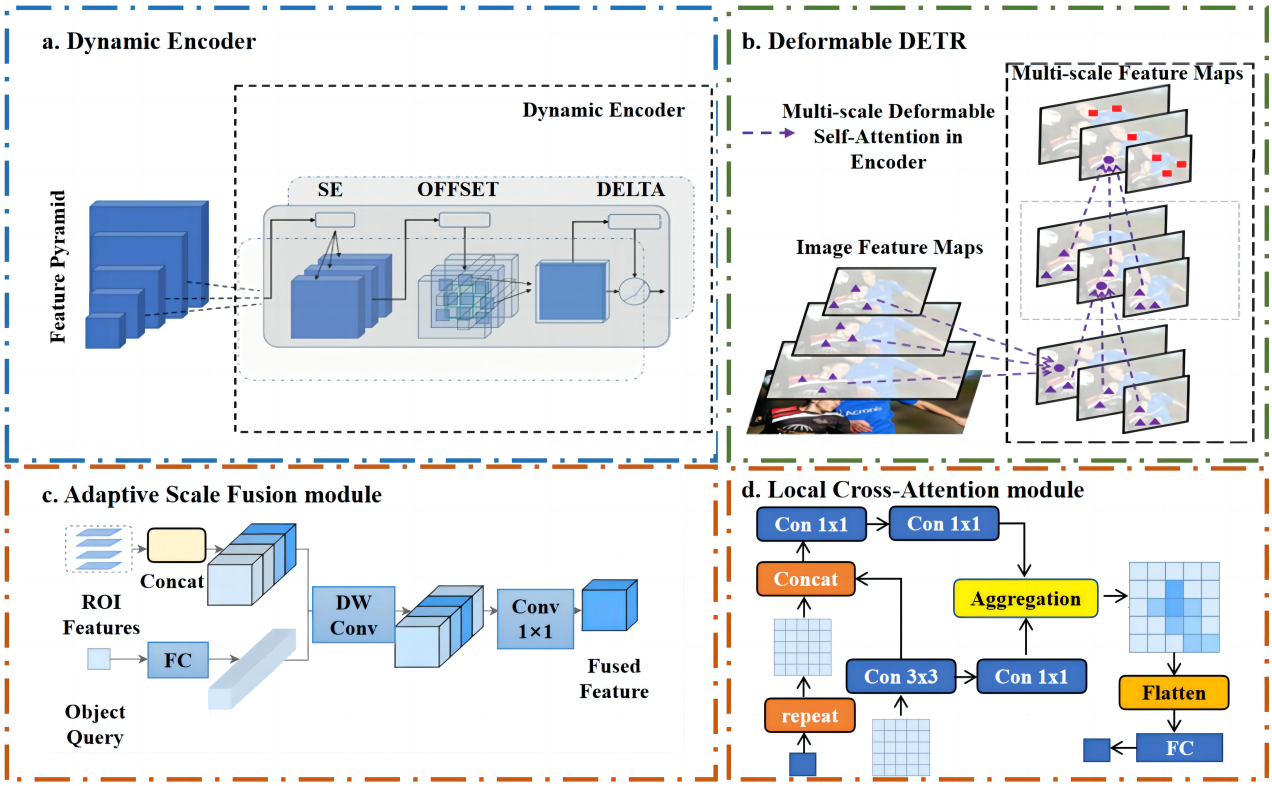


Fig. 4. Three multi-scale feature methods in small object detection. Figure a employs Deformable Convolution to extract multi-level feature maps. Figure b uses a Convolutional Neural Network as the backbone to extract feature maps of various sizes for multi-scale fusion and employs an attention mechanism to achieve a natural integration of multi-scale features.Figure c and Figure d represent the ASF module and the LCA module in CF-DETR, respectively.

map that weights each position in the input feature map. The advantage of spatial attention lies in its ability to help the model deeply understand local features of the image, which is especially important for small object detection, as small objects may occupy only a tiny portion of the image. Conversely, the channel attention mechanism focuses on the depth dimension of the feature map, i.e., different feature channels. It identifies and enhances the most important feature channels for the current task while suppressing less important ones, allowing the model to concentrate more on critical features. Channel attention is particularly advantageous in dealing with complex backgrounds or subtle differences between target categories. In small object detection, combining these two attention mechanisms allows the simultaneous utilization of spatial and channel information, thereby improving the model's understanding of image content and the accuracy of object detection. As shown in Fig 5, there are two typical algorithms.

ViT and DETR, as the first models to successfully apply attention mechanisms in image classification and object detection tasks, marked a significant milestone for this mechanism in the visual domain. ViT's groundbreaking success revealed the immense potential of attention mechanisms in computer vision and spurred a series of studies exploring how to further optimize and extend Transformers in image processing applications. DETR expanded the application scope of the Transformer architecture in the visual domain. DETR redefined the object detection problem with its innovative set prediction method, using a set-based global loss function and bipartite matching strategy, combined with the Transformer encoder-decoder architecture to directly output predictions in parallel. DETR's design not only improved the efficiency and accuracy of object detection but also opened new possibilities for future research and practice.

To accelerate the convergence of the DETR model, researchers have proposed several improvements. Deformable DETR introduced multi-scale deformable attention modules, which focus on processing critical sampling points in the surrounding area, significantly reducing the number of epochs needed for training and surpassing DETR's performance in small object detection. In contrast, Conditional DETR [49] proposed a conditional spatial query mechanism that enhances localization accuracy and speeds up training through conditional cross-attention. This spatial query carries category and bounding box prediction information extracted from the previous decoder layer, focusing on small areas within the endpoints and target boxes, effectively narrowing the content query range, reducing training difficulty, and accelerating model convergence. Anchor DETR [50] introduced a row-column decoupled attention mechanism, decomposing 2D key features into rows and columns and applying row and column attention respectively, which reduced memory consumption and achieved performance comparable to or better than the standard attention in DETR. These methods optimized the DETR model in various aspects, accelerating convergence speed, and making significant progress in small

object detection and memory consumption. Future research can further explore the combination and optimization of these mechanisms to improve model performance and efficiency in a wider range of applications. Despite the introduction of the Transformer architecture, the challenge of small object information loss remains, prompting researchers to propose a series of innovative methods. DeoT [51] introduced the E-OTM module, which achieves global feature representation through deformable multi-head self-attention (DMHSA) and is equipped with TBRS, utilizing channel refinement modules (CRM) and spatial refinement modules (SRM) for channel and spatial attention to further refine the output features of the Transformer module. FEA-Swin [52] pioneered the foreground-enhanced attention block (FEAB) in the Swin-tiny architecture, strengthening the learning of contextual information and feature distinguishability. Combined with the weighted bidirectional feature pyramid network (BiFPN) and carefully designed skip connections, this method effectively preserved detailed information of small objects. Meanwhile, the BiFPN network optimized the balance between detection accuracy and efficiency by eliminating redundant levels. DKTNet [53] proposed the Dual-Key Transformer Network (DKTNet), enhancing feature attention through a dual-K enhancement mechanism and replacing traditional spatial-level attention with a channel-level self-attention mechanism to highlight important feature channels and suppress confusing ones. LPSW [54] integrated the advantages of Transformer and CNN by designing the locally perceptive Swin Transformer [55] backbone network and proposed the Spatial Attention Interlaced Execution Cascade (SAIEC) network framework. Through multi-task learning and an improved spatial attention module, this framework enhanced the network's mask prediction capability and constructed a novel network model using LPSW as the backbone.

These innovative methods provide important insights and practical foundations for addressing the challenge of small object information loss. By introducing global feature representation, foreground-enhanced attention blocks, and dual-key Transformer networks, researchers have made significant progress in the field of small object detection. Future research can further explore the potential and optimization space of these methods to meet the ever-changing demands of practical applications, promoting the development and innovation of small object detection technology.

C. Utilization of Contextual Information

In the process of detecting small objects, various challenges such as specific scene object detection and occluded targets are often encountered. Due to the limited information contained within small objects, contextual information plays a crucial role in their detection. The application of contextual information ranges from global image-level statistics to local image-level neighboring area information.As shown in Fig 6, there are three typical algorithms.

To enhance the accuracy of small object detection, researchers have explored various strategies for integrating local and global contextual information. CF-DETR [36] utilizes its

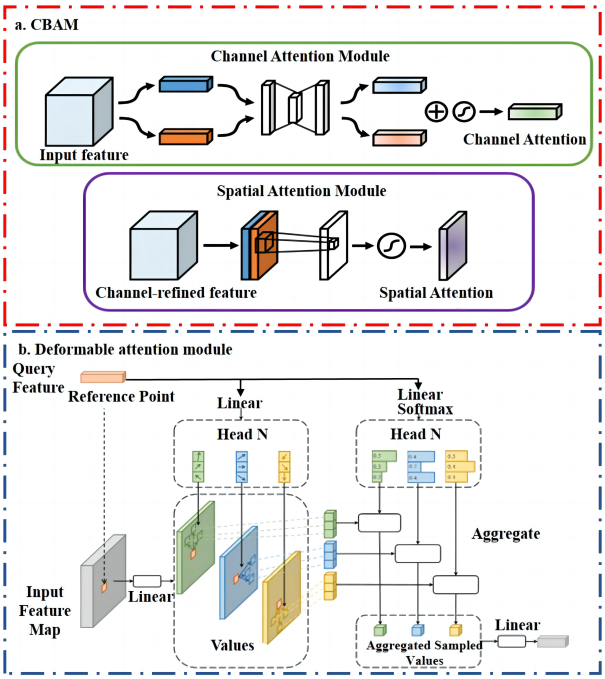


Fig. 5. Two typical models of attention mechanism application.Figure a combines spatial attention and channel attention to enhance detection accuracy.Figure b adopts deformable attention, allowing the model to dynamically select and focus on the positions with the most informative content.

innovative CF decoder layer to combine local and global information, progressively refining features and significantly improving detection accuracy. The Foreground Enhanced Attention Swin Transformer (FEA-Swin) framework and the improved Weighted Bidirectional Feature Pyramid Network (BiFPN) effectively preserve small object details through skip connections. Additionally, researchers have utilized multi-scale feature cascades and attention mechanisms to focus on targets within the image [56]. IENet [57] combines object appearance and contextual information for robust detection, while the Pyramid Context Learning module (PCL) deeply mines contextual information from different feature levels [58]. These methods optimize feature extraction and processing through multi-scale, multi-level feature integration and attention mechanisms, effectively enhancing the accuracy and robustness of small object detection. This results in more precise detection and identification of small objects, providing new directions and insights for the development of the field.

Considering the limitations of computational resources, researchers are also seeking to reduce over-reliance on context detection techniques in small object detection. Given the diversity of small objects and the complexity of their background information, not all scenarios require contextual techniques. For example, O2DETR [37] posits that global feature interaction is unnecessary for small objects and instead uses depthwise separable convolutions for local interaction to reduce computational load, employing contextual techniques only when there is confusion between the target and background. DyCC-Net [59] dynamically adjusts the network structure to adapt to inputs of varying complexity, achieving efficient input-aware inference, significantly improving

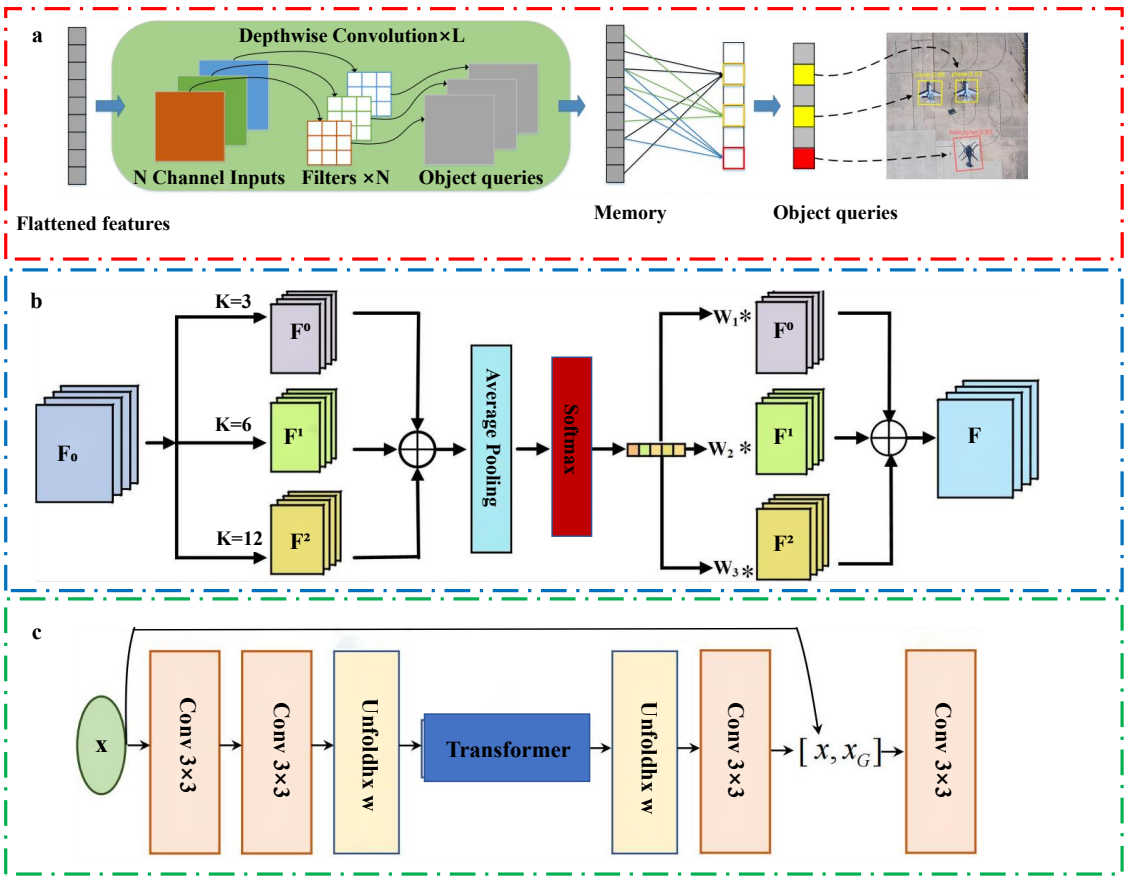


Fig. 6. Three typical methods of context feature processing.[a] O2DETR.[b] MATR.[c] HTDet.

inference efficiency, and reducing computational complexity. MATR [60] introduced a Selective Context Module (SCM), which features a dynamic selection mechanism to enhance high-resolution features with spatial details, distinguishing them more clearly from noisy backgrounds. DKTNet [53] suggests that compared to the fully connected computation of traditional Transformer architectures, 2D convolutions can more effectively capture local details and global contextual information, while 1D convolutions can significantly reduce network complexity, providing new directions for optimization. HTDet [61] proposed a new framework based on hybrid transformers, first utilizing a lightweight hybrid transformer network to extract global contextual information, followed by a fine-grained feature pyramid network to address the issue of weak signal disappearance, offering a new solution for small object detection.

These methods collectively emphasize the importance of flexibly adjusting feature extraction strategies based on the requirements of the detection task, to reduce computational complexity and enhance detection accuracy. They provide new directions and insights for the development of small object detection technology, demonstrating how to effectively utilize contextual information in resource-limited scenarios. Future research can build on these foundations to address challenges in practical applications.

D. Improvement of Training Strategies

In the field of deep learning, improving training strategies by adopting diverse methods and techniques optimizes the model training process, thereby enhancing model performance and generalization ability. These strategies encompass auxiliary decoding/encoding techniques, data augmentation, pre-training, and denoising training. This section will detail these strategies and provide a systematic algorithmic classification.As shown in Fig 7, a portion of the algorithms is illustrated.

Auxiliary decoding/encoding is a feedforward network designed for bounding box regression and object classification, enhancing performance by connecting to independent decoding layers. This method trains the model by combining individual losses at different scales, achieving more precise detection results. B. Singh proposed SNIPER [62], an efficient multi-scale training algorithm for instance-level visual recognition tasks that accelerates the training process by sampling low-resolution regions from a multi-scale image pyramid. To address the scale normalization issue, a Scale-Aware Network (SAN) [63] maps convolutional features of different scales to a scale-invariant subspace, constructing a unique learning method that considers only inter-channel relationships without spatial information for efficient network learning. N. Bodla's Soft-NMS algorithm [64] avoids object elimination by decaying detection scores of other objects as a continuous function of their overlap with the highest-scoring detection box, applying non-maximum suppression in post-processing. Due to

the network's general lack of robustness to scale variations, B. Singh also proposed a detector that trains and tests at the same scale on the image pyramid [65], using a scale-normalized training scheme that only trains on objects within the desired scale range. To better extract features at various scales, VIDT [66] combined multi-layer deformable Transformer decoders, generating new [DET] tokens through multi-scale deformable attention to aggregate critical content from multi-scale feature maps, improving detection performance. TTRD [47] aims to aggregate features of multi-scale global spatial positions and simulate interactions between paired instances through a modified Transformer. To reduce computational resource consumption, MDef-DETR [67] applied multi-scale hierarchical modeling with visual priors in a pure Transformer architecture. As the model deepens, feature resolution decreases while the number of channels increases, effectively reducing Transformer computational resources. CBNet [46] proposed a feature fusion technique where the output features of the previous backbone network (i.e., high-level features) are used as part of the input features for subsequent backbone networks, progressively expanding the receptive field. This iterative feature transmission method helps capture and utilize multi-scale information at different stages for more effective small object detection.

Data augmentation is a technique to expand the dataset through various methods without substantially increasing data, which becomes an effective solution given the scarcity of small object datasets. This technique enriches the detection dataset by applying multiple augmentation techniques such as rotation, flipping, scaling, cropping, translation, and adding noise. Supervised data augmentation methods include SMOTE [68], SamplePairing [69], and mixup [68], while unsupervised methods include GAN [8] and AutoAugment [70]. Sw-YoloX [71] augments training data by randomly scaling, cropping, and arranging four images into one, then generating new training samples by mixing two images, enhancing the model's generalization ability. NLFFTNet [56] introduced a configurable hybrid splicing dynamic data augmentation method to address data imbalance between different categories. To avoid overfitting, T-TRD combined data augmentation with Transformers to improve the detection performance of remote sensing images.

Pre-trained models are an application of transfer learning that learns context-aware representations of each member from large datasets. Pre-trained models can transfer knowledge learned from open domains to downstream tasks to improve the performance of low-resource tasks. Due to significant differences between source datasets (e.g., ImageNet) and target datasets (i.e., remote sensing image datasets), TRD proposed an attention mechanism-based transfer CNN (T-TRD) to adapt pre-trained models for better remote sensing image object detection. FP-DETR [72] explored how to fully pre-train an encoder-only transformer and smoothly fine-tune it for object detection through task adapters. Building on this, Group DETR v2 [73] further improved performance through encoder-decoder pre-training and fine-tuning.

Denoising training introduces noise during the training process and trains the model to reconstruct or predict the original data without noise. This method can improve the model's generalization ability by forcing the model to learn to extract and recover useful information from noisy data. In object detection, particularly in training Detection Transformer (DETR) models, denoising training addresses the issue of slow model convergence. DINO [58] improved the performance and efficiency of previous DETR models by using contrastive denoising training, a mixed query selection method for anchor initialization, and a two-stage look-ahead scheme for box prediction. DN-DETR [74] proposed a novel denoising training method to accelerate DETR training. In addition to the Hungarian loss, their method inputs noisy ground truth bounding boxes into the Transformer decoder and trains the model to reconstruct the original boxes, effectively reducing the difficulty of bipartite matching and speeding up convergence.

These innovative training strategy improvements bring new possibilities to the field of small object detection, enabling models to learn more effectively from complex data and enhance their recognition accuracy. Future research is expected to further optimize algorithms using these strategies, reduce computational costs, and improve application capabilities in various real-world environments, especially in resource-constrained scenarios. This will drive advancements in small object detection technology in critical areas such as autonomous driving, remote sensing image analysis, and medical imaging diagnosis, providing powerful technical support for solving complex real-world problems.

E. Exploitation of Spatio-Temporal Information

The utilization of spatiotemporal information refers to leveraging the spatial and temporal relationships of targets across consecutive frames or scenes to enhance detection accuracy. This strategy is particularly applicable to video or dynamic environments. Early methods for small object detection in videos, such as R-CNN, relied on Region Proposal Networks (RPN) to generate candidate regions and combined them with deep learning feature extraction to improve classification accuracy. However, these methods were slow and involved complex post-processing. Improvements like Fast R-CNN and Faster R-CNN optimized the algorithm structure to increase detection speed. Models like SSD and YOLO accelerated detection by using single-shot detection on multi-scale feature maps and framing detection as a regression problem, though at the cost of some accuracy. The TPN [75] model enhanced video object detection performance by integrating target feature representations through video pipeline proposals, but this increased computational resource demands.

Traditional Video Object Detection (VOD) methods depended on multi-stage processes and manually designed feature aggregation components, such as optical flow models and relational networks, leading to inefficiencies. TransVOD [76] introduced a spatiotemporal Transformer architecture, simplifying the VOD process, reducing model complexity and computational costs, and improving detection accuracy and efficiency by establishing long-term dependencies across video frames. Aerial Trans [77] addressed issues of occlusion, loss,

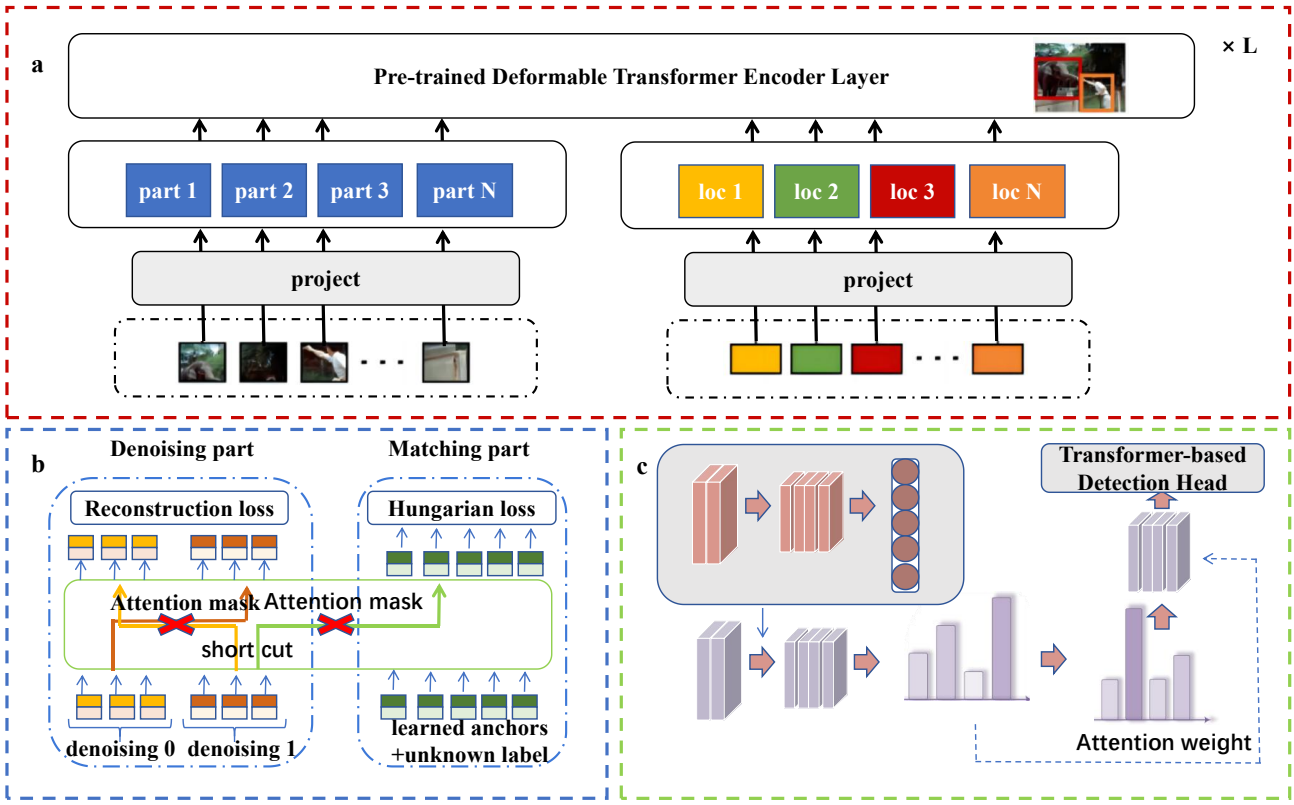


Fig. 7. Three typical methods of improving training strategies.[a] FP-DETR.[b] DN-DETR.[c] T-TRD.

and drift in small object tracking by proposing two trackers that utilize Transformer attention mechanisms to supplement small object contextual information, enhancing tracking performance.As shown in fig 8, SparseVOD [78] proposed an innovative video object detection process that uses Sparse R-CNN to generate detection objects by incorporating temporal information, alleviating the burden of post-processing and enabling end-to-end training. PTSEFormer [79] introduced a new detection method that incrementally incorporates temporal and spatial information to enhance features rather than aggregating them all at once, more effectively utilizing contextual information from adjacent frames. This method uses the Transformer-based detector DETR, avoiding heavy post-processing. VSTAMcitefujitake2022video presented a sparse aggregation per-element feature method to reduce processing time and memory costs, enhancing features per element before object candidate region detection to improve accuracy. This approach maintains high accuracy while reducing computational and memory demands. FAQ [80] addressed the issue of feature degradation in video object detection by proposing a method that utilizes temporal information from adjacent frames and merges their features, demonstrating its effectiveness.

These methods leveraging spatiotemporal information bring unprecedented depth and dimension to small object detection, making recognition and tracking in dynamic environments more precise and reliable. Future research may focus on further refining and optimizing these methods for seamless integration and efficient operation in a broader range of application scenarios.

F. Efficiency and Optimization in Computing

In the classification of small object detection algorithms, computational efficiency and optimization constitute a crucial research area. This field focuses on reducing the computational resources required during algorithm execution while maintaining or enhancing detection performance. The primary objective is to develop algorithms capable of processing large volumes of image data quickly and accurately, particularly in resource-constrained environments such as mobile devices or edge computing platforms.

Traditional Transformer models, due to their use of global self-attention, consume significant computational resources when processing large images. Deformable DETR addresses this issue by introducing a deformable attention mechanism that calculates attention only at a sparse set of positions in the image, effectively reducing computational complexity and accelerating model convergence. Swin Transformer adopts a hierarchical feature map that progressively reduces in size with network depth, similar to traditional CNNs, which helps capture features from coarse to fine granularity while improving efficiency. Swin Transformer significantly reduces computation by performing self-attention calculations within local windows.

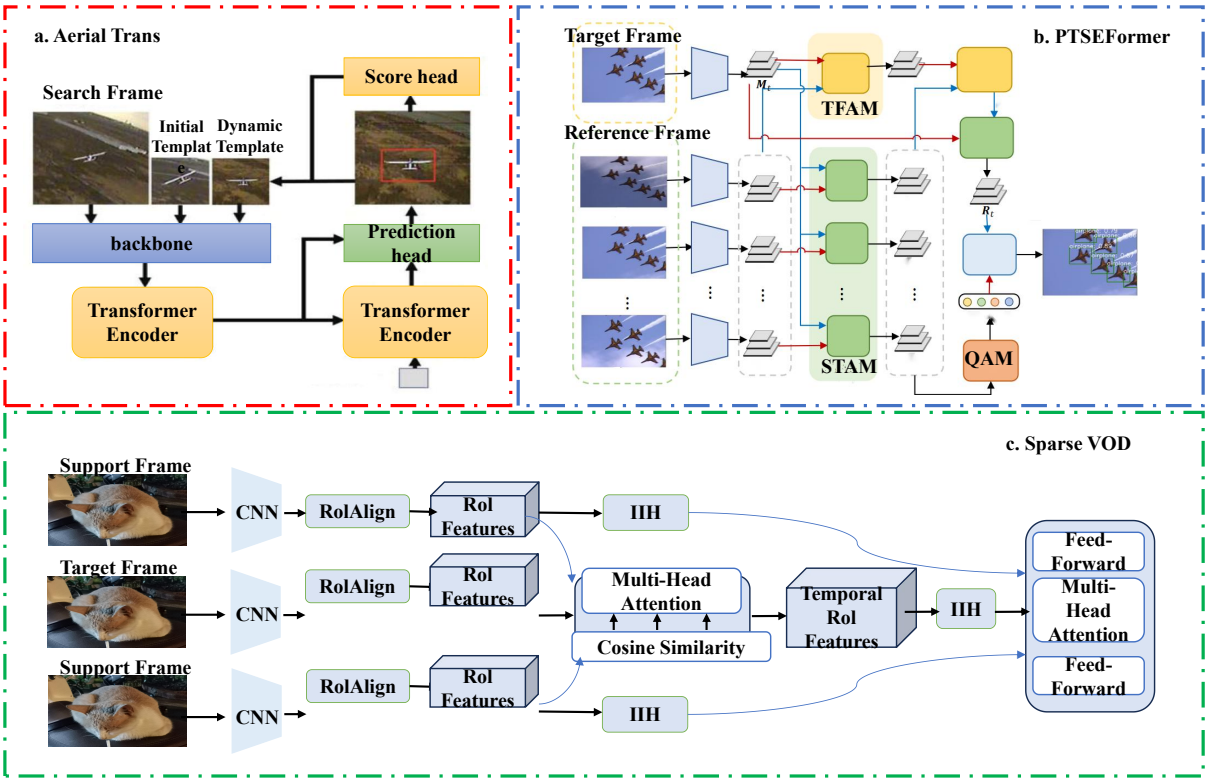


Fig. 8. Three typical methods of utilizing spatiotemporal information.Figure A simplifies the process by introducing a spatial-temporal Transformer architecture. Figure B proposes two types of small target trackers to provide a global response. Figure C presents an innovative video object detection workflow, using Sparse R-CNN combined with temporal information to generate detection objects.

As shown in fig 9,O2DETR [37] suggests that global feature interaction is unnecessary and employs depthwise separable convolutions for local interactions, further reducing computation and speeding up model convergence. Conditional DETR [49] reduces training difficulty by learning conditional spatial queries for decoder multi-head cross-attention, narrowing the spatial range of content queries. DAB-DETR [81] accelerates training convergence by directly using box coordinates as queries for the Transformer decoder and dynamically updating them layer by layer, leveraging positional priors to enhance feature similarity. QueryDet [48] introduces a Cascaded Sparse Query (CSQ) mechanism that efficiently utilizes high-resolution features by first obtaining the rough location of small objects on high-level feature maps and then conducting detailed queries on low-level feature maps, thereby improving small object detection performance while maintaining fast inference speed.

These advanced methods for computational efficiency and optimization in the field of small object detection not only demonstrate technological innovation but also lay a solid foundation for achieving efficient and precise small object detection in various application scenarios. Future research can further explore the potential of these methods, integrating additional optimization techniques and innovative algorithms to meet the demands of different applications. Particularly in resource-constrained environments like mobile devices and edge computing platforms, these efficient small object detection algorithms will play a crucial role.

IV. Datasets

Datasets are the foundation of small object detection and an important basis for evaluating the performance of small object detection methods. Different datasets have distinct characteristics and application scenarios, posing varying requirements and challenges for small object detection methods. Therefore, selecting appropriate datasets is crucial for the research and application of small object detection.As shown in table 2, this section will provide an overview of some publicly available datasets for specific scenarios, introducing their sources, content, scale, and features. It will also compare their difficulty levels and applicability, as well as quantitatively and qualitatively assess and analyze the performance of various small object detection methods on different scenario datasets. This will help determine the most effective small object detection methods for that scenario.

A. General image datasets

General image datasets refer to collections of images that encompass a wide variety of categories and scenes, typically used to assess the universality and robustness of object detection methods. In these datasets, small objects are usually

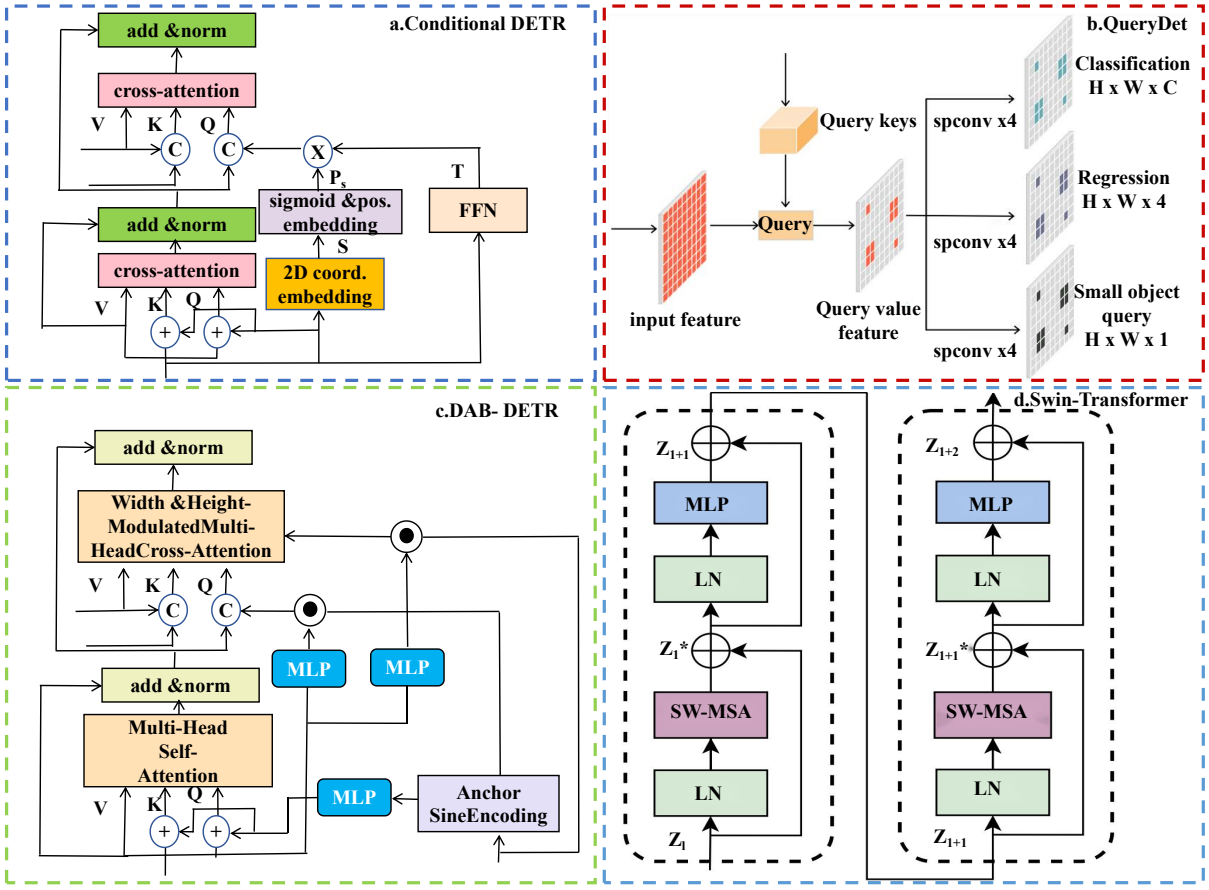


Fig. 9. Four typical methods of optimizing computational efficiency.

TABLE II
SUMMARY OF DATASETS

Dataset	Years	Object Classes	Instances	Country	Application
MS COCO	2015	91	2,500,000	United States	Generic Applications
PASCAL VOC	2012	20	27,450	United Kingdom	
WIDER FACE	2016	61	393,703	China	
SKU-110K	2019	110,712	147.4 per image	The State of Israel	
DAIR-V2X	2021	10	-	China	Self-Driving
Argoverse	2020	15	11,052	United States	
KAIST Multispectral Pedestrian	2015	4	-	Korea	
CQCAR-20	2024	20	11,000	China	
Cityscapes	2016	30	-	Germany	
DOTA	2017	15	188,282	China	Aerial Images
HRSC2016	2016	27	2,976	China	
UCAS-AOD	2014	2	18,895	China	
UAV123	2016	-	-	Saudi Arabia	
AMMW	2024	30	-	China	Active Milli-Meter Wave Images
MM-Fi	2023	40	-	Singapore	
nuScenes	2019	23	-	Singapore	
RadarScenes	2016	5	-	Germany	

defined as targets that occupy a small proportion or resolution in the image, such as pedestrians, vehicles, animals, etc.As shown in fig 10, the following are commonly used general image datasets:

MS COCO: Microsoft Common Objects in Context (MS COCO) is a large-scale dataset built by Microsoft for object detection, segmentation, keypoint detection, and image captioning. It contains 91 categories, 328,000 images, and 2,500,000 object instances. Small objects in this dataset are defined as targets with bounding box sizes smaller than 32×32 pixels, accounting for 23.5% of all objects.

PASCAL VOC: PASCAL Visual Object Classes (PASCAL VOC) is a dataset created by the computer vision group at the Technical University of Munich, Germany, for object detection, segmentation, classification, and action recognition. It includes 20 categories, 11,540 images, and 27,450 object instances. Small objects in this dataset are generally defined as targets where the width and height of the bounding box are less than 10% of the image's width and height, accounting for 14.8% of all objects.

WIDER FACE [82]: WIDER FACE is a dataset for face detection released by researchers at the Hong Kong University of Science and Technology in 2018. It includes 61 categories, 32,203 images, and 393,703 face instances. Small objects in this dataset are generally defined as faces with bounding box areas smaller than 1024 pixels, accounting for 29.7% of all faces.

SKU-110K [83]: The SKU-110K dataset is a dense collection of retail shelf images intended to support research in object detection tasks. Developed by Eran Goldman et al., it contains over 110,000 unique Stock Keeping Unit (SKU) categories. The objects in these categories are densely arranged, often appearing similar or even identical, and are positioned very close to each other. The SKU-110K dataset includes 11,762 images taken in dense scenes, with over 1.7 million bounding box annotations, totaling approximately 1,733,678 instances. These images are collected from thousands of supermarkets, featuring varying scales, perspectives, lighting conditions, and noise levels. All images are adjusted to a resolution of one million pixels.

The advantage of these datasets lies in the diversity of categories and scenes, which helps demonstrate the universality and robustness of small object detection. However, they also have disadvantages: the number and proportion of small objects are low, leading to sample imbalance and difficulty in fully training and evaluating small object detection methods. Additionally, since the definition of small objects varies across different datasets, cross-dataset performance comparison is challenging.

Considering these characteristics and limitations of general image datasets, there are several key aspects to focus on for future research and applications. Firstly, developing datasets containing more small object samples or using synthetic augmentation techniques to balance datasets will be an important research direction to address the issue of low quantity and proportion of small objects. Secondly, unifying or standardizing the definition of small objects will help achieve fairer and more consistent performance evaluations across different datasets. Lastly, as general image datasets cover a rich variety of categories and scenes, they provide researchers with valuable opportunities to explore and test the generalization capabilities of object detection methods and their adaptability to different environments. In summary, general image datasets play a crucial role in the research and application of small object detection, and in-depth study and optimization of these datasets will have a profound impact on advancing the technology in this field.

B. Datasets for Intelligent Vehicles

Datasets for intelligent vehicles are large-scale collections specifically designed for intelligent vehicle technologies, such as autonomous driving and vehicle perception systems. These datasets typically contain a vast amount of images, videos, sensor data, and other types of information collected from real-world scenarios, used to train and test intelligent vehicle algorithms. As shown in fig 11,the following are commonly used datasets in the field of intelligent vehicles:

DAIR-V2X [84]: DAIR-V2X (Vehicle-to-Everything Autonomous Driving Dataset) is released jointly by Tsinghua University's Institute for AI Industry Research (AIR), Beijing High-level Autonomous Driving Demonstration Area, Beijing CarNet Technology Development Co., Baidu Apollo, and Beijing Academy of Artificial Intelligence. It is the world's first large-scale, multimodal, multi-view dataset for vehicle-to-everything autonomous driving research. It contains a series of data based on real autonomous driving scenarios, aimed at promoting smarter decision-making and enhancing the safety of autonomous driving. The dataset allows the simultaneous use of infrastructure and vehicle-end information to track and predict the behavior of surrounding traffic participants. Specifically, the DAIR-V2X dataset includes the following parts: DAIR-V2X-C (Collaborative Dataset): Contains 38,845 frames of image data and 38,845 frames of point cloud data for studying vehicle-to-everything autonomous driving. These data are collected from real scenarios and include both 2D and 3D annotations. DAIR-V2X-I (Infrastructure Dataset): Contains 10,084 frames of image data and 10,084 frames of point cloud data, focusing more on roadside monocular 3D object detection tasks. DAIR-V2X-V (Vehicle-end Dataset): Contains 22,325 frames of image data and 22,325 frames of point cloud data for studying vehicle-end perception and prediction. These datasets achieve vehicle-to-everything spatio-temporal synchronized annotation for the first time, with a rich variety of sensor types, including vehicle-end cameras, vehicle-end LiDAR, roadside cameras, and roadside LiDAR. The 3D annotation attributes of obstacle targets are comprehensive, annotating 10 types of common road obstacle targets.

Argoverse [85]: The Argoverse dataset is released by Argo AI, Carnegie Mellon University, and Georgia Institute of Technology, containing two parts: 3D Tracking and Motion Forecasting. Argoverse is somewhat different from Waymo; although it also includes LiDAR and camera data, it only covers 113 scenes recorded in Miami and Pittsburgh. Its uniqueness lies in being the first dataset to include high-definition map data, mainly featuring 290 kilometers of lane

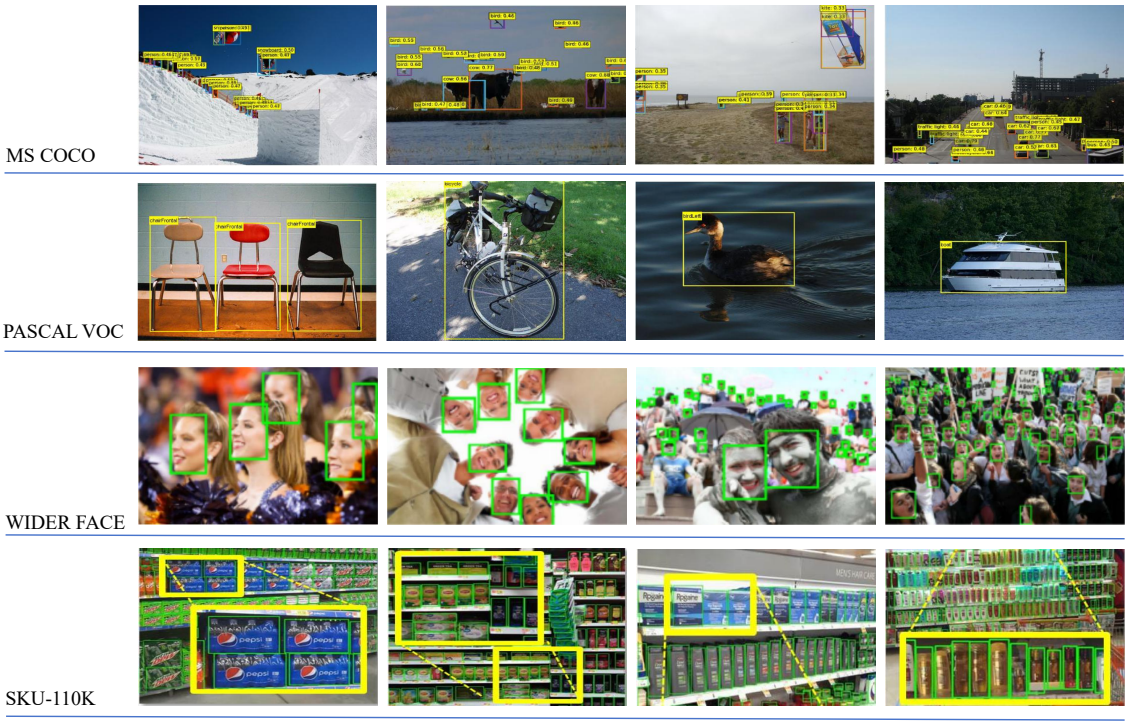


Fig. 10. Examples of detection results on Generic Applications Dataset.

maps in Pittsburgh and Miami, with information such as location, connections, traffic signals, elevation, etc. Its sensors consist of 2 LiDARs, 7 high-resolution ring cameras (1920 × 1200), and 2 stereo cameras (2056 × 2464). In Argoverse 3D tracking, it includes 3D tracking annotations for 113 scenes, each segment lasting 15-30 seconds, totaling 11,052 tracked objects, annotating objects within 5 meters, with 15 labels, 70% of which are vehicles, and 30% pedestrians, bicycles, motorcycles, etc. In Argoverse Motion Forecasting, it includes 324,557 scenes, each lasting 5 seconds, and contains a 2D bird's-eye view of each tracked object sampled at 10 Hz.

KAIST Multispectral Pedestrian [86]: The KAIST Multispectral Pedestrian dataset is released by the Korea Advanced Institute of Science and Technology, providing a multispectral pedestrian detection dataset with color-thermal image pairs during day and night. The dataset improves the accuracy of pedestrian detection by complementing the advantages of color images and thermal imaging, overcoming previous issues such as pedestrian occlusion, cluttered backgrounds, and unclear nighttime imaging. It provides 95,328 pairs of color-thermal images during day and night, aligning images through a beam splitter to eliminate image parallax. The data is collected in Seoul, South Korea, with an image resolution of 640x480, 103,128 manual 2D box annotations, 1,182 pedestrians, and four different types of annotations: person, people (unclear human figures), cyclist, person? (uncertain if a pedestrian).

Cityscapes [87]: Cityscapes is released by the Mercedes-Benz Research and Development Center and is recognized as one of the most authoritative and professional semantic segmentation evaluation sets in the field of autonomous driving. The dataset collects urban scenes from 50 cities in Germany and neighboring countries across three seasons: spring, summer, and autumn, and captures stereoscopic vision video sequences using binocular cameras. After conversion, this dataset can be used for object detection tasks, providing 2,975 and 500 data for model training and validation, respectively. It includes categories such as people, riders, cars, trucks, buses, trains, motorcycles, and bicycles, totaling 30.

CQCAR-20:As shown in fig 12, this dataset is meticulously produced by the Safety-AI lab team at Chongqing University, with materials sourced from road traffic scenes in Chongqing City. The dataset is specifically designed for traffic road object detection in the intelligent vehicle field, covering nearly 20 categories, including but not limited to traffic lights, traffic signs, pedestrians, and bicycles, which we consider to fall within the category of small objects. In terms of quantity, the dataset contains about 3,000 training set images, 500 validation set images, and 500 test set images, with approximately 11,000 target boxes annotated in total. By providing such a specialized dataset, the Safety-AI lab at Chongqing University aims to support researchers in the field of intelligent vehicles, especially those teams dedicated to training and detection in small object scenarios, thereby enhancing the perception and response capabilities of intelligent vehicle systems in real road environments.

C. Aerial Image Datasets

Aerial image datasets, collections of images captured from a high-altitude perspective, are commonly used to evaluate object detection methods in the field of remote sensing. In these datasets, small objects refer to entities that appear as tiny in size or low in resolution in the images, such as vehicles,

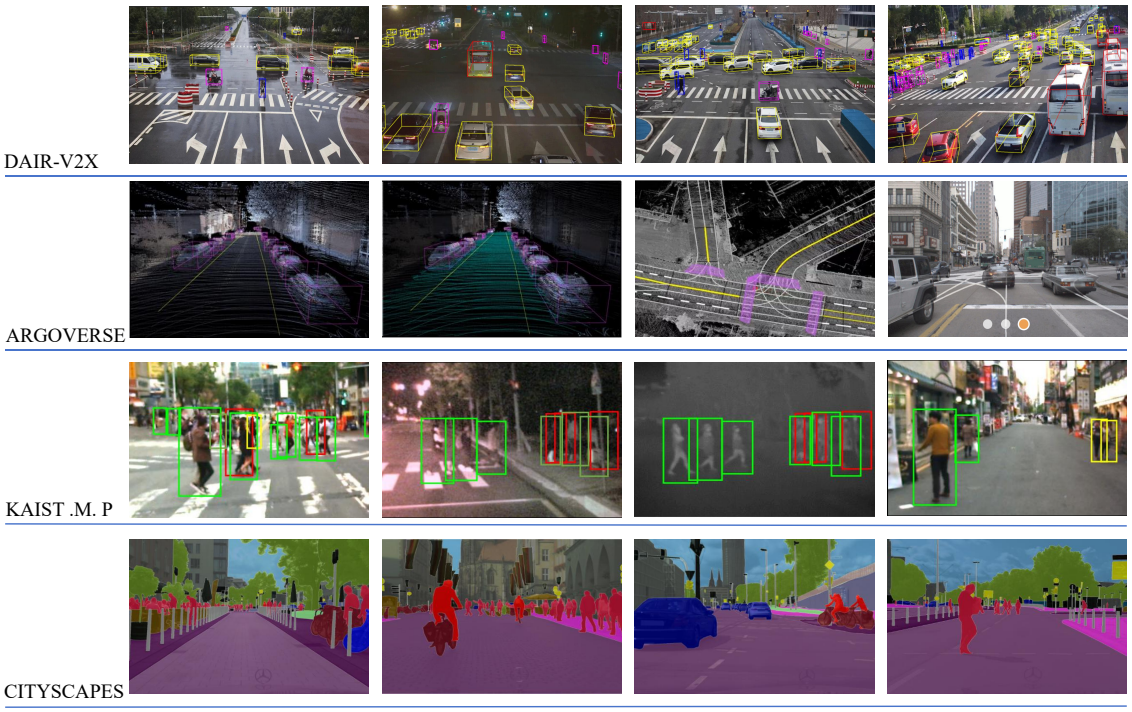


Fig. 11. Examples of detection results on intelligent Vehicle Domain Datasets.

ships, and airplanes. As shown in fig 13,there are some of the main aerial image datasets:

DOTA [88]: DOTA (A Large-scale Dataset for Object Detection in Aerial Images) is a large-scale dataset for object detection in aerial images, containing 15 categories, 2,806 images, and 188,282 object instances. Small objects in this dataset are defined as targets with both width and height less than 30 pixels, accounting for 13.7% of all objects. The evaluation metric for this dataset is mean Average Precision (mAP), calculated at an IoU threshold of 0.5.

HRSC2016 [89]: HRSC2016 (High Resolution Satellite Image Ship Detection Dataset) is a dataset for ship detection in aerial images, containing 27 categories, 1,061 images, and 2,976 ship instances. Small objects in this dataset are defined as ships with both width and height less than 20 pixels, accounting for 10.8% of all ships. The evaluation metric for this dataset is mAP, calculated at an IoU threshold of 0.5. UCAS-AOD [90]: UCAS-AOD (University of Chinese Academy of Sciences Aerial Object Detection) is a dataset for object detection in aerial images, containing 2 categories, 910 images, and 18,895 object instances. Small objects in this dataset are defined as targets with both width and height less than 16 pixels, accounting for 11.2% of all objects. The evaluation metric for this dataset is mAP, calculated at an IoU threshold of 0.5.

UAV123 [91]: This dataset contains 123 videos captured by drones, one of the largest object tracking datasets with over 110K frames. These videos are shot by drones at low altitudes, and each video comes with corresponding ground truth annotations.

The notable advantages of these aerial image datasets are that they provide high-resolution images and a rich variety of target types, effectively demonstrating the practical application potential of small object detection in the field of remote sensing. However, they also have clear limitations: the singularity of categories and simplicity of scenes limit the assessment of the generalization ability and robustness of small object detection methods. Additionally, due to the lack of uniformity in the definition of small objects, performance comparisons between these datasets are also constrained. In aerial images, objects appear particularly small due to their relative distance from the camera, and the bird's-eye view perspective means that objects can appear anywhere in the image, bringing additional challenges to the object detection task.

When evaluating the performance of different algorithms in such application scenarios, we have chosen the DOTAcitexia2018dota dataset as the testing benchmark because it has become a widely applied standard in the field of object detection. Among CNN-based methods, ReDet [92] has shown excellent performance, especially in terms of rotational invariance. This method uses a rotation-equivariant network and a rotation-invariant RoI alignment mechanism to extract features and accurately predict the orientation and position of targets. With data augmentation techniques such as multi-scale training, testing, random rotation cropping, and random rotation flipping, ReDet has improved the model's robustness in detecting small objects and multi-directional targets. In a hybrid framework, ReDet achieves the highest accuracy value (80.89%) with only 12 training epochs, thanks to its multi-scale learning strategy and pre-training on the ImageNet dataset.

Fig. 12. Examples of detection results on CQCAR-20 dataset for YOLOv5 based on CNN and DETR based on Transformer.

D. Active Millimeter Wave Datasets

Active Millimeter Wave (AMMW) datasets,as shown in fig 14,typically composed of data collected using active millimeter-wave radar sensors, are crucial resources for evaluating object detection methods in scenarios that require privacy protection, adaptation to harsh weather, and low-light conditions. In these datasets, small objects generally refer to those that produce weak reflections in radar signals or have a low signal-to-noise ratio, such as human bodies, metal items, and animals. Here are some of the primary active millimeter-wave datasets:

AMMW [93]: AMMW (Active Millimeter Wave) is a dataset for active millimeter-wave imaging, containing two subsets: AMMW-A and AMMW-B. AMMW-A includes 58k single-channel millimeter-wave images taken from different angles, each with a human subject possibly carrying one or more concealed items. AMMW-B consists of 2k three-dimensional millimeter-wave images taken from the front and back, each with a human subject possibly carrying one or more concealed items. Small objects in this dataset are defined as targets with both width and height less than 20 pixels, accounting for 12.4% of all objects. The evaluation metrics for this dataset are Detection Rate (DR) and False Alarm Rate (FAR), calculated at different confidence thresholds.

MM-Fi [94]: MM-Fi (Multi-Modal Non-Intrusive 4D Human Dataset for Versatile Sensing) is a dataset for multi-modal non-intrusive human perception, containing 320,000 synchronized frames of 40 human subjects across 5 modalities (active millimeter-wave radar, RGB camera, depth camera, thermal imager, and IMU). Small objects in this dataset are defined as targets with both width and height less than 10 pixels, accounting for 9.6% of all objects. The evaluation metric for this dataset is mean Average Precision (mAP), calculated at different IoU thresholds (0.5, 0.75, 0.5:0.95).

nuScenes [95]: nuScenes (A multimodal dataset for autonomous driving) is a multi-modal dataset for autonomous driving, containing 1.4 million images (from 6 cameras), 390,000 frames of LiDAR data (from 1 LiDAR), 1.4 million frames of millimeter-wave radar data (from 5 radars), and 1.4 million object annotations in 40,000 keyframes collected from 1,000 driving scenes in Boston and Singapore. Small objects in this dataset are defined as targets with both width and height less than 32 pixels, accounting for 15.2% of all objects. The evaluation metric for this dataset is mAP, calculated at different scales (small, medium, large) and IoU thresholds (0.5, 0.75, 0.5:0.95).

RadarScenes [96]: RadarScenes (A Real-World Radar Point Cloud Data Set for Automotive Applications) is a real-world radar point cloud dataset for automotive applications, containing data from 10 scenes collected by 4 77GHz autonomous driving radar sensors and 1 documentary camera, including various road conditions in urban, rural, and highway settings. Small objects in this dataset are defined as targets with both width and height less than 20 pixels, accounting for 12.4% of all objects. The evaluation metric for this dataset is mAP, calculated at an IoU threshold of 0.5.

The AMMW datasets, with their multi-modal data and complex scenarios, fully demonstrate the application potential and challenges faced by active millimeter-wave radar in various perception tasks. However, these datasets are limited by their relatively small data volume and limited annotation information, insufficient to support the full training and evaluation of deep learning algorithms. Additionally, due to the lack of uniformity in the definition of small objects, performance comparisons between different datasets are challenging. In our research, we chose the AMMW dataset for experimentation, as it is currently the largest-scale active millimeter-wave imaging dataset, offering a wealth of data and diversity conducive to testing the generalization ability and robustness of small object detection algorithms.

Our research focuses on two types of methods: one based on CNN algorithms and the other on hybrid architectures

Fig. 13. Examples of detection results on Aerial Image Dataset.

combining CNN and Transformer. Among CNN methods, Yolov5 stands out, combining techniques such as anchor box adaptation, multi-scale prediction, label smoothing, and multi-modal fusion to effectively adapt to the detection needs of small and irregular objects and handle the noise and blurry millimeter-wave images in the AMMW dataset. At the same time, Yolov5 leverages the advantages of multi-view millimeter-wave images to improve detection accuracy and efficiency. In hybrid architectures, the MATR algorithm performs excellently, enhancing the perception and localization accuracy of small objects through multi-head self-attention mechanisms and multi-scale feature fusion. MATR's multi-head self-attention mechanism captures global dependencies between targets of different scales and orientations, thus enhancing the detection capability for irregular targets in the AMMW dataset. Additionally, MATR employs anchor box adaptation and multi-modal fusion techniques to adapt to the characteristics of small and irregular targets in the AMMW dataset, further improving detection accuracy and efficiency.

E. Summary and Analysis of Datasets

In the analysis of current datasets for small object detection, several key characteristics can be identified. Firstly, there is a disparity in the definition of small objects; some are based on the size of the bounding box, others on the proportion of the object to the image, and still others on indicators such as the object's reflectivity or signal-to-noise ratio. This diversity in definitions makes it challenging to directly compare and evaluate small objects across different datasets. Secondly, the number of samples of small objects is generally low and their proportion within datasets is small, leading to a pronounced issue of sample imbalance. This poses a challenge for the effective training and evaluation of small object detection methods. Therefore, there is a need to develop detection methods that can effectively extract features from limited data and enhance the ability to distinguish small objects. Furthermore, the categories and scenarios involving small objects are very diverse, including but not limited to general images, road scenes, aerial images, active millimeter-wave images, and videos. This requires small object detection technology to have good universality and robustness, capable of adapting to a variety of data types and environmental conditions. Lastly, the difficulty of detecting small objects encompasses various factors, such as small size, low resolution, rapid movement, occlusion, and low signal-to-noise ratio. This further demands that detection technologies effectively address these challenges to improve detection accuracy and stability.

In summary, datasets for small object detection are not only important resources for research but also an area in need of further refinement. Future research should focus on establishing a unified standard for the definition of small objects, building larger datasets for small object detection, and providing a fair evaluation system to promote the development and optimization of small object detection technology.

V. Discussion

Small object detection, as one of the key areas of computer vision research, has attracted considerable attention in recent years. Section 2 of this paper systematically outlines the trends and transformations in technology development within this field in recent years. Subsequently, Section 3 delves into various strategies for small object detection, providing

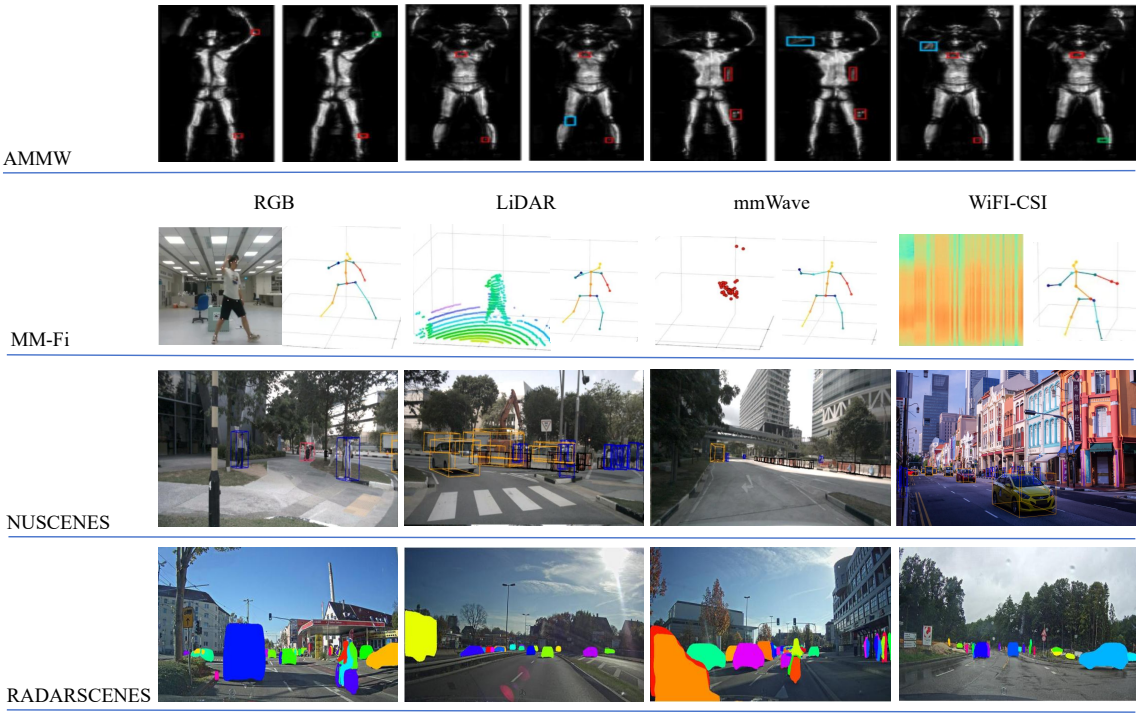


Fig. 14. Examples of detection results on Aerial Image Dataset.

a detailed classification and comparative analysis, and on this basis, offers strategic suggestions for the future direction of technology development. In addition, the article also comprehensively reviews small object detection datasets across multiple domains. Although this study provides a comprehensive overview of the field related to small object detection, there are still significant research gaps compared to the expected goals. Therefore, we have much work to do, mainly focusing on the following aspects:

1) Combination of Various Detection Methods: In this paper, we mainly mentioned six solutions for small object detection, including multi-scale and resolution methods, attention mechanisms, utilization of contextual information, training strategy improvements, architecture optimization, spatio-temporal information utilization, and computational efficiency and optimization. Through the overview of the methods used, we found that most methods are applied in isolation, like [97] which only used super-resolution technology to improve the ability of small object detection, and there are many similar cases. Of course, a few researchers combined multiple detection methods [32] , crossing technologies in hopes of achieving better results. We hope that such cross-technology phenomena will become more common in the future, with the potential to achieve better detection results.

2) New Frameworks for Small Object Detection: Generally, the models we employ are primarily based on bounding box regression networks, incorporating optimizations and enhancements derived from both traditional Convolutional Neural Networks (CNNs) and Transformer architectures.However, when facing different detection and classification tasks, this cannot serve as a universal solution. In this paper, we mentioned the Grouped Corner Detection Network [98] and the Bottom-Up Object Detection Network [99], both of which use new detection methods and produce competitive results. Therefore, developing new frameworks for detecting small objects is a research direction with great development prospects.

3) Domain-Specific Small Object Datasets: During the training process of models, datasets play a crucial role in object detection, especially those for specific domains. These datasets can compare the effectiveness of different algorithms and distinguish their strengths and weaknesses. Over the past years, researchers from different fields have contributed a series of effective object detection datasets. However, for our research needs, this is not enough. We need larger general-purpose datasets for small object detection, domain-specific small object datasets, and standards for judging the effectiveness of small object detection.

4) Small Object Detection Under Complex Conditions: So far, we are still in the stage of training algorithms on standard datasets. The photos in the datasets are static and always have corresponding deviations from real scenarios, such as real-time changes in lighting and dynamic complex environments like rain, fog, snow, etc. Here, we make a bold assumption about the future training process, hoping that in the future, algorithm models can be trained in dynamic datasets, and even undergo synchronous training in real scenarios.

5) Improvement of Attention Mechanisms: Attention mechanisms have been proven to be very effective in multiple domains, especially when dealing with complex backgrounds and highlighting important features. Applying attention mechanisms to small object detection can help the model focus more on key areas, thereby improving detection accuracy.

6) Generalization Across Domains and Datasets: Improving the model's generalization ability across different domains and datasets is another important research direction. This includes developing small object detection technologies that can adapt to different environments, backgrounds, and target features.

In summary, the future development of small object detection will involve innovation in deep learning architectures, the integration of various detection methods, improvements in attention mechanisms, and enhanced capabilities under complex conditions. The development of these technical routes will help overcome existing challenges and improve the accuracy and efficiency of small object detection.

VI. Conclusion

In this review, we have comprehensively collated nearly a hundred academic papers on small object detection. The paper first defines small objects in various datasets and application domains, followed by an in-depth analysis of the technical performance and characteristics of datasets in small object detection. Section 2 and 3 discuss in detail the applications of CNNs, Transformers, and their hybrid architectures across different technical fields, covering multi-scale and resolution methods, attention mechanisms, utilization of contextual information, improvements in training strategies, utilization of spatio-temporal information (especially in video images or dynamic scenes), and optimization of computational efficiency. These technical solutions each have their unique features and have shown exciting progress. In Section 4, we reviewed datasets for small object detection in four different domains. However, it is noteworthy that the number of publicly available datasets specifically designed for small objects is still relatively small, leading many studies to rely on general public datasets. To address this issue, we have developed a traffic road dataset specifically for the intelligent vehicle domain. The purpose of this dataset is to provide a more targeted experimental environment for research in small object detection. Overall, the aim of this paper is to provide researchers with a comprehensive perspective, helping them understand and evaluate the characteristics and complementarity of these different architectures, thereby inspiring new innovative thinking and driving the continuous progress of small object detection technology.

Acknowledgment

The authors thank the financial support of National Natural Science Foundation of China (Grant No: 51605054), Key Technical Innovation Projects of Chongqing Artificial Intelligent Technology (Grant No. cstc2017rgznzdyfX0039), Chongqing Social Science Planning Project (No:2018QNJJ16), Fundamental Research Funds for the Central Universities (No: 2019CDXYQC003).

References

[1] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[2] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[5] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[6] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2011.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[8] A. Aggarwal, M. Mittal, and G. Battineni, "Generative adversarial network: An overview of theory and applications," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100004, 2021.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[10] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[11] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," *arXiv preprint arXiv:2012.09958*, 2020.

[12] A. M. Rekavandi, S. Rashidi, F. Boussaid, S. Hoefs, E. Akbas, *et al.*, "Transformers in small object detection: A benchmark and survey of state-of-the-art," *arXiv preprint arXiv:2309.04902*, 2023.

[13] D. Wahyudi, I. Soesanti, and H. A. Nugroho, "Toward detection of small objects using deep learning methods: a review," in *2022 14th International Conference on Information Technology and Electrical Engineering (ICITEE)*. IEEE, 2022, pp. 314–319.

[14] K. Tong and Y. Wu, "Deep learning-based detection from the perspective of small or tiny objects: A survey," *Image and Vision Computing*, vol. 123, p. 104471, 2022.

[15] Y. Liu, P. Sun, N. Wergeles, and Y. Shang, "A survey and performance evaluation of deep learning methods for small object detection," *Expert Systems with Applications*, vol. 172, p. 114602, 2021.

[16] G. Chen, H. Wang, K. Chen, Z. Li, Z. Song, Y. Liu, W. Chen, and A. Knoll, "A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal," *IEEE Transactions on systems, man, and cybernetics: systems*, vol. 52, no. 2, pp. 936–953, 2020.

[17] K. Tong, Y. Wu, and F. Zhou, "Recent advances in small object detection based on deep learning: A review," *Image and Vision Computing*, vol. 97, p. 103910, 2020.

[18] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[19] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[20] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[21] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, *et al.*, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[23] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[24] B. Hou, X. Chen, S. Zhou, H. Jiang, and H. Wang, "Sr-yolo: Small objects detection based on super resolution," in *International Conference on Intelligence Science*. Springer, 2022, pp. 352–362.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[26] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[29] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[31] L. R. Medsker, L. Jain, *et al.*, "Recurrent neural networks," *Design and Applications*, vol. 5, no. 64-67, p. 2, 2001.

[32] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, and X. Liu, "A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[33] Z. Liu, J. Du, F. Tian, and J. Wen, "Mr-cnn: A multi-scale region-based convolutional neural network for small traffic sign recognition," *IEEE Access*, vol. 7, pp. 57 120–57 128, 2019.

[34] Z. Liu, D. Li, S. S. Ge, and F. Tian, "Small traffic sign detection from large image," *Applied Intelligence*, vol. 50, no. 1, pp. 1–13, 2020.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[36] X. Cao, P. Yuan, B. Feng, and K. Niu, "Cf-detr: Coarse-to-fine transformers for end-to-end object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 1, 2022, pp. 185–193.

[37] T. Ma, M. Mao, H. Zheng, P. Gao, X. Wang, S. Han, E. Ding, B. Zhang, and D. Doermann, "Oriented object detection with transformer," *arXiv preprint arXiv:2106.03146*, 2021.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] G. X. Hu, Z. Yang, L. Hu, L. Huang, J. M. Han, *et al.*, "Small object detection with multiscale features," *International Journal of Digital Multimedia Broadcasting*, vol. 2018, 2018.

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

[41] D. W. Ma, X. J. Wu, and H. Yang, "Efficient small object detection with an improved region proposal networks," in *IOP Conference Series: Materials Science and Engineering*, vol. 533, no. 1. IOP Publishing, 2019, p. 012062.

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

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large," 2014.

[44] Z. Liang, J. Shao, D. Zhang, and L. Gao, "Small object detection using deep feature pyramid networks," in *Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part III 19*. Springer, 2018, pp. 554–564.

[45] P. Du, X. Qu, T. Wei, C. Peng, X. Zhong, and C. Chen, "Research on small size object detection in complex background," in *2018 Chinese Automation Congress (CAC)*. IEEE, 2018, pp. 4216–4220.

[46] T. Liang, X. Chu, Y. Liu, Y. Wang, Z. Tang, W. Chu, J. Chen, and H. Ling, "Cbnet: A composite backbone network architecture for object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 6893–6906, 2022.

[47] Q. Li, Y. Chen, and Y. Zeng, "Transformer with transfer cnn for remote-sensing-image object detection," *Remote Sensing*, vol. 14, no. 4, p. 984, 2022.

[48] C. Yang, Z. Huang, and N. Wang, "Querydet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 13 668–13 677.

[49] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, "Conditional detr for fast training convergence," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3651–3660.

[50] Y. Wang, X. Zhang, T. Yang, and J. Sun, "Anchor detr: Query design for transformer-based detector," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2567–2575.

[51] T. Ding, K. Feng, Y. Wei, Y. Han, and T. Li, "Deot: an end-to-end encoder-only transformer object detector," *Journal of Real-Time Image Processing*, vol. 20, no. 1, p. 1, 2023.

[52] W. Xu, C. Zhang, Q. Wang, and P. Dai, "Fea-swin: Foreground enhancement attention swin transformer network for accurate uav-based dense object detection," *Sensors*, vol. 22, no. 18, p. 6993, 2022.

[53] S. Xu, J. Gu, Y. Hua, and Y. Liu, "Dktnet: dual-key transformer network for small object detection," *Neurocomputing*, vol. 525, pp. 29–41, 2023.

[54] X. Xu, Z. Feng, C. Cao, M. Li, J. Wu, Z. Wu, Y. Shang, and S. Ye, "An improved swin transformer-based model for remote sensing object detection and instance segmentation," *Remote Sensing*, vol. 13, no. 23, p. 4779, 2021.

[55] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[56] K. Zeng, Q. Ma, J. Wu, S. Xiang, T. Shen, and L. Zhang, "Nlfftnet: A non-local feature fusion transformer network for multi-scale object detection," *Neurocomputing*, vol. 493, pp. 15–27, 2022.

[57] Y. Lin, P. Feng, J. Guan, W. Wang, and J. Chambers, "Ienet: Interacting embranchment one stage anchor free detector for orientation aerial object detection," *arXiv preprint arXiv:1912.00969*, 2019.

[58] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022.

[59] Y. Xi, W. Jia, Q. Miao, X. Liu, X. Fan, and J. Lou, "Dycc-net: Dynamic context collection network for input-aware drone-view object detection," *Remote Sensing*, vol. 14, no. 24, p. 6313, 2022.

[60] P. Sun, T. Liu, X. Chen, S. Zhang, Y. Zhao, and S. Wei, "Multi-source aggregation transformer for concealed object detection in millimeter-wave images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 6148–6159, 2022.

[61] G. Chen, Z. Mao, K. Wang, and J. Shen, "Htdet: a hybrid transformer-based approach for underwater small object detection," *Remote Sensing*, vol. 15, no. 4, p. 1076, 2023.

[62] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," *Advances in neural information processing systems*, vol. 31, 2018.

[63] Y. Kim, B.-N. Kang, and D. Kim, "San: Learning relationship between convolutional features for multi-scale object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 316–331.

[64] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms–improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5561–5569.

[65] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3578–3587.

[66] H. Song, D. Sun, S. Chun, V. Jampani, D. Han, B. Heo, W. Kim, and M.-H. Yang, "Vidt: An efficient and effective fully transformer-based object detector," *arXiv preprint arXiv:2110.03921*, 2021.

[67] M. Maaz, H. B. Rasheed, S. H. Khan, F. S. Khan, R. M. Anwer, and M.-H. Yang, "Multi-modal transformers excel at class-agnostic object detection," *arXiv*, 2021.

[68] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[69] H. Inoue, "Data augmentation by pairing samples for images classification," *arXiv preprint arXiv:1801.02929*, 2018.

[70] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

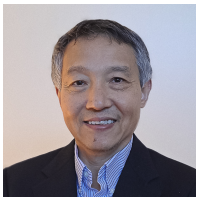
[71] J. Ding, W. Li, L. Pei, M. Yang, C. Ye, and B. Yuan, "Sw-yolox: An anchor-free detector based transformer for sea surface object detection," *Expert Systems with Applications*, vol. 217, p. 119560, 2023.

[72] W. Wang, Y. Cao, J. Zhang, and D. Tao, "Fp-detr: Detection transformer advanced by fully pre-training," in *International Conference on Learning Representations*, 2021.
[73] Q. Chen, J. Wang, C. Han, S. Zhang, Z. Li, X. Chen, J. Chen, X. Wang, S. Han, G. Zhang, *et al.*, "Group detr v2: Strong object detector with encoder-decoder pretraining," *arXiv preprint arXiv:2211.03594*, 2022.
[74] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "Dn-detr: Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 619–13 627.
[75] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2778–2788.
[76] Q. Zhou, X. Li, L. He, Y. Yang, G. Cheng, Y. Tong, L. Ma, and D. Tao, "Transvod: end-to-end video object detection with spatial-temporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
[77] C. Liu, S. Xu, and B. Zhang, "Aerial small object tracking with transformers," in *2021 IEEE International Conference on Unmanned Systems (ICUS)*. IEEE, 2021, pp. 954–959.
[78] K. A. Hashmi, D. Stricker, and M. Z. Afzal, "Spatio-temporal learnable proposals for end-to-end video object detection," *arXiv preprint arXiv:2210.02368*, 2022.
[79] H. Wang, J. Tang, X. Liu, S. Guan, R. Xie, and L. Song, "Ptseformer: Progressive temporal-spatial enhanced transformer towards video object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 732–747.
[80] Y. Cui and L. Yang, "Faq: Feature aggregated queries for transformer-based video object detectors," *arXiv preprint arXiv:2303.08319*, 2023.
[81] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "Dab-detr: Dynamic anchor boxes are better queries for detr," *arXiv preprint arXiv:2201.12329*, 2022.
[82] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525–5533.
[83] E. Goldman, R. Herzig, A. Eisenschtat, J. Goldberger, and T. Hassner, "Precise detection in densely packed scenes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5227–5236.
[84] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, *et al.*, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370.
[85] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8748–8757.
[86] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.
[87] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.
[88] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3974–3983.
[89] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *International conference on pattern recognition applications and methods*, vol. 2. SciTePress, 2017, pp. 324–331.
[90] 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 *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 3735–3739.
[91] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 445–461.
[92] J. Han, J. Ding, N. Xue, and G.-S. Xia, "Redet: A rotation-equivariant detector for aerial object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2786–2795.
[93] Y. Yu, L. Qiao, Y. Wang, and Z. Zhao, "Active millimeter wave three-dimensional scan real-time imaging mechanism with a line antenna array," *arXiv preprint arXiv:2102.04878*, 2021.
[94] J. Yang, H. Huang, Y. Zhou, X. Chen, Y. Xu, S. Yuan, H. Zou, C. X. Lu, and L. Xie, "Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[95] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
[96] O. Schumann, M. Hahn, N. Scheiner, F. Weishaupt, J. F. Tilly, J. Dickmann, and C. Wöhler, "Radarscenes: A real-world radar point cloud data set for automotive applications," in *2021 IEEE 24th International Conference on Information Fusion (FUSION)*. IEEE, 2021, pp. 1–8.
[97] F. Xiaolin, H. Fan, Y. Ming, Z. Tongxin, B. Ran, Z. Zenghui, and G. Zhiyuan, "Small object detection in remote sensing images based on super-resolution," *Pattern Recognition Letters*, vol. 153, pp. 107–112, 2022.
[98] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 734–750.
[99] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 850–859.

Ke Wang received the B.S., M.S. and Ph.D degrees in vehicle engineering from the Hunan University, Hunan, China, in 2007, 2009 and in 2013. He finished his Postdoctoral research at College of Engineering, Michigan University Ann Arbor, USA, in 2016 and 2017. Since 2017, he has been an Associate Professor with the State Key Laboratory of Mechanical Transmission, Chongqing University. He is the author of one book, more than 30 articles, and more than 15 inventions. He won the first prize of the China Automotive Industry Science and Technology Progress Award twice (in 2021 and in 2022). His research interests are the intelligent vehicle, environment perception and Safety AI.

Yang Chen received the B.S. degree with distinction in vehicle engineering from Northeastern University, China in 2020. He is currently pursuing the M.S. degree in vehicle engineering at Chongqing University, Chongqing, China. His research interests include intelligent automotive perception systems, small object detection, and 3D scene recognition.

Sheng Li received the B.S. degree with distinction in Robotics Engineering from China University of Mining and Technology in 2022. He is currently pursuing the M.S. degree in Vehicle Engineering at Chongqing University, Chongqing, China. His research interests include intelligent vehicle environmental perception, small object detection, and artificial intelligence.



Jianbo Lu (F'20) received a Ph.D. in Aerospace Engineering from Purdue University and has 20+ years of industry experience including technical leadership positions. With 90+ papers and 150+ US patents in vehicle controls, automation, safety, performance, robotics, and mechatronics, his contributions can be found in millions of vehicles on the road. He received the highest corporate award HFTA at Ford Motor Company twice. He was an AE for IEEE TCST and IFAC J. of Control Practice Engineering. He is on the editorial board of ASME J. of Autonomous Vehicles and Systems and a Fellow of IEEE, SAE, ASME, and AAIA.