

Received 4 August 2022, accepted 31 August 2022, date of publication 12 September 2022, date of current version 21 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3205593

## APPLIED RESEARCH

# Anchor-Free Weapon Detection for X-Ray Baggage Security Images

YAN HUANG<sup>1</sup>, XINSHA FU<sup>1</sup>, AND YANJIE ZENG<sup>2</sup>

<sup>1</sup>School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510640, China

<sup>2</sup>Guangdong Provincial Transport Planning and Research Center (GTPRC), Guangzhou 510055, China

Corresponding author: Xinsha Fu (fuxinsha@163.com)


This work was supported in part by the National Science Foundation of China under Grant 41571397, Grant 41501442, and Grant 51778242; and in part by the National Natural Science Foundation of China under Project 51978283.

**ABSTRACT** Considering the real-time and high-precision requirements of image processing in X-ray baggage security screening; and problems such as the inflexibility and complex computation of anchor-based object detection, this paper introduces an anchor-free mode convolutional neural network object detection method for detecting weapons (knives and handguns) in X-ray baggage security images. The advantage of the anchor-free method over the anchor-based method is that the size of the anchor box does not have to be set, and the generalization ability is strong; the absence of the anchor box reduces the number of computations, and solves the problem of unbalanced positive and negative samples in the anchor-based method. To fully evaluate the effectiveness of the anchor-free method for X-ray baggage screening image detection, a large number of images containing knives and handguns were collected and annotated in the early stages of this work to produce a dataset that could be used for training. Six mainstream anchor-free methods (CornerNet, CenterNet, CornerNet-Lite, ExtremeNet, Objects as Points and You Only Look Once (YOLOx)) are introduced. For experimental integrity, this paper adds an anchor-based comparison experiment, using Faster-RCNN, YOLOv3 and YOLOv5 to perform the same work. The experimental results show that the YOLOx, Objects as Points and ExtremeNet anchor-free methods used in this paper have excellent performance in weapon detection in X-ray baggage security images. Among them, the mean average precision (mAP) of YOLOx combined with the CSPDarknet53 network reached 0.905, and the mAP of ExtremeNet combined with the Hourglass-104 network reached 0.900; the performance of the Objects as Points method was also good. All these methods performed better than the anchor-based methods compared in this paper. Therefore, we believe that the anchor-free method has a practical effect in weapon detection for X-ray luggage images.

**INDEX TERMS** Object detection, X-ray baggage security images, anchor-free.

## I. INTRODUCTION

X-ray inspection equipment, as a widely used means of detecting security risks, has been installed increasingly often in key locations in crowded areas such as train stations and airports, as an important protective barrier against terrorist attacks. At present, the detection of dangerous goods still relies on the human eye to identify pictures, which not only consumes time and manpower, but also makes it easy to

The associate editor coordinating the review of this manuscript and approving it for publication was Jiju Poovancheri .

misidentify and miss detection when the operation task is difficult. Therefore, automatic detection in X-ray images is a topic that is challenging and worthy of research.

Deep learning-based image object detection techniques have shown very competitive performance in recent years, and after convolutional neural networks achieved great success in classification tasks with ImageNet [1] in 2012, Girshick *et al.* [2] were the first to propose a framework for object detection in region-based convolutional networks. Since then, a new phase of object detection has begun. Akcay *et al.* [25], for example, considered the use of convolutional neural

networks with migration learning applied to X-ray baggage images, and divided their study into two parts: classification and object detection. They proposed using the AlexNet [1] network to extract image features, and a support vector machine (SVM) classifier was trained to achieve a classification accuracy of 0.994 within the image object region. Sliding window-based convolutional neural networks (CNNs), faster region based CNNs (F-R CNNs) [1], region-based fully convolutional networks (R-FCNs) [4] and You Only Look Once (YOLOv2) [6] were explored for X-ray luggage object detection in images, and the object detection results of X-ray baggage security images based on the CNN were good.

References [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], and [35] also proposed using convolutional neural networks to detect objects in X-ray baggage security images. However, all of these methods generate a large number of anchors during the detection process; when using anchors, they need to be densely tiled at each feature scale, and only a small fraction of the samples are positive, so the proportions of positive and negative samples varies greatly. Ultimately, computing resources are spent on useless samples, and the general use of anchors requires preprocessing to mine difficult negative cases. Therefore, this paper, inspired by the anchor-free idea, aims to determine the location and size of the detection frame by eliminating anchors and directly looking for key point information in the feature image, and the possibility of consuming fewer computational resources to obtain more accurate detection results in X-ray baggage security screening scenarios is explored.

We collected a large number of X-ray baggage security images, labelled the knives and handguns that needed to be detected, and created a dataset for object detection. Unlike common reflected images, X-ray images [2] are greyscale images formed by X-ray generators projecting the remaining energy generated by a beam of low-energy X-rays through the object onto a sensor or detector; the greyscale values are affected by the thickness, density and atomic number of the material. According to the review of the detection of aviation safety explosives in [2], in recent years, with the development of detectors, computers, image processing and other related technologies, the imaging quality of X-ray security equipment has been continuously improved. The imaging mode has developed from traditional single-energy to dual-energy X-ray imaging [10], and the detection purpose has expanded from simple shape recognition to exploring the essential properties of substances. Because dual-energy X-ray technology for object detection is based on the chemical composition (atomic number) of an object rather than only on the density change as in single-energy X-ray technology, the dual-energy X-ray measurement method can distinguish between organic and inorganic materials, basically eliminating the changes in most of the thickness of the material and displaying the image density differences according to the chemical composition (atomic number). To improve the recognition of image content, we will use the density difference of the grey image according to the atomic number to fill for in the colour of

pseudo colour image [11]; the equivalent of an atomic number less than 10 is organic and will be coloured orange, the equivalent of an atomic number greater than 18 is inorganic and will be coloured blue, and material with an atomic number between these two values or that is a mixture of the two types will be coloured green.

All images used in this experiment were provided by a model of dual-energy X-ray detector, manufactured by UNICOMP, which provides two energy images simultaneously. It means that two sets of data can be obtained during a radiography to generate two images corresponding to high-energy and low-energy rays respectively. The dual-energy detector has two scintillators, gadolinium sulfide (GOS) (153mg/cm<sup>2</sup>) at low energy and cesium iodide CsI (TI) at high energy. The measured object is moved by the conveyor belt at a speed of 22cm/s. The maximum width of the scanned object is 650 mm, and the height is 500 mm. We collected a large number of pistol and knife models, mixed with ordinary objects and other interference objects into the suitcase. After output the raw image by X-ray scanning equipment, the image was coloured according to the atomic number, and the image was compressed to 960 × 640 resolution, 24 bits depth, and no other post-processing was done.

Unlike the anchor-based method, the anchor-free method is based on finding the key object points to determine the object location, and the key point generation strategy has a direct impact on the accuracy and speed of detection. This experiment introduces six anchor-free methods, namely, CornerNet [40], CornerNet-Lite [42], CenterNet [41], ExtremeNet [44], Objects as Points [43] and YOLOx [45], all of which have different combinations of methods for selecting key points and can have different detection results. In this paper, key points are classified into three types, corner points, centre points, and extreme points, and the locations of these key points are based on the mapping from the backbone network output of the feature heatmap to the location of the object. In addition to the YOLOx method, which uses the CSPDarknet53 network structure (a fusion of CSPNet and Darknet53), there are several other anchor-free methods that adopt the Hourglass network as the backbone network. Hourglass is a network model similar to encoding and decoding. It can capture local and global information, which is helpful for key point prediction. To compare anchor-based methods, this paper also performs the same experiments on several classic anchor-based methods, such as Faster-RCNN, YOLOv3 and YOLOx and compares the experimental results with those of the anchor-free methods.

The main contributions of this paper are as follows. (1) This paper analyses the hashrate deficiency of the traditional anchor-based object detection algorithm, and introduces the latest anchor-free object detection algorithm for the task of detecting X-ray baggage security knife and handgun images to address the abovementioned problems. (2) In this paper, several recent anchor-free object detection algorithms are investigated, the advantages and disadvantages of the respective methods are analysed, and comparative experiments are

conducted. (3) Given the paucity of knife and handgun detection data in X-ray luggage images, this paper collects and labels a large number of X-ray luggage images containing these two items to construct a new X-ray image-based detection dataset. Based on this dataset, a comprehensive evaluation of each of the above algorithms is carried out. Experimentally, we conclude that anchor-free methods have better practicability than the anchor-based methods introduced for the task of weapon detection in X-ray baggage security images.

## II. RELATED WORKS

Research on X-ray baggage security imagery has been continuously updated with the development of computer vision, and previous research has undergone several phases: image enhancement [12], [13], [14], [15], [16], traditional image handcrafted feature extraction [17], [18], [19], [20], [21], [22], [23], [24] and end-to-end neural network object detection [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35].

By enhancing the visibility and edge contrast of objects in an image and removing noise, the efficiency of identifying objects in X-ray baggage screening images by staff can be improved. Maneesha Singh *et al.* [12] used neural networks with cross-validation methods to select the best image enhancement algorithm based on the visibility characteristics of X-ray luggage images, and the experimental results showed that the system played a positive role in enhancing X-ray baggage images 93.04% of the time, no role 6.22% of the time and a negative role 0.73% of the time. Liang *et al.* [13] used an image hash algorithm to enhance the visibility of hidden low-density items in X-ray baggage scan images, resulting in a 62% increase in the speed of manual detection of low-density items and a 58% increase in manual detection accuracy. Zhiyu Chen *et al.* [14] used a background-subtraction-based method for image noise reduction and then an image enhancement algorithm based on histograms. Abidi B *et al.* [15] used RGB and hyperspectral image (HSI) based colour conversion to colour code X-ray greyscale images of weapons to improve image visualization and increase the manual detection rate of weapons to 97%.

The implementation of image handcrafted feature extraction and a classifier enables automated, high-precision object detection of in X-ray luggage images. The segmentation part of the object detection algorithm mentioned by Mery *et al.* [17] used a fusion of multiple methods: first binarizing the image, then extracting interest regions and scale-invariant feature transform (SIFT) key point matches through a Laplace transformation of Gaussian edges. Automatic detection was performed in experiments with 18 samples, which showed a true positive rate of 94.3% and a false positive rate of 5.6%. Bastan M *et al.* [18] investigated the applicability of a bag of words (BoW) in X-ray image classification by comparing multiple feature extraction methods and showed that difference of Gaussian (DoG) features, Hessian Laplace features, Harris features and features from accelerated segment test

(FAST) performed more competitively, it was also concluded that the SIFT descriptors performed best, but not as well as on conventional images, and that the main problem was the lack of texture information in X-ray images. Franzel T *et al.* [19] performed object detection of X-ray luggage images from multiple viewpoints, where a combination of a histogram of oriented gradient (HOG) features and an SVM classifier was used for supervised learning to construct a classification model, and the experimental results showed that the average of the single-view detection accuracy (AP) increased from 49.7% to 64.5%, with multiple views able to detect approximately 80% of handguns. Schmidt-Hackenberg *et al.* [20] used four methods for the feature extraction of X-ray baggage images for comparison (SLF-HMAX, V1-like, SIFT, PHOW), using linear binary SVM kernels as classifiers, and the experimental results showed that SLF-HMAX and V1-like visual cortical elicitation were superior to the bag-of-visual-words (BoVW) approach. Turcsany D *et al.* [21] proposed a novel BoW representation scheme for the X-ray baggage image object detection task, which was implemented in the SVM classifier framework using a speeded-up robust features (SURF) detector and descriptors, and it achieved a true positive rate of 99.07% and a false positive rate of 4.31% in the firearm detection scenario. Muhammet Bastan *et al.* [22] used rotation invariant texture, SIFT, and colour descriptors; used SPIN and its extended versions ESPIN and CSPIN as point descriptors; and incorporated all these features into a regular bag-of-features (BoF) framework. The detection method used the original efficient subwindow search (ESS) algorithm combined with the SVM linear structure. The results showed that the object detection performance on X-ray images greatly helped to extend the features and provide multiple views. M. E. Kundegorski *et al.* [23] comprehensively compared the combination of feature extraction and descriptors in the BoVW technique to build classifiers using SVM, and showed that SURF feature extraction and descriptors have the highest accuracy and high execution rates.

Entering the developmental period of deep learning object detection, S. Akcay *et al.* [26] used deep CNNs to study the image classification problem in the context of X-ray baggage security and achieved a detection accuracy of 98.92%. S. Akcay *et al.* [25] studied the application of deep neural networks for classification and object detection in X-ray baggage security imagery and achieved an accuracy of 0.994 for the classification task by combining AlexNet network structures [1] and SVM classifiers. In addition, they used SW-CNNs, F-RCNNs [1], and YOLOv2 [6] for object detection and achieved a mean average precision (mAP) of 0.885 for six-class object detection and 0.974 for two-class object detection; the detection efficiency reached 100 ms per sheet, which shows that the deep convolutional neural network has very good performance in the X-ray baggage security imagery detection task. Galvez *et al.* [27] used a YOLO [5] object detector to detect threat objects in X-ray images to address the problems of occlusion and rotation in X-ray

baggage security images, and the mAP of this method reached 45.89% in  $416 \times 416$  images. It reached 51.48% in  $608 \times 608$  images and 52.40% in multiscale images. On the other hand, transfer learning achieved a mAP of only 29.54%, and a mAP of 29.17% was achieved for multiscale images. Koçi *et al.* [28] used X-ray images to check for threatening items in baggage, and concluded that the best detection was achieved by a combination of the Faster R-CNN [1] detection models and the ResNet101 [37] feature extractor, which yielded an accuracy of 87.58% ( $\pm 0.75\%$  error margin). Ponnusamy *et al.* [29] used the deep convolutional neural network of YOLO [5] to classify luggage images on a field programmable gate array (FPGA) platform. The results showed that with less resource occupancy, the YOLO [5] implementation provide a maximum accuracy of 98.9% in classifying X-ray baggage images and identifying hazardous materials. Saavedra *et al.* [30] proposed a framework that simulates a large number of X-ray images, using a combination of PGGAN [38] and superimposition strategies [34]; this method was tested in the detection of four types of threatening objects in real X-ray images: guns, knives, razor blades and shuriken (ninja stars). The experiments showed that YOLOv3 [7] obtained the best mAP, with 96.3% for guns, 76.2% for knives, 86.9% for razor blades and 93.7% for shuriken, while the average mAP for all threat objects was 80.0%. Chang, An, *et al.* [31] proposed a two-stage prohibited object detection network that can identify prohibited objects in heavily cluttered X-ray baggage images to reduce the false positives caused by neglecting the actual physical sizes of items. Extensive experimentation demonstrated that the proposed method outperformed state-of-the-art object detection methods. Altındağ *et al.* [32] introduced a publicly available single-view dual-channel X-ray dataset called the HUMS X-ray dataset, and three popular object detection algorithms namely the Faster RCNN [1], YOLOv3 [7], and the single-shot detector (SSD) [39] were applied to the X-ray dataset. The HUMS X-ray dataset is publicly available and includes low-energy, high-energy and false-coloured images. Ma *et al.* [33] proposed an effective anomalous object detection network to improve the detection accuracy of anomalous objects in X-ray images. The experimental results showed that the method achieved a mAP of 85.9% on the SIXray dataset and a mAP of 85.8% on OPIXray dataset.

According to these previous studies, it is clear that deep convolutional neural networks have good effects in X-ray baggage security image detection, but the object detection methods in [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], and [35] are based on generating a large number of anchors, many of which are useless; we would like to achieve reduce effort and obtain better detection results. Therefore, several methods [40], [41], [42], [43], [44], [45], which are currently emerging in anchor-free object detection, are introduced and applied to weapon detection in X-ray baggage security images, and the applicability of these anchor-free methods in this scenario is evaluated.

### III. METHODOLOGY

To find the object region, the anchor base extracts the bounding box for the region in which the object is located via the region proposal network (RPN), while the anchor-free method achieves the same end by generating a keypoint for the object region. The generation of key points should be based on the heatmaps generated by the image attention mechanism, similar to the way humans observe images, where the global image is quickly scanned to obtain the object area that needs to be focused on, and then more attention resources are devoted to this area to obtain more details about the object while suppressing useless information; this is the difference between anchor-free and anchor-based mechanisms.

#### A. BACKBONE NETWORK

The deep feature images extracted by convolutional neural networks have an attentional effect upon activation, which responds to regions of interest but easily loses deep features. To capture information from multiscale feature maps, Newell *et al.* [46] proposed the Hourglass network structure, motivated by the need to capture information at each scale. The network structure is hourglass-shaped, using a residual module as the basic network unit, with repeated top-down and bottom-up structures to infer the locations of key points of the object. Hourglass network used by anchor-free in this paper has made some modifications on this basis. Before entering Hourglass module, image through a  $7 \times 7$  convolution module with stride 2 and 128 channels reducing the resolution by 4 times. After the hourglass module is modified, the max pooling downsampling method is removed and the downsampling method with step 2 is used instead. The feature resolutions are reduced 5 times, and the channel is increased to (256, 384, 384, 384, 512). This Hourglass module is named Hourglass-52, show in Figure1, and a stack of two modules is called Hourglass-104.

YOLOx's backbone network, CSPDarknet53, is a combination of Darknet53 and the cross stage partial network(CSPNet) [47]. CSPNet breaks up the feature map into two parts, one of which carries out a convolution operation, and the other of which concatenates the results of the previous part of the convolution operation, as shown in Figure 2. CSPNet respects the variability of gradients by integrating the feature maps at the beginning and end of the network stage, to reduce the amount of computation and ensure accuracy.

#### B. KEY POINT STRATEGY

The CornerNet [40] method designs the corner pooling module to locate the upper left and lower right corners of the object. Corner pooling provides a prior prediction of the corner points, which makes corner point location more accurate and solves the problem that bounding box corner points often appear outside the ground truth. Taking the top-left corner point as an example, max pooling is performed from left to right for each row pixel of the feature map matrix; at the same

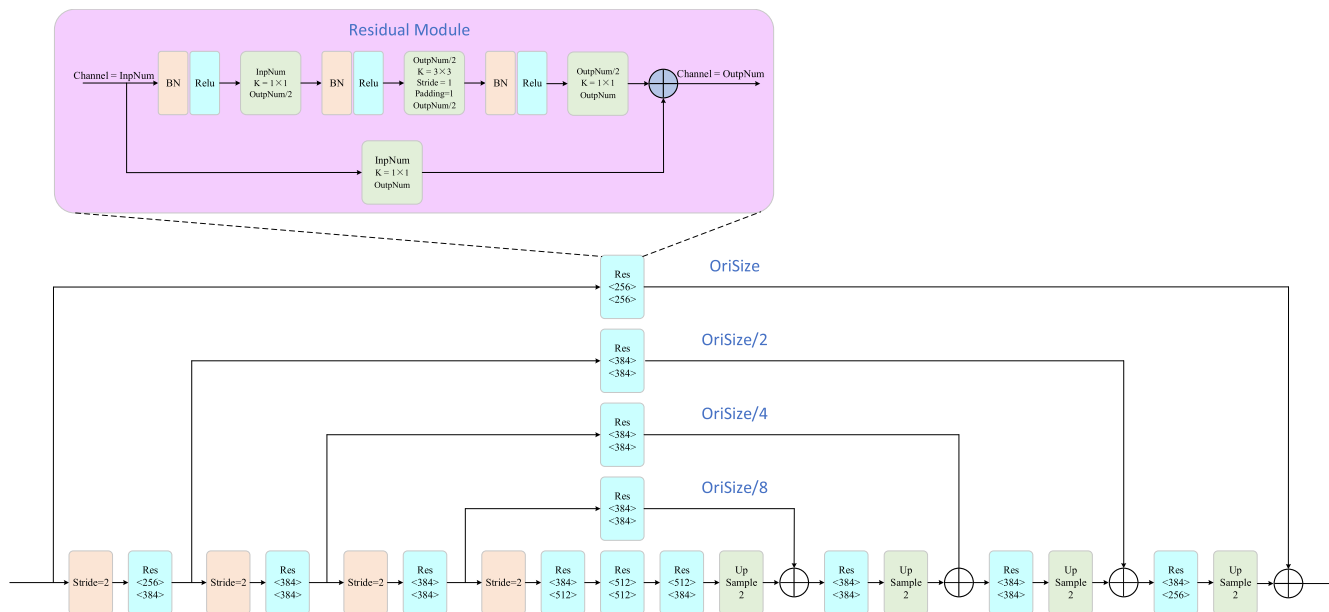


FIGURE 1. Hourglass network structure.

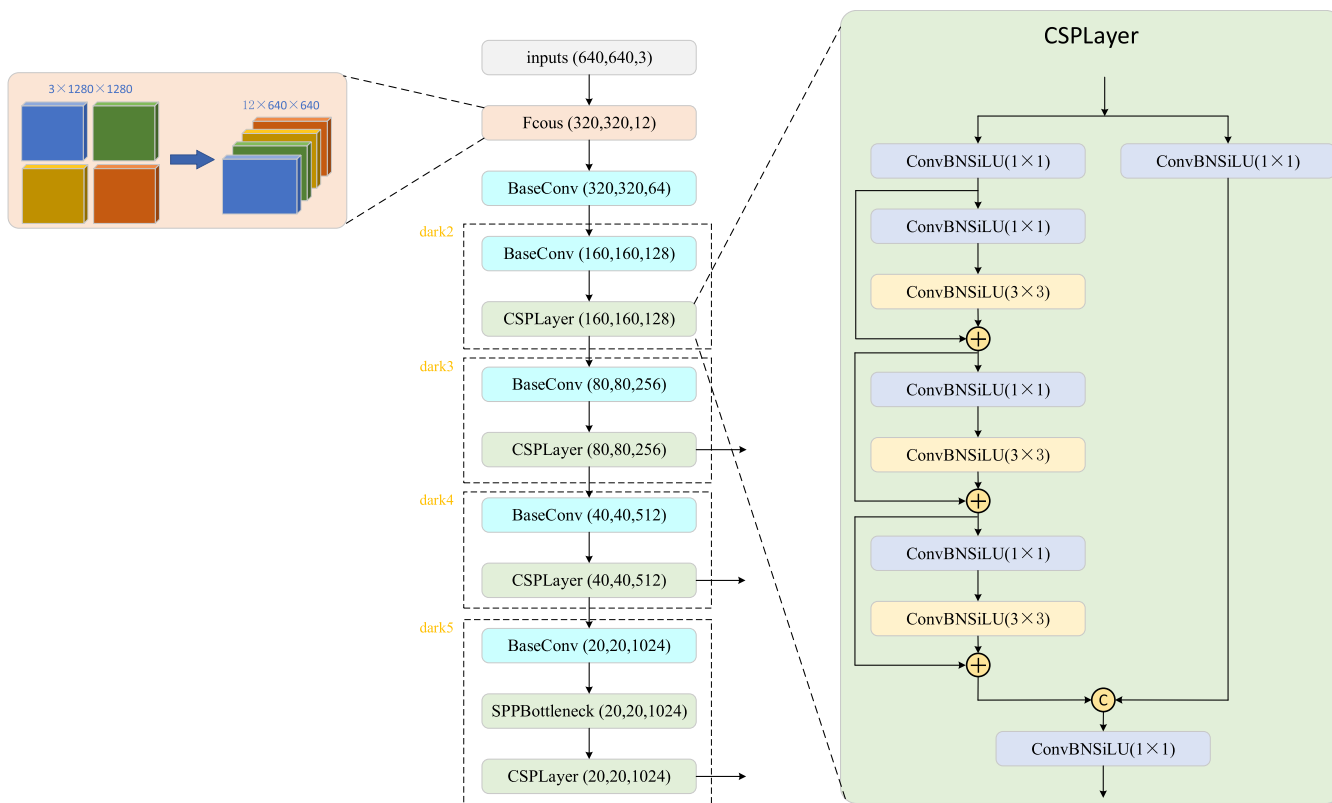


FIGURE 2. CSPDarknet53 network structure.

time, max-pooling is performed from the top down for each column; The two feature maps that performed max pooling are added, and the maximal area is the predicted top-left corner point, as shown in Figure 3.

CenterNet [41] obtains the centre heatmap and corner heatmaps from centre pooling and cascade corner pooling, respectively, which are used to predict the location of the key points. Similar to max pooling in CornerNet, centre pooling

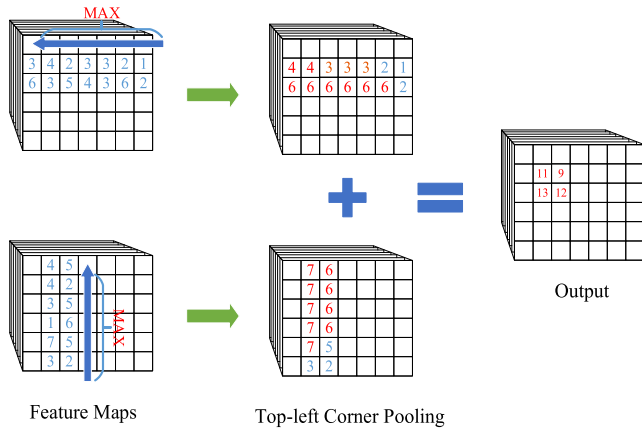


FIGURE 3. Schematic diagram of Corner pooling.

obtains the centre point by concatenating max pooling operations in different directions, up and down. Cascade corner pooling first extracts the boundary maxima of the feature image object region, and then continues to extract the maxima internally at the boundary maxima and sums them with the boundary maxima. Compared to CornerNet’s corner points, CenterNet’s corner points have richer semantics concerning the associated objects, as shown in Figure 3.

Extreme points: The ExtremeNet [44] method uses four-point annotation to annotate the extreme points of the object in each of the four directions when annotating data. The key point is inferred from the peak on the heatmap  $\hat{Y}^{(C)} \in (0, 1)^{H \times W}$  output by the backbone network, which is the value of a position on the heatmap that exceeds a certain threshold  $\tau_p$  and is the maximum value within its  $3 \times 3$  grid. The fully convolutional encoder-decoder network is then used to predict a multichannel heatmap, and each channel corresponds to a key point.

C. KEY POINT LOSS FUNCTION

The bounding box generation of the anchor-free method is based on the location of the key points, but points around the ground truth location can also generate bounding boxes that satisfy the intersection-over-union (IoU) condition. Thus, CornerNet gives an unstandardized penalty factor consisting of a two-dimensional Gaussian function that reduces the penalty for negative samples within a certain radius around the ground truth; the closer a sample is to the ground truth, the smaller the penalty.

$$y_{cij} = \begin{cases} e^{-\frac{x^2+y^2}{2\sigma^2}} & x^2 + y^2 \leq r^2 \\ 0 & o.w. \end{cases} \quad (1)$$

Here,  $y_{cij}$  is denoted as the value when the heatmap is a positive sample at position  $(i, j)$  and is classified as Class  $c$ .  $y_{cij} = 1$  and  $y_{cij} = 0$  represent positive samples and negative samples, respectively, where  $x$  and  $y$  denote the position of the negative sample  $(i, j)$  relative to the coordinates of the positive sample (centre of the circle).  $\sigma = r/3$ , where  $r$

denotes the radius of the circle. The  $y_{cij}$  value decreases more slowly as the negative sample moves away from the positive sample. To maintain the consistency of the penalty and the increase/decrease in distance, the penalty factor is set to  $(1 - y_{cij})$ , so the loss function for key point detection is: (2), as shown at the bottom of the next page.

$N$  the number of objects in the image,  $\alpha$  and  $\beta$  are hyper-parameters that control the contribution to the loss,  $P_{cij}$  is the predicted value on the prediction heatmap, and the predicted location  $(i, j)$  is the probability of the corner being classified as  $c$ .

The method of this paper extracts the key point map (shown in Figure 5) during the detection process and predicts the corner, centre, and extreme points of the object. With the prediction of the key points, the anchor points generated at the anchor base are eliminated, and the object is ensured to have response in the feature map.

D. MAIN PROCESS

To gain an overall understanding of these anchor-free methods, this subsection summarizes the overall flow of the methods to better understand their processes for handling data and the differences between them.

CornerNet [40] first inputs the images to the backbone network, Hourglass Network-104, and the output feature images are then input to two prediction modules for the top-left and bottom-right corners of the bounding box. To determine that the top-left corner point and the bottom-right corner point belong to the same object, drawing on the Newell [46] associative embedding method, an embedding is generated for the corner points while detecting them, and the corner point is grouped by calculating the distance between the top-left and bottom-right embedding, with the smaller distance indicating that the two corner points belong to the same group. The offsets module is used to predict the offset of the corner position, adjust the corner position, and map it back to the input resolution.

CenterNet [41] uses a fully convolutional network to directly obtain a 4-fold downsampled heatmap, with the number of channels of the heatmap equal to the number of object categories to be detected, and then uses centre pooling and cascade corner pooling to obtain the centre heatmap and corner heatmaps, respectively, which are used to predict the position of key points. After obtaining the corner position and category, the offsets map the corner position to the corresponding position in the input image, and then the embeddings determine which two corners belong to the same object to form a bounding box. For more accurate detection, CenterNet predicts not only the corners but also the centre point. CenterNet defines a centre area for each bounding box and determines whether the centre area of each bounding box contains a centre point. If it does, the prediction box is retained, and the confidence of the box is the average of the confidence of the centre, top-left and bottom-right points. If it does not, the bounding box is removed, which gives the network the ability to perceive the information inside the

object area and can effectively remove incorrect bounding box.

Objects as Points [43] passes the image into the backbone network, and a quadruple-downsampled heatmap is obtained; then, a  $3 \times 3$  maxpool layer is used to extract the peak point of the heatmap, i.e., the centre point, and the peak point position of each feature map predicts the width and height information of the object as well as the offset of the centre point and the bounding box size.

CornerNet-Lite [42] is a combination of two effective variants of CornerNet: CornerNet-Saccade, which uses an attention mechanism to avoid processing all pixels of the image, and CornerNet-Squeeze, which introduces a new compact backbone architecture.

CornerNet-Saccade finds the correct size of the foreground area with an attention map and then crops it out for the next stage of the fine inspection image. Therefore, CornerNet-Saccade is divided into two stages: object location estimation and object detection. The first stage of CornerNet-Saccade predicts three different sizes of attention maps and some coarse bounding boxes from the downsized images to obtain the positions and rough sizes of the objects in the images, which need to be evaluated later. The second stage of CornerNet-Saccade crops out the object region on the original map based on the object location predicted by the attention maps and the coarse bounding box. The final bounding box is generated in the cropping region by the corner point detection mechanism, exactly as in CornerNet.

To reduce CornerNet's computing resources on Hourglass-104, CornerNet-Squeeze was proposed, inspired by SqueezeNet [48], to replace residual blocks with Fire modules in SqueezeNet; inspired by MobileNet [49], CornerNet-Squeeze replaced the layer  $2 \times 3 \times 3$  standard convolution with  $3 \times 3$  deep separable convolution.

ExtremeNet [44] uses Hourglass to detect 5 key points (4 extreme points and 1 centre point) for each classification. Since there are many key points for predicting the outputs of the four channels of the extreme points and there are  $n^4$  ways to combine them, to make it easier to group them, these keypoints are recorded as  $t$ ,  $b$ ,  $r$ , and  $l$ ; then, the resulting geometric centroids are:  $c = \left( \frac{t_x+r_x}{2}, \frac{t_y+b_y}{2} \right)$ . If the value of this geometric centre point on the centre heatmap is greater than a certain threshold,  $\hat{Y}_{C_x, C_y}^{(C)} \geq \tau_p$ , then the set of key points is valid, i.e., the key points belong to the same object.

YOLOx [45] switches YOLO [5] to an anchor-free strategy, it reduces the predictions for each location from 3 to 1 and directly predicts four values (two offsets in terms of the left-top corner of the grid, as well as the height and width of the predicted box). Through this modification, the

detector's parameters and giga floating-point operations per second (GFLOPs) are reduced to make it faster.

#### IV. EXPERIMENTS

The experiments used the six anchor-free object detection algorithms described above, the CornerNet method and the CornerNet-Lite method based on corner point detection, the CenterNet method based on a combination of corner and centre points, the Objects as Points method based on the centre point, and the ExtremeNet method using extreme point detection. The anchor-based methods—Faster-RCNN, YOLOv3 and YOLOv5—were also compared for experimental completeness.

**Dataset:** Since X-ray baggage security images are unconventional images with few sources of data acquisition and even fewer datasets for object detection of knives and guns, the data for this experiment were obtained from a X-ray machine manufacturer, and several different types of knives, handguns, and other items were combined for X-ray scanning. From the tens of thousands of pictures, 10,233 X-ray pictures of knives and pistols were selected as the main material for the experiment. To obtain a more complete dataset, we carried out extensive image annotation work using an annotation tool to create the labels required for the experiment from the positions of the knives and handguns in the image. On average, each image contains two to three labels.

**Training Details:** We trained the methods using the PyTorch framework with an image input size of  $511 \times 511$  and an output size of  $128 \times 128$ . To reduce overfitting, standard data augmentation was used, including random horizontal flipping, random scaling, random cropping, and random colour dithering, which included adjusting the brightness, saturation, and contrast of the image; the training loss was optimized using Adam. The number of training iterations was 100,000, the learning rate was  $2.5 \times 10^{-4}$ , and the batch size varied depending on the network size and number of stacks; the more parameters there were, the smaller the batch size. Training was performed on a single Nvidia GeForce Titan 1080 GPU, and each network training took approximately two days to complete.

**Evaluation:** We evaluated the performance of the anchor-free methods in the X-ray baggage security image object detection task using the mAP and average recall (AR), which were averaged across multiple IoUs using 3 IoU thresholds  $\text{IoU} \in [0.5:0.75:0.95]$ , which could enable a better location and position of the object detector. To test the performance of these models, we divided part of the dataset into images and labels for testing, from which we selected 1,000 for the validation set and 1,000 for the test set.

$$L_{\text{det}} = -\frac{1}{N} \sum_{c=1}^C \sum_{i=1}^H \sum_{j=1}^W \begin{cases} (1 - p_{cij})^\alpha \log(p_{cij}) & y_{cij} = 1 \\ (1 - y_{cij})^\beta (p_{cij})^\alpha \log(1 - p_{cij}) & o.w. \end{cases} \quad (2)$$

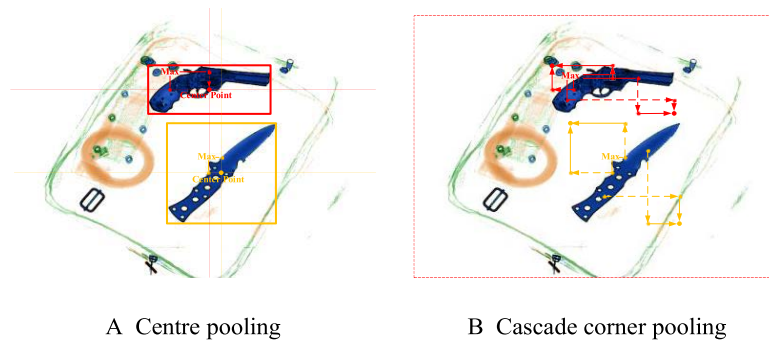


FIGURE 4. Schematic diagram of CenterNet.

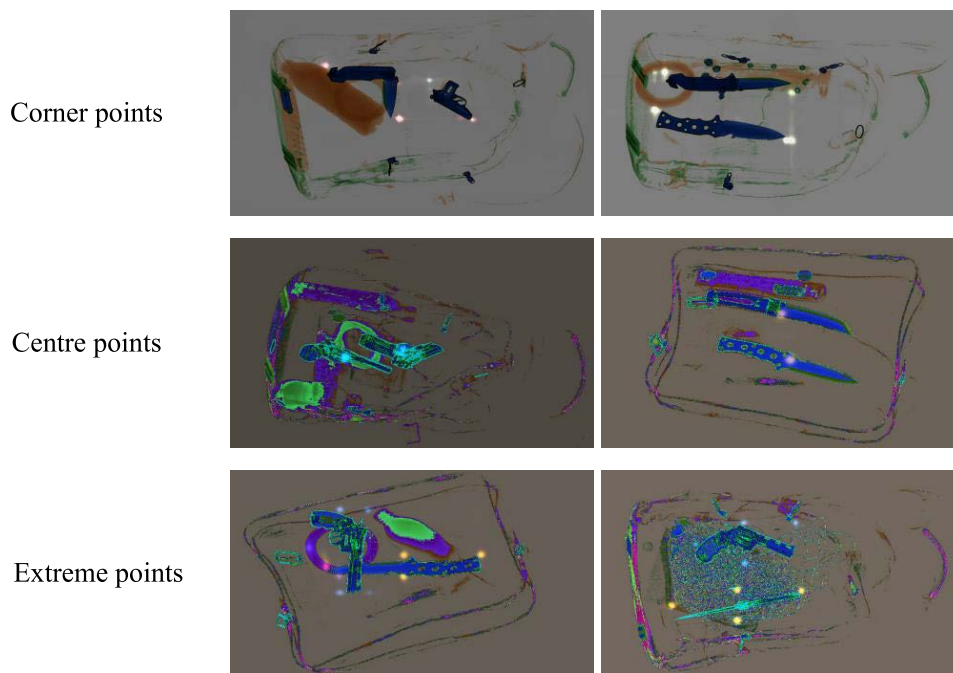


FIGURE 5. Key points map.

TABLE 1. Detection results of anchor-based and anchor-free.

	Methods	Backbone	FPS	AP50	AP75	AP50:95
Anchor-based	Faster-rcnn	VGG16	14.2	0.879	—	—
	YOLOv3	Darknet-53	29.6	0.882	0.865	0.665
	YOLOv5	CSPDarknet53	66.98	0.971	0.880	0.806
	ComerNet	Hourglass-104	2.2	0.860	0.800	0.761
		Hourglass	2.79	0.874	0.810	0.770
	ComerNet-Squeeze	Hourglass-104	16.8	0.932	0.899	0.849
	ComerNet-Saccade	Hourglass	16.6	0.612	0.550	0.530
Anchor-free	CenterNet	Hourglass-104	1.78	0.907	0.836	0.790
		Hourglass	2.29	0.943	0.895	0.843
	ExtremeNet	Hourglass-104	2.3	0.966	0.956	0.900
		Hourglass	13.3	0.952	0.946	0.865
	Objects as points	Hourglass-104	18.6	0.979	0.951	0.878
		Resnet-18	<b>71.1</b>	0.969	0.927	0.816
		DLA-34	55.6	0.978	0.950	0.881
YOLOX	CSPDarknet53	40.5	<b>0.996</b>	<b>0.966</b>	<b>0.905</b>	

Table 1 presents the final experimental results, with the addition of three classical anchor-based methods as comparison experiments in addition to the six anchor-

free methods, each based on its own applicable network structure, and evaluation metrics including the number of frames per second (FPS) and three IoU



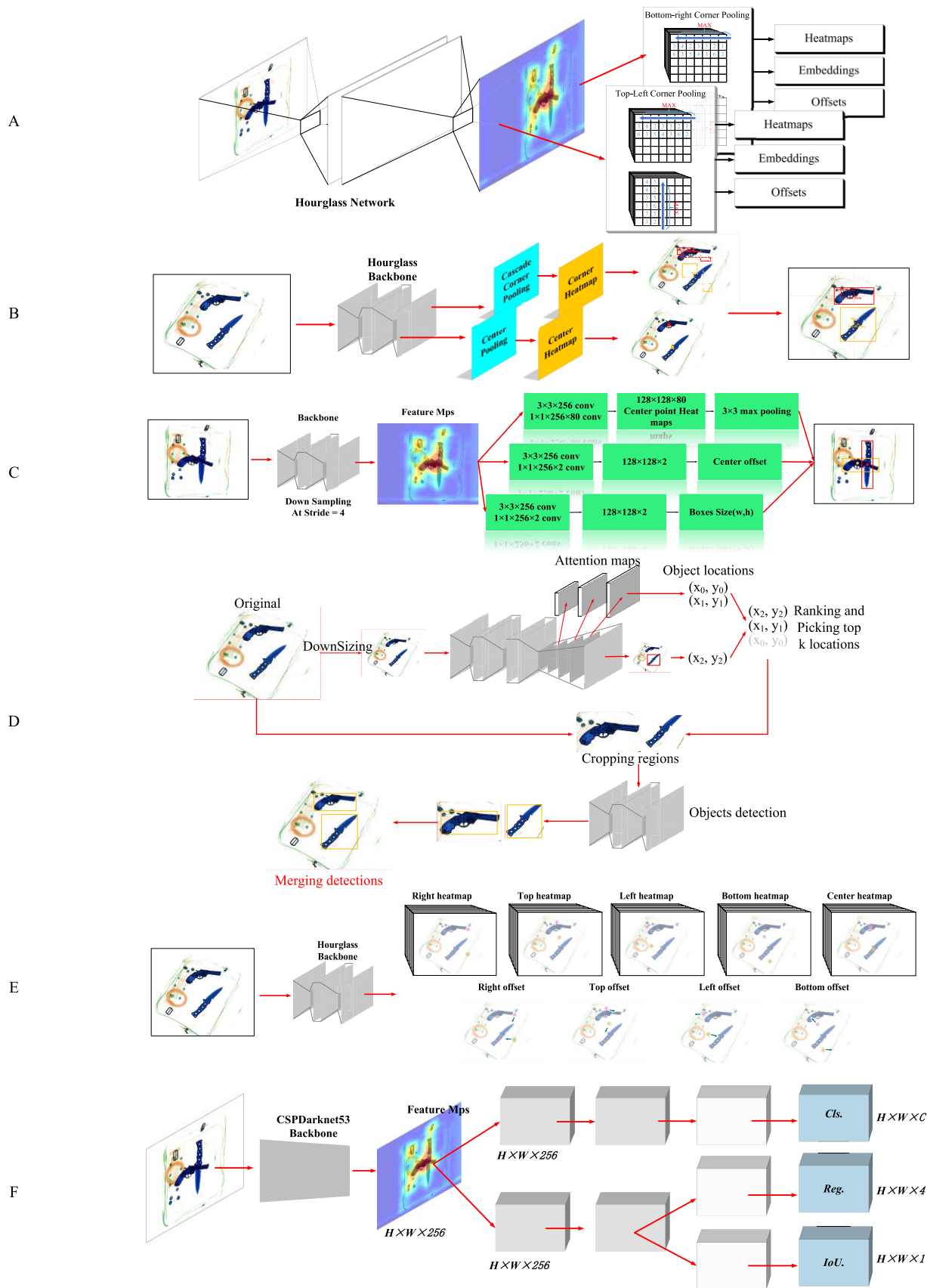


FIGURE 6. Method flowchart: A. CornerNet, B. CenterNet, C. Objects as Points, D. CornerNet-Lite, E. ExtremeNet, F. YOLOx.

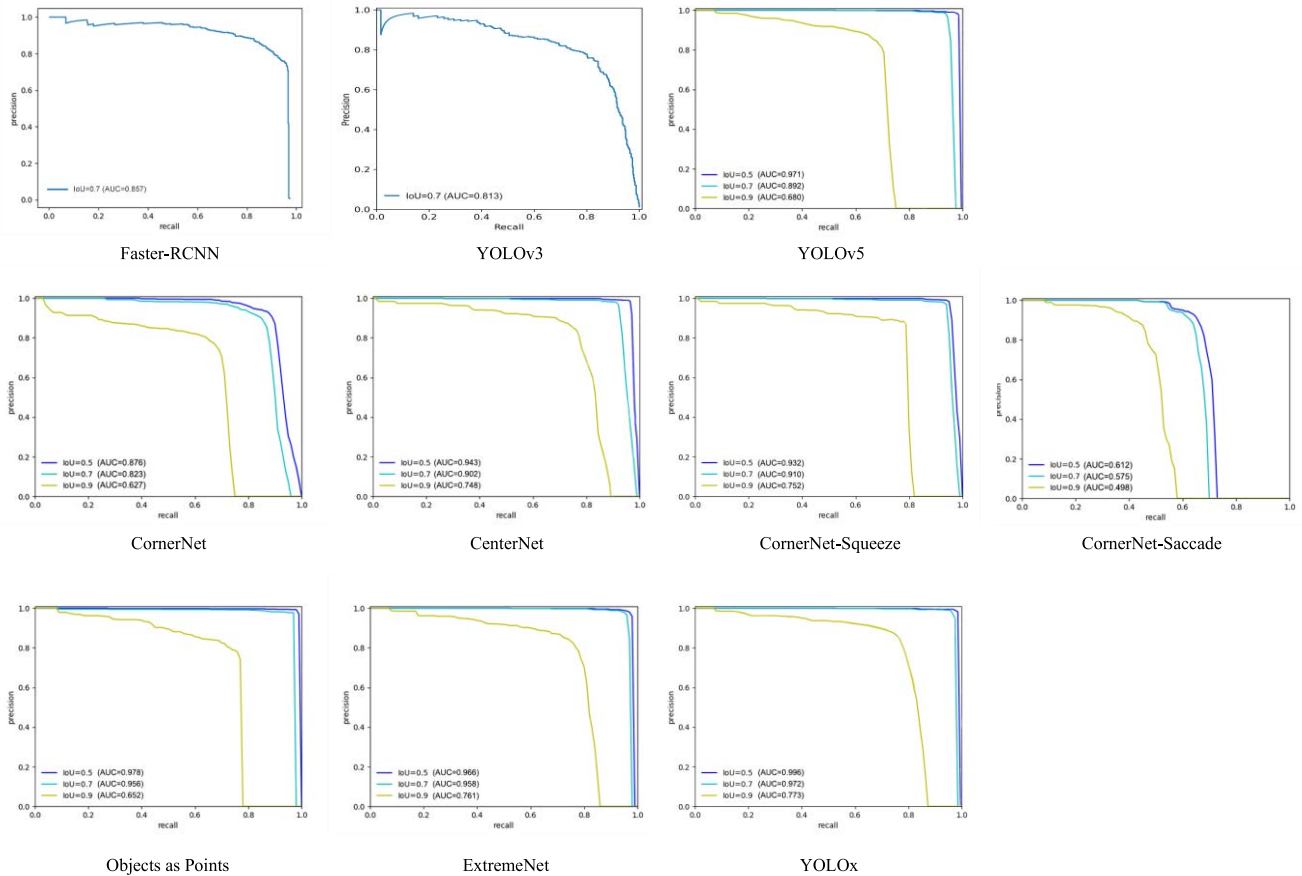


FIGURE 7. P-R (precision recall) diagram.

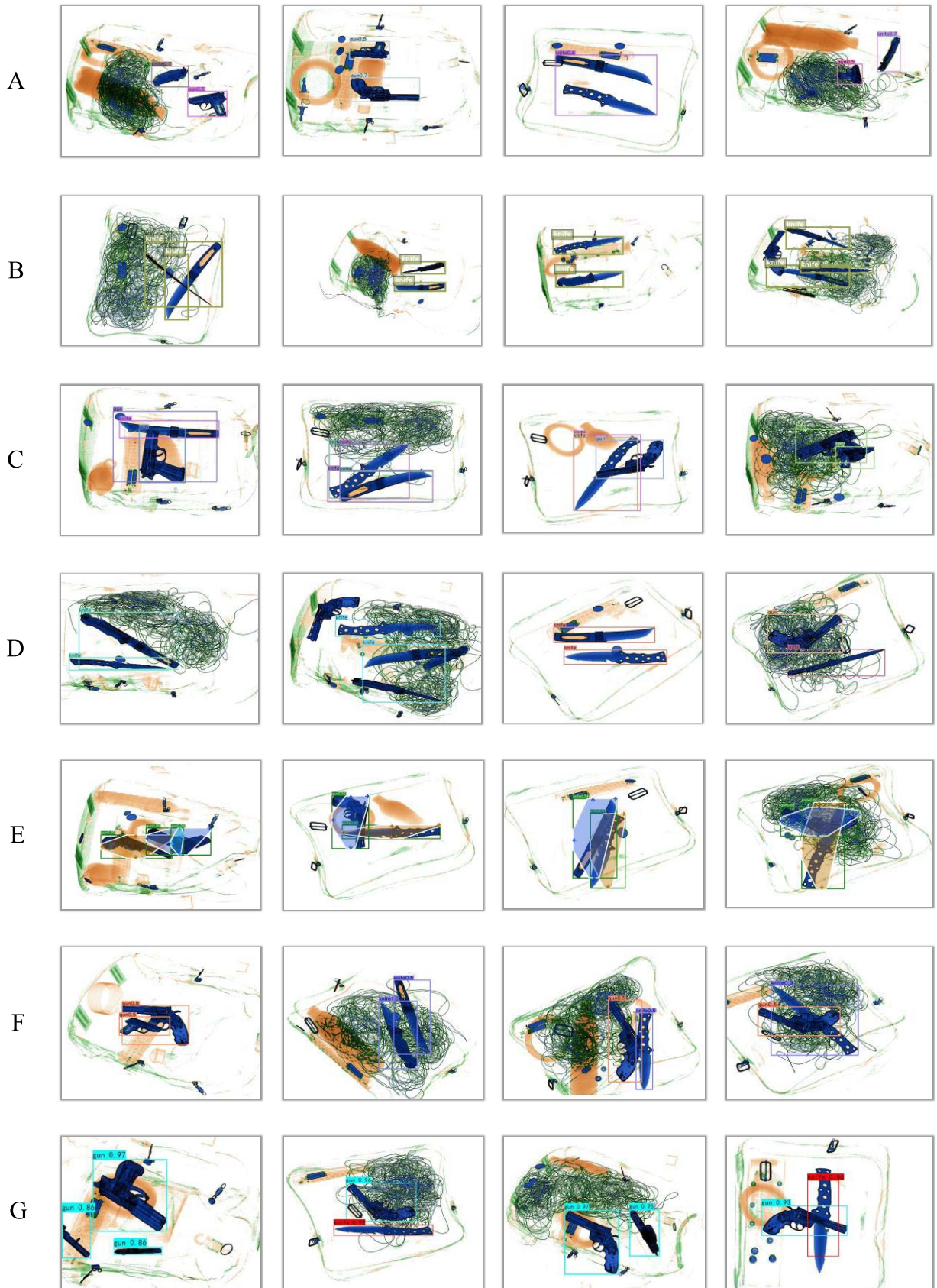
threshold. The precision-recall (P-R) curve is shown in Figure 7.

Since the anchor-free method is more in line with the way the human eye’s attention mechanism locates objects, the goal of this paper is to determine whether the anchor-free method is feasible and superior in the field of object detection in X-ray baggage security images; from the experimental results, the performances of the Objects as Points, ExtremeNet and YOLOx methods were impressive. Their detection accuracy almost reached or even surpassed that of YOLOv5, a new anchor-based method. Objects as Points reached 18.6 FPS with Hourglass-104, surpassing Faster-RCNN in speed and performing similarly to it in accuracy; with DLA-32, it reached 55.6 FPS with 0.881 accuracy. ExtremeNet had an FPS of 2.3 and an accuracy of 0.900 with Hourglass-104. YOLOx had an FPS of 40.5 on the CSPDarknet53 network, which was not as high as that of YOLOv5, but it showed the best accuracy across all thresholds of IoU. Therefore, these three anchor-free methods have advantages over the anchor-based methods used in this paper for X-ray baggage screening images, as shown in Figure 7.

V. DISCUSSION

The experiment introduces six anchor-free methods for the detection of knives and handguns in X-ray baggage security

images. There is some research continuity between these methods. The CornerNet method locates an object through corner points. Due to the absence of anchor restrictions, combining the corner points into an accurate bounding box requires a very high-level corner point combination algorithm because the assistance of global information is not available in determining whether two corner points belong to the same object; therefore, it is easy to combine two corner points of different objects into a bounding box. Therefore, in determining whether the top-left corner and bottom-right corner belong to the same object, CenterNet considers adding centre point information to further determine whether the centre of the box consisting of these two points contains a centre point with a high response value. Likewise, ExtremeNet predicts four extreme points and predicts a central point to increase the confidence level of the extreme point combination. From the results, CenterNet is more accurate than CornerNet, and the ExtremeNet method has the highest accuracy of all methods, verifying that the centre point is indeed effective in improving detection accuracy. The YOLOx method assigns a 3 × 3 area in the centre location of each object as a positive sample, which means that YOLOx also adopts the anchor-free strategy of the centre point but expands this point to a certain range, which further verifies the importance of the centre point strategy for anchor-free methods.



**FIGURE 8.** Detection effect demonstration A. CornerNet, B. CenterNet, C. CornerNet- Saccade, D. CornerNet-Squeeze, E. ExtremeNet, F. Objects as Points, G. YOLOx.

In addition, we find that the detection speed is affected when considering the combination of two types of key points. The fastest FPS of CornerNet, CenterNet and ExtremeNet on the Hourglass-104 network is only 2.3, which poses a great challenge for meeting real-time requirements. The Objects as Points method considers the centre point as an important positioning tool and aims to maintain the detection speed, so it simplifies, by focusing only on the centre position and its offset prediction, and it does not consider other types of key point combinations, which greatly improves the detection speed. What is more surprising is that the detection accuracy is still excellent.

To further exploit the performance advantages of the Objects as Points method, we also aimed to merge the deep-layer aggregation (DLA) [50] network and the ResNet-18 network, which resulted in a very significant performance improvement, especially on the basis of the DLA network, along with satisfactory speed and accuracy. As a result, the accuracy and FPS were 0.881 and 55.6, respectively. Both the DLA network and Hourglass are neural network structures with feature fusion functions, while the DLA network can fuse semantic and spatial information for recognition and localization by extending the common method of allowing skip connections and using the aggregation structure of multi-level skip connections. With improved model performance and a reduced number of model parameters compared to that of hourglass, DLA is able to support the Objects as Points method to dramatically increase detection speed while maintaining high accuracy.

## VI. CONCLUSION

In the task of weapon detection in X-ray baggage images, computer vision-aided detection requires high accuracy and real-time performance, but existing anchor-based methods are not very generalizable and require anchors with different sizes and aspect ratios to be set for different datasets; such settings can be considered hyperparameters, which have an impact on the average accuracy. In addition, to improve detection recall, it is generally necessary to densely flatten a large number of anchors, which on the one hand makes the matching computation IoU larger and on the other hand leads to an extreme imbalance between positive and negative samples. To address these problems, we introduced the anchor-free method in an attempt to improve the accuracy and speed of weapon detection in X-ray baggage images. For this purpose, we obtained a large number of X-ray scan images, mainly including guns and knives, from X-ray equipment manufacturers and carried out extensive data annotation work to produce the datasets used for the experiments. Then, exhaustive comparative experiments were conducted between anchor-based and anchor-free methods, and the experimental results were analysed.

ExtremeNet, Objects as Points and YOLOx, anchor-free methods outperformed anchor-based methods used in this paper in the detection of weapons in X-ray baggage security images. YOLOx had the highest overall accuracy of 0.905 on

the CSPDarknet53 network. ExtremeNet achieved a detection accuracy of 0.900 on the Hourglass-104 skeleton network, and Objects as Points achieved an accuracy of 0.881 on the DLA-34 skeleton network. Additionally, given the real-time nature of the detection task, Objects as Points worked well with a lighter-weight network structure. Overall, the anchor-free approach is simpler and more flexible and can be improved and developed further.

In the future, more classes of datasets can be constructed to further enrich the object detection dataset of X-ray baggage security images; in addition, with the emergence of better skeleton network structures, the anchor-free method can achieve improved detection accuracy and speed accordingly.

## REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and E. G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.
- [2] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
- [3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.
- [4] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.
- [5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
- [6] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
- [7] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
- [8] C. A. Carlsson and G. A. A. Carlsson. *Basic Physics of X-Ray Imaging*. Stockholm, Sweden: Linköping Univ. Electronic Press, 1973, [Online]. Available: <https://www.diva-portal.org/smash/get/diva2:276160/FULLTEXT02.pdf>
- [9] S. Singh and M. Singh, "Explosives detection systems (EDS) for aviation security," *Signal Process.*, vol. 83, no. 1, pp. 31–55, 2003.
- [10] G. M. Blake and I. Fogelman, "Technical principles of dual energy X-ray absorptiometry," *Semin. Nucl. Med.*, vol. 27, no. 3, pp. 210–228, Jul. 1997.
- [11] B. R. Abidi, Y. Zheng, A. V. Gribok, and M. A. Abidi, "Improving weapon detection in single energy X-ray images through pseudocoloring," *IEEE Trans. Syst., Man C, Appl. Rev.*, vol. 36, no. 6, pp. 784–796, Nov. 2006.
- [12] M. Singh, S. Singh, and D. Partridge, "A knowledge-based framework for image enhancement in aviation security," *IEEE Trans. Syst., Man B, Cybern.*, vol. 34, no. 6, pp. 2354–2365, Dec. 2004.
- [13] B. R. Abidi, J. Liang, M. Mitkes, and M. A. Abidi, "Improving the detection of low-density weapons in X-ray luggage scans using image enhancement and novel scene-decluttering techniques," *J. Electron. Imag.*, vol. 13, no. 3, pp. 523–538, 2004.
- [14] Z. Chen, Y. Zheng, B. R. Abidi, D. L. Page, and M. A. Abidi, "A combinational approach to the fusion, de-noising and enhancement of dual-energy X-ray luggage images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Sep. 2005, p. 2.
- [15] B. R. Abidi, Y. Zheng, A. V. Gribok, and M. A. Abidi, "Improving weapon detection in single energy X-ray images through pseudocoloring," *IEEE Trans. Syst., Man C, Appl. Rev.*, vol. 36, no. 6, pp. 784–796, Nov. 2006.
- [16] S. Nercessian, K. Panetta, and S. Aghaian, "Automatic detection of potential threat objects in X-ray luggage scan images," in *Proc. IEEE Conf. Technol. Homeland Secur.*, May 2008, pp. 504–509.
- [17] D. Mery, "Automated detection in complex objects using a tracking algorithm in multiple X-ray views," in *Proc. CVPR Workshops*, Jun. 2011, pp. 41–48.
- [18] M. Bastan, M. R. Yousefi, and T. M. Breuel, "Visual words on baggage X-ray images," in *Proc. Int. Conf. Comput. Anal. Images Patterns*. Cham, Switzerland: Springer, 2011, pp. 360–368.

- [19] T. Franzel, U. Schmidt, and S. Roth, "Object detection in multi-view X-ray images," in *Proc. Joint DAGM (German Assoc. Pattern Recognit.) OAGM Symp.* Cham, Switzerland: Springer, 2012, pp. 144–154.
- [20] L. Schmidt-Hackenberg, M. R. Yousefi, and T. M. Breuel, "Visual cortex inspired features for object detection in X-ray images," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 2573–2576.
- [21] D. Turcsany, A. Mouton, and T. P. Breckon, "Improving feature-based object recognition for X-ray baggage security screening using primed visualwords," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Feb. 2013, pp. 1140–1145.
- [22] M. Baştan, "Multi-view object detection in dual-energy X-ray images," *Mach. Vis. Appl.*, vol. 26, nos. 7–8, pp. 1045–1060, Nov. 2015.
- [23] M. E. Kundegorski, S. Akcay, M. Devereux, A. Mouton, and T. P. Breckon, "On using feature descriptors as visual words for object detection within X-ray baggage security screening," in *Proc. 7th Int. Conf. Imag. Crime Detection Prevention (ICDP)*, 2016, p. 12.
- [24] Q. Lu and R. W. Conners, "Using image processing methods to improve the explosive detection accuracy," *IEEE Trans. Syst., Man C, Appl. Rev.*, vol. 36, no. 6, pp. 750–760, Nov. 2006.
- [25] S. Akcay, M. E. Kundegorski, C. G. Willcocks, and T. P. Breckon, "Using deep convolutional neural network architectures for object classification and detection within X-ray baggage security imagery," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2203–2215, Sep. 2018.
- [26] S. Akcay, M. E. Kundegorski, M. Devereux, and T. P. Breckon, "Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1057–1061.
- [27] R. L. Galvez, E. P. Dadios, A. A. Bandala, and R. R. P. Vicerra, "YOLO-based threat object detection in X-ray images," in *Proc. IEEE 11th Int. Conf. Humanoid, Nanotechnol., Inf. Technol., Commun. Control, Environ., Manage. (HNICEM)*, Nov. 2019, pp. 1–5.
- [28] J. Koci, A. O. Topal, and M. Ali, "Threat object detection in X-ray images using SSD, R-FCN and faster R-CNN," in *Proc. Int. Conf. Comput., Netw., Telecommun. Eng. Sci. Appl. (CoNTESA)*, Dec. 2020, pp. 10–15.
- [29] V. Ponnusamy, D. R. Marur, D. Dhanaskodi, and T. Palaniappan, "Deep learning-based X-ray baggage hazardous object detection—An FPGA implementation," *Revue d'Intell. Artificielle*, vol. 35, no. 5, pp. 431–435, Oct. 2021.
- [30] D. Saavedra, S. Banerjee, and D. Mery, "Detection of threat objects in baggage inspection with X-ray images using deep learning," *Neural. Comput. Appl.*, vol. 33, pp. 7803–7819, Jul. 2021.
- [31] A. Chang, Y. Zhang, S. Zhang, L. Zhong, and L. Zhang, "Detecting prohibited objects with physical size constraint from cluttered X-ray baggage images," *Knowl.-Based Syst.*, vol. 237, Feb. 2022, Art. no. 107916.
- [32] E. E. Altındağ and S. E. Yuksel, "Threat detection in X-ray baggage security imagery using convolutional neural networks," *Proc. SPIE*, vol. 12104, pp. 113–125, Jun. 2022.
- [33] C. Ma, L. Zhuo, J. Li, Y. Zhang, and J. Zhang, "EAOD-N et: Effective anomaly object detection networks for X-ray images," *IET Image Process.*, vol. 16, no. 10, pp. 2638–2651, Aug. 2022.
- [34] D. Mery and A. K. Katsaggelos, "A logarithmic X-ray imaging model for baggage inspection: Simulation and object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 57–65.
- [35] D. Mery, E. Svec, M. Arias, V. Riffio, J. M. Saavedra, and S. Banerjee, "Modern computer vision techniques for X-ray testing in baggage inspection," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 47, no. 4, pp. 682–692, Apr. 2017.
- [36] S. Akcay, M. E. Kundegorski, C. G. Willcocks, and T. P. Breckon, "Using deep convolutional neural network architectures for object classification and detection within X-ray baggage security imagery," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2203–2215, Sep. 2018.
- [37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [38] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.
- [39] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2016, pp. 21–37.
- [40] H. Law and J. Deng, "CornerNet: Detecting objects as paired key-points," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 734–750.
- [41] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.
- [42] H. Law, Y. Teng, O. Russakovsky, and J. Deng, "CornerNet-lite: Efficient keypoint based object detection," 2019, *arXiv:1904.08900*.
- [43] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.
- [44] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 850–859.
- [45] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding Yolo series in 2021," 2021, *arXiv:2107.08430*.
- [46] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2016, pp. 483–499.
- [47] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 390–391.
- [48] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*.
- [49] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
- [50] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.



**YAN HUANG** received the master's degree in transportation engineering from the Changsha University of Science and Technology, Changsha, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Civil and Transportation Engineering, South China University of Technology, Guangzhou, China. His research interests include object detection and LiDAR simultaneous localization and mapping.



**XINSHA FU** received the State Council Special Allowance in 1993 and was exceptionally promoted as a Professor, in 1998. He has published over 60 articles and five monographs. His research interests include highway planning and design, computer aided engineering and design of highways, transportation infrastructure management systems, intelligent transportation systems, 3S technology, and teaching and research of traffic information. He was awarded two Second Prizes and seven Third Prizes of provincial and ministry-level awards.



**YANJIIE ZENG** received the Ph.D. degree from the School of Civil and Transportation Engineering, South China University of Technology, Guangzhou, China, in 2021. He is currently with the Guangdong Provincial Transport Planning and Research Center (GTPRC), Guangzhou. His research interests include object detection and visual tracking.