Array 19 (2023) 100305

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array

2D and 3D object detection algorithms from images: A Survey

Wei Chen [a,b,c], Yan Li [a], Zijian Tian [a,*], Fan Zhang [a]

[a] School of Mechanical, Electrical & Information Engineering, China University of Mining and Technology (Beijing), Beijing, 100083, China
[b] Key Laboratory of Intelligent Mining and Robotics, Ministry of Emergency Management, China
[c] School of Computer Science & Technology, China University of Mining and Technology, Xuzhou, 221116, China

ARTICLE INFO

Keywords:
Image
Object detection
CNNs
Transformer
3D

ABSTRACT

Object detection is a crucial branch of computer vision that aims to locate and classify objects in images. Using deep convolutional neural networks (CNNs) as the primary framework for object detection can efficiently extract features, which is closer to real-time performance than the traditional model that extracts features manually. In recent years, the rise of Transformer with powerful self-attention mechanisms has further enhanced performance to a new level. However, when it comes to specific vision tasks in the real world, it is necessary to obtain 3D information about the spatial coordinates, orientation, and velocity of objects, which makes research on object detection in 3D scenes more active. Although LiDAR-based 3D object detection algorithms have excellent performance, they are difficult to popularize in practical applications due to their high price. Hence, we summarize the development process, different frameworks, contributions, advantages, disadvantages, and development trends of image-based 2D and 3D object detection algorithms in recent years to help more researchers better understand this field. Besides, representative datasets , evaluation metrics , related techniques and applications are introduced, and some valuable research directions are discussed.

1. Introduction

As the foundation of the computer vision field, object detection aims to identify and classify objects and regress boundary boxes from the input image, making it a multi-tasking operation that includes object recognition and localization. Based on deep learning, object detection is a prerequisite for other downstream vision tasks such as object tracking and human pose estimation, making it imperative to research high-performance and efficient object detectors. In recent years, research on object detection has become increasingly popular, with more than 37,000 literature references related to object detection according to Google Researcher, covering various aspects such as proposing, improving, and implementing object detection models, highlighting its growing importance.

Earlier traditional object detection algorithms generally included three stages: region proposal, feature extraction, classification, and regression. The region proposal stage usually extracts regions of interest (RoI) from the input image using a sliding window approach. However, considering that the objects to be detected may have different sizes, different window sizes are needed to obtain candidate regions through multiple iterations. The second stage commonly relies on hand-designed methods to extract features from the RoI, including Local Binary Pattern (LBP) [1], Histogram of Oriented Gradient (HOG) [2], Scale Invariant Feature Transform (SIFT) [3], and others. Finally, the extracted features are used for classification and boundary box regression. These traditional algorithms are significantly flawed, such as slow speed, low accuracy, and high computational overhead, which have gradually been replaced by deep convolutional neural networks (CNNs).

The AlexNet [4] proposed by Hinton et al. won the championship in the ImageNet image recognition competition in 2012, which gradually made deep learning algorithms based on CNNs the mainstream of object detection. Fig. 1 shows the basic process framework of object detection. For the input image into the network, features generally need to be extracted through the backbone (AlexNet [4], ResNet [5], Swin Transformer [6], PVT [7], etc.), and then fused and enhanced through the neck (FPN [8], PAN [9], YOLOF [10], DETR [11], etc.), finally input into the head (YOLO [12], SSD [13], *R*–FCN [14], ET-Head [15], etc.) to predict classes and location coordinates of objects. Researchers have successively proposed anchor based RCNN series, YOLO series, and SSD series algorithms. The detection performance and speed of these methods have greatly improved compared to the earlier traditional algorithms. Subsequently, some researchers proposed using some key

* Corresponding author.
E-mail addresses: chenwdavior@163.com (W. Chen), 18600873522@163.com (Y. Li), Tianzj0726@126.com (Z. Tian), zf@cumtb.edu.cn (F. Zhang).

https://doi.org/10.1016/j.array.2023.100305
Received 18 April 2023; Received in revised form 14 June 2023; Accepted 24 June 2023
Available online 27 June 2023

points or anchor points to regress boundary boxes, eliminating the limitations imposed by preset anchor boxes. Compared to the stable parameter learning in CNN, Transformer [16], which has risen in popularity in recent years, has shown stronger adaptability to large datasets with its dynamic parameter learning mechanism, leading to a huge wave of development in the field of natural language processing (NLP). With the successful application of Transformer to the computer vision field with ViT [17], some researchers have proposed using Transformer as the Neck or Backbone in some models, which has become a hot research topic in recent years.

2D object detection can only regress the 2D boundary box of an object, which cannot meet the practical needs of real-world 3D space. The 3D object detection task aims to use sensor data to predict a detailed set of information about an object's 3D size, coordinates, speed, and heading angle, enabling intelligent applications such as unmanned cars and smart robots to sense real-time traffic conditions and plan reasonable driving routes. However, 3D object detection is often not a simple extension of 2D algorithms in 3D space, which is reflected in the following aspects.

- The lack of geometric constraints makes depth estimation more difficult, especially for distant objects and those obscured by the environment.
- The huge multidimensional data presents a significant challenge for real-time detection.
- Different modalities of data need certain fusion mechanisms to be processed.
- A specific evaluation system is required to provide criteria for model improvement.

Therefore, 3D object detection algorithms have become increasingly important in the industrial and academic communities in recent years.

This article collects top research papers on image-based object detection from top international conferences (ICCV, CVPR, ECCV, etc.) and journals (IEEE TPAMI, IJCV, etc.) in the computer vision field in recent years, analyzing, comparing, and summarizing them. Fig. 2 shows the algorithm classification of this article. Specifically, this article's main contributions are summarized below.

(1) We present the task of image-based object detection, its development overview, 2D and 3D datasets, and evaluation metrics (Section 1 and 5).
(2) The 2D object detection algorithms are divided into three categories: Anchor-based, Anchor-free, and Transformer-based. The 3D object detection algorithms are divided into four categories: monoculor-based, stereo-based, pseudo-LiDAR-based, and multi-view-based. The process, advantages and disadvantages of each category are summarized, and the key contributions of each algorithm are analyzed in detail (Section 2 and 3).
(3) We introduce related techniques in the field of object detection and the latest progress of four application scenarios (Section 4 and 6).

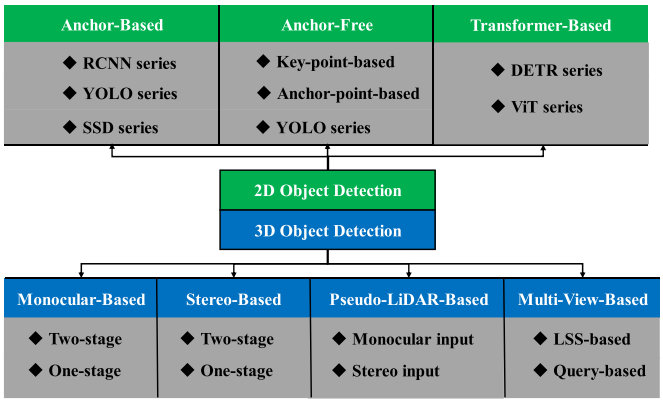

Fig. 2. Image-based object dtection algorithm classification.

(4) We summarize the full paper and prospect several valuable research directions (Sections 7 and 8).

2. 2D object detection algorithms

Fig. 3 illustrates the development of image-based 2D object detection algorithms, which can roughly be divided into two categories: Anchor-Based and Anchor-Free methods. The Anchor-Based methods are more mature in terms of technology development, but they have drawbacks such as sensitivity to hyperparameters and poor multi-scale detection performance due to the need for manual design of prior anchor boxes. The other approach is to abandon anchor boxes and predict densely at the pixel-level through setting up prior anchor points or using some key feature points to regress the bounding box. Both types of methods use convolutional neural networks for the entire detection process, while another type of method uses the Transformer architecture, which has been popular in recent years, to participate in feature extraction and fusion. Although this method also belongs to the anchor-free methods, it has a more concise and efficient end-to-end detection framework, and therefore, it is introduced separately as a category.

2.1. Anchor-based methods

In object detection tasks, a useful strategy is to predefine a set of dense anchor boxes of different scales and to differentiate these anchor boxes into positive and negative samples based on a certain labeling assignment strategy. Finally, the category and position offset for positive samples are predicted, and the final prediction results are obtained. This is anchor-based object detection method, which can be divided into three branches: RCNN series, YOLO series, SSD series (Fig. 4).

2.1.1. RCNN series

The most basic model in object detection is the *R*–CNN [18] (Girshick et al., 2014), which generates a large number of candidate boxes for the input image through a selective search algorithm [19], inputs them into the convolution backbone to extract features, and then uses SVM

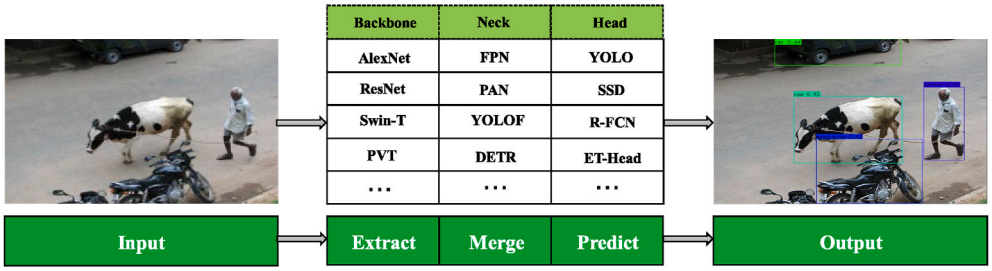

Fig. 1. The basic process framework of the object detection.

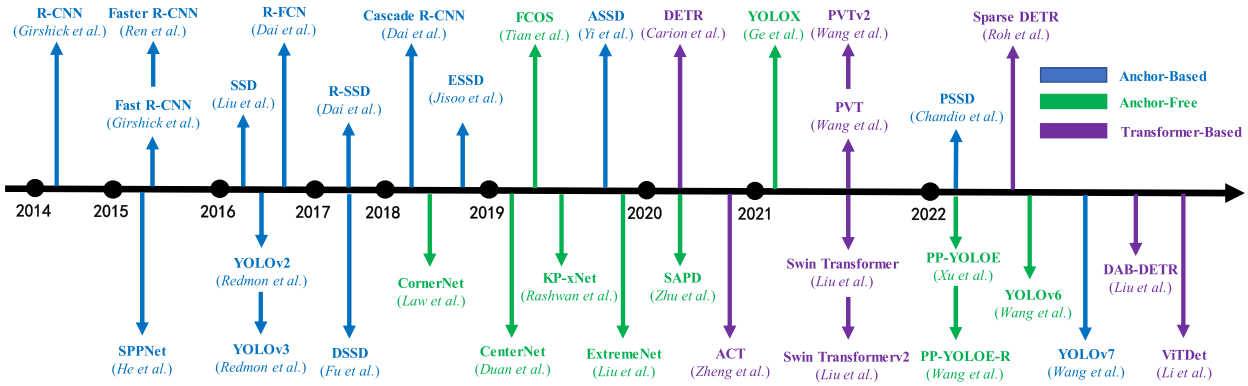

Fig. 3. Chronological overview of the 2D detection methods (author information in parentheses).

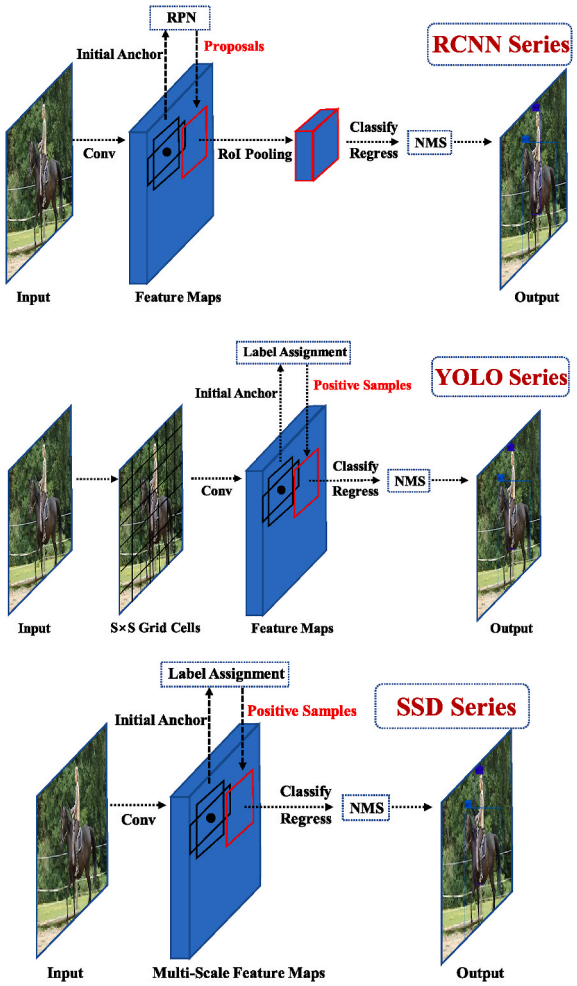

Fig. 4. The flow of the anchor-based methods. For two-stage methods represented by the RCNN series, these anchor boxes are used to generate regions of interest, and then classification and regression are conducted on these sparse local candidate regions. Single-stage methods represented by the YOLO and SSD series directly use the screened positive samples to predict the bounding box.

(Support Vector Machine) [20] and linear regressor for classification and location fine-tuning respectively. In addition, the problem of insufficient data is solved to some extent by the idea of pre-training on large data sets and then fine-tuning on small specific data sets. Compared with traditional algorithms, *R*–CNN is the first to use CNNs to extract features from images. Both accuracy and generalization ability of the model are greatly improved, but there are many aspects that need to be improved.

On the one hand, *R*–CNN generates many candidate boxes of different sizes with the help of a selective search algorithm, which not only needs to calculate the results through the convolutional network repeatedly, but also needs to fit them to a fixed size by scaling and distorting to input into the fully connected layer, which can easily lead to image distortion and affect the detection. To this end, He et al. (2015) proposed SPPNet [21], which solves the above problems by adding a spatial pyramid pooling layer before the fully connected layer. SPPNet can integrate the size of the candidate box by only performing one convolution operation on the image, which greatly improves the speed of the model. In addition, SPPNet supports multi-size image training, and improves the robustness to object deformation by multi-level pooling.

On the other hand, *R*–CNN divides the whole detection process into four network modules to complete, which is not only slow in training, but also requires high resources of computer equipment. To this end, Girshick et al. (2015) proposed a faster version: Fast *R*–CNN [22], which uses a region of interest (RoI) pooling layer to integrate the extracted image features before feeding them into the fully connected layer for classification and regression. Compared with *R*–CNN, Fast *R*–CNN not only improves the accuracy by 11.5%, but also runs more than 200 times faster. However, because it still uses selective search algorithm to obtain candidate boxes, the real-time detection effect on large-scale datasets is not ideal. To break through this bottleneck, Ren et al. (2017) proposed Faster *R*–CNN [23] to generate anchor candidates through a Region Candidate network (RPN) that shares features with the detection network, and then the RoI pooling layer maps the foreground anchor (RoI) to a fixed-size feature map. Finally, the prediction results are generated through the fully connected layer. Faster *R*–CNN puts candidate generation, feature extraction, and final classification and regression into the same convolutional neural network, which realizes "end-to-end" object detection for the first time, so it has a speed closer to real-time object detection. However, there are several problems: However, there are several problems.

(1) After RoI Pooling processing, all RoI will go through the fully connected layer with many parameters, and excessive redundant calculation restricts the detection speed.
(2) When RoI Pooling is rounded twice, some pixels will be discarded or reused, which will restrict the detection accuracy.

Therefore, some researchers have proposed corresponding improvement ideas.

In response to the above problem (1), the *R*–FCN [14] proposed by Dai et al. (2016) adopts a fully Convolutional network (FCN) design without the fully connected layer. However, due to the translation invariance of convolutional neural network, it is difficult to effectively achieve object localization, so the *R*–FCN model adds a

location-sensitive score map to solve the contradiction. Compared with Faster *R*–CNN, *R*–FCN greatly improves the speed at the cost of a small performance loss. To solve the problem (2), Mask *R*–CNN [24] proposed by He et al. (2017) replaces the RoI pooling layer with the RoI alignment layer. It uses bilinear interpolation when mapping RoIs to fixed-size feature maps, which can better preserve the spatial information on the feature maps. At the same time, Mask *R*–CNN uses ResNet-FPN as the backbone network, which performs better in multi-scale object detection tasks. Cai et al. (2018) proposed Cascade *R*–CNN [25], which focuses on improving the object localization ability of Faster-RCNN. All previous works generally set the IoU threshold to 0.5 to determine the positive and negative samples. This is because simply raising the IoU threshold may lead to a mismatch between the quality of proposals in the training stage and the inference stage, and the model will fall into overfitting. However, since the sample output is always closer to the real box than the input after each regression, Cascade *R*–CNN proposes a resampling mechanism of cascade regression to increase the IoU threshold corresponding to the proposals in each stage, so that the detector in each stage will have higher quality proposals to train without overfitting.

2.1.2. YOLO series

The YOLO (you only look once) family of methods is one of the most representative works in the field of object detection, and there have been many versions so far (Fig. 5).

The early YOLOv1 [12] (Redomon et al., 2016) algorithm did not use an anchor box mechanism. Instead, it divided the input image into S × S grid cells and output the class probabilities and bounding boxes on this grid cell. Since the feature extraction, fusion and final classification and regression prediction are carried out in the same CNNs, it has high processing speed and is easier to optimize the model end-to-end. However, YOLOv1 only predicts one category and two bounding boxes in each grid cell, which is difficult to deal with overlapping and multi-scale object prediction.

YOLOv2 [26] (Redomon et al., 2017) reuses the anchor design and improves the overall YOLOv1 model by Batch Normalization (BN) on the input of each layer of the network to speed up the model convergence. To improve the recall rate of the model's small object detection, YOLOv2 sets five anchor boxes with different aspect ratios at the same position with the help of k-means clustering. In the prediction stage, YOLOv2 replaces the fully connected layer with a fully convolutional network for classification and regression. These improvements make the detection performance of YOLOv2 model greatly improved under the premise of high processing speed, and the mAP value of YOLOv2 exceeds Faster RCNN at 40FPS. In 2018, the YOLOv3 [27] model launched by Redomon et al. to strengthen the feature extraction ability of the network, the backbone network uses DarkNet-53 composed of residual modules to efficiently stack the deep network structure. In addition, the model sets a total of nine anchor boxes of three scales on the same position. Moreover, the activation function is replaced by sigmoid from softmax to effectively handle multi-scale object detection and multi-label classification tasks respectively. In the case of the same detection accuracy, YOLOv3 model beats any other model at the same time in speed.

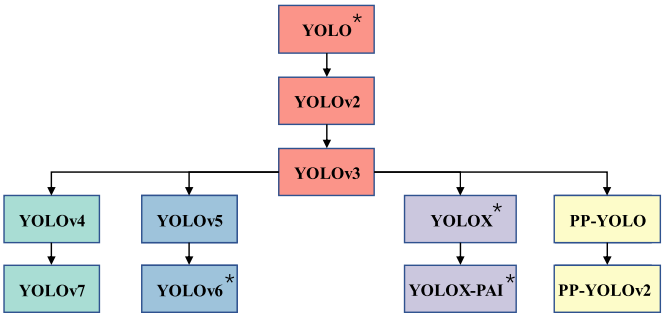

Fig. 5. The development branches of YOLO series methods. There are four main derivative branches after the YOLOv3 model. Including anchor-based YOLOv4, YOLOv5, *P*P-YOLO and anchor-free YOLOX (marked with "*" to indicate the anchor-free mechanism algorithm, mainly introduced in section 2.2).

To further improve the performance of YOLO model on real-time object detection tasks, Alexey Bochkovskiy et al. (2020) proposed a new generation of YOLOv4 [9] model. It uses the Cross Stage Partial Connection (CSP) structure in the Backbone training phase to reduce the redundancy of feature maps, and improves the generalization ability of the model by CutMix and Mosaic data augmentation methods. In addition, YOLOv4 uses an improved PANet (path aggregation network), SPP (spatial pyramid pooling) module and CBAM (convolutional block attention module) on the neck side to bidirectional fusion of multi-scale features and increase the receptive field, so as to improve the spatial positioning accuracy. These improvements make YOLOv4 have 10% higher AP than YOLOv3, but the number of parameters is relatively large. Therefore, Wang et al. (2022) proposed a more lightweight YOLOv7 [28] algorithm. YOLOv7 focuses on the optimization of modules and methods in the training process, and proposes several trainable bag-of-freebies methods, including the planned re-parameterized model and the coarse-to-fine dynamic label assignment strategy, which reduces the model parameters by 40% compared with YOLOv4 and greatly improves the detection accuracy. In order to learn features efficiently, YOLOv7 proposes an extended ELEN architecture, and this extension realizes channel expansion and regularization of computational blocks through group convolution, without reducing parameter utilization.

In 2020, YOLOv5 [29] proposed by Jocher is another improved branch of the YOLOv3 model. It is basically the same as the detection network of YOLOv4, but an adaptive anchor box screening and image scaling mechanism is proposed in the initial stage of model training, which has faster detection speed. In order to promote the deployment of YOLO models on the industrial equipment, Long et al. (2020) proposed *P*P-YOLO [30], which uses a number of techniques at various stages of training and prediction, including larger batch size to improve training stability, DropBlock regularization to improve robustness, and Matrix NMS to quickly filter redundant boxes. *P*P-YOLO achieves the optimal balance between speed and accuracy at that time without increasing the number of parameters. The next year, the *P*P-YOLOv2 [31] proposed by Huang et al. improved the detection accuracy better with a small training cost by improving the PANet and the Mish activation function.

2.1.3. SSD series

Similar to YOLO series, SSD [13] proposed by Liu et al. (2016) is another one-stage method of dense prediction, which directly regresses the class confidence and bounding box offset of the object according to the selected positive sample anchor box. It mainly solves the problem that the early YOLOv1 model is difficult to detect small object detection. In the backbone network, SSD uses convolutional layers of different sizes to extract multi-scale features, and abandones the fully connected layer in the prediction stage, and directly uses the fully convolutional network to reduce the model parameters. In addition, SSD also uses data augmentation and hard sample mining techniques to further improve the detection accuracy. Therefore, compared with YOLOv1, SSD has a higher recall rate in small object detection tasks, but there are two main problems.

(1) For the lack of utilization of context information between feature maps of different layers, the performance of small object detection is still not ideal.
(2) Only a small part of the generated anchor boxes is distributed near the real box, which leads to the problem of class imbalance between positive and negative samples in the training process.

Therefore, some improved versions of SSD have been gradually proposed.

DSSD [32] (Fu et al., 2017), *R*-SSD [33] (Jisoo et al., 2017), ESSD [34] (Zheng et al., 2018), FSSD [35] (Yang et al., 2019) by improving the feature fusion module, the above problem (1) is solved. DSSD uses a deeper ResNet-101 as the backbone network and adds a deconvolution module after the auxiliary convolutional layer to build an "hourglass" network architecture, which can better obtain high-level semantic information and low-level spatial information. *R*-SSD uses both pooling and deconvolution modules to effectively enhance the connection between feature maps of different layers. Due to their deep network structure with larger parameters, the speed of these two models is much slower than that of SSD models. To this end, ESSD only adds extension modules to the first three convolutional layers to efficiently capture semantic features, and FSSD directly concatenates the feature maps of each layer, which is more lightweight and efficient. ASSD [36] proposed by Yi et al. (2019) adds a self-attention module between the feature map and the prediction module, so that the model can pay attention to the pixel-level feature relationship. The PSSD [37] proposed by Chandio et al. (2022) uses dilated convolution to effectively expand the receptive field and uses bidirectional FPN design to obtain more spatial and semantic information, which greatly improves the real-time detection performance.

To solve the problem (2), RetinaNet [38] proposed by Lin et al. (2017) adopts a focal loss function, which can increase the weight of hard samples in the cross-entropy loss during the training process, so that the positive and negative samples for training are maintained in the appropriate proportion. Zhang et al. (2018) believe that RCNN series methods can effectively solve the class imbalance problem by refining the anchor box, and then propose a sparse prediction model: RefineDet [39], which uses the anchor box refinement module to filter out negative samples and correct the position and size of positive samples, and further iterates and optimizes through the alignment module in the following. However, the multi-scale features obtained by RefineDet in a top-down manner cannot align and optimize the anchor box well. Therefore, Nie et al. (2019) proposed the EFGRNet [40] model, which fully obtains the multi-scale context features through the bottom-up feature pyramid structure and uses these features to guide the refinement of the anchor box. Therefore, more accurate prediction results can be obtained.

2.2. Anchor-free methods

To detect objects from images, the anchor mechanism can provide selected positive samples for the final classification and regression, but it also has some limitations.

- The design of hyperparameters such as the size and aspect ratio of the anchor box has a great impact on the performance of the detection model, and the anchor box designed solely by artificial prior is easy to lead to poor generalization ability of the model.
- Most of the generated dense anchor boxes are negative samples, which will cause the problem of class imbalance between positive and negative samples.
- The square anchor box of positive samples still has the problems of positioning ambiguity and background feature interference.

Therefore, some researchers question the rationality of the "box" design relying on artificial prior, and have proposed some anchor-free algorithms, which can be roughly divided into two types of methods: key-point-based and anchor-point-based methods. Fig. 6 illustrates the flow of these two types of methods.

2.2.1. Key-point-based

Law et al. (2018) proposed CornerNet [41], which innovatively achieves anchor-free object detection by representing prediction boxes with sparse key corners. CornerNet generates a heatmap of the top-left corner and bottom-right corner for an object through a single convolutional network, and groups a pair of corners belonging to the object together by generating connection vectors for the corners. In addition, considering that most of the objects to be detected are verty-free, CornerNet also adds a corner pooling layer to improve the reliability of corner localization. To further improve the efficiency of the model, Law et al. (2019) proposed two improved versions [42] in the following year, including CornerNet-Saccade model which introduced the attention mechanism and CornerNet-Squeeze model which simplified the pixel

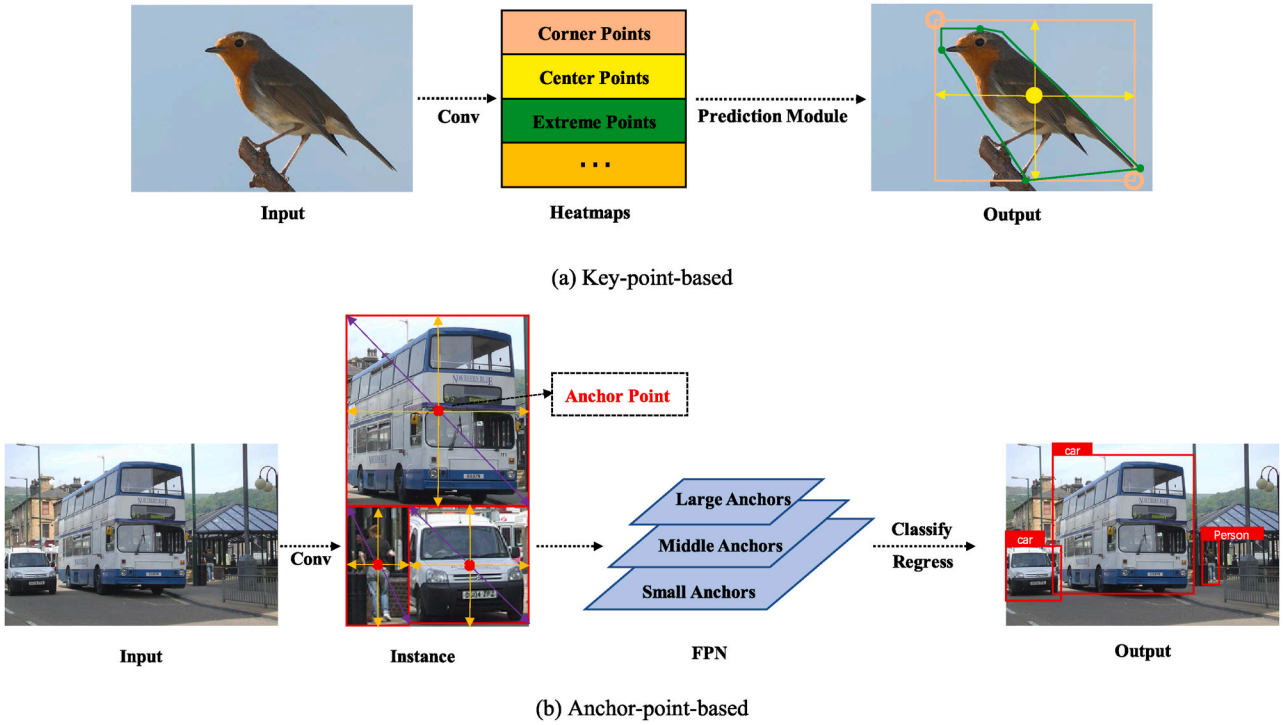

Fig. 6. The flow of the anchor-free methods.

processing. However, there are two problems in CornerNet.

(1) The corner pooling layer belongs to the Max pooling operation, which will lose feature information to a certain extent.
(2) The top-left and bottom-right corners belonging to the same object are grouped together using a feature embedding operation, which makes the grouping difficulty of the model grow squared with the number of objects in the image.

To this end, KP-xNet [43] proposed by Rashwan et al. (2019) uses convolutional layers of different sizes to perform corner detection on multi-scale objects, such that each convolution layer is only responsible for detecting objects in a certain size range. With the design of automatic perception of object size, KP-xNet has a more lightweight network architecture, and the detection speed is greatly improved.

CenterNet [44] proposed by Duan et al. (2019) employs an alternative predictive box representation that reproduces bounding boxes by generating a heatmap of the center point for the object along with the corresponding width and height. CenterNet can directly filter out redundant prediction boxes on the heatmap, which is more convenient than the time-consuming NMS post-processing operation. In addition, CornerNet has strong versatility in downstream tasks such as 3D object detection and human pose estimation. Zhou et al. (2019) believe that the rectangular prediction box may contain too many irrelevant background pixels, which cannot represent the visual information of the object well. Therefore, he proposed an anchor-free box detection model based on pure geometric appearance: ExtremeNet [45]. Through the hourglass backbone network, ExtremeNet detects the top, bottom, left and right four extreme point heatmaps of the object, and selects the appropriate combination of five extreme points according to the scores of the geometric centers of the four extreme points on the central point heatmap. Since the high scoring boxes obtained by this central grouping method may contain the extreme points of adjacent objects, the authors use soft NMS to deal with the potential spurious boxes. In addition, considering that objects may have vertical or horizontal edges, which may lead to low scores of extreme points on these edges and be missed, the authors use an edge aggregation method to enhance the intermediate pixel responses at such edges.

2.2.2. Anchor-point-based

Another class of methods perform dense pixel-by-pixel detection with preset anchor points. The FCOS [46] model proposed by Tian et al. (2019) is a representative work, which uses a multi-level FPN prediction structure like RetinaNet and implements positive and negative anchor point classification and anchor point to prediction box distance regression through two branches respectively. In addition, a centrality prediction branch is added to the classification branch to suppress the interference of anchor points far from the object center on the regression. The FCOS model is not only efficient, but also has good scalability on other vision tasks.

However, there is lack of theoretical basis for assigning feature layers with different resolutions only according to the size of the object in FPN network. Therefore, Zhu et al. (2019) proposed FSAF (feature selective anchor-free) [47], which dynamically assigns the optimal FPN layer for each instance to learn features. Specifically, FASF is based on RetinaNet, which adds an anchorless branch and the original branch to each layer of the FPN for parallel weighted training. Finally, the prediction results of the two branches are combined in the inference stage to obtain the result. In contrast to FCOS and FSAF, which assign each instance to a separate feature layer, Kong et al. (2019) proposed FoveaBox [48], which presets the size acceptance range of each FPN layer, then assigns the instance to the appropriate multiple feature layers to learn according to the size of the ground truth, and outputs the prediction box by regressive the distance from the pixel point to the top left and bottom right corner.

Compared with the key-point-based method, the anchor-point-based method has a lighter and more efficient network architecture, but its detection accuracy is lower. As for the reason, Zhu et al. (2020) believe that this kind of algorithm gives equal weight to the loss of all anchor points in the training process, which leads to filtering out the prediction boxes with low confidence but high IoU in the post-processing. Assigning a single feature layer only according to the instance size will inevitably lead to the waste of other feature information. To this end, SAPD [49] proposed by Zhu et al. uses soft-weighted anchor points and soft-selected pyramid levels training strategies to consciously reduce the participation of anchor points close to the instance border in the training process, so that the model achieves a better balance between performance and speed.

In both key-point-based and anchor-point–based methods, discrete points are used to form the prediction box, which is not only easy to lose adjacent information, but also requires additional branches to group discrete points belonging to the same object, which is cumbersome. To this end, CrossDet [50] proposed by Qiu et al. (2021) adopts a pair of adaptive cross lines of horizontal and vertical axes to represent the position of the object. Specifically, CrossDet first predicts the rough position of the crossing lines through a regression branch, then adaptively aggregates line features from the horizontal and vertical dimensions through a crossline extraction module (CEM), and finally accurately regresses the position of the crossing lines through a decoupling mechanism. This object representation provides a novel idea for the anchor-free object detection model and shows comparable performance with other state-of-the-art models on large benchmark datasets such as COCO.

2.2.3. YOLO series

Some researchers try to use the above two methods to improve the YOLO series algorithms for anchor removal. YOLOX [51] proposed by Ge et al. (2021) takes YOLOv3 as the baseline, divides the input image into multiple grids, and predicts the two offsets in the left-top corner and the width and height of the prediction box at once on each grid. By decoupling classification from regression tasks, YOLOX makes it easy to optimize network parameters end-to-end. In addition, YOLOX proposed a new label assignment strategy (SimOTA), which can find the globally optimal confidence assignment for all instances in an image. Based on this, Zou et al. (2022) proposed an improved version of it: YOLOX-PAI [52]. In the head, YOLOX-PAI coordinates the tasks on the two branches of regression and classification through an attention mechanism and enhances features by an adaptive spatial feature fusion scheme. In addition, the PAI-Blade proposed by the model is an inference optimization framework that integrates multiple techniques and has the best performance at that time in 1 ms inference time. In 2022, YOLOv6 [53] proposed by Li et al. takes YOLOv5 as the baseline, strengthens feature extraction by replacing the backbone network with EfficientRep, and adopts a decoupled detection head and label assignment strategy like YOLOX to optimize the model. To be easier to deploy on industrial equipment, YOLOv6 also uses the self-distillation method to help the network learn richer features and improve the detection accuracy, speed, and generalization ability.

Similar to FCOS, *P*P-YOLOE [15] proposed by Xu et al. (2022) performs pixel-level prediction by presetting anchor points, which greatly reduces the number of model parameters and optimizes the inference speed. At the same time, considering that residual connections and dense connections have good performance in solving gradient disappearance and aggregating multi-scale receptive fields respectively, they propose a RepResBlock that combines the advantages of both to construct the backbone network and the neck side, and further optimize the feature extraction process by cross-stage partial connections and adding channel attention mechanism. In addition, a label assignment strategy called TAL (task alignment learning) is proposed in *P*P-YOLOE to bring the optimal anchor points of classification and regression tasks closer, to obtain high confidence and high localization accuracy prediction results at the same time. In the same year, *P*P-YOLOE-R [54]

proposed by Wang et al. (2022) is an improved model for rotating object detection. It introduces the ProbIoU loss to handle the problem of object boundary discontinuity and realizes the efficient prediction of object Angle through a decoupling detection head.

2.3. Transformer-based methods

Convolutional neural networks extract features by sliding independent convolutional windows, which to some extent makes the model unable to effectively obtain global feature information. The object detection algorithm based on Transformer can capture long-range dependencies on the object and easily obtain global effective information, which has gradually become a hot research direction in this field. In this section, Transformer-based object detection algorithms are divided into two categories to introduce: DETR series methods and ViT series methods (Fig. 7).

2.3.1. DETR series

In 2020, DETR [11] proposed by Carion et al. was the first to apply Transformer to the field of object detection. DETR still uses the convolution module to extract features in the backbone network and encodes the image features into a set of object queries with the help of the Transformer structure in the neck side. The category and location information of each object are obtained by decoding and predicting the set. DETR can effectively use the global context relationship between the object and the image, and only predict one bounding box for each object by using the Hungarian matching algorithm, eliminating the post-processing operation for redundant prediction boxes in traditional methods, which is a more concise and efficient end-to-end detection model. However, when DETR processes high-resolution images, the calculation amount of the attention module will be greatly increased, which leads to slow convergence speed of the model and difficulty in small object detection. To solve this problem, some researchers have proposed improvement methods including improving attention module and queries.

The computational complexity of DETR self-attention module increases squarely with the increase of input image resolution. For this reason, ACT [55] proposed by Zheng et al. (2020) uses locality sensitive hashing to adaptively cluster similar query features. By replacing the self-attention module in DETR with ACT, FLOPs can be greatly reduced without loss of performance. Gao et al. (2021) proposed a plug-and-play SMCA [56] module to replace the co-attention module in DETR. SMCA reduces the search range of feature aggregation to near the center of the object and improves the convergence speed by combining the learnable co-attention weights with the spatial prior of the query. In 2021, Dai et al. proposed an improved model based on dynamic attention mechanism: Dynamic DETR [57], which simulates the encoder of Transformer through convolutional structure, combines space, scale, and other factors to dynamically adjust the attention module, so as to improve the sparsity of feature representation, and accelerates model convergence through RoI-based dynamic attention module at the decoder. Deformable DETR [58] proposed by Zhu et al. (2020) designed a multi-scale deformable attention module, which effectively combines the sparse spatial sampling ability of deformable convolution and the global relational modeling capability of Transformer to improve the performance of small object detection. Moreover, since the model only focuses on a small number of sampling points (keys), the convergence speed is also greatly improved. However, Deformable DETR inputs multi-scale features into the encoder without distinction, which will undoubtedly increase the complexity of the model. To this end, Roh et al. (2022) propose Sparse DETR [59], which uses a scoring network to filter the input units (tokens) of the encoder and encode the features with the selected high-quality tokens. In this way, sparser queries can be obtained, and inference speed can be greatly improved.

Sun et al. [60] attributed the slow convergence of DETR to the cross-attention module of the decoder and found that directly removing the decoder structure did not affect the overall performance but lacked a deep exploration of the decoder. Liu et al. (2022) believe that the lack of a positional prior for learnable queries in the decoder is the main factor for slow convergence. To this end, the authors proposed DAB-DETR [61] to accelerate model convergence by using dynamic anchor boxes updated layer by layer as queries to provide location priors. In 2022, the Anchor DETR [62] proposed by Wang et al. encodes the query by presetting the anchor points and solves the problem of overlapping

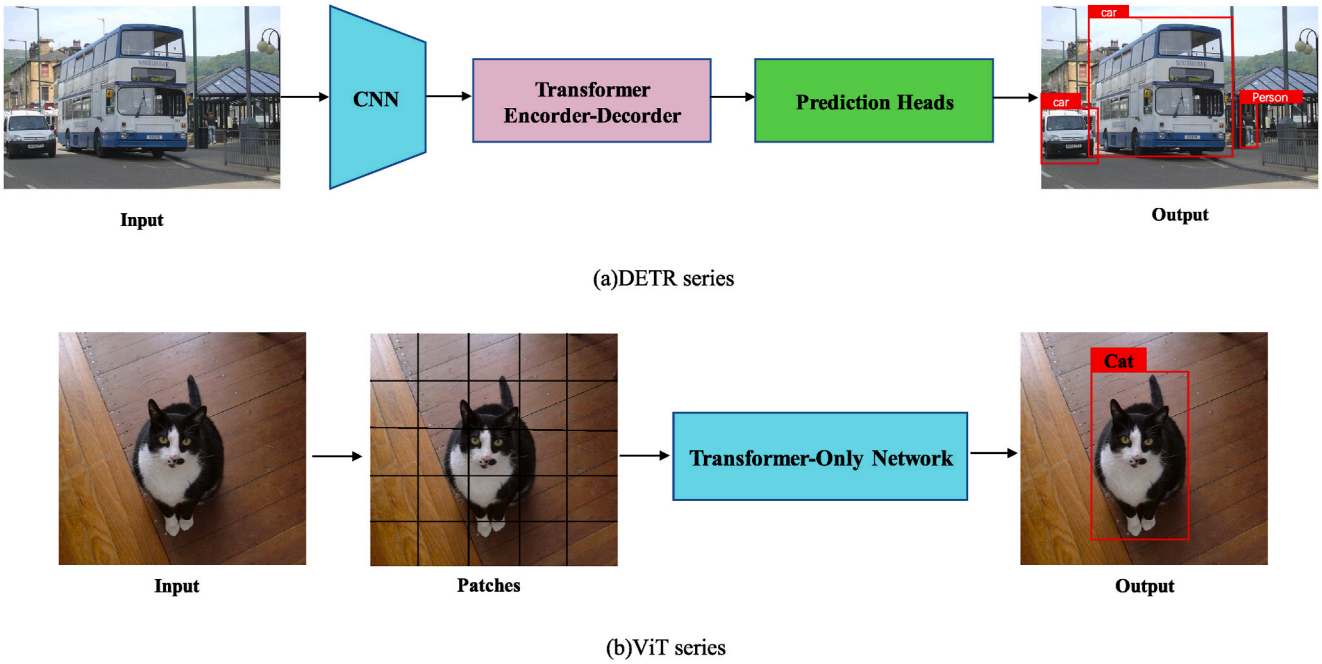

Fig. 7. The flow of theTransformer-based methods. The DETR series method still extracts features through convolutional neural network in the backbone network, and the ViT series method first segments the image into sparse patches, and then inputs them into the Transformer-only structure network in the form of sequence to predict the result.

occlusion of the object by adding multiple prediction branches around the anchor points. In addition, Anchor DETR also introduces a RCDA (Row-Column Decouple Attention) module to decouple the 2D feature map into row-column features, which reduces the computational cost. In addition to the above two kinds of improvement methods, other researchers have also improved the overall speed and accuracy of the model by combining dense priors or using new denoising training techniques.

2.3.2. ViT series

Due to the limited computer resources, the Transformer used as the backbone network for image tasks must first solve the problem of too long pixel sequence. To this end, ViT [17] proposed by Dosovitskiy et al. (2020) is successfully applied to image classification tasks by expanding the pixel selection range and dividing the input image into sparse patches. ViT-FRCNN [63] proposed by Josh et al. (2020) directly uses ViT as a Backbone network for object detection, but it exposes some problems.

(1) the feature maps output by ViT are of single scale and low resolution.
(2) The computational cost will increase squared with the increase of image resolution.

To solve these problems, some researchers have proposed variant models of ViT.

PVT [7] proposed by Wang et al. (2021) empowers a progressive shrinking pyramid structure into the Transformer backbone to flexibly learn multi-scale high-resolution features. The multi-head attention in ViT is replaced by the spatial reduction attention (SRA), which ensures that high-resolution feature maps and global receptive fields can be processed at a small computational cost. As the first pure Transformer backbone network, PVT has better performance than convolutional backbones such as ResNeXt and is compatible with many dense prediction models. However, PVT considers the image as a sequence of non-overlapping patches, which will affect the local continuity of features. Therefore, Wang et al. (2021) proposed PVTv2 [64], which enlarging and overlapping the patch window to strengthen the connection with each other and using zero-padding convolution to maintain the resolution of the feature map. Compared with the fixed size position encoding in PVT, it can deal with images with different resolutions more flexibly. In addition, PVTv2 also proposes an improved linear spatial reduction attention (LSRA), which ensures the linear growth of computational complexity by adding pooling operation based on SRA. Compared with PVT, it reduces GFLOPs by 22%.

CNNs can generate multi-resolution feature maps compared to original ViT, which has a natural advantage in object detection tasks. Therefore, another idea is to continue to use the hierarchical, laye-by-layer downsampling design of CNNs. The Swin Transformer [6] proposed by Liu et al. (2021) ensures the linear growth of computational complexity by limiting the self-attention calculation to non-overlapping local Windows, and gradually increases the receptive field in the downsampling process with the help of a hierarchical structure, so as to extract features from local to global more efficiently. It is worth mentioning that the Swin Transformer does not divide the image into low-resolution patches (16×16 for ViT and 4×4 for PVT). Instead, the original image is directly fed into the model to avoid information loss caused by subsequent resizing.

In 2021, Liu et al. proposed an improved version with up to 3 billion parameters: Swin TransformerV2 [65], which mainly adapted large-scale models from three aspects: (1) The pre-normalization technique adopted by Swin Transformer leads to the activation amplitude of the residual module accumulating layer by layer with the deepening of the network, which affects the stability of the model training. For this reason, the author proposes a post-normalization technique (res-post--norm) to avoid the activation amplitude of the model being too high. (2) The window size needs to be changed when the pre-trained model is fine-tuned to the downstream task, which easily leads to the problem of model misfit. To solve this problem, the authors introduce a log-spaced continuous position bias (Log-CPB) to realize the smooth transfer of the model between different resolution versions. (3) A self-supervised pre-training method (SimMIM) is used to reduce the dependence on labeled data. Swin Transformerv2 with such large parameters enables it to train images with a resolution of 1536×1536, but it also increases the memory occupation of GPU.

However, the design of the hierarchy and shifted windows is not concise enough. Li et al. (2022) propose ViTDet [66], which can build a complete FPN by upsampling the last feature map of the ViT. This allows the backbone network design to be independent of the object detection task and can be fine-tuned for different downstream tasks. In addition, ViTDet splits the pre-trained backbone network into four sub-modules, and only needs to apply window attention once in the final stage to realize the information interaction of different patches, which is not only concise, but also compatible with the above fine-tuning operations.

2.4. Summary

Table 1 summarizes some representative 2D algorithms. RCNN series algorithms first generate proposals using anchor boxes and then detect them. This more refined two-step process has higher detection accuracy. The YOLO series algorithm directly predicts the filtered anchor box, which is simpler and more efficient. SSD series algorithms can achieve a better trade-off between speed and accuracy and have unique advantages in small object detection tasks with multi-size feature maps. To eliminate some limitations of anchor boxes, the methods represented by FCOS perform pixel-level dense prediction by presetting anchor points, while the methods represented by CornerNet represent the bounding box with some key points. DETR series algorithms still use convolutional networks to extract features and try to use the powerful self-attention mechanism of Transformer network to realize the correlation modeling of global features, while Swin Transformer and PVT directly use pure Transformer structure network model for object detection. Although the performance of such Transformer-Based methods is not as good as CNNs-based methods on some small datasets, they can give full play to their advantages on large-scale datasets.

Table 1
Performance comparison of representative 2D detection models on the MS COCO test.

Classification	Method	Backbone	AP $_{[0.5,0.95]}$ (%)	FPS
Anchor-based	Faster *R*–CNN	VGG-16	21.9	7
	R–FCN	ResNet-101	29.9	9
	Mask *R*–CNN	ResNeXt-101	39.8	11
	YOLO v2	VGG-16	21.6	40
	YOLO v3	DarkNet-53	31.0	35
	YOLO v4	CSPDarkNet-53	43.5	23
	YOLO v7	E-ELAN	51.4	161
	SSD	VGG-16	28.8	19.3
	DSSD	ResNet-101	33.2	5.5
	ASSD	ResNet101	34.5	6.1
	PSSD	VGG-16	33.8	45
Anchor-free	CornerNet	Hourglass-104	42.2	0.9
	CenterNet	Hourglass-104	47.0	2.9
	FCOS	ResNeXt-101	44.7	8.9
	FoveaBox	ResNeXt-101	43.9	7.25
	YOLOX	DarkNet-53	47.4	90.1
	YOLO v6	CSPStackRep Block	52.5	98
Transformer-based	DETR	ResNet-101	43.5	20
	SMCA	ResNet-50	43.7	10
	Deformable DETR	ResNet-50	46.0	18.2
	Sparse-DETR	ResNet-50	46.3	20.5
	Anchor DETR	ResNet50-DC5	44.2	19

By horizontally comparing different series of algorithms, it can be found that the model gradually achieves a higher sense of "end-to-end" detection. The early RCNN divides training, fine-tuning and final classification and regression into multiple stages, which is cumbersome and difficult to optimize. The subsequent Fast *R*–CNN, YOLOv3 and other models put all stages in the same convolutional network, and all network layers can be updated through training, which is faster and more efficient. However, these algorithms need to use post-processing operations such as NMS to filter out redundant prediction results, and NMS is difficult to optimize due to its non-derivable characteristics. Therefore, the subsequent models such as DETR and DeFCN omit the NMS step through the "one-to-one" sample prediction strategy and achieve a higher sense of "end-to-end" detection. This "end-to-end" upgrade makes the detection process simpler, which is of great significance for the optimization of the overall structure of the model, the expansion of downstream tasks, and the improvement of the interpretability of the model.

For the same model, through longitudinal observation of the evolution process of the version, it can be found that we need not only high accuracy to avoid false and missed detection, but also as lightweight as possible to facilitate industrial deployment and real-time detection. Take the YOLO series models as an example: YOLOv1 first proposes the idea of directly predicting the object category and location on the grid cells divided by the feature map, which lays the development foundation and its own advantages of the YOLO series algorithms. YOLOv2 and YOLOv3 add some new technologies, such as Anchor, FPN, BN, etc., which makes the performance and speed of the model more excellent. YOLOv4, v5 and v7 algorithms still use the detection process of anchor box mechanism, but the structure of the detection network and the training and reasoning process are optimized in an all-around way. YOLOX, YOLOv6, *P*P-YOLOE and other algorithms adopt anchor-free, decoupling head mechanism, and lightweight means such as knowledge distillation, which have faster convergence speed, stronger generalization ability, and can be easily deployed on low-resource edge devices.

3. 3D object detection algorithms

2D object detection can only perform classification and localization on the plane, while 3D object detection can use the input sensor data to predict the 3D coordinates, direction, speed and other spatial information of the object, which is the basis of the perception system of autonomous vehicles or intelligent robots. As important components of the perception layer of autonomous driving, the most common on-board sensors include camera and LiDAR. Generally, the point cloud data collected by LiDAR contains rich spatial information and geometric shape information, but it has the following drawbacks: (1) expensive price; (2) sensitive to weather environment; (3) Limited ranging range; (4) The color and texture details of the object cannot be obtained. The image data collected by the camera contains rich semantic features such as color and texture, which can easily be competent for road scene detection such as lane lines and traffic signs, and the price is relatively low. This chapter mainly focuses on 3D object detection algorithms based on image data, which are divided into three categories: monocular-based, stereo-based and pseudo-LiDAR-based, and their chronological overview is shown in Fig. 8.

3.1. Monocular-based

Monocular-based 3D object detection aims to use the mature 2D object detection framework and geometric priors to predict the 3D information of an object from a single image. Generally, it can be divided into two-stage methods and one-stage anchor-based and anchor-free series methods (Fig. 9). Although the detection accuracy of this kind of algorithm is not high, it is favored by industry and academia because of its low cost and high potential research value.

3.1.1. Two-stage

Mono3D [67] proposed by Chen et al. (2016) is a representative work. It uses the assumption that the object is on the ground-plane as a constraint condition, proposes an energy minimization approach to obtain 3D candidate boxes that fit the real size, and then uses multiple geometric priors such as shape and position to filter these candidate boxes. Finally, a convolutional network is used to predict the final 3D result. However, the method of obtaining proposals by exhaustive search in 3D space in Mono3D will inevitably increase the computational cost. Therefore, GS3D [68] proposed by Li et al. (2019) provides constraints to reduce the search space by using 2D bounding boxes predicted by Faster *R*–CNN. The appearance information projected by the 3D bounding box on the plane is used to improve the accuracy of the final classification and regression. Manhardt et al. (2019) proposed ROI-10D [69], which uses the ResNet-FPN network to obtain the 2D detection box of the input image and uses the SuperDepth network to obtain the per-pixel depth value, and then uses the RoI alignment to obtain the fused feature map to participate in the 3D bounding box prediction. RoI-10D also uses a differentiable loss function to align 3D bounding boxes with ground truth boxes, resulting in higher 3D AP compared to Mon3D, and also achieves very accurate results on tasks such as pose estimation and object shape recovery.

Since 3D parameters such as position, size, and rotation angle have different dimensions, the stability of training will be affected if they are put into the same regression loss. For this reason, MonoDIS [70] proposed by Simonelli et al. (2019) decouples the regression loss into multiple parts for training and defines the "sIoU" to further optimize the

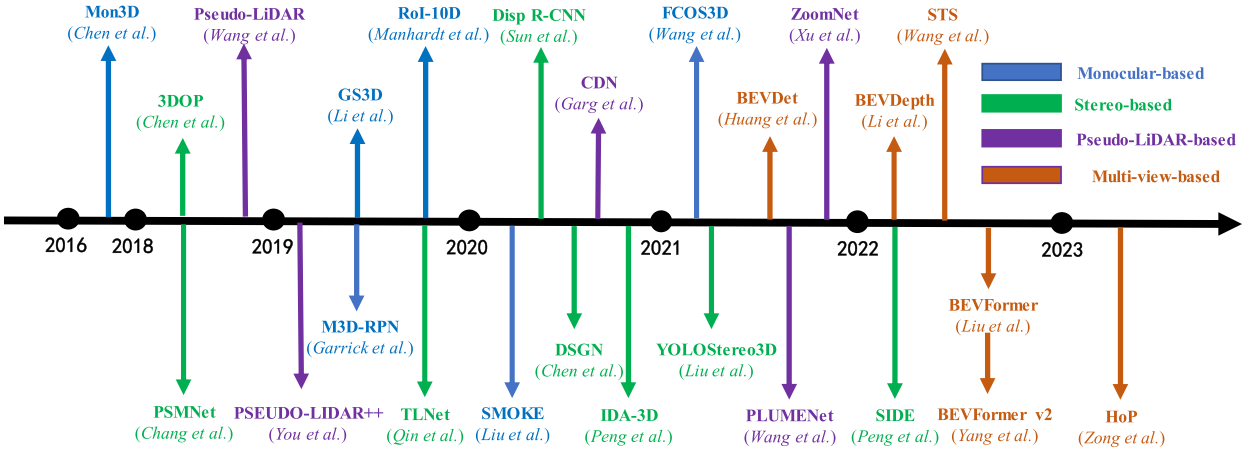

Fig. 8. Chronological overview of the 3D detection methods (author information in parentheses).

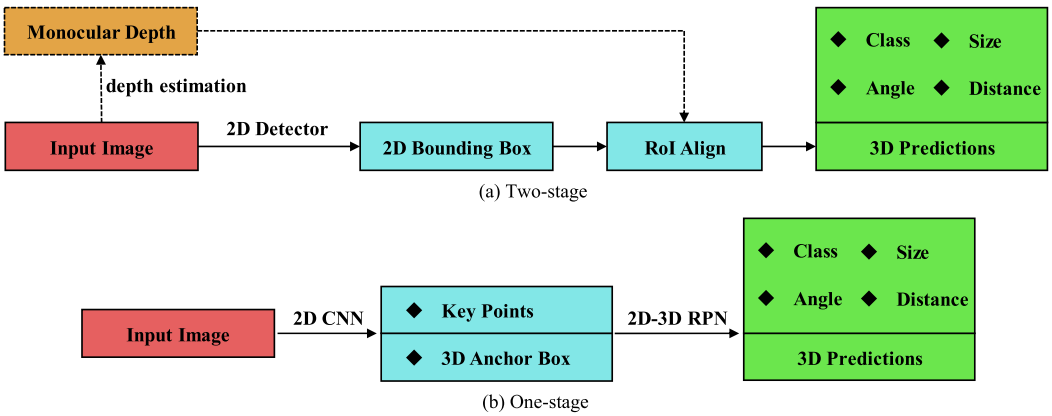

Fig. 9. The flow of the monoculor-based methods. **(a)** The two-stage method first utilizes bounding boxes generated by a 2D object detection model to provide a prior, and then utilizes a specialized 3D prediction head to output the results in the second stage. **(b)** The one-stage method first uses convolutional network to extract features, and then uses 3D, depth anchor boxes, or some key points to output results.

training process. MonoGRNet [71] proposed by Qin et al. (2019) divides the entire detection task into four progressive subtasks to complete, which not only makes full use of the rich semantic features in 2D images, but also greatly improves the inference speed by eliminating the need for generating proposals and post-processing operations. In addition, MonoGRNet also uses an instance depth estimation (IDE) module to accurately estimate the center depth of the object, which is more lightweight than the above SuperDepth network adopted by RoI-10. MonoRCNN [72] proposed by Shi et al. (2021) adopts a geometry-based distance separation method. It improves the accuracy, robustness, and interpretability of distance prediction by decomposing the object distance into physical height and projection height. To improve the reliability of the model in the inference stage, GUPNet [73] proposed by Lu et al. (2021) adopts a geometric uncertainty projection (GUP) module. GUP estimates depth based on the distribution of 3D heights, thus providing reliable confidence for each depth. In addition, GUPNet also adopted a training strategy of hierarchical task learning to effectively cope with the error amplification phenomenon in the projection process, which improved the stability of model training.

3.1.2. One-stage

The M3D-RPN [74] proposed by Garrick et al. (2019) is the first to perform 2D detection and 3D detection in a unified framework. Specifically, M3D-RPN uses two parallel branches to obtain global and local features of an image and generates both 2D and 3D proposals for objects by sharing anchor boxes, which effectively exploits the correlation between 2D scale and 3D depth. To achieve a higher degree of "end--to-end" one-stage detection, Kumar et al. (2021) proposed a differentiable NMS method: GrooMeD-NMS [75], which can continuously update its parameters during training, eliminating the mismatch problem between 3D localization and classification. Luo et al. (2021) proposed M3DSSD [76], which tries to improve the inherent problems of anchor-based methods. It uses the shape alignment and center alignment to ensure the coincidence of the anchor center and the 3D center of the object, which better improves the model accuracy. In addition, an Asymmetric Non-local Attention Block (ANAB) is used to effectively utilize global features and long-range dependencies, which is more lightweight than the early non-local module.

As a typical anchor-free 2D detector, CenterNet [44] does not consider the interference of 2D detectors on 3D detection when expanding on 3D space. To this end, SMOKE [77] proposed by Liu et al. (2020) replaces the 2D detector by estimating the projection of 3D keypoints onto the 2D plane and improves inference speed by parallel prediction branches. Similar to MonoDIS, SMOKE also employs a multi-step decoupling strategy to predict 3D parameters, making model training more efficient. To achieve accurate prediction of truncated objects, Zhang et al. (2021) proposed MonoFlex [78], which decouples the truncated objects and normal objects by using the offset distribution difference between them and uses an edge fusion module to decouple the feature learning and prediction part of truncated objects separately. These decoupling operations facilitate the discriminative treatment of truncated objects by using their special properties, and flexibly predict object depth by uncertainty weighted average in the following. FCOS3D [79] proposed by Wang et al. (2021) is an extension of FCOS algorithm on 3D tasks. It defines the 3D center-ness with a 2D Gaussian distribution based on the 3D-center and improves the recall rate of the model in large vehicle detection scenarios by decoupling 2D and 3D features and adding variability convolution in the backbone network. In addition, some researchers have proposed monocular 3D detection models in application scenarios such as lane line detection [80] and indoor panoramic detection [81], which have achieved good detection results.

3.2. Stereo-based

Stereo-based 3D object detection aims to detect 3D information of an object from two views. Compared with monocular-based 3D detection, these algorithms can obtain additional geometric constraints, to have stronger depth perception and anti-background interference ability. Stereo-based 3D detection algorithms can also be divided into two-stage methods and one-stage methods (Fig. 10).

3.2.1. Two-stage

3DOP [82] proposed by Chen et al. (2018) first generates a pair of 3D proposals from the left and right images, and then uses Fast *R*–CNN that combines context and depth information to predict the result. These authors also try to use LiDAR to generate proposals, which can handle object detection tasks in complex scenarios such as long range and occlusion. However, 3DOP doesn't take full advantage of the sparsity and dense constraints in stereo images, for which Li et al. (2019) proposed another two-stage method: Stereo *R*–CNN [83], which can be regarded as an extension of the Fater *R*–CNN model for 3D tasks, and thus has a faster network architecture than 3DOP. Moreover, Stereo *R*–CNN adds an additional keypoint prediction branch after the stereo RPN module and performs photometric alignment for the left and right images to provide sparse constraints and dense constraints for subsequent 3D prediction respectively, which has higher 3D positioning accuracy.

The above methods use dense pixel-level depth maps to estimate the depth, which does not focus on the specific instance, so it is very redundant and time-consuming. Qin et al. (2019) proposed TLNet [84], which explicitly constructs the inst-level geometric correlation of stereo images by introducing 3D anchor boxes and uses a feature reweighting strategy to make the network pay more attention to the key parts of the

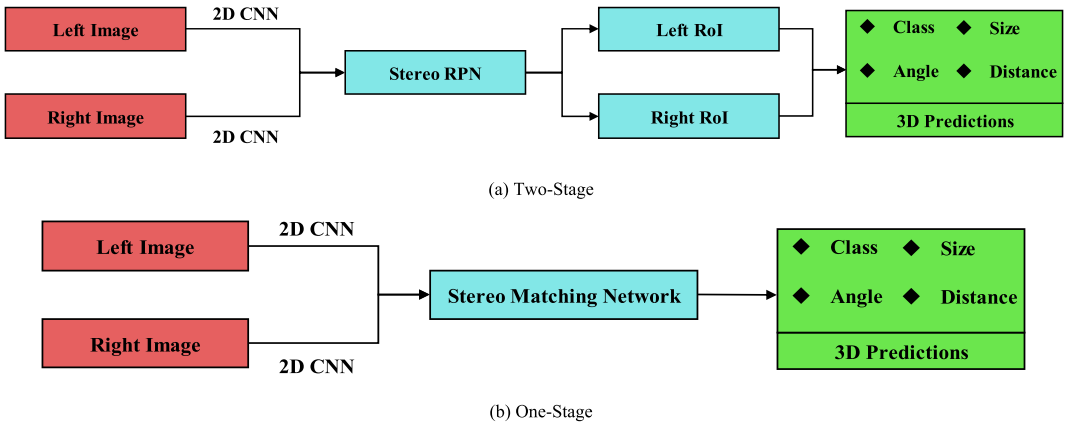

Fig. 10. The flow of the stereo-based methods. Firstly, the convolutional backbone network is used to extract features from the left and right stereo images. For the two-stage method, the left and right RoIs are obtained by means of the stereo RPN module, and then they are fused for prediction. For the one-stage method, the stereo matching network is directly used to fuse the left and right image features for prediction, which is more concise and efficient.

object, so that the model can be more efficient. Similarly, the Disp *R*–CNN [85] proposed by Sun et al. (2020) only estimates the depth of the instance level pixels, and with the help of the shape prior as a guide, it can predict the 3D information of the object more accurately and efficiently.

In 2020, Peng et al. proposed an "end-to-end" 3D inspection model without post-processing: IDA-3D [86], which designs an Instance-Depth-Perception (IDA) module to make the network pay more attention to the global spatial information of the object, reduce the sensitivity of depth error to distance by adaptively adjusting the disparity range, and make the depth estimation result of the instance more reliable by matching cost reweighting. IDA-3D not only has a very lightweight structure, but also has strong robustness in some complex scenes such as long distance and fuzzy occlusion. However, Peng et al. (2022) believe that IDA-3D ignores the sparsity and locality of depth information, which will cause the network to spend too much calculation on the depth estimation of the background region. Therefore, the SIDE [87] proposed by these authors regards the 2D bounding box area as the focus of depth estimation, and only estimates the depth of the object center, which greatly improves the speed of the model. Then, the structure-aware attention module and the matching reweighting strategy are used to make the depth information in the local cost volume more compact, which improves the performance of the model in the case of long distance and occlusion.

3.2.2. *One-stage*

To reduce the computational cost, some researchers have proposed more lightweight one-stage models. The PSMNet [88] proposed by Chang et al. (2018) introduces global context features into images through SPP [21] and dilated convolution modules and improves the utilization of these context features by stacking hourglass convolution modules, which effectively improves the robustness of the model in complex scenes. Based on PSMNet, DSGN [89] proposed by Chen et al. (2020) attempts to further bridge the performance gap between image-based and LiDAR-based 3D models by using stereo features extracted from a Siamese network to construct a plane-sweep volume (PSV). Then, the PSV is transformed into a 3D geometry volume (3DGV) by differentiable warping to encode both high-level semantic features and pixel-level 3D geometric features. Using deep point cloud information as supervision, DSGN outperforms PSMNet by 10% AP in performance. In 2022, these authors also proposed an improved version: DSGN++ [90], which proposes a depth-wise plane sweeping (D-PS) module that can effectively expand the size of the 2D feature map, and adopts a key component called "cyclic slicing" to ensure the consistency of the left and right channels, without increasing the calculation of the 3D feature volume. Similar to PSV, DSGN++ also proposes a Dual-view Stereo Volume (DSV) module to enhance the connection between front view and top view, which can encode semantic, and depth features more effectively. In addition, the Stereo-LiDAR Copy-Paste (SLCP) strategy proposed by DSGN++ solves the problem of class imbalance in the actual detection scene by aligning multi-modal data and improves the utilization of data.

By using a more powerful LiDAR-based detector to guide the learning of stereo-based detectors, the deployment of high-performance detectors on low-resource devices can be accelerated. For this purpose, Guo et al. (2021) proposed LIGA-Stereo [91]. Different from the traditional knowledge distillation, which guides the detection results at the level of detection results, LIGA-Stereo uses feature consistency constraints to guide the learning of geometric features from student to teacher at the intermediate level and enhances the learning of semantic features by adding 2D detection heads. This results in superior performance of LIGA-Stereo over DSGN.

YOLOStereo3D [92] proposed by Liu et al. (2021), abandonds the computationally expensive disparity estimation strategy and attempts to use powerful stereo features to 3D extend the 2D detection framework for more efficient end-to-end 3D detection. To this end, YOLOStereo3D calculates the pixel-level correlation on the feature map by normalized dot product, which can construct a more lightweight cost-volume, and uses a densely connected ghost module and a hierarchical fusion structure to balance the number of channels between stereo features and monocular features to avoid losing features in the fusion stage. On a single GPU, YOLOStereo3D can match the performance of other state-of-the-art models at 10FPS. In the same year, StereoCenterNet proposed by Shi et al. which utilizes some sparse keypoints to regress accurate 3D bounding boxes, has a more streamlined network structure, and improves the recall of the model for occluded objects by introducing a geometric constraint module for photometric alignment.

3.3. *Pseudo-LiDAR-based*

The pseudo-LiDAR-based method aims to use image features to imitate LiDAR point cloud data, and simultaneously use the advanced algorithms of monocular or stereo depth estimation and LiDAR-based, which is a pioneering idea. Pseudo-LiDAR-based 3D detection algorithms can also be divided into monocular input methods and stereo input methods (Fig. 11).

3.3.1. *Monocular input*

Traditional image-based 3D detectors use the depth map in the form of additional channels or simple superposition to assist in predicting 3D information. Wang et al. (2018) believe that this interactive representation of image and depth map does not consider the discontinuities and

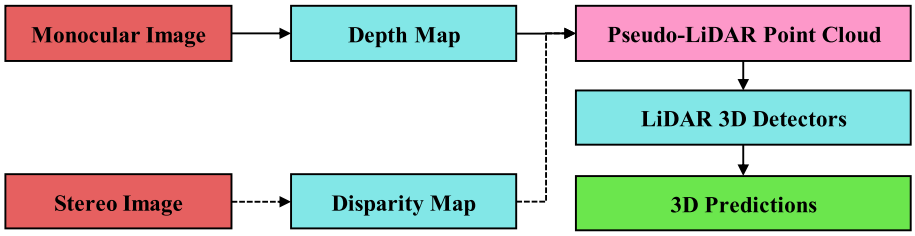

Fig. 11. The flow of the pseudo-LiDAR-based methods. The depth map obtained from the monocular image, or the disparity map obtained from the stereo image are used to generate the pseudo-LiDAR point cloud data, and then the data are input into the LiDAR-based detector to realize 3D object detection.

scale changes of objects in space, so they first proposed Pseudo-LiDAR [93]. It enhances the reliability of image-based depth estimation by mimicking high-quality LiDAR signals. On KITTI's vehicle detection task, Pseudo-LiDAR achieves 45.3% accuracy, which is more than three times better than traditional monocular or stereo3D algorithms. On this basis, Ma et al. (2019) proposed AM3D [94] to further improve the performance of the model by embedding RGB information into the point cloud data. To better integrate the two types of modal information of color and space, these authors also introduced an attention mechanism in the embedding process.

ForeSeE [95] proposed by Wang et al. (2020) focused on the interactive representation of foreground and background pixels. They used different branches to predict the depth of the foreground and background regions respectively and utilized the similarity of each other with the foreground-background sensitivity loss to promote the prediction effect on each other's branches.

To further narrow the performance gap between pseudo-LiDAR-based and LiDAR-based detector, Ye et al. (2020) proposed DA-3Ddet [96], which inputs LiDAR and pseudo-LiDAR data into two branches of a Siamese network and uses domain adaptation techniques to narrow the difference in feature representation between them. In addition, DA-3Ddet introduces a context aware foreground segmentation module (CAFS) to further improve the quality of pseudo-LiDAR data.

3.3.2. Stereo input

To improve the performance of pseudo-LiDAR methods on long-distance detection tasks, You et al. (2019) proposed Pseudo-LiDAR++ [97], which directly predicts depth end-to-end by replacing the disparity cost volume with the depth cost volume, preventing excessive focus on close objects in stereo depth estimation. And a graph-based depth correction algorithm (GDC) was used to correct the quantization error of depth prediction to further improve the detection accuracy.

Similar to ForeSeE, CG-Stereo [98] proposed by Li et al. (2020) also uses separate decoders to predict the depth values of foreground and background pixels and uses confidence estimation as a soft attention mechanism to guide the network to focus on foreground regions with richer depth information, which improves the detection accuracy. OC-Stereo [99] proposed by Pan et al. (2020) uses RoIs and instance segmentation modules to guide the network to focus on foreground object pixels while consciously ignoring background pixel differences, which greatly improves the processing speed of the model. In addition, OC-Stereo also introduces point cloud loss to balance the number of pixels of near and far objects, which is more friendly for distant object detection.

The discrete average disparity estimation leads to the instability of depth estimation, which is more significant on the fuzzy pixels at the object boundary. Therefore, the CDN [100] proposed by Garg et al. (2020) uses a loss function based on Wasserstein distance to replace the traditional regression loss, so that the model can output arbitrary disparity value distribution. The depth value is estimated explicitly, which improves the accuracy and reliability of depth estimation. CDN can be well compatible with some depth estimation networks without increasing the training cost and can also be easily extended to some downstream vision tasks. Also, to improve the disparity estimation network, ZoomNet [101] proposed by Xu et al. (2021) introduces an adaptive instance scaling module to enhance the utilization of texture information in the original image, and a partial position learning module to supplement the shape prior to improve the robustness of the model in the case of occlusion.

All the above pseudo-LiDAR methods train the depth estimation network and the 3D detection network separately, which is not good for optimizing the global performance of the model. Qian et al. (2020) proposed PL-E2E [102], which proposes a depth differentiable Change of Representation (CoR) module to integrate the two sub-networks into an end-to-end framework. However, PL-E2E does not unify the feature representation used by the two networks, and PLUMENet [103] proposed by Wang et al. (2021) solves this problem well. It unifies the two tasks into a metric space by constructing a pseudo-LiDAR feature volume (PLUME), which greatly improves the speed of the model with its shared-based feature computation.

3.4. Multi-view-based methods

Multi-view-based 3D object detection aims to predict object bounding boxes from multi-view camera inputs. BEV has the following advantages.

(1) It is a complete global scene representation, which can clearly present the location and scale of objects, thus overcoming occlusion problems.
(2) It is an ideal link of temporal and spatial information, which is convenient for temporal multi-frame fusion and multi-modal fusion.
(3) It is beneficial to extend to downstream tasks such as object tracking and trajectory prediction.

Therefore, the mainstream idea of such algorithms is to use multi-view images to generate BEV features and then complete 3D prediction, which can be mainly divided into two branches: LSS-based and query-based methods (Fig. 12).

3.4.1. LSS-based

The process of LSS-based method is shown in Fig. 12 (a). This kind of algorithm adopts the Lift-Splat-Shoot [104] paradigm. For the extracted multi-view 2D features, it first lifts them into 3D voxels under the supervision of pixel depth, and then obtains the BEV feature map with the help of the internal and external parameters of the camera, and finally inputs them into the task head to predict the result. BEVDet [105] proposed by Huang et al. (2021) is a representative work of this kind of methods. To prevent overfitting and improve the adaptation of the model in 3D scenes, the authors also propose a matching data augmentation strategy and an upgraded NMS strategy. To improve model performance, Li et al. (2022) proposed BEVDepth [106] by introducing robust explicit depth supervision to optimize depth estimation, and by aligning the coordinates of cone features of different frames with the current coordinate system, GPU is used to parallelize

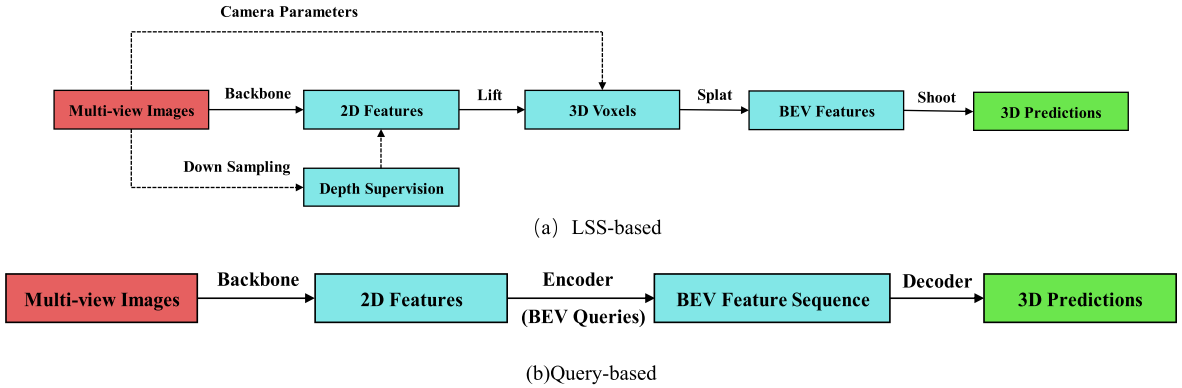

Fig. 12. The flow of the multi-view-based methods.

these features to achieve efficient motion speed estimation. To save computational cost, BEVStereo [107] proposed by Li et al. (2022) not only predicts a monocular depth, but also uses the search space provided by the depth center and range of each pixel to obtain the stereo depth. To optimize depth estimation, STS [108] proposed by Wang et al. (2022) uses the geometric correspondence of adjacent frames to construct stereo matching to estimate depth and fuses it with the original depth estimation result. In addition, STS also improves the problem of too sparse sampling points in close range with a spacing-increasing discretization strategy.

3.4.2. *Query-based*

The process of query-based method is shown in Fig. 12 (b), which uses learnable queries to encode 2D image features into BEV feature sequences, and then outputs 3D prediction results through the decoder. This algorithm can effectively fuse spatio-temporal information, and then accurately judge the motion state of the object. BEVFormer [109] proposed by Li et al. (2022) is a representative work, which enables each BEV query to interact only with part of the associated regions on the image through deformable sparse attention, avoiding the high dependence on camera parameters in LSS-based methods. In addition, BEVFormer also uses a temporal self-attention module to extract temporal information from historical BEV features, but simply adopting the inheritance fusion method is easy to lead to forgetting. To this end, BEVFormer v2 [110] proposed by Yang et al. (2022) leverages longer time series information by aligning history with current BEV features and cascading channel dimensions. In addition, BEVFormer v2 adds a perspective 3D detection head to generate high-quality proposals and use these proposals to assist subsequent object queries to accelerate model convergence.

The above methods only fuse a few time series frames, and the use of low-resolution feature maps affects the fusion effect of multi-view stereo matching. To this end, SOLOFusion [111] proposed by Park et al. (2022) maximally balances the impact of spatial resolution and temporal differences on localization potential by exploiting the complementary effects of high-resolution short-term frames and low-resolution long-term frames. In addition, these authors also proposed a Gaussian-SpacedTop-k deep sampling method to save computational overhead. Zong et al. (2023) proposed HoP [112], which uses long and short-term temporal decoders to generate pseudo-BEV features for adjacent frames to predict object position and motion information in historical frames, thereby improving the existing BEV feature learning effect. As a plug-and-play improvement of the training phase, HoP incurs no additional inference overhead but can bring significant performance gain.

3.5. *Summary*

Table 2 summarizes some representative 3D algorithms. As a downstream task of 2D object detection, image-based 3D object detection can be regarded as its extended application in 3D space, so the design paradigms including anchor box, keypoint and NMS post-processing continue to be used in 3D algorithms. Moreover, for monocular-based 3D detection, an intuitive idea is to first use a mature 2D detector to obtain high-quality candidate boxes. The 3D information of the object is then predicted on these candidate regions of interest. In

Table 2
Performance comparison of representative 3D detection model performance on KITTI car and nuScenes tests.

Classification	Method	Inference Time (ms)	AP_{3D} (%)			NDS (%)
			Easy	Moderate	Hard	
Monocular-based	MonoRCNN	70	18.36	12.65	10.03	–
	MonoDIS	100	10.37	7.94	6.40	38.4
	FCOS3D	–	–	–	–	42.8
	SMOKE	30	14.03	9.76	7.84	–
Stereo-based	Stereo *R*–CNN	300	47.58	30.23	23.72	–
	DSGN	670	7.64	4.37	3.74	–
	YOLOStereo3D	80	65.68	41.25	30.42	–
	SIDE	260	61.22	44.46	37.15	–
Pseudo-LiDAR-based	Pseudo-LiDAR	400	54.53	34.05	28.25	–
	DA-3Ddet	–	11.5	16.8	8.9	–
	Pseudo-LiDAR++	400	61.11	42.43	36.99	–
	CDN	600	74.52	54.22	46.36	–
Multi-view-based	BEVDet	–	–	–	–	48.2
	BEVDepth	–	–	–	–	60.9
	BEVFormer	25	–	–	–	56.9
	BEVFormer v2	–	–	–	–	63.4
	HoP	–	–	–	–	68.5

recent years, the research on monocular-based 3D algorithms mainly focuses on how to unify the 2D detector and the 3D detector into a framework for end-to-end training. In addition, since many 3D parameters are not in the same metric space, some researchers try to decouple their regression losses to improve the training effect.

However, there are inherent defects in estimating the depth of an object from a single image, especially for long-distance and occlusion detection, the effect is not satisfactory. Therefore, in recent years, a mainstream idea in industry and academia is stereo-based 3D algorithm. For this kind of algorithms, the predicted depth information is more accurate and reliable because the photometric calibration of the left and right images can be relied on to provide dense geometric constraints. Although the additional view provides more foreground information, it also brings more redundant information. Therefore, the research on stereo-based 3D algorithms in recent years mainly focuses on how to make the network focus on the foreground area with richer depth information and improve the accuracy and efficiency of 3D prediction by training the model end-to-end.

However, both monocular-based and stereo-based 3D detectors still have a large performance gap with LiDAR-based 3D detectors, which is caused by the differences in the characteristics of the data itself. Therefore, some researchers have proposed pseudo-LiDAR methods to improve the representation of depth estimation, using pseudo-LiDAR point cloud data generated by depth or disparity maps, to express the original distribution of information more explicitly. In recent years, research on pseudo-LiDAR methods has mainly focused on how to distribute and interact with the data, including input images and depth maps, spatial and semantic, foreground and background, and so on. This kind of method bridges the performance gap between image-based and LiDAR-based 3D detection to the greatest extent and is a novel idea with great research value.

Since temporal information is closely related to spatial and motion information, introducing temporal information into the 3D perception framework, and promoting the fusion of temporal and spatial information are the main research contents of multi-view-based methods in recent years. How to better associate LSS-based and Query-based methods and fuse long-term and short-term temporal information to capture complete object motion information is the key focus.

4. Related techniques

The implementation of object detection faces many challenges.

- Relying on manually designed detection networks cannot balance detection accuracy and speed well.
- High-performance detectors often have huge parameters, which cannot be well deployed on low-resource devices.
- The distribution of data domain is different due to different geographical environment and weather.

To solve these problems, some researchers have tried to use some related technologies to solve them, including neural network search (NAS), knowledge distillation (KD), domain adaptation (DA), and graph neural networks (GNN) and generative adversarial networks (GAN), which are widely used in recent years. This section will introduce them.

4.1. Neural network search

The detection network of traditional models often relies on manual design, which cannot balance accuracy and speed well. In recent years, the rise of neural network search (NAS) is expected to solve this problem by designing a search strategy in the specified search space, which can automatically iterate to obtain the most efficient network model.

DetNAS [113] proposed by Chen et al. (2019) was the first to apply NAS technology to object detection, but it was limited to searching the backbone network. OPANAS [114] proposed by Liang et al. (2021) uses six heterogeneous information paths to build the search space and finds the optimal path aggregation architecture through an evolutionary algorithm. Compared with the previous NAS-FPN [115], it can search the FPN structure with stronger performance faster. The Hit-detector [116] proposed by Guo et al. (2020) firstly selects appropriate sub-search spaces for backbone, neck, and head by group sparsity regularization. Then, the optimal structure is obtained in the respective sub-search space. This hierarchical search method can not only reduce the search time, but also reduce the memory consumption. *PP*-PicoDet [117], proposed by Yu et al. (2021), uses the more lightweight CSP-PAN as the neck and obtains a more streamlined backbone network through a simple channel-wise search.

With the expansion of the scale of the object detection model, the introduction of NAS can avoid manual intervention in hyperparameter design, which is of great significance for the improvement of model generalization ability and other aspects. How to design more efficient search strategies and low-cost evaluation strategies will be the main research content of NAS applied to object detection in the future.

4.2. Knowledge distillation

Knowledge distillation (KD), which is popular in recent years, is a new method of model lightweight, which transfers the knowledge of the trained teacher network to the student network, so that the lightweight student network has the performance close to the teacher network. This technique facilitates the deployment of high-performance object detectors on low-resource mobile devices.

Hinton et al. [118] first applied KD to object detection to improve the performance of the student detection network by using three loss functions to extract the backbone network, classification head and regression head respectively. Wang et al. (2019) believed that learning all the features on the feature map without distinction would affect the performance of the student model and proposed a method of fine-grained feature limitation (FGFI) [119]. Only the knowledge-intensive positions of anchors near the ground truth are distilled. However, FGFI simply believes that background information contains too much noise and does not consider the importance of background information for learning features of student networks. Guo et al. (2021) found that using background information to learn features can also distill student networks with excellent performance, and proposed DeFeat [120] for this purpose. It decouples FPN and RoI alignment features into foreground and background information and can adaptively embed this more valuable information into the detector to improve performance. Similarly, FGD [121] proposed by Yang et al. (2022) separates foreground and background information through local distillation, enabling students to focus on key pixels and channels in the teacher network, and complete the lost global information through global distillation. To consider the valuable instance-level information, Chen et al. (2021) [122] designed a graph structure to represent the relationship between objects and used an adaptive background loss weight to treat objects and background information differently. GID [123] proposed by Dai et al. (2021), uses an instance selection module (GISM) to integrate knowledge based on features, relations, and responses for distillation.

Using the attention mechanism to guide the student network to pay attention to more valuable knowledge in the teacher network, and making full use of the relationship between the foreground and background in the image are the main ideas of applying knowledge distillation technology to object detection tasks in recent years.

4.3. Domain adaptation

Domain adaptation (DA) belongs to a kind of transductive transfer learning, which is suitable for the situation that the data distribution of the source domain and the target domain are different. In fact, DA is very important in object detection tasks, because in many industrial scenes (especially disaster accidents), image data is not only scarce, but also

often too different from the data used for model training. Therefore, DA technology is needed to improve the cross-domain generalization performance of object detection models.

Chen et al. (2018) [124] first applied unsupervised domain adaptation to object detection models. They used Fast *R*–CNN as a baseline, trained a domain classifier on feature maps and proposals, and used adversarial training to minimize the domain adaptation loss. The consistency regularization loss term is added to further align the image-level and instance-level domain classification results. However, the proposals generated by RPN may be inaccurate, and blindly aligning these regions will only be countereffective. Zhu et al. (2019) [125] believe that domain adaptation focuses on local feature alignment, so they first use a region mining module to identify high-quality local proposal regions. These regions are then aligned with the help of GAN. The SAPNet [126] proposed by Li et al. (2020) utilizes a task-guided spatial attention mechanism to capture multi-scale contextual information, enabling the model to focus on the most discriminative semantic regions. Similarly, Inoue et al. (2018) [127] and Xu et al. (2022) [128] adopted progressive weak supervision and hybrid supervision strategies to reduce domain shift, respectively, and achieved good performance in a variety of vision task scenarios.

Therefore, using GAN, attention mechanism and unsupervised, weak supervision and other technical means to reduce the distribution shift between the source domain and the target domain is a mainstream research idea of domain adaptation object detection in recent years.

4.4. Graph neural networks

In recent years, Graph Neural Networks (GNN) are the combination of graph data and deep neural networks. They have powerful reasoning ability and interpretability, which are very suitable for processing point cloud data for 3D object detection and have considerable research prospects.

Point-GNN [129] proposed by Shi et al. (2020) is the first to encode 3D Point cloud features with graph structure and uses an automatic registration mechanism to reduce the sensitivity to geometric transformation. In addition, these authors propose a box merging and scoring operation to deal with redundant prediction boxes, which has high detection accuracy. Najibi et al. (2020) proposed another GNN-based one-stage 3D detector: DOPS [130], which extracts features from a voxelized point cloud through a 3D sparse U-Net. DOPS exploits the latent shape information on the existing dataset to obtain the true 3D shape of the object in a weakly supervised learning manner, which has good versatility in both indoor and outdoor scenes. Similarly, Man et al. (2021) [131] and Ansari et al. (2022) [132] introduce distance perception and angle perception respectively when encoding point clouds and have excellent performance on KITTI.

GNN architecture can encode 3D point clouds more compactly, and it has excellent performance when applied to 3D object detection models. However, the high computational cost is a great challenge for real-time detection, so promoting the implementation of such algorithms in industrial applications will be a hot research content in the future.

4.5. Generative adversarial networks

As one of the most promising methods of unsupervised learning, generative adversarial networks (GAN) has been widely used in the field of object detection in recent years by gradually competing between two components: generator and discriminator to obtain optimized output results.

Wang et al. (2017) [133] took Fast *R*–CNN as the baseline, introduced the idea of GNN to better use data to improve the generalization ability of the model, and cascaded two adversarial networks (ASDN and ASTN), to make the network learn occlusion and deformation samples respectively, to perform more robust in these complex scenes. Li et al. (2017) proposed a perceptual GAN [134] model that uses a generator to generate super-resolution representations for small objects, showing more accurate results for small object detection on traffic sign and pedestrian detection datasets. Similarly, SOD-MTGAN [135] proposed by Zhang et al. (2018) uses the upsampling of the generator to generate super-resolution representations for blurred images, while the authors propose a multi-task discriminator mechanism to backpropagate the classification and regression losses to the generator to facilitate its better recovery of image details.

Using GAN to solve the problems of occlusion, deformation, blur, and other problems that may exist in the image from the data level is the main means of applying this deep learning model to object detection, and it also provides a valuable research idea for researchers.

5. Datasets and evaluation metrics

5.1. Datasets

Due to the fact that the training and testing evaluation of object detection networks need to be conducted on some large datasets, Table 3 provides a brief introduction to some common 2D datasets, including PASCAL VOC [136], ImageNet [137], MS COCO [138], and Open Images [139]. Pascal VOC (Pascal Visual Object Classes) is one of the most significant computer vision competitions, held annually from 2005 to 2012, with two versions of dataset for object detection: VOC 2007 and VOC 2012. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) was held annually from 2010 to 2017, using ImageNet dataset for object detection. MS COCO is a large-scale dataset covering multiple tasks such as object detection and segmentation. It is the most authoritative dataset in the field of object detection, and has been held every year since the first version was released in 2014. The competition for Open Images Detection (OID) began in 2018, with a larger dataset and a more diverse scene and labeled classes.

In recent years, 3D object detection datasets that have been publicly released can be classified into two categories: indoor and outdoor scenes. Table 4 provides a brief introduction to some of the datasets. The earliest KITTI [140] dataset collected a large amount of real data in German cities, rural areas, and on highways using a vehicle-mounted camera and LiDAR and provided corresponding 3D annotation paradigm. The subsequent Waymo Open [141] and nuScenes [142] datasets are also large-scale 3D datasets for autonomous driving scenes but have greatly improved in the diversity of collection scenes and overall data size. In 2021, Carnegie Mellon University introduced AIODrive [143], which uses the simulator CARLA [144] to generate a large amount of multimodal data for autonomous driving scenes at extremely low cost. Some researchers have also introduced 3D object detection datasets for indoor scenes, including ScanNet [145], SUN RGB-D [146], NYU Depth V2 [147], etc. Since the detection range is limited to a regular and narrow room, class annotation is easier and data quality is higher.

Table 3
The introduction to partial 2D object detection datasets (The numbers in parentheses represent the number of labeled instances).

Dataset	Classes	Train	Val	Test
PASCAL VOC 2007	20	2501 (6301)	2510 (6307)	4952 (14,976)
PASCAL VOC 2012		5717 (13,609)	5823 (13,841)	10,991
ILSVRC 2014	200	456,567 (478,807)	20,121 (55,502)	40,152
ILSVRC 2017		456,567 (478,807)	20,121 (55,502)	65,500
MS COCO 2015	80	82,783 (604,907)	40,504 (291,875)	81,434
MS COCO 2017		118,287 (860,001)	5000 (36,781)	40,670
OID 2018	>600	1,743,042 (14,610,229)	41,620 (204,621)	125,436 (625,282)

Table 4
The introduction to partial 3D object detection datasets.

Dataset	Indor	Size					Diversity			Benchmark	Locations
		Images	Train	Val	Test	3D Boxes	Scenes	Classes	Night/Rain		
KITTI 3D	No	15 K	7418 × 1	–	7518 × 1	200 K	–	8	No/No	Yes	Germany
nuScenes	No	1.4 M	28,130 × 6	6019 × 6	6008 × 6	1.4 M	1000	23	Yes/Yes	Yes	SG, USA
Waymo Open	No	1 M	122,200 × 5	30,407 × 5	40,077 × 5	12 M	1150	4	Yes/Yes	Yes	USA
AIODrive	NO	250 K	–	–	–	26 M	250	–	Yes/Yes	No	–
NYU Depth V2	Yes	1449	795	–	654	35,064	464	26	No/No	No	USA
SUN RGB-D	Yes	10,335	5285	–	5050	64,595	–	–	No/No	Yes	USA
ScanNet	Yes	2.5 M	749,768	–	150,000	36,213	1513	21	No/No	Yes	USA

Compared to the complex and variable outdoor detection, this visual task has lower difficulty.

5.2. Evaluation metrics

To measure the performance of algorithms, many datasets in the early days launched corresponding 2D evaluation metrics, generally including two major categories: accuracy metrics, model complexity and speed metrics. For 3D object detection scenes, it is necessary to detect attributes such as 3D position, size, orientation, and speed, and therefore a more complex evaluation system is needed. Currently, KITTI, NuScenes and other datasets have launched corresponding evaluation metrics. Here, we briefly introduce some commonly used 2D object detection evaluation metrics and use the most widely used KITTI and NuScenes datasets as examples to introduce some commonly used 3D evaluation metrics.

5.2.1. Accuracy metrics

The basic metrics for measuring detection accuracy include Precision (P), Recall(R), and Intersection over Union (IoU). The comprehensive metrics mainly include Average Precision (AP) and Overall Average Precision (mAP). Fig. 13 shows a schematic diagram of common statistical metrics and IoU.

(1) *Basic metrics*: IoU denotes the ratio of the intersection and union of the estimated bounding box P_a and the real box G_a (Equation (1)); P denotes the ratio of true positive samples in the positive samples predicted by the model (Equation (2)); R denotes the ratio correctly predicted by the model in the total positive samples (Equation (3)).

$$\mathrm{IoU} = \frac{P_a \cap G_a}{P_a \cup G_a} \tag{1}$$

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

If the IoU is higher than a predefined threshold, the prediction box will be judged as a positive sample (foreground) by the model, otherwise it will be judged as a negative sample (background); P and R are generally negatively correlated, which can be intuitively expressed as a *P*–R curve.

(2) *Comprehensive metrics*: AP denotes the average precision value of the model under each recall value, that is, the area under the *P*–R curve, which can comprehensively measure the detection accuracy of the model in a certain category. mAP denotes the average AP value of all categories, which is used to measure the detection accuracy of the model in multi-category detection tasks. The equation is as follows:

$$\mathrm{AP} = \int_0^1 \mathrm{P}(r)\mathrm{d}r \tag{4}$$

$$\mathrm{mAP} = \frac{\sum_{s=1}^{S} \mathrm{AP}(s)}{S} \tag{5}$$

Where P(r) denotes the precision value when the recall rate is r, S denotes the total number of classes detected, and AP(s) denotes the AP value of the model on the *S*-th class. The calculation of mAP needs to specify the threshold of IoU. A higher threshold of IoU means that the model has stricter requirements on positioning, which can improve the quality of positive samples to a certain extent, but also increase the missed detection rate and further aggravate the imbalance of positive and negative samples. However, low threshold will also increase the false detection rate due to the decrease of the overall quality of positive samples. Therefore, the threshold is set to 0.5 in VOC and other datasets, but this fixed threshold setting lacks universality for multi-scale target detection. Therefore, in COCO dataset, the threshold value of IoU ranges from 0.5 to 0.95 at an interval of 0.05, and the mAP obtained under all thresholds is averaged to obtain the final result (see Fig. 14).

5.2.2. Complexity and speed metrics

In many practical application scenarios, it is necessary for the model to realize real-time object detection to avoid possible security risks and data processing delays. The detection speed is not only directly related to the computing power of the hardware device, but also closely related to the design of the network structure, the number of layers and the parallelism. The more complex the model is, the lower the detection speed will be.

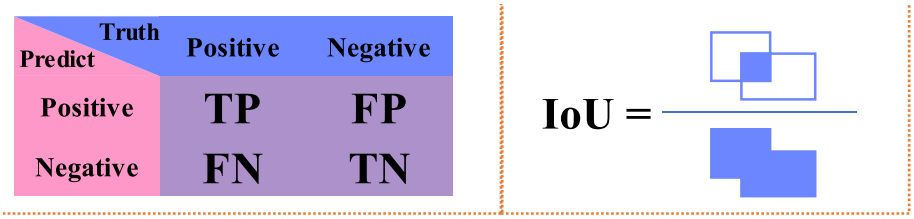

Fig. 13. On the left is the confusion matrix, which is used for statistical objective predictions. When the prediction is a positive sample, TP and FP represent the correct and wrong prediction, respectively. When the prediction is a negative sample, TN and FN represent the cases where the prediction is correct and wrong, respectively. On the right is the calculation diagram of IoU.

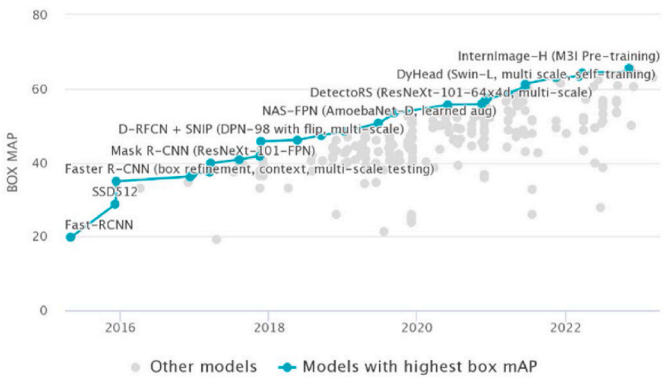

Fig. 14. Models with highest box mAP on COCO test-dev (From https://paperswithcode.com/sota/object-detection-on-coco).

(1) *Parameters/FLOPs*: Parameters can visually represent the volume of a model; FLOPs are the floating-point operations. The size of FLOPs is related not only to Parameters but also to the size of the input image, both of which are used to measure the complexity of the model. FLOPs are easily confused with FLOPS, which stands for floating-point operations per second and is a common measure of hardware power.
(2) *Latency/FPS*: Latency is the network forward propagation time, and FPS is the number of frames processed per second, both of which are used to measure the detection speed of the model. Among them, FPS is the most used metric to measure the detection speed in target detection. The larger the FPS value, the faster the detection speed of the model.

Improving the detection accuracy of the algorithm is of great significance to prevent missed detection and false detection. However, the high accuracy of the model often requires huge model parameters, which not only occupies computer resources, but also greatly affects the operating efficiency of the model. Therefore, the lightweight of object detection models is of great significance for improving model deployment efficiency, detection speed, and enhancing model interpretability.

5.2.3. KITTI metrics

In KITTI dataset, IoU, AP, and mAP in 3D object detection metrics are calculated in a similar way as in the above 2D scene. However, because three ways of calculating IoU are proposed, the calculation of AP and mAP are also changed accordingly. In addition, KITTI also proposes an index (AOS) to measure the directional consistency between the estimated 3D bounding box and the ground truth.

(1) The three definitions of IoU include IoU_{BBOX}, IoU_{BEV} and IoU_{3D}, which respectively represent the IoU of the predicted box and the ground truth in 2D, BEV (bird's eye view) and 3D cases.
(2) Based on the above three kinds of IoU, there are three kinds of AP: AP_{BBOX}, AP_{BEV}, AP_{3D}. The AP is calculated using the following formula :

$$AP|_R = \frac{1}{|R|}\sum_{r\in R}\max_{\tilde{r}:\,\tilde{r}\geq r}\rho_{\text{interp}} \tag{6}$$

where the interpolation function ρ_{interp} is $\max_{\tilde{r}:\,\tilde{r}\geq r}\rho(\tilde{r})$ and $\rho(r)$ represents the precision value when the recall is r.

Initially KITTI3D sets eleven equally spaced recall levels, that is, R_{11} = {0,0.1, 0.2, …, 1}. When r = 0, it is 1, which is obviously not a reasonable measure of the quality of the algorithm on tasks with different difficulty levels. Simonelli et al. [70] proposed R_{40} = {1/40, 2/40, 3/40, …, 1}, which can average the precision results over 40 recall locations, eliminating the above problem.

(3) *AOS*: AOS is the average orientation similarity, the larger the value is, the closer the rotation direction of the estimated 3D bounding box is to the ground truth, and the equation is as follows:

$$AOS = \frac{1}{|R|}\sum_{r\in R}\max_{\hat{r}:\,\hat{r}>r}s(\hat{r}) \tag{7}$$

where r is the recall rate, s(r)∈[0,1] is the orientation similarity, which is the cosine distance normalized between all the estimated results and the ground truth, and the equation is as follows:

$$s(r) = \frac{1}{|D(r)|}\sum_{i\in D(r)}\frac{1+\cos\Delta_\theta^{(i)}}{2}\delta_i \tag{8}$$

Where D(r) is the set of all positive samples under recall r, $\Delta_\theta^{(i)}$ is the angle difference between the estimated 3D bounding box and the ground truth of object i, δ_i is used to avoid repeated detection of object i (When i is a positive sample (IoU >0.5), $\delta_i = 1$, otherwise $\delta_i = 0$).

5.2.4. NuScenes metrics

To better measure the localization accuracy, nuScenes uses the center distance d of the 2D plane instead of the previous IoU for the threshold matching of AP and mAP. In addition, the nuScenes dataset defines a set of TP metrics to measure the position, size, orientation, attributes, and speed of the predicted target, and proposes a comprehensive index (NDS) based on this.

(1) *mAP*: To reduce the impact of noise in low recall and low precision regions, points with P and R values less than 10% are ignored in the calculation of AP values, and mAP is calculated as follows:

$$mAP = \frac{1}{|C||D|}\sum_{c\in C}\sum_{d\in D}AP_{c,d} \tag{9}$$

where C denotes the class set and D = {0.5, 1,2,4} is the matching threshold of the 2D center distance d.

(2) *mTP*: TP (true positive) metrics are defined as follows:
- Average Translation Error (ATE) is the 2D Euclidean distance of the object center (units in meters).
- Average Scale Error (ASE) is the 3D IoU error (1-IoU) after alignment of the object center and orientation.
- Average Orientation Error (AOE) is the smallest yaw angle difference between prediction and ground truth (units in radians).
- Average Velocity Error (AVE) is the absolute velocity error as the L_2 norm of the velocity differences in 2D (units in meter/second).
- Average Attribute Error (AAE) is defined as 1 minus attribute classification accuracy (1−acc).

These five TP metrics are calculated when the center distance d = 2 m. And if the recall rate is less than 10%, the TP metric is set to 1. Based on the above TP metrics, the mean TP metric (mTP) over all object categories is formulated as follows:

$$mTP = \frac{1}{|C|}\sum_{c\in C}TP_c \tag{10}$$

where TPc denotes the five TP metrics of category c.

(3) *NDS*: NDS (nuScenes detection score) is a comprehensive index that integrates the above mAP and mTP, and the formula is as follows:

$$NDS = \frac{1}{10}\left[5mAP + \sum_{i=1}^{5}(1-\min(1,mTP_i))\right] \tag{11}$$

where $mTP_i \in [0,1]$ denotes the above five TP metrics (see Fig. 15).

6. Applications

Object detection technology has a wide range of applications in the real world, but according to different task scenarios, there are different difficulties and challenges. This section will introduce the latest research on object detection in four application scenarios: face detection, pedestrian detection, remote sensing image detection, and medical image detection.

6.1. Face detection

Face detection is designed to recognize and locate the corresponding face from the input image or video frame, which is a key component of face recognition task. Difficulties encountered in this task scenario include facial expression and skin color differences, scale changes, occlusion, etc. Dooley et al. [148] (2022) investigated these robustness differences and proposed a corresponding benchmark face detection system.

Wang et al. (2023) proposed EfficientFace [149], which combines multilevel features through a symmetrical bi-directional feature pyramid network (SBiFPN), and enhances robustness in the case of face scale difference and occlusion respectively through a receptive field enhancement module and an attention module. In the task of synthetic face detection, CYBORG [150] proposed by Aidan et al. (2023) effectively assists the model to distinguish synthetic images by introducing human saliency awareness. Piland et al. [151] (2023) proposed a directed region of interest diminution (DROID) method to automatically reduce saliency entropy by combining it with CYBORG to ensure low entropy of the model focus. The authors conducted experiments to verify the effectiveness of this low entropy for the model generalization performance.

Although the development of face detection technology has brought convenience to mobile payment, check-in, access control and other services, there are also huge security vulnerabilities. Muhammad et al. [152] (2023) introduced 2D image morphing to generate large amounts of data and improved the performance of the face presentation attack detection (PAD) models through an ensemble stack-based domain generalization technique. Similarly, Dwivedi et al. [153] (2023) proposed an ensemble Xai method based on stacked ensemble learning techniques to achieve more efficient and interpretable morphing face detection. Due to privacy and legal constraints, real facial PAD data is scarce. Fang et al. (2023) proposed a large-scale synthetic facial PAD dataset: SynthASpoof [154]. Kong et al. [155] (2023) optimize the Low-Rank Adaptive (LoRA) module under the joint supervision of cross-entropy and single-center loss with ViT backbone network, which improves the generalization performance of the model while reducing the number of parameters.

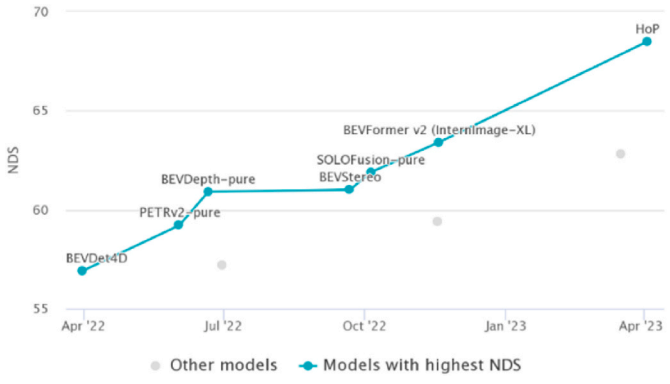

Fig. 15. Models with highest NDS on nuScenes Camera Only (From https://paperswithcode.com/sota/3d-object-detection-on-nuscenes).

6.2. Pedestrian detection

Pedestrian detection has a wide range of applications in the fields of automatic driving, security and criminal investigation. The difficulties faced by it include: pedestrian appearance changes, occlusion and posture changes, and high real-time requirements. Due to the lack of strict definition of different occlusion conditions, the performance of different algorithms is highly subjective. Therefore, Gilroy et al. [156] proposed an objective benchmark to evaluate the comprehensive performance of pedestrian detectors under different occlusion degrees.

In view of the shortcomings of current pedestrian detectors that overfit a specific dataset, Hasan et al. [157] (2022) proposed a progressive fine-tuning strategy to improve the cross-dataset generalization performance of the model. For pedestrian detection with different appearances, Lin et al. [158] (2022) adopted a contrastive learning training strategy to reduce the semantic feature distance of pedestrians with different appearances. Booster-SHOT [159] proposed by Hwang et al. (2022) is an efficient end-to-end multi-view pedestrian detector, which uses a homography attention module to improve the detection effect under occlusion. However, the training data used for such multi-view detectors are usually from a single scene, fixed position camera. To this end, Vora et al. [160] (2022) proposed a Generalized MVD dataset that contains data collected from multiple scenes and multiple cameras, which can make the model have stronger generalization performance.

For pedestrian detection in heavily occluded environments in the wild, Zhang et al. (2022) proposed FC-Net [161], which uses a self-activation module and a feature calibration module to enhance the features of visible parts and suppress the features of occluded regions. Liu et al. (2023) proposed VLPD [162], which adopts a vision-semantic segmentation method to obtain accurate context information, and uses contrastive learning to make the model further distinguish it from pedestrian information, and performs well on the detection task of heavy occlusion.

6.3. Remote sensing image detection

Remote sensing image detection is widely used in military reconnaissance, urban planning, traffic navigation and other fields. At present, there are difficulties in remote sensing image detection such as huge resolution, occlusion and domain adaptation.

To address the lack of small object detection datasets in remote sensing images, Wang et al. [163] (2021) proposed a large-scale Remote Sensing Super-Resolution Object Detection (RSSOD) dataset. RSSOD highlights small target samples and provides advanced performance benchmarks. Shamsolmoali et al. (2022) proposed a high-performance small object detector: IPSSD [164], which utilizes an image pyramid network (IPN) and a rotational pooling layer to extract and process horizontal and vertical region proposals respectively, and a feature fusion network (FFN) to fuse the multi-scale IPN layers.

Change detection (CD) of remote sensing images aims to identify the changed regions of remote sensing images at the same location and at different times. Yan et al. (2023) proposed a fully Transformer network (FTN) [165], which exploits the powerful extraction ability of transformers for global features to obtain complete CD regions and boundaries. Ye et al. (2023) proposed AFCF3D-Net [166], which utilizes 3D convolution and adjacent-level feature cross-fusion (AFCF) modules to fuse bi-temporal image features and multi-level features, respectively, with high accuracy on CD tasks. Liu et al. proposed LSNet [167], which drastically reduces the model parameters at the cost of minor accuracy loss by combining depthwise separable atrous convolution with refined Siamese feature fusion.

Saliency detection (SOD) of remote sensing images aims to identify the most compelling objects or regions in remote sensing images. Lin

et al. (2022) proposed AGNet [168], which utilizes semantic and contextual attention modules to pinpoint the location of salient objects, and further optimizes the prediction results through a self-refinement module (SRM). Li et al. (2023) proposed SeaNet [169], which uses a dynamic semantic matching module (DSMM) to process high-level features and obtain the location information of salient objects, and uses an edge self-alignment module to optimize the edge detail information. SeaNet not only has the most advanced detection accuracy, but also has only an amazing 2.76 M parameters, which is very friendly for practical application deployment.

6.4. Medical image detection

The detection of tumors, nodules, fractures and other medical images can help doctors make more accurate and efficient diagnosis, and then improve the level and quality of medical treatment. At present, the main difficulties of medical image detection are the large variation of appearance factors such as size and shape, and the cumbersome data annotation.

Zhang et al. [170] (2020) proposed an object detector for human pancreatic tumors, which effectively solves the problem of small tumor size by enhancing low-level features with the FPN and utilizing adaptive feature fusion and dependency computing modules to obtain contextual features and global features respectively. Shuvo et al. (2023) proposed an end-to-end cascaded tri-stage framework (CTSF) [171] for lung cancer detection. It uses an improved 3D Res-Unet algorithm for segmentation to obtain lung regions, then uses YOLOv5 to detect lung nodules, and finally uses an improved ViT network to classify cancerous nodules. To address the lack of medical image annotations, Bai et al. [172] (2023) proposed an end-to-end teacher-student detection framework that can efficiently train high-precision models by mining unlabeled lesions on small batches.

7. Conclusion

This paper focuses on the deep learning algorithms of image object detection in 2D and 3D scenes in recent years, and summarizes the advantages, disadvantages and development trends of each algorithm series. In addition, this paper also introduces related technologies in object detection, datasets and evaluation metrics, and the latest progress of application scenarios. At present, the research on image-based 2D object detection is in the ascending-trend, and Transformer opens a new page for it. Moreover, high-performance real-time 3D object detection models have been successfully applied to many industrial applications such as autonomous vehicles, unmanned aerial vehicles, and intelligent inspection robots. In the future, using attention mechanism, knowledge distillation, domain adaptation and other technical means to improve the detection performance, efficiency, and generalization ability of the network to complex scenes is still a big development trend of object detection algorithm research.

8. Prospect

In this part, the paper prospects six valuable research directions.

(1) *Large datasets*: Data and algorithms complement each other, and the establishment of larger data sets can facilitate researchers to develop object detectors with better performance and stronger generalization ability. Although some enterprises and open challenges have successively launched many large-scale datasets, in some specific industrial scenarios, it is very difficult to obtain datasets, and adding annotations to the data is also quite tedious and complex. Therefore, establishing a large-scale dataset with multiple scenes, multiple categories, and covering multiple modalities is a challenging and valuable work.

(2) *Perfect evaluation system*: The evaluation index is the ruler of the algorithm and provides the direction for the improvement of the algorithm. The mAP formula in COCO dataset is more complex than VOC, and it can better reflect the positioning accuracy of the model in real scenes. However, for 3D object detection tasks, it is obviously not applicable to simply project the 3D bounding box onto the plane and use the 2D AP value to evaluate the algorithm. AP_{3D} proposed by KITTI is more friendly for LiDAR detection and stereo vision tasks because it can directly measure the 3D bounding box detection accuracy. In addition, for indoor panoramic detection [81], video detection [173], and driving route planning [174,175], it is necessary to measure performance, speed, and safety indicators in a targeted manner. Tasks at different levels have different requirements for evaluation indicators, so it is imperative to establish a comprehensive evaluation system covering a variety of indicators with task orientation.

(3) *Semi-supervised and weakly supervised object detection*: Mature 2D and 3D object detection models generally rely on a large amount of labeled data for learning, and labeling image data is a very time-consuming work. Semi-supervised and weakly supervised training strategies proposed in recent years are expected to break this bottleneck. Semi-supervised learning can use a small amount of labeled data and a large amount of unlabeled data to train models. In recent years, some researchers have tried to use dense learning, knowledge distillation, data augmentation and other methods to improve the utilization of unlabeled data. Weakly supervised learning can use easily obtained image-level labeled data to train models. In recent years, researchers have mainly tried to solve the problems of local focus and slow convergence caused by the lack of instance-level constraints.

(4) *Small object detection*: According to the definition of MS COCO, the absolute size of small objects is less than 16×16 pixels, so the feature information that can be used is limited, and the detail information will be further lost with the deepening of the network, which is a major difficulty in the field of object detection. Some researchers have tried to improve the detection performance of small objects at the detection network level [176,177], the data preprocessing level [178], and the post-processing level [179], but this is often at the cost of high computational cost. The GAN model mentioned above can more efficiently approximate the detection performance of large and mesoscale objects by generating super-resolution representations for small objects. In addition to algorithm improvement, the establishment of a large-scale small object detection dataset and the corresponding evaluation indicators is another corresponding focus.

(5) *Image-only 3D object detection*: Thanks to the unremitting efforts of researchers, camera-only 3D detection algorithms from multi-view images have been successfully implemented in the assisted driving systems of Tesla and other enterprises. Image-based algorithms are gradually catching up with Lidar-based algorithms in performance. Taking NuScenes dataset as an example, the top-ranked image-only algorithm is HoP, which achieves 68.5% NDS, only 8.5% behind the top-ranked LiDAR-based algorithm YZL-Fusion (NDS = 77.0%). In the foreseeable future, multi-camera-based detectors will be more competitive than LiDAR detectors and become the optimal choice for autonomous vehicle perception systems.

(6) *Open-world object detection*: Traditional object detection methods generally predefine which categories to detect. However, when the algorithm is deployed in an open world oriented multi-dynamic scene such as robot or self-driving car, many unknown categories of things may appear at any time. To this end, Joseph et al. (2021) proposed the Open-World Object Detection (OWOD) [180]: first label new classes as “unknown”, and then re-label these new classes through incremental learning after gradually

obtaining the class labels. These authors also proposed the corresponding evaluation system and a new object detector: ORE, which learns unknown objects through a classification head based on energy grading and an RPN sensing unknown data and fine-tunes the model after each incremental learning to mitigate forgetting. Subsequently, OW-DETR [181] proposed by Gupta et al. (2022) utilized the Transformer self-attention mechanism to improve model performance, and MLUC [182] proposed by Cen et al. (2022) first applied 3D object detectors to open-world scenes. OWOD requires labeling all new class objects as "unknown", which leads to the need for human intervention during incremental learning and limits the generalization ability of the model. Zheng et al. (2022) proposed a new task: Open Set Object Detection and Discovery (OSODD) [183] aims to detect unknown objects with minimal supervision and their latent classes. At present, the research on OWOD and OSODD is still in the exploratory stage, and there is still much room for improvement in the comprehensive description of these problems, the establishment of comprehensive evaluation system and the improvement of the corresponding algorithms.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.52274160,52074305,51874300); The National Natural Science Foundation of China and the People's Government of Shanxi Province Coal-Based Low Carbon Joint Fund (No. U1510115).

References

[1] Lienhart R, Maydt J. An extended set of Haar-like features for rapid object detection. In: Proceedings international conference on image processing; 2002. I [I].
[2] Dalal N, Triggs B. Histograms of oriented gradients for human detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)2005 2005;1:886–93.
[3] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004;60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.
[4] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. Neural Information Processing Systems 2012;25. https://doi.org/10.1145/3065386.
[5] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR); 2016. p. 770–8. https://doi.org/10.1109/CVPR.2016.90.
[6] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF international conference on computer vision (ICCV); 2021. p. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.
[7] Wang W, Xie E, Li X, Fan DP, Song K, Liang D, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. IEEE/CVF International Conference on Computer Vision (ICCV); 2021. p. 548–58. https://doi.org/10.1109/ICCV48922.2021.00061. 2021.
[8] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR); 2017. p. 936–44. https://doi.org/10.1109/CVPR.2017.106.
[9] Bochkovskiy A, Wang C-Y, Liao H-y. YOLOv4: optimal speed and accuracy of object detection. 2020.
[10] Chen Q, Wang Y, Yang T, Zhang X, Cheng J, Sun J. You only look one-level feature. 2021.
[11] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-End object detection with transformers. Computer Vision – ECCV 2020;2020:213–29.
[12] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR); 2016. p. 779–88. https://doi.org/10.1109/CVPR.2016.91.
[13] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. SSD: single shot MultiBox detector. Computer Vision – ECCV 2016 2016:21–37.
[14] Dai J, Li Y, He K, Sun JR-FCN. Object detection via region-based fully convolutional networks. 2016.
[15] Xu S, Wang X, Lv W, Chang Q, Cui C, Deng K, et al. PP-YOLOE: an evolved version of YOLO. 2022.
[16] Wu H, Zhao H, Zhang M. Not all attention is all you need. 2021.
[17] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. 2020.
[18] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE conference on computer vision and pattern recognition; 2014. p. 580–7.
[19] Uijlings JRR, van de Sande KEA, Gevers T, Smeulders AWM. Selective search for object recognition. Int J Comput Vis 2013;104(2):154–71. https://doi.org/10.1007/s11263-013-0620-5.
[20] Chopra D, Khurana R. Support vector machine. 2023. p. 58–73.
[21] He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 2015;37(9):1904–16. https://doi.org/10.1109/TPAMI.2015.2389824.
[22] Girshick R, Fast R-CNN. IEEE international conference on computer vision (ICCV) 2015. 2015. p. 1440–8.
[23] Ren S, He K, Girshick R, Sun J, Faster R-CNN. Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39(6):1137–49. https://doi.org/10.1109/TPAMI.2016.2577031.
[24] He K, Gkioxari G, Dollár P, Girshick R, Mask R-CNN. IEEE international conference on computer vision (ICCV)2017. 2017. p. 2980–8.
[25] Cai Z, Vasconcelos N. Cascade R-CNN: delving into high quality object detection. In: 2018 IEEE/CVF conference on computer vision and pattern recognition; 2018. p. 6154–62.
[26] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. 2017 IEEE conference on computer vision and pattern recognition (CVPR). 2017. p. 6517–25.
[27] Redmon J, Farhadi A. YOLOv3: an incremental improvement. 2018.
[28] Wang C-Y, Bochkovskiy A, Liao H-y. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. 2022.
[29] Jocher G. YOLOv5 2020. https://doi.org/10.5281/zenodo.4154370.
[30] Long X, Deng K, Wang G, Zhang Y, Dang Q, Gao Y, et al. PP-YOLO: an effective and efficient implementation of object detector. 2020.
[31] Huang X, Wang X, Lv W, Bai X, Long X, Deng K, et al. PP-YOLOv2: a practical object detector. 2021.
[32] Fu C-Y, Liu W, Ranga A, Tyagi A, Berg A. Dssd : deconvolutional single shot detector. 2017.
[33] Jeong J, Park H, Kwak N. Enhancement of SSD by concatenating feature maps for object detection. 2017.
[34] Zheng L, Fu C, Zhao Y. Extend the shallow part of single shot multibox detector via convolutional neural network. 2018. p. 141.
[35] Yang J, Wang L. Feature fusion and enhancement for single shot multibox detector. 2019 Chinese automation congress. CAC; 2019. p. 2766–70.
[36] Yi J, Wu P, Metaxas D. ASSD: attentive single shot multibox detector. 2019.
[37] Chandio A, Gui G, Kumar T, Ullah I, Ranjbarzadeh R, Roy AM, et al. Precise single-stage detector. 2022.
[38] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: 2017 IEEE international conference on computer vision (ICCV); 2017. p. 2999–3007.
[39] Zhang S, Wen L, Bian X, Lei Z, Li SZ. Single-shot refinement neural network for object detection. In: 2018 IEEE/CVF conference on computer vision and pattern recognition; 2018. p. 4203–12.
[40] Nie J, Anwer RM, Cholakkal H, Khan FS, Pang Y, Shao L. Enriched feature guided refinement network for object detection. In: 2019 IEEE/CVF international conference on computer vision (ICCV); 2019. p. 9536–45.
[41] Law H, Deng J. CornerNet: detecting objects as paired keypoints. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer vision – ECCV 2018. Cham: Springer International Publishing; 2018. p. 765–81.
[42] Law H, Teng Y, Russakovsky O, Deng J. CornerNet-lite: efficient keypoint based object detection. 2019.
[43] Rashwan A, Kalra A, Poupart P. Matrix nets: a new deep architecture for object detection. In: 2019 IEEE/CVF international conference on computer vision workshop (ICCVW); 2019. p. 2025–8.
[44] Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q. CenterNet: keypoint triplets for object detection. In: 2019 IEEE/CVF international conference on computer vision (ICCV); 2019. p. 6568–77.
[45] Zhou X, Zhuo J, Krähenbühl P. Bottom-up object detection by grouping extreme and center points. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2019. p. 850–9.
[46] Tian Z, Shen C, Chen H, He T. FCOS: fully convolutional one-stage object detection. IEEE/CVF International Conference on Computer Vision (ICCV); 2019. p. 9626–35. 2019.
[47] Zhu C, He Y, Savvides M. Feature selective anchor-free module for single-shot object detection. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2019. p. 840–9.
[48] Kong T, Sun F, Liu H, Jiang Y, Li L, Shi J. FoveaBox: beyound anchor-based object detection. IEEE Trans Image Process 2020;29:7389–98. https://doi.org/10.1109/TIP.2020.3002345.

[49] Zhu C, Chen F, Shen Z, Savvides M. Soft anchor-point object detection. 2020. p. 91–107.
[50] Qiu H, Li H, Wu Q, Cui J, Song Z, Wang L, et al. CrossDet: crossline representation for object detection. In: 2021 IEEE/CVF international conference on computer vision (ICCV); 2021. p. 3175–84.
[51] Ge Z, Liu S, Wang F, Li Z, Sun J. YOLOX: exceeding YOLO series in 2021. 2021.
[52] Zou X, Wu Z, Zhou W, Huang J. YOLOX-PAI: an improved YOLOX version by PAI. 2022.
[53] Li C, Li L, Jiang H, Weng K, Geng Y, Li L, et al. YOLOv6: a single-stage object detection framework for industrial applications. 2022.
[54] Wang X, Wang G, Dang Q, Liu Y, Hu X, Yu D. PP-YOLOE-R: an efficient anchor-free rotated object detector. 2022.
[55] Zheng M, Gao P, Wang X, Li H, Dong H. End-to-End object detection with adaptive clustering transformer. 2020.
[56] Gao P, Zheng M, Wang X, Dai J, Li H. Fast convergence of DETR with spatially modulated Co-attention. In: 2021 IEEE/CVF international conference on computer vision (ICCV); 2021. p. 3601–10.
[57] Dai X, Chen Y, Yang J, Zhang P, Yuan L, Zhang L. Dynamic DETR: end-to-end object detection with dynamic attention. In: 2021 IEEE/CVF international conference on computer vision (ICCV); 2021. p. 2968–77.
[58] Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end-to-end object detection. 2020.
[59] Roh B, Shin J, Shin W, Kim S. Sparse DETR: efficient end-to-end object detection with learnable sparsity. 2021.
[60] Sun Z, Cao S, Yang Y, Kitani K. Rethinking transformer-based set prediction for object detection. In: 2021 IEEE/CVF international conference on computer vision (ICCV); 2021. p. 3591–600.
[61] Liu S, Li F, Zhang H, Yang X, Qi X, Su H, et al. DAB-DETR: dynamic anchor boxes are better queries for DETR. 2022.
[62] Wang Y, Zhang X, Yang T, Sun J. Anchor DETR: query design for transformer-based detector. Proc AAAI Conf Artif Intell 2022;36:2567–75. https://doi.org/10.1609/aaai.v36i3.20158.
[63] Beal J, Kim E, Tzeng E, Park D, Zhai A, Kislyuk D. Toward transformer-based object detection. 2020.
[64] Wang W, Xie E, Li X, Fan D-P, Song K, Liang D, et al. PVT v2: improved baselines with pyramid vision transformer, vol. 8. Computational Visual Media; 2022. p. 1–10. https://doi.org/10.1007/s41095-022-0274-8.
[65] Liu Z, Hu H, Lin Y, Yao Z, Xie Z, Wei Y, et al. Swin transformer V2: scaling up capacity and resolution. 2021.
[66] Li Y, Mao H, Girshick R, He K. Exploring Plain vision transformer backbones for object detection. 2022. p. 280–96.
[67] Chen X, Kundu K, Zhang Z, Ma H, Fidler S, Urtasun R. Monocular 3D object detection for autonomous driving. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR); 2016. p. 2147–56.
[68] Li B, Ouyang W, Sheng L, Zeng X, Wang XGS3D. An efficient 3D object detection framework for autonomous driving. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2019. p. 1019–28.
[69] Manhardt F, Kehl W, Gaidon A. ROI-10D: monocular lifting of 2D detection to 6D pose and metric shape. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2019. p. 2064–73.
[70] Simonelli A, Bulò SR, Porzi L, Lopez-Antequera M, Kontschieder P. Disentangling monocular 3D object detection. IEEE/CVF International Conference on Computer Vision (ICCV)2019 2019:1991–9.
[71] Qin Z, Wang J, Lu Y. MonoGRNet: a geometric reasoning network for monocular 3D object localization. Proc AAAI Conf Artif Intell 2019;33:8851–8. https://doi.org/10.1609/aaai.v33i01.33018851.
[72] Shi X, Ye Q, Chen X, Chen C, Chen Z, Kim TK. Geometry-based distance decomposition for monocular 3D object detection. In: IEEE/CVF international conference on computer vision (ICCV)2021; 2021. p. 15152–61.
[73] Lu Y, Ma X, Yang L, Zhang T, Liu Y, Chu Q, et al. Geometry uncertainty projection network for monocular 3D object detection. In: 2021 IEEE/CVF international conference on computer vision (ICCV); 2021. p. 3091–101.
[74] Brazil G, Liu X. M3D-RPN: monocular 3D region proposal network for object detection. In: 2019 IEEE/CVF international conference on computer vision (ICCV); 2019. p. 9286–95.
[75] Kumar A, Brazil G, Liu X. GrooMeD-NMS: grouped mathematically differentiable NMS for monocular 3D object detection. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2021. p. 8969–79.
[76] Luo S, Dai H, Shao L, Ding Y. M3DSSD: monocular 3D single stage object detector. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2021. p. 6141–50.
[77] Liu Z, Wu Z, Tóth R. SMOKE: single-stage monocular 3D object detection via keypoint estimation. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)2020 2020:4289–98.
[78] Zhang Y, Lu J, Zhou J. Objects are different: flexible monocular 3D object detection. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2021. p. 3288–97.
[79] Wang T, Zhu X, Pang J, Lin D. FCOS3D: fully convolutional one-stage monocular 3D object detection. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)2021 2021:913–22.
[80] Yan F, Nie M, Cai X, Han J, Xu H, Yang Z, et al. ONCE-3DLanes: building monocular 3D lane detection. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2022. p. 17122–31.
[81] Shen Z, Liao K, Nie L, Zheng Z, Zhao Y. PanoFormer: panorama transformer for indoor 360$$^{\circ }$$ depth estimation. 2022. p. 195–211.
[82] Chen X, Kundu K, Zhu Y, Ma H, Fidler S, Urtasun R. 3D object proposals using stereo imagery for accurate object class detection. IEEE Trans Pattern Anal Mach Intell 2018;40(5):1259–72. https://doi.org/10.1109/TPAMI.2017.2706685.
[83] Li P, Chen X, Shen S. Stereo R-CNN based 3D object detection for autonomous driving. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2019. p. 7636–44.
[84] Qin Z, Wang J, Lu Y. Triangulation learning network: from monocular to stereo 3D object detection. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2019. p. 7607–15.
[85] Sun J, Chen L, Xie Y, Zhang S, Jiang Q, Zhou X, et al. Disp R-CNN: stereo 3D object detection via shape prior guided instance disparity estimation. 2020.
[86] Peng W, Pan H, Liu H, Sun Y. IDA-3D: instance-depth-aware 3D object detection from stereo vision for autonomous driving. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2020. p. 13012–21.
[87] Peng X, Zhu X, Wang T, Ma Y. SIDE: center-based stereo 3D detector with structure-aware instance depth estimation. In: 2022 IEEE/CVF winter conference on applications of computer vision (WACV); 2022. p. 225–34.
[88] Chang JR, Chen YS. Pyramid stereo matching network. In: 2018 IEEE/CVF conference on computer vision and pattern recognition; 2018. p. 5410–8.
[89] Chen Y, Liu S, Shen X, Jia J. DSGN: deep stereo geometry network for 3D object detection. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2020. p. 12533–42.
[90] Chen Y, Huang S, Liu S, Yu B, Jia J. DSGN++: exploiting visual-spatial relation for stereo-based 3D detectors. IEEE Trans Pattern Anal Mach Intell 2023;45(4):4416–29. https://doi.org/10.1109/TPAMI.2022.3197236.
[91] Guo X, Shi S, Wang X, Li H. LIGA-stereo: learning LiDAR geometry aware representations for stereo-based 3D detector. In: 2021 IEEE/CVF international conference on computer vision (ICCV); 2021. p. 3133–43.
[92] Liu Y, Wang L, Liu M. YOLOStereo3D: a step back to 2D for efficient stereo 3D detection. IEEE International Conference on Robotics and Automation (ICRA) 2021 2021:13018–24.
[93] Wang Y, Chao W-L, Garg D, Hariharan B, Weinberger K. Pseudo-LiDAR from visual depth estimation: bridging the gap in 3D object detection for autonomous driving. 2018.
[94] Ma X, Wang Z, Li H, Zhang P, Ouyang W, Fan X. Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving. IEEE/CVF International Conference on Computer Vision (ICCV)2019; 2019. p. 6850–9.
[95] Wang X, Yin W, Kong T, Jiang Y, Li L, Shen C. Task-aware monocular depth estimation for 3D object detection. Proc AAAI Conf Artif Intell 2020;34:12257–64. https://doi.org/10.1609/aaai.v34i07.6908.
[96] Ye X, Du L, Shi Y, Li Y, Tan X, Feng J, et al. Monocular 3D object detection via feature domain adaptation. 2020. p. 17–34.
[97] You Y, Wang Y, Chao W-L, Garg D, Pleiss G, Hariharan B, et al. Pseudo-LiDAR++: accurate depth for 3D object detection in autonomous driving. 2019.
[98] Li C, Ku J, Waslander S. Confidence guided stereo 3D object detection with split depth estimation. 2020.
[99] Pon AD, Ku J, Li C, Waslander SL. Object-centric stereo matching for 3D object detection. 2020 IEEE International Conference on Robotics and Automation (ICRA) 2020:8383–9.
[100] Garg D, Wang Y, Hariharan B, Weinberger K, Chao W-L. Wasserstein distances for stereo disparity estimation. 2020.
[101] Xu Z, Zhang W, Ye X, Tan X, Yang W, Wen S, et al. ZoomNet: Part-aware adaptive zooming neural network for 3D object detection. Proc AAAI Conf Artif Intell 2020;34:12557–64. https://doi.org/10.1609/aaai.v34i07.6945.
[102] Qian R, Garg D, Wang Y, You Y, Belongie S, Hariharan B, et al. End-to-End pseudo-LiDAR for image-based 3D object detection. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2020. p. 5880–9.
[103] Wang Y, Yang B, Hu R, Liang M, Urtasun R. PLUMENet: efficient 3D object detection from stereo images. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2021:3383–90.
[104] Philion J, Fidler S. Lift, splat, shoot: encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. 2020.
[105] Huang J, Huang G, Zheng Z, Du D. BEVDet: high-performance multi-camera 3D object detection in bird-eye-view. 2021.
[106] Li Y, Ge Z, Yu G, Yang J, Wang Z, Shi Y, et al. BEVDepth: acquisition of reliable depth for multi-view 3D object detection. 2022.
[107] Li Y, Bao H, Ge Z, Yang J, Sun J, Li Z. BEVStereo: enhancing depth estimation in multi-view 3D object detection with dynamic temporal stereo. 2022.
[108] Wang Z, Min C, Ge Z, Li Y, Li Z, Yang H, et al. STS: surround-view temporal stereo for multi-view 3D detection. 2022.
[109] Li Z, Wang W, Li H, Xie E, Sima C, Lu T, et al. BEVFormer: learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. 2022. p. 1–18.
[110] Yang C, Chen Y, Tian H, Tao C, Zhu X, Zhang Z, et al. BEVFormer v2: adapting modern image backbones to bird's-eye-view recognition via perspective supervision. 2022.
[111] Park J, Xu C, Yang S, Keutzer K, Kitani K, Tomizuka M, et al. Time will tell: new outlooks and A baseline for temporal multi-view 3D object detection. 2022.
[112] Zong Z, Jiang D, Song G, Xue Z, Su J, Li H, et al. Temporal enhanced training of multi-view 3D object detector via historical object prediction. 2023.
[113] Chen Y, Yang T, Zhang X, Meng G, Pan C, Sun J. DetNAS: neural architecture search on object detection. 2019.
[114] Liang T, Wang Y, Tang Z, Hu G, Ling H. OPANAS: one-shot path aggregation network architecture search for object detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)2021 2021:10190–8.

[115] Ghiasi G, Lin TY, Le QV. NAS-FPN: learning scalable feature pyramid architecture for object detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019:7029–38.
[116] Guo J, Han K, Wang Y, Zhang C, Yang Z, Wu H, et al. Hit-detector: hierarchical trinity architecture search for object detection. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2020. p. 11402–11.
[117] Yu G, Chang Q, Lv W, Xu C, Cui C, Ji W, et al. PP-PicoDet: a better real-time object detector on mobile devices. 2021.
[118] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. 2015.
[119] Wang T, Yuan L, Zhang X, Feng J. Distilling object detectors with fine-grained feature imitation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019:4928–37.
[120] Guo J, Han K, Wang Y, Wu H, Chen X, Xu C, et al. Distilling object detectors via decoupled features. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2021:2154–64.
[121] Yang Z, Li Z, Jiang X, Gong Y, Yuan Z, Zhao D, et al. Focal and global knowledge distillation for detectors. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2022. p. 4633–42.
[122] Chen Y, Chen P, Liu S, Wang L, Jia J. Deep structured instance graph for distilling object detectors. In: IEEE/CVF international conference on computer vision (ICCV)2021; 2021. p. 4339–48.
[123] Dai X, Jiang Z, Wu Z, Bao Y, Wang Z, Liu S, et al. General instance distillation for object detection. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2021. p. 7838–47.
[124] Chen Y, Li W, Sakaridis C, Dai D, Gool LV. Domain adaptive faster R-CNN for object detection in the wild. In: 2018 IEEE/CVF conference on computer vision and pattern recognition; 2018. p. 3339–48.
[125] Zhu X, Pang J, Yang C, Shi J, Lin D. Adapting object detectors via selective cross-domain alignment. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2019. p. 687–96.
[126] Li C, Du D, Zhang L, Wen L, Luo T, Wu Y, et al. Spatial attention pyramid network for unsupervised domain adaptation. In: Vedaldi A, Bischof H, Brox T, Frahm J-M, editors. Computer vision – ECCV 2020. Cham: Springer International Publishing; 2020. p. 481–97.
[127] Inoue N, Furuta R, Yamasaki T, Aizawa K. Cross-domain weakly-supervised object detection through progressive domain adaptation. In: IEEE/CVF conference on computer vision and pattern Recognition2018; 2018. p. 5001–9.
[128] Xu Y, Sun Y, Yang Z, Miao J, Yang Y, H2Fa R-CNN. Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2022. p. 14309–19.
[129] Shi W, Rajkumar R. Point-GNN: graph neural network for 3D object detection in a point cloud. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2020. p. 1708–16.
[130] Najibi M, Lai G, Kundu A, Lu Z, Rathod V, Funkhouser T, et al. DOPS: learning to detect 3D objects and predict their 3D shapes. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2020. p. 11910–9.
[131] Man Y, Weng X, Sivakumar PK, O'Toole M, Kitani K. Multi-echo LiDAR for 3D object detection. In: 2021 IEEE/CVF international conference on computer vision (ICCV); 2021. p. 3743–52.
[132] Ansari M, Meraz M, Chakraborty P, Javed DM. Angle-based feature learning in GNN for 3D object detection using point cloud. 2022. p. 419–32.
[133] Wang X, Shrivastava A, Gupta A. A-Fast-RCNN: hard positive generation via adversary for object detection. 2017.
[134] Li J, Liang X, Wei Y, Xu T, Feng J, Yan S. Perceptual generative adversarial networks for small object detection. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR); 2017. p. 1951–9.
[135] Bai Y, Zhang Y, Ding M, Ghanem B. SOD-MTGAN: Small Object Detection via Multi-Task Generative Adversarial Network: 15th European Conference 2018: 210–26. Munich, Germany, September 8-14, 2018, Proceedings, Part XIII.
[136] Everingham M, Van Gool L, Williams C, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. Int J Comput Vis 2010:303–38.
[137] Deng J, Dong W, Socher R, Li LJ, Kai L, Li F-F. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition; 2009. p. 248–55.
[138] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer vision – ECCV 2014. Cham: Springer International Publishing; 2014. p. 740–55.
[139] Krasin I, Duerig T, Alldrin N, Veit A, Abu-El-Haija S, Belongie S, et al. OpenImages: a public dataset for large-scale multi-label and multi-class image classification. 2016.
[140] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition; 2012. p. 3354–61.
[141] Sun P, Kretzschmar H, Dotiwalla X, Chouard A, Patnaik V, Tsui P, et al. Scalability in perception for autonomous driving: Waymo open dataset. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2020. p. 2443–51.
[142] Caesar H, Bankiti V, Lang AH, Vora S, Liong VE, Xu Q, et al. nuScenes: a multimodal dataset for autonomous driving. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2020. p. 11618–28.
[143] Weng X, Man Y, Cheng D, Park J, O'Toole M, Kitani K. All-in-one drive: a large-scale comprehensive perception dataset with high-density long-range point clouds. 2020.
[144] Dosovitskiy A, Ros G, Codevilla F, López A, Koltun V. CARLA: an open urban driving simulator. 2017.
[145] Dai A, Chang AX, Savva M, Halber M, Funkhouser T, Nießner M. ScanNet: richly-annotated 3D reconstructions of indoor scenes. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR); 2017. p. 2432–43.
[146] Song S, Lichtenberg SP, Xiao J, Sun RGB-D. A RGB-D scene understanding benchmark suite. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR); 2015. p. 567–76.
[147] Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from RGBD images. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C, editors. Computer vision – ECCV 2012. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 746–60.
[148] Dooley S, Wei G, Goldstein T, Dickerson J. Robustness disparities in face detection. 2022. https://doi.org/10.48550/arXiv.2211.15937.
[149] Wang G, Li J, Wu Z, Xu J, Shen J, Yang W. EfficientFace: an efficient deep network with feature enhancement for accurate face detection. 2023. https://doi.org/10.48550/arXiv.2302.11816.
[150] Boyd A, Tinsley P, Bowyer K, Czajka A. CYBORG: blending human saliency into the loss improves deep learning. 2021.
[151] Piland J, Czajka A, Sweet C. Improving model's focus improves performance of deep learning-based synthetic face detectors. 2023.
[152] Muhammad U, Romaissa B, Oussalah M. Domain generalization via ensemble stacking for face presentation attack detection. 2023.
[153] Dwivedi R, Kumar R, Chopra D, Kothari P, Singh M. An efficient ensemble explainable AI (XAI) approach for morphed face detection. 2023.
[154] Fang M, Huber M, Damer N. SynthASpoof: developing face presentation attack detection based on privacy-friendly synthetic data. 2023. https://doi.org/10.48550/arXiv.2303.02660.
[155] Kong C, Li H, Wang S. Enhancing general face forgery detection via vision transformer with low-rank adaptation. 2023.
[156] Gilroy S, Mullins D, Jones E, Parsi A, Glavin M. The impact of partial occlusion on pedestrian detectability. 2022.
[157] Hasan I, Liao S, Li J, Akram S, Shao L. Pedestrian detection: domain generalization, CNNs, transformers and beyond. 2022.
[158] Lin Z, Pei W, Chen F, Zhang D, Lu G. Pedestrian detection by exemplar-guided contrastive learning. IEEE Trans Image Process 2022:1. https://doi.org/10.1109/TIP.2022.3189803.
[159] Hwang J, Benz P, Kim T-h. Booster-SHOT: boosting stacked homography transformations for multiview pedestrian detection with attention. 2022.
[160] Vora J, Dutta S, Jain K, Karthik S, Gandhi V. Bringing generalization to deep multi-view pedestrian detection. 2023. p. 110–9. https://doi.org/10.1109/WACVW58289.2023.00016.
[161] Zhang T, Ye Q, Zhang B, Liu J, Zhang X, Tian Q. Feature calibration network for occluded pedestrian detection. 2022. https://doi.org/10.48550/arXiv.2212.05717.
[162] Liu M, Jiang J, Zhu C, xu Y. VLPD: context-aware pedestrian detection via vision-language semantic self-supervision. 2023.
[163] Wang Y, Bashir SMA, Khan M, Ullah Q, Wang R, Song Y, et al. Remote sensing image super-resolution and object detection: benchmark and state of the art. Expert Syst Appl 2022;197:116793. https://doi.org/10.1016/j.eswa.2022.116793.
[164] Shamsolmoali P, Zareapoor M, Granger E, Chanussot J, Yang J. Enhanced single-shot detector for small object detection in remote sensing images. 2022.
[165] Yan T, Wan Z, Zhang P. Fully transformer network for change detection of remote sensing images. Computer Vision – ACCV 2022;2023:75–92.
[166] Ye Y, Wang M, Zhou L, Lei G, Fan J, Qin Y. Adjacent-level feature cross-fusion with 3D CNN for remote sensing image change detection. 2023. https://doi.org/10.48550/arXiv.2302.05109.
[167] Liu B, Chen H, Wang Z. LSNet: extremely light-weight siamese network for change detection in remote sensing image. 2022.
[168] Lin Y, Sun H, Liu N, Bian Y, Cen J, Zhou H. Attention guided network for salient object detection in optical remote sensing images. Artificial Neural Networks and Machine Learning – ICANN 2022:25–36. 2022.
[169] Li G, Liu Z, Bai Z, Lin W, Ling H. Lightweight salient object detection in optical remote sensing images via feature correlation. IEEE Trans Geosci Rem Sens 2022;60:1–12. https://doi.org/10.1109/TGRS.2022.3145483.
[170] Zhang Z, Li S, Wang Z, Lu Y. A novel and efficient tumor detection framework for pancreatic cancer via CT images. In: 42nd annual international conference of the IEEE engineering in medicine & biology society. EMBC; 2020. p. 1160–4. https://doi.org/10.1109/EMBC44109.2020.9176172. 2020.
[171] Shuvo S. An automated end-to-end deep learning-based framework for lung cancer diagnosis by detecting and classifying the lung nodules. 2023.
[172] Bai X, Xia Y. An end-to-end framework for universal lesion detection with missing annotations. 2022 16th IEEE International Conference on Signal Processing (ICSP) 2022;1:411–5. https://doi.org/10.1109/ICSP56322.2022.9965335.
[173] Mao H, Yang X, Dally B. A delay metric for video object detection: what average precision fails to tell. In: 2019 IEEE/CVF international conference on computer vision (ICCV); 2019. p. 573–82.
[174] Philion J, Kar A, Fidler S. Learning to evaluate perception models using planner-centric metrics. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2020. p. 14052–61.
[175] Deng B, Qi C, Najibi M, Funkhouser T, Zhou Y, Anguelov D. Revisiting 3D object detection from an egocentric perspective. 2021.
[176] Liu S, Huang d, Wang Y. Receptive field block net for accurate and Fast object detection. 2017.

[177] Najibi M, Samangouei P, Chellappa R, Davis LS. SSH: single stage headless face detector. In: 2017 IEEE international conference on computer vision (ICCV); 2017. p. 4885–94.
[178] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. SSD: single shot MultiBox detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer vision – ECCV 2016. Cham: Springer International Publishing; 2016. p. 21–37.
[179] Chang X, Wang J, Yang W, Yu H, Yu L, Xia G-S. RFLA: Gaussian receptive field based label assignment for tiny object detection. 2022. p. 526–43.
[180] Joseph KJ, Khan S, Khan FS, Balasubramanian VN. Towards open world object detection. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2021. p. 5826–36.
[181] Gupta A, Narayan S, Joseph KJ, Khan S, Khan FS, Shah M. OW-DETR: open-world detection transformer. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2022. p. 9225–34.
[182] Cen J, Yun P, Cai J, Wang MY, Liu M. Open-set 3D object detection. International Conference on 3D Vision 2021;(3DV):869–78. 2021.
[183] Zheng J, Li W, Hong J, Petersson L, Barnes N. Towards open-set object detection and Discovery. In: 2022 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW); 2022. p. 3960–9.