IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 19, 2020, accepted February 29, 2020, date of publication March 6, 2020, date of current version March 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2978912

Efficient Foreign Object Detection Between PSDs and Metro Doors via Deep Neural Networks

YUAN DAI[1], **WEIMING LIU**[1], **HAIYU LI**[2], **AND LAN LIU**[2]
[1]School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510641, China
[2]Guangzhou Metro Group Company, Ltd., Guangzhou 510335, China

Corresponding author: Weiming Liu (mingweiliu@126.com)

This work was supported by the China's 13th Five-Year Key Research and Development Plan for the Safety Guarantee Technology of Urban Rail Transit System under Grant 2016YFB1200402.

ABSTRACT Platform Screen Doors (PSDs) have been widely used in modern Asian and European metro systems due to the advantages of safety, comfort for passengers. Unfortunately, someone or something will be caught by PSDs and metro doors occasionally, which may lead to serious accidents. Therefore, the foreign object detection between PSDs and metro doors is a burning problem. Moreover, this problem is a challenging and still largely under-explored topic. In recent years, we have seen significant improvements in generic object detection built on deep learning techniques. Accordingly, this paper adopts deep learning technologies to address the problem of foreign object detection between PSDs and metro doors. To the best of our knowledge, this is the first attempt to use deep learning to solve the problem. To realize this, a dataset including 984 real-world images (with 600 × 480 pixels) labeled for six types of foreign objects (*bag*, *bottle*, *person*, *plastic bag*, *umbrella*, *other*) is developed. Then, we compared the performance of some state-of-the-art object detection algorithms (such as You Only Look Once -YOLOv3, Single Shot MultiBox Detector -SSD, and CenterNet) on the dataset. Experimental results demonstrate that the foreign object detection algorithms based on deep neural networks have achieved excellent results, which not only improves the accuracy of detection but also give the categories of foreign objects. YOLOv3 - tiny can achieve the fastest detection speed, up to 200 Frame Per Second (FPS); CenterNet can achieve the best detection results, up to 99.7% mean Average Precision (mAP).

INDEX TERMS PSDs, metro doors, foreign object detection, deep learning, computer vision.

I. INTRODUCTION

PSD is fast becoming a key instrument in modern metro systems, due to its comfort and safety. According to the Metro Industry Standards, a certain gap (as shown in Fig. 1) must be kept between PSDs and metro doors to ensure the safety of passengers. Occasionally, people or objects are clamped in the gap by PSDs and metro doors, which usually leads to serious accidents. Hence, it holds great practical significance to solve the problem of foreign object detection between PSDs and metro doors. For this reason, this paper is focused on how to detect foreign objects rapidly and accurately.

Over the years, many metro corporations still use manual observation to determine whether there exists foreign objects between PSDs and metro doors. In a consequence of the complex metro systems and the influence of various external environments, the false detection rate of manual observation is high. Moreover, the driverless metro is the inevitable trend of urban rail transit in the future, so it is no longer possible to detect foreign objects by manual observation. On one hand, there are little research on detecting foreign objects between PSDs and metro doors. On the other hand, with the development of sensor technology, there has been a lot of research on the area of foreign objects invasion in general sense. Thus, some researchers are trying to solve the problem of foreign object detection between PSDs and metro doors with traditional sensor technology. Foreign object detection system based on sensor technology mainly includes the following three ways: infrared multi-beam detector [1], laser-based detector [2] and video-sensors-based detector [3].

Infrared multi-beam detector and the laser-based detector work similarly in principle, but the former has a certain emission angle, and the laser focusing performance is strong. Infrared multi-beam detector and the laser-based detector are

The associate editor coordinating the review of this manuscript and approving it for publication was Jihwan P. Choi.

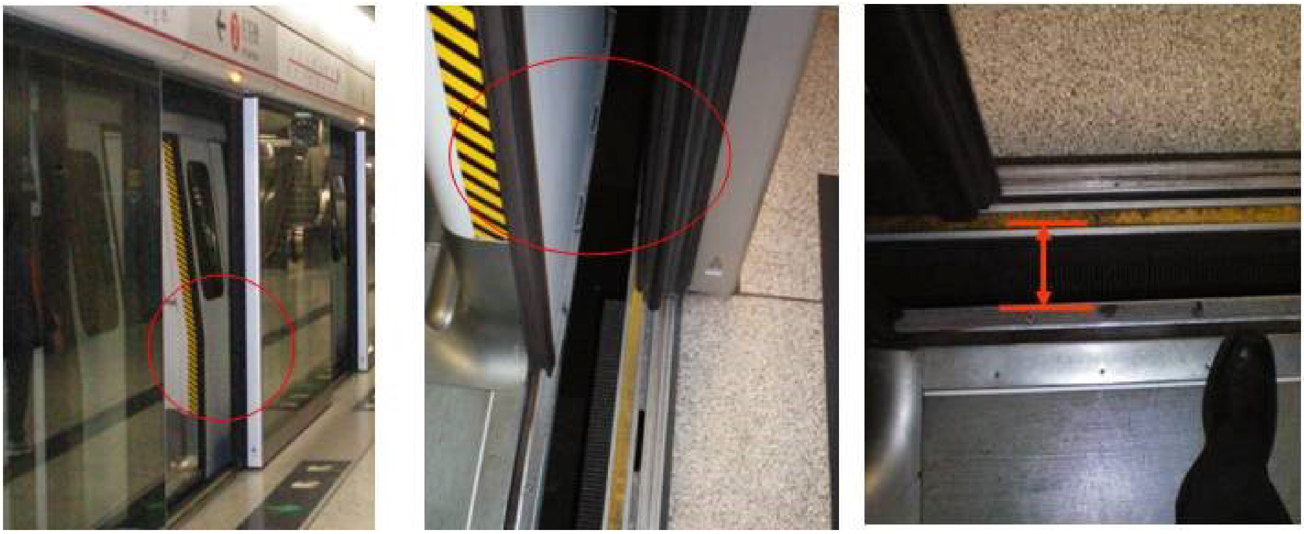


FIGURE 1. The gap between PSD and metro door.

point and area detectors, which may cause high false alarm rate due to vibration. In addition, for large gap platforms, there will be a huge blind area of detection, which does not guarantee metro safety.

The video-sensors-based detector is mainly by setting up the obstacle detection device based on traditional computer vision between PSDs and metro doors. However, a large number of features need to design manually in traditional computer vision [4]–[7]. It is a time-consuming and labor-intensive task. Moreover, traditional computer vision can only give the results of the existence of foreign objects, but the specific categories of foreign objects cannot be known. Thus, people are needed to see the detection results then make decisions. This can not reduce labor costs to some extent.

With the improvement of computing power, many scholars have studied the possibility of using deep neural networks to solve object detection. There have been many excellent works, such as YOLOv1 [8], SSD [9], Regien-based Convolutional Neural Network (R-CNN) [10], Fast-RCNN [11], Faster-RCNN [12], Mask-RCNN [13], CenterNet [14], etc. Deep neural networks have strong representation and modeling ability. They can learn the feature representation of the object automatically through supervised or unsupervised method layer by layer. Then generate high-level abstract representation through a series of nonlinear transformation of the original data, avoiding the tedious and inefficient manual design features. Object detection algorithms based on deep learning can be divided into two classes: one-stage and two-stage. One-stage methods directly regress the class confidence and coordinates of objects (no region proposals), which is relatively fast. Two-stage methods first generate the region proposals (possibly containing the object) and then classify each region proposal (which also correct the location). Two-stage methods are slower than one-stage methods but have better performance. The aforementioned YOLO, SSD belong to one-stage methods. The representative works of two-stage methods include Fast-RCNN, Mask-RCNN, and so on.

In this paper, for the consideration of detecting speed, we choose the one-stage methods to solve the problem of foreign object detection between PSDs and metro doors. Firstly, a dataset containing various foreign objects from real-world is developed. (As we all know, the metro system is a very complicated system, we tried our best to collect these data. We will release this dataset for research as soon as possible and hope it can lay the foundation for future research.) Then, some one-stage methods are implemented. These deep-learning-based models are trained, validated, and tested using the dataset. The experimental results demonstrate that the deep-learning-based methods can detect the presence of foreign objects efficiently and give the classes of foreign objects. Therefore, deep-learning-based methods are helpful to assist the driver in safety detection before driving, which can not only effectively guarantee the safety of passengers, but also help to reduce the operation cost. In addition, reducing the detection time before the driver drives will help relieve the pressure of people flow and improve the efficiency of metro operation.

The contributions of this paper can be summarized as follows:

- To the best of our knowledge, this paper is the first research to adopt deep learning to solve the problem of foreign object detection between PSDs and metro doors. We hope our works can supply new insights for similar problems.
- Experimental results demonstrate that deep neural networks perform excellent. On one hand, deep neural networks can efficiently detect the presence of foreign objects. On the other, they can also give the categories of foreign objects that are present. Therefore, the workload

of employees can be greatly reduced, and traffic pressure will be eased.

The rest of the paper is organized as follows. Section II introduces some important related works. Section III introduces the methodologies of the YOLO series, SSD, and anchor-free models. Section IV presents the procedure for generating dataset. In Section V, we present all experimental details and results of the experiments performed on the dataset. Finally, the paper is summarized in section VI. We note that a shorter conference version of this paper accepted by the 2019 6th International Conference on Systems and Informatics (ICSAI 2019). The shorter conference paper has not been published yet, so we can not cite it. Our initial conference paper only experimented with one type (anchor-based) of deep learning algorithm. This manuscript attempts a new type (anchor-free) of deep learning algorithms and adds a lot of comparative experiments.

II. RELATED WORKS

A. TRADITIONAL METHODS IN FOREIGN OBJECT DETECTION

1) INFRARED MULTI-BEAM DETECTOR

The infrared multi-beam is essentially an optoelectronic technology. It has a pair of transmitter and receiver. The transmitter emits infrared with a certain Angle to ensure that the receiver can receive it accurately. Thus, a light curtain is formed between the transmitter and receiver. When an object enters the area which infrared passes, the infrared is blocked, which causes the receiver's internal circuitry to produce an alarm signal indicating an obstruction.

2) LASER-BASED DETECTOR

The laser-based detector also using optoelectronic technology, but a laser signal is emitted. Compared with infrared light, the laser has the characteristics of strong light, strong signal, and less signal loss in the process of propagation. The laser-based detector emits a beam of laser that bounces off the surface of an object. By reflecting the light, the distance between the object and the detector can be measured, so the size and shape of the object can be measured.

Since the infrared multi-beam detector and the laser-based detector belong to the point and area detector, there is a large blind area, which will greatly affect metro safety. On one hand, at least three pairs of transmitters and receivers need to be installed at different heights to ensure the accuracy of detecting the presence of foreign objects. At the same time, transmitter and receiver must be accurately connected to ensure the accurate reception of the signal. More transmitters and receivers are installed at different heights, more accurate the detection will be. On the other hand, the rate of false alarm will increase also due to the actual field environment (vibration). The maximum detection distance of an infrared multi-beam detector can reach 15 meters. Thus, for long-distance detection, the whole system needs a lot of equipment, which greatly increases the costs and reduces the stability of the system.

3) VIDEO-SENSORS-BASED DETECTOR

A vision sensor is a machine vision system integrating image collecting, processing, and transmitting. The video-sensors-based detector is widely used for detecting the shape, size, quality of industrial equipment. In traditional machine vision, the design of good features is the key and bottleneck of model performance. Manual design features require a lot of experience, special knowledge of the field and data, and a lot of debugging. Another difficulty is that the traditional classifier algorithm belongs to the general classifier, and has not made special optimization for manual-design features.

B. DEEP NEURAL NETWORKS IN OBJECT DETECTION

1) TWO-STAGE METHODS

The two-stage methods divide the detection problem into two stages, first extracts a series of proposals, and then classifies the proposals.

R-CNN applied CNN to extract features, due to the advantage of CNN's good performance for feature extraction. Compared with the traditional detection methods, the detection accuracy of R-CNN had been greatly improved. Fast-RCNN was based on VGG16 [15], and the training was end-to-end. At the same time, Fast-RCNN was nearly 9 times faster than RCNN in training and 213 times faster than R-CNN in testing. The biggest innovation of Faster-RCNN lied in the propose of Region Proposal Network (RPN), which used the anchor mechanism to combine the proposal generate and CNN. Region-based Fully Convolutional Network (R-FCN) used a well-designed position-sensitive score maps to realize position sensitivity and used fully convolutional networks, which greatly reduced the network parameters. Mask-RCNN not only solved the object detection efficiently but also achieved high-quality instance segmentation. Cascade-RCNN continuously optimized the results by cascading multiple detectors. Each detector defined positive and negative samples based on different IoU thresholds. The output of the former detector was the input of the latter, and the later the detector was, the higher the threshold value of IoU was.

2) ONE-STAGE METHODS

The one-stage methods remove the step for proposal generation. They usually take the whole picture as input, and carry out classification and regression at the same time, realizing the end-to-end detection process and significantly improving the computing speed.

Inspired by the idea of regression, YOLOv1 used the one-stage network to directly detect objects, which is fast but not very accurate. SSD borrowed the ideas of Faster RCNN and YOLOv1, usesd fixed bounding boxes to generate proposals on the basis of one-stage networks, and utilized multi-layer feature information to improve the speed and

TABLE 1. The deep neural networks used in this paper.

Methods		Highlights

Anchor-based	SSD	Multi-scale feature maps;
		one-stage network;
		elegant network and simple implementation

	YOLOv3	DarkNet-53;
		predictions across scales;
		logistic classifiers instead of softmax

	Gaussian YOLOv3	Gaussian modeling; reconstruction of loss function
Anchor-free	CenterNet	Model an object as a single point; achieves good speed-accuracy trade-off;
	FCOS	Anchor box free and proposal free; faster training and testing as well as less training memory footprint; can be used as a RPN in two-stage detectors; can be immediately extended to solve other vision tasks with minimal modification

detection accuracy. In YOLOv2 [16], better anchor priors were defined by performing k-means [17] clustering on the training data instead of setting them manually. For the imbalance problems in object detection, RetinaNet [18] tried to use focal loss to solve it. Combining other good ideas, YOLOv3 [19] improved prediction accuracy, especially for small objects, while maintaining its speed advantage. Based on the network of YOLOv3, Gaussian YOLOv3 [20] improved the performance of the model by increasing the output of the network and improving the loss function of the network.

For the object detection methods based on anchor-free, YOLOv1 was one of the early representatives, and then ushered in the blowout period of anchor-free methods. The main idea of CenterNet was to regress to other bounding box properties through the information of the key point, such as the distance between the key point and the four sides, posture, direction, and other information. Unlike CenterNet regressing from the key point to the boundary distance to get the bounding box, CornerNet [21] goes the other way. CornetNet defined the bounding box directly using two corner points: top-left, bottom-right. It used a set of corners to identify a target. FCOS [22] used the idea of semantic segmentation, the problem of object detection is solved by means of per-pixel prediction.

III. METHODOLOGY

To detect and localize multiple types of foreign objects between PSDs and metro doors, deep-learning-based methods are used for quasi real-time processing of images. In this section, we will give an introduction to some state-of-the-art object detection deep-learning-based methods, including SSD, YOLOv3, CenterNet and so on. Table 1 was a simple comparison of deep neural networks that used in this paper. From the simple summary in Table 1, we can see that the methods used in this paper basically have the characteristics of fast detection speed and high accuracy. This meets the requirements of real-time detection and can be applied to practical projects well.

A. MODELS IN THE YOLO FAMILY

In this section, four models in YOLO family are introduced. YOLOv3 and Gaussian YOLOv3 are the two methods used in this paper.

1) YOLOv1

YOLOv1 creatively combined the two phases of region proposals step and object recognition step into one. By eliminating the time-consuming step (region proposal), YOLOv1 was extremely fast, almost in real-time. YOLOv1 utilized GoogleNet [23]'s structure. It replaced GoogleNet's inception modules with 1 × 1 reduction layers followed by 3 × 3 convolutional layers. YOLOv1 was known for its speed, which can reach 45 FPS. Fast YOLO (with a smaller network) can even reach 155 FPS. The fast speed was due to YOLOv1's network design, which combined recognition with location. This unified design allows training and prediction to be done in an end-to-end way, which was very efficient. The disadvantages of YOLOv1 were that the detection effect of small-scaled objects was not good (especially some small objects clustered together), the prediction accuracy of the borders was not very high, and the overall prediction accuracy was slightly lower than two-stage methods.

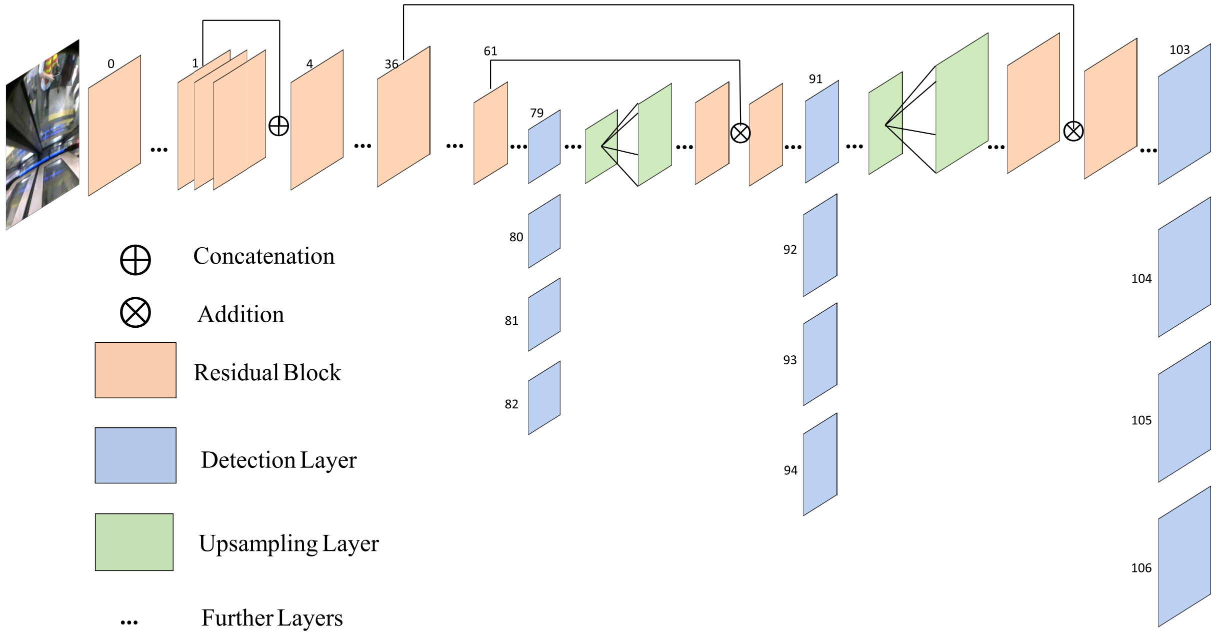


FIGURE 2. Network architecture of YOLOv3.

2) YOLOv2/YOLO9000

Compared with YOLOv1, YOLOv2 had been improved in three aspects: more accurate prediction, faster speed and more object recognition on the basis of maintaining the processing speed. YOLO9000 was an attempt to use a very large number of classification samples in ImageNet [24] to train together with COCO [25]'s object detection dataset. This made YOLO9000 able to detect many objects even if it has not learned many samples. To further improve the speed, YOLOv2 proposed the Darknet-19. Darknet-19 was smaller than the VGG-16 and no less accurate than VGG-16, but the floating-point computation was reduced to about 1/5 to ensure faster speed. Inspired by the idea of Network In Network [26], Darknet-19 used global average pooling to make prediction, placing the convolution kernel of 1 × 1 between the convolution kernel of 3 × 3 to compress features. In addition, the Batch Normalization [27] was used to train and accelerate the convergence and regularization of the model.

3) YOLOv3

YOLOv3 improved prediction accuracy, especially for small-scale objects, while maintaining its speed advantage. The main improvements of YOLOv3 included: adjusted the network structure; multi-scale features were used for object detection; object classification used multiple independent logistic instead of softmax. In parts of basic image feature extraction, YOLOv3 adopted a new network structure called darknet-53. It learned from the residual network [28] and set shortcut connections between some layers. YOLOv2 used a passthrough structure to detect fine-grained features. In YOLOv3, three feature maps of different scales were further used for object detection. As the number and scale of the output feature maps changed, the size of the prior bounding box also needed to be adjusted accordingly. YOLOv2 had started to use k-means clustering to obtain the size of the prior bounding box. And YOLOv3 continued this method by setting three prior bounding boxes for each downsampling scale, and a total of nine prior bounding boxes with different sizes can be obtained by clustering. The nine prior bounding boxes in the COCO dataset were: (10 × 13), (16 × 30), (33 × 23), (30 × 61), (62 × 45), (59 × 119), (116 × 90), (156 × 198), (373 × 326). Moreover, using logistic for prediction was helpful to support the detection of multi-label objects. The network architecture of YOLOv3 was shown in Fig. 2.

4) GAUSSIAN YOLOv3

In the project of object detection, the trade-off of speed-accuracy is very important, and YOLOv3 is outstanding. Gaussian YOLOv3 improved the network architecture of YOLOv3 by taking advantage of the Gaussian distribution so that the network can output the uncertainty of each bounding box, thus improving the accuracy of the network. Specifically, Gaussian YOLOv3 implementd the output to the bounding box reliability by increasing the output of the network and reconstructing the loss function of the network. Thus, Gaussian YOLOv3 can output the reliability of each bounding box without changing the network structure of YOLOv3 or increasing the amount of computation, which improved the overall performance of the algorithm.

B. SSD

Some of the basic concepts of SSD are presented in this section. More details about SSD can be found in original paper. In view of the respective shortcomings and advantages

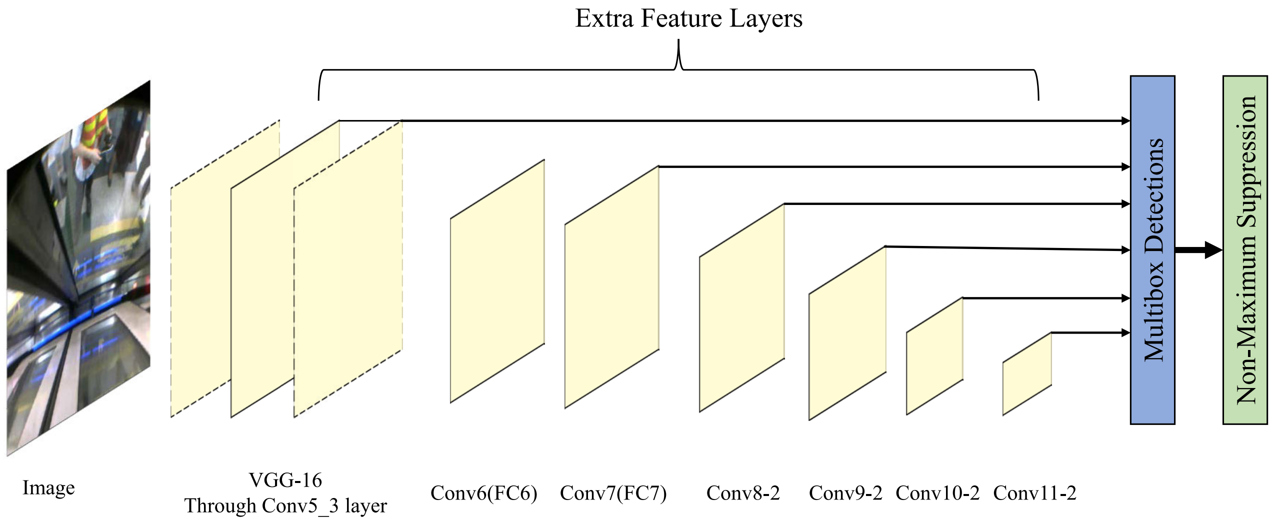


FIGURE 3. Network architecture of SSD.

of YOLOv1 and Faster-RCNN, SSD was proposed. The entire network of SSD adopted the idea of one-stage to improve the detection speed. Moreover, SSD incorporated the idea of anchors from Faster-RCNN. Then, SSD performed layered feature extraction, successively calculates border regression and classification operations. Hence, SSD can adapt to the training and detection tasks of multiple scale objects. The backbone network of SSD was VGG-16. Researchers have made some fine-tuning to make it useful for detection tasks, including using convolution layer to replace the fully connected layer, removing the dropout layer, and replacing the final max-pooling layer with an expanded convolution. However, the main disadvantage of SSD was that it still has poor recognition of small objects. This was mainly because small-scale objects are mostly trained with anchors of lower levels, but the features of lower levels are not sufficiently nonlinear and cannot be trained with sufficient accuracy. SSD's architecture was shown in Fig. 3.

C. ANCHOR-FREE METHODS

1) CenterNet

Different from the SSD and YOLOv3, CenterNet is a typical representative of anchor-free object detection algorithms. In general, most of the object detection algorithms usually identify the objects as axis-aligned boxes in the image. Many excellent algorithms will exhaustivity potential object locations and then classify them, which is time-consuming and inefficient. CenterNet simplified the complexity and directly predicted the object as a point, completely dropping post-processing operations such as Non Maximum Suppression (NMS), and applied this method to pose estimation and 3D object detection. In the original paper, four architectures (ResNet-18, ResNet-101, DLA-34 [29], and Hourglass-104 [30]) were experimented. Experimental results show that DLA-34 was the best-performing architecture. Therefore, DLA-34 was used as our architecture in this paper. The network architecture of DLA-34 was shown in Fig. 4.

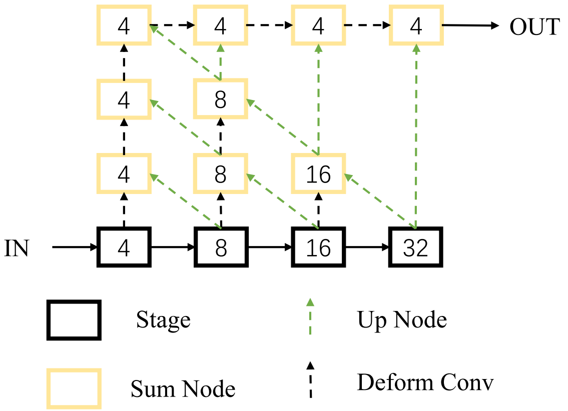


FIGURE 4. Network architecture of DLA-34.

2) FULLY CONVOLUTIONAL ONE-STAGE OBJECT DETECTION

FCOS is another representative of anchor-free object detection algorithm. FCOS solved the problem of object detection by means of per-pixel prediction, similar to semantic segmentation. FCOS implemented the proposal free and anchor free, significantly reducing the number of hyperparameters. Designing hyperparameters usually requires heuristic adjustments, and there are many tricks. In addition, by eliminating the anchor, FCOS completely avoided complex IoU calculations and matches between the anchor and the ground truth during training, reducing the total training memory footprint by about two times. FCOS can also be used as RPN of the two-stage detector, and its performance is obviously better than other RPN algorithms based on anchor. Finally, with a few modifications, FCOS can be extended to other

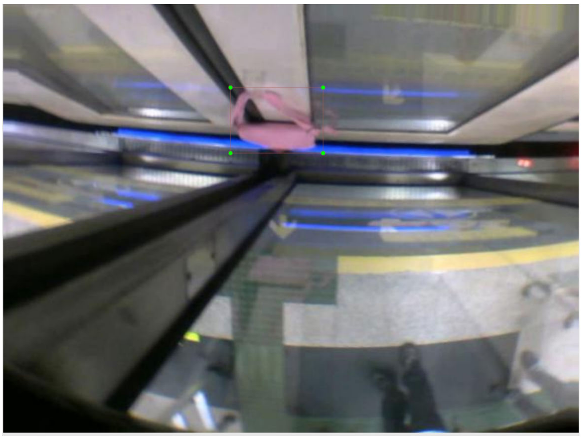


FIGURE 5. An example of annotated image.

visual tasks, including instance segmentation and key point detection.

IV. DATASET

In this section, the acquisition of the dataset are illustrated in detail.

A. DATASET CONSTRUCTION

To develop a dataset containing foreign objects as many as possible, 164 images are collected using High-Definition (HD) cameras. We first place all kinds of foreign objects manually between the metro door and the PSD on the premise of ensuring safety. Then we use the HD camera to take pictures. The metro station where we collect data is in Guangzhou, China. Due to the complexity and high safety requirements of the metro system, our time and conditions for collecting data are limited. Therefore, the number of images is small. Since the same fixed light source was used to collect the data, the visual-angle, the occlusion, and the illumination have hardly changed. We will release the dataset for research as soon as possible.

B. DATA AUGMENTATION

In general, a successful neural network needs a large number of parameters, and many neural networks have millions of parameters. To make these parameters work correctly requires a large number of data to train, but the actual situation is that the dataset we got is very small (just as mentioned before, we only 164 images). Therefore, we performed a series of flips and crop operations on the images to obtain 984 images in this paper.

C. TRAINING, VALIDATION, AND TEST DATASETS

The labelImg is labeled with the picture frame, a labeled sample is shown in Fig. 5. An XML file is automatically generated for each picture, recording various information of the picture (file name, storage path, width, height, depth, object information) and bounding box coordinate information (top left, bottom right), etc. A total of 984 pictures, divided into seven classes (*normal*, *bag*, *bottle*, *person*, *other*, *umbrella*, *plasticbag*). To generate the testing dataset, 30% of the images are randomly selected from annotated images. The remaining images not selected for the testing dataset are used to generate a training and validation dataset. Table 2 shows the details of the dataset.

V. EXPERIMENTS

A. MODELS' IMPLEMENTATION DETAILS

All experiments are performed on a computer with a Intel Core i7-6820HK Central Processing Unit (CPU), 32 GB DDR4 memory, and two GeForce 1070 Graphics Processing Units (GPUs). The hyperparameters of all deep neural networks used in this paper were set according to the original papers.

1) SSD

The total loss function of SSD as follows:

$$L(x,c,l,g)=\frac{1}{N}(L_{conf}(x,c)+\alpha L_{loc}(x,l,g)). \tag{1}$$

As can be seen, the loss function of SSD contains two items: localization loss (L_{loc}) and the confidence loss (L_{conf}). N is the number of matched default boxes. (l) and (g) represent the predicted box and the ground truth box, respectively. (c) is the classes confidences, and the weight term α is setted to one follow the original paper.

2) YOLOv3

YOLOv3 made more subtle design adjustments to YOLOv2 and redesigned a new network with a slightly more complex structure, improving accuracy while maintaining speed. The loss function of YOLOv3 as follows:

$$\begin{aligned}Loss=&\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}I_{ij}^{obj}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2]\\&+\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}I_{ij}^{obj}(2-w_i\times h_i)[(w_i-\hat{w}_i)^2+(h_i-\hat{h}_i)^2]\\&-\sum_{i=0}^{S^2}\sum_{j=0}^{B}I_{ij}^{obj}[\hat{C}_i log(C_i)+(1-\hat{C}_i)log(1-C_i)]\\&-\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B}I^{noobj}[\hat{C}_i log(C_i)+(1-\hat{C}_i)log(1-C_i)]\\&-\sum_{i=0}^{S^2}I_i^{obj}\sum_{c\in classes}[\hat{p}_i(c)log(p_i(c)\\&+(1-\hat{p}_i(c))log(1-p_i(c))].\end{aligned} \tag{2}$$

where λ_{coord} is the weight of the loss of the coordinates of a boundary box. λ_{noobj} is the weight of the loss when detecting background. I_{ij}^{obj} is the jth bounding box predictor that is in the ith cell is valid in prediction. s^2 is the number of grid cells.

TABLE 2. The proportion of training, validation, and test sets. *Normal* means there are no foreign objects.

Foreign objects class	Training and validation		Testing	
	Objects	Number of images	Objects	Number of images
Bag	145	145	65	65
Bottle	57	55	24	20
Other	29	29	14	14
Person	255	255	107	107
Plasticbag	103	103	41	41
Umbrella	54	54	25	25
Normal	47	47	24	24
Total	690	688	300	296

B is the number of bounding boxes to be predicted in each grid cell. x_i and y_i denote the coordinates of the base point of the ith bounding box. w_i and h_i denote the weight and height of the ith bounding box. C_i is the confidence score of the jth bounding. $p_i(c)$ is the conditional class probability for category c.

3) GAUSSIAN YOLOv3

Gaussian YOLOv3 estimates the uncertainty of the bounding box by the following Gaussian model.

$$p(y|x) = N(y; \mu(x), \sum(x)) \tag{3}$$

where x is a given test input, y is the output. And $\mu(x)$ is the mean functions, $\sum(x)$ is the variance functions.

The reconstructed loss function is as follows:

$$L_x = -\sum_{i=1}^{W}\sum_{j=1}^{H}\sum_{k=1}^{K}\gamma_{ijk}\log(N(x_{ijk}^{G}|\mu_{t_x}(x_{ijk}), \sum\nolimits_{t_x}(x_{ijk}))+\varepsilon). \tag{4}$$

where L_x is the NLL (Negative Log-Likelihood) loss of t_x. W and H are the number of grids in horizontal and vertical directions respectively. K is the number of anchors. $\mu_{t_x}(x_{ijk})$ and $\sum_{t_x}(x_{ijk})$ are prediction values and variance from the detection layer respectively. x_{ijk}^{G} is the GT of t_x.

4) CenterNet

In original paper, CenterNet can be used to do object detection, 3D detection and pose estimation. In this paper, we only focus on object detection. More details about CenterNet, can be found in the original paper. The loss function of CenterNet as follows:

$$L_k = \frac{-1}{N}\sum_{xyc}\begin{cases}(1-\hat{Y}_{xyc})^{\alpha}log(\hat{Y}_{xyc}) & \text{if } Y_{xyc}=1\\ (1-Y_{xyc})^{\beta}(\hat{Y}_{xyc})^{\alpha}log(1-\hat{Y}_{xyc}) & \text{otherwise.}\end{cases} \tag{5}$$

Equation (5) is the loss function of the keypoint prediction. α and β are hyper-parameters of the focal loss, and N is the number of keypoints in image I.

$$L_{off} = \frac{1}{N}\sum_{p}|\hat{O}_{\tilde{p}} - (\frac{p}{R} - \tilde{p})|. \tag{6}$$

Equation (6) is the loss function of the local offset. CenterNet conducted downsampling on the image, and accuracy error would be caused when the obtained feature map was remolding to the original image. Therefore, an additional local offset was used for each center point.

$$L_{size} = \frac{1}{N}\sum_{k=1}^{N}|\hat{S}_{pk} - s_k|. \tag{7}$$

Equation (7) is the object size loss.

$$L_{det} = L_k + \lambda_{size}L_{size} + \lambda_{off}L_{off}. \tag{8}$$

Equation (8) is the total loss, which is the sum of keypoint prediction loss, local offset loss and object size loss.

5) FCOS

The total loss function in FCOS is as follows:

$$\begin{aligned}L(\{\mathbf{p}_{x,y}\}, \{\mathbf{t}_{x,y}\}) = &\frac{1}{N_{pos}}\sum_{x,y}L_{cls}(\mathbf{p}_{x,y}, c_{x,y}^{*})\\ &+\frac{\lambda}{N_{pos}}\mathbb{1}_{\{c_{x,y}^{*}>0\}}\sum_{x,y}L_{reg}(\mathbf{t}_{x,y}, \mathbf{t}_{x,y}^{*}).\end{aligned} \tag{9}$$

where L_{cls} and L_{reg} are focal loss and IoU loss respectively.

B. EVALUATION METRICS

The evaluation metrics often used in object detection is mAP, which is the average of AP (Average Precision) over all categories. In this paper, we follow the calculation method prior to VOC2010 [31]: AP is computed by sampling the monotonically decreasing curve at a fixed set of uniformly-spaced recall values 0, 0.1, 0.2 . . . 1. In addition to detection accuracy, another important evaluation metric of object detection algorithm is speed. Only fast speed can realize real-time detection. A common measure of speed is FPS, the number of images that can be processed per second. Higher value of mAP and FPS were, better algorithm would be.

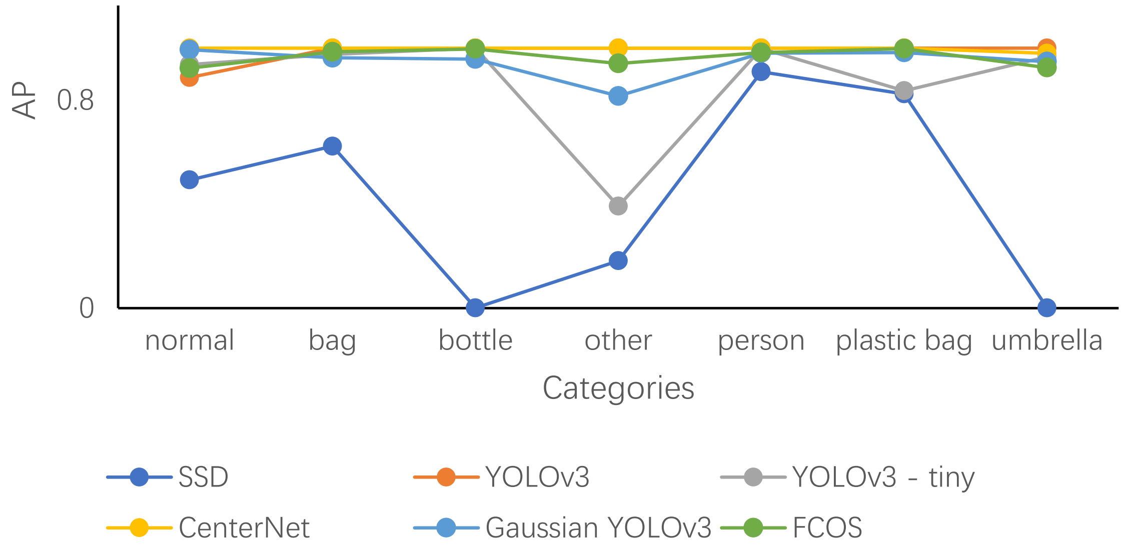


FIGURE 6. The performance of the networks for the testing set.

TABLE 3. Experiments results. The best results are indicated in bold rows.

Method	mAP	FPS
SSD	0.433	12
YOLOv3	0.984	45
YOLOv3 - tiny	0.872	**200**
Gaussian YOLOv3	0.950	44
CenterNet	**0.997**	25
FCOS	0.965	5

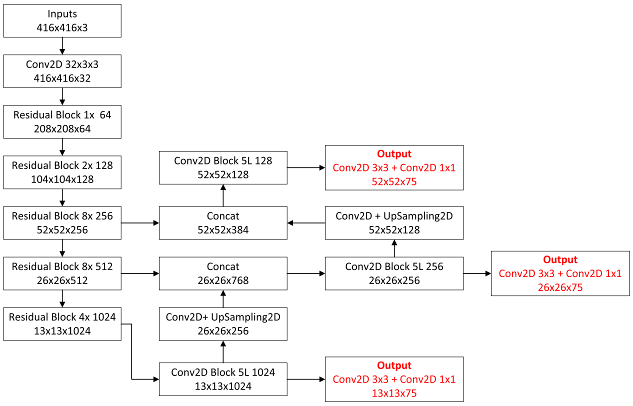

Inputs 416x416x3
Conv2D 32x3x3 416x416x32
Residual Block 1x 64 208x208x64
Residual Block 2x 128 104x104x128
Residual Block 8x 256 52x52x256
Residual Block 8x 512 26x26x512
Residual Block 4x 1024 13x13x1024
Conv2D Block 5L 128 52x52x128
Concat 52x52x384
Concat 26x26x768
Conv2D+ UpSampling2D 26x26x256
Conv2D Block 5L 1024 13x13x1024
Output Conv2D 3x3 + Conv2D 1x1 52x52x75
Conv2D + UpSampling2D 52x52x128
Conv2D Block 5L 256 26x26x256
Output Conv2D 3x3 + Conv2D 1x1 26x26x75
Output Conv2D 3x3 + Conv2D 1x1 13x13x75

FIGURE 7. Major modules in YOLOv3.

C. COMPARISON OF METHODS

As can be seen from Table 3, the methods based on deep learning can achieve impressive results while maintaining a fast detection speed. YOLOv3 - tiny is the fastest foreign object detection (200 FPS) method on the dataset, with 87.2% mAP. Its mAP value is twice than SSD, and the FPS value is more than three times than SSD. As can be seen from Fig. 7 and Fig. 8, YOLOv3 - tiny is a simplified version of YOLOv3. The backbone of YOLOv3 - tiny removes the residual layer and is shallower (only 7 layers, similar to Darknet-19), which makes YOLOv3 - tiny unable to extract higher-level semantic features. In addition, compared with YOLOv3, YOLOv3 - tiny removes some feature layers, only 2 independent prediction branches are retained, and the scale information is not sufficient as YOLOv3. YOLOv3 pushes mAP to 98.4% while still maintaining fast performance (45 FPS). The FPS of Gaussian YOLOv3 is similar to that of YOLOv3 (44 vs. 45), but the detection is slightly less effective than that of YOLOv3 (95% vs. 98.4%). However, CenterNet was the best performer in the mAP value, with 99.7%. This somewhat compromises its speed, which is only 25 FPS. As the anchor free method, FCOS is inferior to Centernet (96.5% vs. 99.7%), especially in FPS (5 vs. 25). Fig. 6 and Fig. 9 show the prediction performance of deep neural networks in various classes in a more intuitive and detailed way. It can be seen that almost all foreign objects

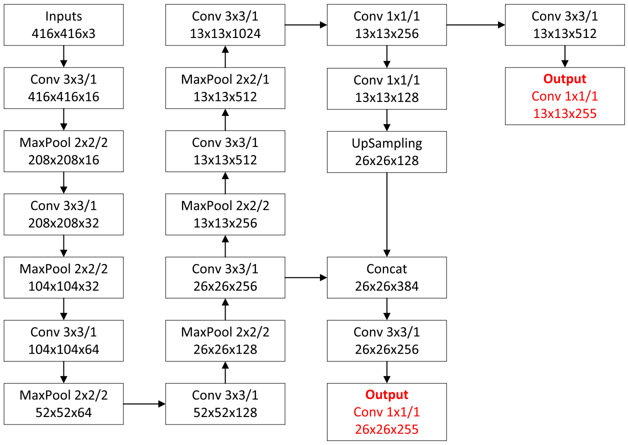
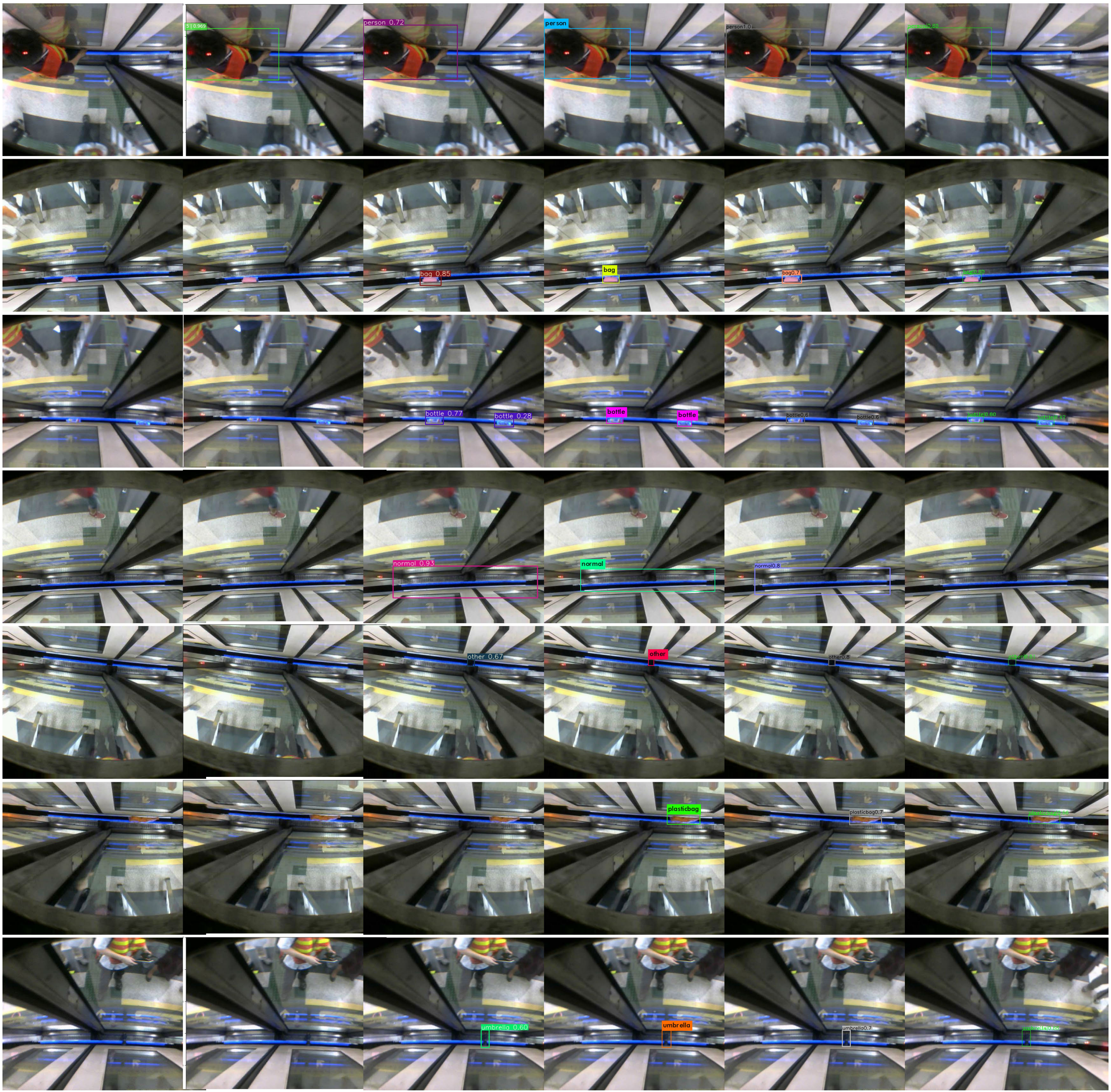


FIGURE 8. Major modules in YOLOv3 - tiny.



(a) Input (b) SSD (c) YOLOv3 (d) Gaussian YOLOv3 (e) CenterNet (f) FCOS

FIGURE 9. Detection results. The categories from top to bottom are: *person, bag, bottle, normal, other, plastic bag, umbrella*.

can be detected by other deep neural networks except SSD. Whether it's a small foreign object like a bag or a very obvious foreign object like a person. In these methods, CenterNet and Gaussian YOLOv3 detect all foreign objects. YOLOv3 and FCOS did not detect plastic bag and normal, respectively. In addition, as we aforementioned, these methods based on deep learning can not only efficiently detect the presence of foreign objects but also clearly inform the classes of foreign objects. These advantages can effectively reduce the detection time before driving, improve the efficiency of metro operation, and further ensure the safety of passengers.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we compared the performance of state-of-the-art deep-learning-based object detection algorithms on foreign object detection between PSDs and metro doors. First, we used HD cameras to develop a dataset in a real metro station. Then, some deep neural networks were implemented, we trained and tested them on the dataset. The experimental results show that deep neural networks can not only detect the existence of foreign objects but also inform the types of foreign objects. Deep-learning-based methods can effectively assist drivers to conduct safety inspections before driving, which can effectively protect the safety of passengers, and help reduce operating costs and ease traffic pressure. In addition, deep-learning-based methods are one of the important directions of intelligent transportation in the future.

In the future, following directions are to be explored:

- More diverse foreign object images will be collected in more environments to further increase the robustness of the algorithms. We will release the dataset used in this paper for research as soon as possible.
- Deep learning requires high computational power. We try to transplant these algorithms to low computational power devices such as mobile devices.
- Limited by the existing conditions, Few-shot learning is an important research direction in the future.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their careful reading of our works and their many insightful comments.

REFERENCES

[1] R. Wang, Z. Yang, and W. Kong, "Research on infrared light screen in obstacle detection of subway platform screen doors," *Transducer Microsyst. Technol.*, vol. 32, no. 3, pp. 25–28, 2013.

[2] Z. Li, "Discussion on installation scheme of laser detection device in PSDS," *Chin. Hi-Tech Enterprises*, vol. 19, pp. 46–47, 2009.

[3] F. Tan and J. Liu, "Foreign object detection algorithm for subway platform based on computer vision," *Urban Rail Transit Inf. Technol.*, vol. 26, no. 1, pp. 67–69, 2017.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, Jun. 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177.

[5] P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Anchorage, AK, USA, Jun. 2008, pp. 24–26, doi: 10.1109/CVPR.2008.4587597.

[6] Z. Tu, Z. Guo, W. Xie, M. Yan, R. C. Veltkamp, B. Li, and J. Yuan, "Fusing disparate object signatures for salient object detection in video," *Pattern Recognit.*, vol. 72, pp. 285–299, Dec. 2017, doi: 10.1016/j.patcog.2017.07.028.

[7] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*. [Online]. Available: http://arxiv.org/abs/1905.05055

[8] 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.

[9] 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.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[11] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[12] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.: Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2015, pp. 91–99. [Online]. Available: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[14] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: http://arxiv.org/abs/1904.07850

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[16] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.

[17] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1. Berkeley, CA, USA: Univ. of California Press, 1967, pp. 281–297. [Online]. Available: https://projecteuclid.org/euclid.bsmsp/1200512992

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2999–3007, doi: 10.1109/ICCV.2017.324.

[19] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[20] J. Choi, D. Chun, H. Kim, and H.-J. Lee, "Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 502–511.

[21] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 765–781, doi: 10.1007/978-3-030-01264-9_45.

[22] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[24] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, Miami, FL, USA, 2009, pp. 248–255, doi: 10.1109/CVPRW.2009.5206848.

[25] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zürich, Switzerland, Sep. 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[26] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. 2nd Int. Conf. Learn. Representations, (ICLR)*, Banff, AB, Canada, Apr. 2014, pp. 1–10, [Online]. Available: http://arxiv.org/abs/1312.4400

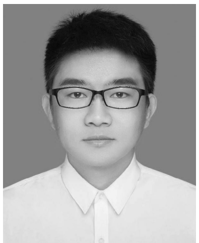
[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, Jul. 2015, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html

[28] 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.

[29] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 2403–2412.

[30] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 483–499, doi: 10.1007/978-3-319-46484-8_29.

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



YUAN DAI was born in Loudi, Hunan, China, in 1995. He received the B.S. and M.S. degrees from the Changsha University of Science and Technology, Changsha, Hunan, in 2016 and 2018, respectively. He is currently pursuing the Ph.D. degree in traffic information engineering and control with the South China University of Technology. His research interests include intelligent transportation, computer vision, and deep learning.

WEIMING LIU received the Ph.D. degree from the National University of Defense Technology, China, in 2004. He is currently a Professor with the School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China. His research interests include digital image monitoring and identification, intelligent traffic system engineering theory and application, and intelligent road traffic control and management.

HAIYU LI received the B.S. degree from the College of Communication Engineering, Jilin University, and the M.B.A. degree from Sun Yat-sen University. She is currently a Senior Engineer with the National Engineering Laboratory, Guangzhou Metro Group Company, Ltd. Her research interests include communication, computer networks, and image processing.

LAN LIU received the B.Sc. degree from Southwest Jiaotong University, in 1998. She is currently the Deputy Senior Engineer with Guangzhou Metro Group Company, Ltd. Her research interest is rail pulling power supply systems.