Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 2085876, 13 pages
https://doi.org/10.1155/2021/2085876

Research Article

A Deep Pedestrian Tracking SSD-Based Model in the Sudden Emergency or Violent Environment

Zhihong Li, Yang Dong, Yanjie Wen, Han Xu, and Jiahao Wu

Beijing Advanced Innovation Center for Future Urban Design, Beijing University of Civil Engineering and Architecture, Beijing 100044, China

Correspondence should be addressed to Zhihong Li; lizhihong@bucea.edu.cn

Received 20 May 2021; Revised 16 June 2021; Accepted 8 July 2021; Published 15 July 2021

Academic Editor: Xinqiang Chen

Public security monitoring is a hot issue that the government and citizens pay close attention to. Multiobject tracking plays an important role in solving many problems for public security. Under crowded scenarios and emergency places, it is a challenging problem to predict and warn owing to the complexity of crowd intersection. There are still many deficiencies in the research of multiobject trajectory prediction, which mostly employ object detection and data association. Compared with the tremendous progress in object detection, data association still relied on hand-crafted constraints such as group, motion, and spatial proximity. Emergencies usually have the characteristics of mutation, target diversification, low illumination, or resolution, which makes multitarget tracking more difficult. In this paper, we harness the advance of the deep learning framework for data association in object tracking by jointly modeling pedestrian features. The proposed deep pedestrian tracking SSD-based model can pair and link pedestrian features in any two frames. The model was trained with open dataset, and the results, accuracy, and speed of the model were compared between normal and emergency or violent environment. The experimental results show that the tracking accuracy of mAP is higher than 95% both in normal and abnormal data sets and higher than that of the traditional detection algorithm. The detection speed of the normal data set is slightly higher than that of the abnormal data set. In general, the model has good tracking results and credibility for multitarget tracking in emergency environment. The research provides technical support for safety assurance and behavior monitoring in emergency environment.

1. Introduction

Public security is one of the important contents concerned by both official and private departments. Many researchers also made great efforts in this area and obtained good research results to improve the level of social security [1–4]. At the same time, with the development of video detection equipment, video detection technology plays an important role in traffic safety management. Video detection technology is widely used in traffic field. Through detecting different traffic tools such as cars [5, 6] and ship [7, 8], people can take corresponding measures and deploy the early warning schemes. As one of the most basic and important tasks of video detection technology, object detection requires not only the classification of objects, but also the marking of the current position of objects. In a sudden emergency or violent environment, facing the urgent evacuation needs and the flustered large flow of people, the difficulty lies in the fact that pedestrians are different from rigid objects but want to escape from the possible exits in the shortest time. Because of the changes of their joint posture, clothing, lighting, and background, the appearance of pedestrians in a sudden emergency or violent environment has irregular escape spaces and constantly change positions, which directly leads to a very difficult judgment of their position detection and escape routes. Fortunately, in recent years, the idea of data-driven methods was put forward [9], and it has made a great breakthrough in related algorithms for monitoring and judgment of escape route for the large-scale panic crowd.

The most advanced target detection algorithms are based on the assumption of boundary boxes, the resampling of pixels or features for each box, and the application of high-

quality classifier for target detection. The algorithm was then applied to PASCAL Visual Object Classes (VOC) [10], Common Objects in Context (COCO) [11], and ImageNet large-scale visual recognition challenge (ILSVRC) datasets for validation. The most advanced algorithms mainly include Faster R-CNN [12], You Only Look Once (YOLO) [13], Single Shot MultiBox Detector (SSD) [14], etc., where Faster R-CNN (convolutional neural network) is developed on the basis of Fast R-CNN [15]. The detection accuracy of the method is very high, but the demand for hardware is correspondingly high, and the detection speed is limited. Even the fastest high-precision detectors are only 7 frames per second (FPS). Many scholars tried to improve the performance of the Fast R-CNN. However, the cost of improving the speed means to reduce the detection accuracy. So, it cannot get to the ideal state. YOLO and SSD are both the single-shot detection models. The YOLO model treats object detection as a spatially separated boundary box and a related probability-like regression problem. A single neural network predicts boundary boxes and quasiprobabilities directly from the full image in one evaluation. The SSD model adds several feature layers to the end of a base network, which predict the offsets to default boxes of different scales. The default boxes in SSD are similar to the *anchor boxes* used in Faster R-CNN [14].

In real situations, occlusion is one of the most important challenges of pedestrian detection in complex environments, especially in crowded scenes under abnormal and emergency conditions. Therefore, it makes the detector sensitive to the threshold of non-maximum inhibition (NMS) in the crowded scene. Such dense occlusion problem cannot be solved by simply adjusting the NMS threshold. A large NMS threshold leads to many false checks, while a small NMS threshold leads to many missed checks.

In order to enhance the accuracy and efficacy of detecting algorithm for identify multiobject behavior in a sudden or violent environment, this paper adopts the SSD algorithm to detect the object and determine the fast-moving multiple panic targets. The remainder of this paper is divided into the following parts: the next part introduces the related work in target detection in recent years and the application in the field of evacuation environments. The third part introduces the core algorithm of SSD for the multitarget detection. The fourth part displays the topology of the SSD network model and network data set in this work. In the fifth part, the trained model is applied to the abnormal and normal data sets, and the corresponding error analysis and efficiency analysis are carried out. The final part is the summary of the research work.

2. Related Work

The task of object detection is to find all interested objects in the image and determine their location and size, which are the core issues in the field of machine vision. The key problems to be solved in target detection processes are as follows: (1) the target may appear in the task location in the graph; (2) there are different sizes of goals; (3) the target may have different shapes; (4) objects block each other, especially dense crowd.

2.1. Object Localization Detection Algorithm. In recent years, great progress has been made in the object localization detection algorithm. The cost of traditional work by using of "window plus image scaling" to determine the accuracy target positions and types in the complex or chaos environment was relatively high. Totally, in the past decade, the target detection algorithm could be divided into two periods, i.e., the period based on traditional manual features and the period based on deep learning.

As for of the typical work during the period based on traditional manual features, Viola and Jones [16] used VJ detector and Haar feature and integrated classifier (AdaBoost) to realize real-time face detection for the first time under extremely limited resources. Dalal and Triggs [17] put forward the HOG feature combined with SVM (large interval classifier) pedestrian detector. Felzenszwalb and McAllester [18] proposed the variable component model (Deformable Part Model: DPM), which used the strategy of "from the whole to the part, then from the part to the whole" to detect. Afterwards, related work during the period made corresponding improvements in the structure, speed, and algorithm of the model for object detection. However, limited to the computational capacity of the hardware, the accuracy of the detection algorithm during this period was still need to be improved [19].

With the development of GPU, deep learning is widely concerned in the field of computer vision. The target detection algorithm based on the convolutional neural network (CNN) was proposed, as well as the R-CNN-based algorithm [19]. Due to that the training of R-CNN is multistage and time efficiency is low, Girshick [15] proposed Fast-RCNN to train classification and regression tasks synchronously. They stated that the training speed was 9 times of R-CNN, and the detection speed was 200 times of R-CNN. Similarly, Ren and He [12] put forward Faster R-CNN, which firstly revealed the end-to-end deep learning detection algorithm, which is measured in seconds per frame (SPF), and even the fastest high-accuracy detector, Faster R-CNN, operates at 7 frames per second. Moreover, Huang et al. [20] proposed a detection framework with full convolution and anchor free, and Yu et al. [21] proposed a *IoU* (Intersection-over-Union) loss to maximize the value between the truth box and predicted box. There have been many attempts to build faster models by attacking each stage, but so far, significantly increased speed comes only at the cost of significantly decreased detection accuracy [14].

Overall, all of the above methods need to form a candidate region box on the feature map and then regress and classify the attributes in images. Therefore, they are also called two-stage detection algorithms. The most improvement of these types of algorism was that the candidate box can accurately describe the subject target, while the disadvantage was that the time efficiency was low. For the corresponding one-stage process, the representative algorithms are YOLO [13], YOLO-v2 [22], and SSD. Typically, only one convolution operation is needed to identify the location and category of the targets in images. At the same time, many scholars have built different improved models based on the SSD model, such as FSSD [23], DSSD [24], and ASSD [25].

2.2. Pedestrian Detection in Extremely Crowded Scenarios. It is quite complex of dealing with the heavy traffic flow characteristics such as multiobjects and group dynamic. The dynamic changes of pedestrian flow in dense traffic network aggravate the uncertainty of traffic state. However, pedestrian detection in dense or evaluation traffic situations is greatly affected by many factors such as multiple objects and their attitudes, shapes, the occlusion, illumination, and background. Especially, in the sudden emergency or violent environment, the crowd density is large, the pedestrian behavior is complex, and the detection accuracy and speed are greatly affected.

Relying on the strong detection ability of the deep convolution neural network (DCNN), Wu et al. [26] put forward an improved Fast-RCNN, which uses small face targets as recognition units to identify pedestrians. Roy and Rahman [27] use the deep convolution neural network to identify emergency vehicles on the road, which alleviates the traffic congestion caused by emergency vehicles. Maksymiv et al. [28] use two different convolution networks to detect the possible emergency situation, and the performance of the two neural networks is also discussed. Wang [29] uses the deep convolution neural network to identify road traffic signs. Du et al. [30] use YOLO V3 to monitor the vehicles and signal lights in the road network in real time. Zhang et al. [31] proposed the OR-CNN model which is based on Faster-RCNN to optimize the mutual occlusion of dense crowds. Wang et al. [32] proposed repulsion loss function for crowd scenes to solve the occlusion problem.

Compared with Faster R-CNN and the previous algorithms, the SSD-based algorithm successfully solves the problem of computing speed and accuracy, and it has been proved being suitable for the dense or evaluation traffic situation detection. The SSD-based algorithm uses multiscale prediction to improve the detection accuracy and uses convolution operation instead of full connection layer to predict the target category and location to improve the detection speed. For instance, Simonyan and Zisserman compared five convolution networks with different depths and concluded that the network with 16 layers of convolution depth has the best performance, and results show that the SSD-based method can be well applied to multitarget detection in complex violence environment.

Therefore, we harness the representation advance of the deep neural network for multiobject tracking. We propose an SSD-based method to not only extract the trajectory and behavior parameters of abnormal targets from the environment of sudden emergency or violence but also to assure the tracking association of objects in different frames. Based on the similarity of known conventional trajectories, the trajectories of unknown abnormal objects are classified, and the motion behaviors of different objects are analyzed and classified, which are shown by their trajectories. As a case study, we show how our model can be applied to emergency violence scenarios. The following sections describe our approach in detail.

The contributions of this paper could be stated as follows:

(a) A deep association tracking network (DATN) for multiobject was proposed for crowded scenarios.

(b) The method includes two parts: multiobject detection and object association.

(c) The proposed method is applied to both normal and emergent states. The accuracy and speed of the model were improved in both two states.

(d) The detection accuracy of the proposed method is higher than that of the traditional detection algorithm. The detection speed of normal data set is slightly higher than that of abnormal data set.

3. Data Set

3.1. Data Set Selection. Stable and large quantities of data are an important part of model accuracy assurance. For supervised learning, the data set mainly consists of training set and test set. Especially for deep learning, in the face of a large number of parameters to be learned, if there is no strong training set, it will lead to overfitting of the model and reduce the generalization ability of the model.

In the sudden emergency or violent environment, the data have complex characteristics, such as (a) crowd with high density and serious blocking by persons, cars, and buildings; (b) the abnormal behavior of the crowd is complex, with various state mutations; and (c) there are great differences in individual behaviors among different categories of people. So, we chose parts of video of Hong Kong chaos.

The training set in this paper adopts the public data set Pascal VOC07 + 12 data set (see Figure 1, in which the training set contains over 3000 person-images with almost 9000 person type targets. Combining with self-made dataset, the ultimate data set is enough for the task of pedestrian tracking with sudden emergency or violent environment.

In order to verify the generalization performance of the model, the test set consists of abnormal video and normal video. Through video cutting technology, the video in MP4 format is divided into multiple pictures, including 625 samples in normal state and 200 in abnormal state. Some data samples are shown in Figure 2, and we can see that people are in different postures, clothing with complex environment and lightening of background. These conditions will obviously increase the difficulty of detection.

3.2. Data Preprocess. The origin data often fail to meet the requirements in some aspects due to angle and image quality. So, we preprocess the data. The specific simple steps are as follows. Apply the two steps to multiple pairs of frames in a 30% overview. The resulting data are used as input data:

(1) Each pixel of the frame video is scaled by random value between 0.7 and 1.5. The resulting image is converted to HSV format. The same method and metrics are then scaled at random values and converted back to RGB format. Finally, it is recalibrated by a random value in the same range.

Figure 1: Partial training data set samples.

(a) (b)

Figure 2: Some test set data samples. (a) Normal state. (b) Abnormal state.

(2) The random value of 1–1.2 was used to expand the sample, and the background color was used to fill the enlarged image area. Then, we clipped the image with random values in the range of [0.8, 1], and finally, we retained only those targets that contained all the center points of the detection box in the original frame.

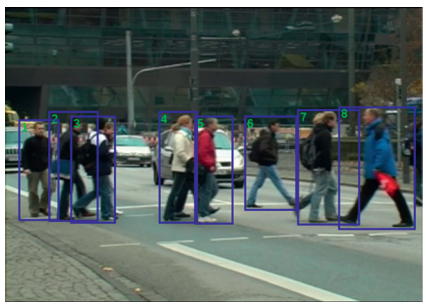
We set a gap of 10–15 frames, to compare three gap frames. As shown in Figure 3, a pedestrian (id:6) in the middle frame is occluded, while the pedestrian can be detected in the before and after frames. The test video is from a public dataset of TUD-Crossing, in which pedestrians cross the road and occlusion occurs many times. So, we can use the interpolation method with the before and back frames to locate the occluded pedestrian in the middle frame. For example, we compare the frames 1 and 13, and 13 and 26, and a row with a value of 0 is added to the matrix for a target that does not appear in either frame. After repeated data preprocessing, each frame finally contains N objects, and the matrix size is $N \times N$. In Figure 3(d), the X is for dummy targets. Parameters have been shared between multiple frames, and then the speed and acceleration changes have been calculated.

4. Model and Algorithm

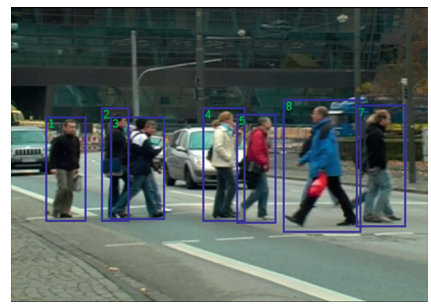
As a typical representative of one-stage target detection algorithm, unlike the process of two stages without candidate frame generation, it has been proved that the SSD-based method has excellent performance in the accuracy of target recognition and detection speed [14].

In a sudden emergency or violent environment, there will be a lot of panic stricken, fast, and irregular moving groups in the scene, and the edge of the building may be filled with individuals trying to squeeze out from the exit. In these cases, our method is very suitable for object recognition in these scenes because of the following advantages: it takes Visual Geometry Group Network (VGG-16) [33] as the backbone network structure, uses multiscale feature map for tracking, to ensure that the image can also have a certain accuracy at a lower resolution. Moreover, it sets a priori box for cells of multiple feature maps, which reduces the difficulty of model training to a certain extent, and finally computes the association matrix to link the object in previous multiframes for reliable trajectory. In this section, we provide a detail discussion on our framework, and the central to our technique is a CNN-based deep association tracking network (DATN). The method is introduced from three aspects: (a) design concept, (b) detection backbone, and (c) tracking section. The architecture of the model is shown in Figure 4.

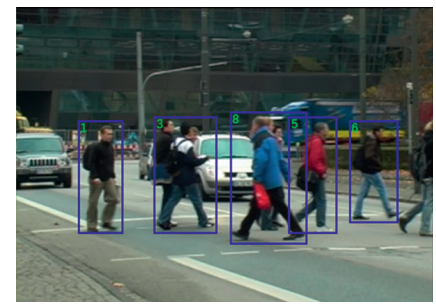
4.1. Design Concept. We track the multiobject in normal and abnormal scenarios' video and link the objects of different frames using the deep association tracking network (DATN). The proposed network was presented as two stages: object detector and association extractor. The object detector is used to detect the target objects. In this part, we can get the types and coordinates of the objects. Then, in the association extractor part, we can match the ID of the same object in different frames. And through these, we can finally get the



(a)



(b)



(c)

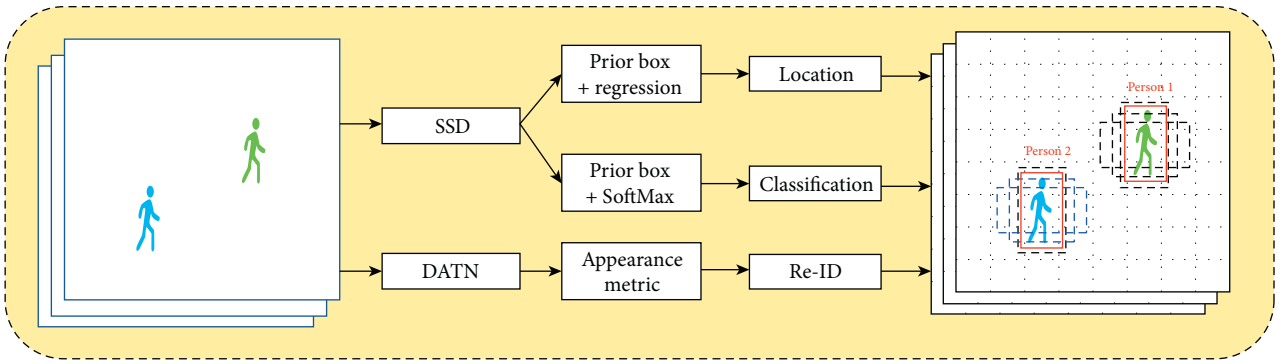
	1	2	3	4	5	6	7	8	X
1	1	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	1	0	0
8	0	0	0	0	0	0	0	1	0
X	0	0	0	0	0	1	0	0	0

(d)

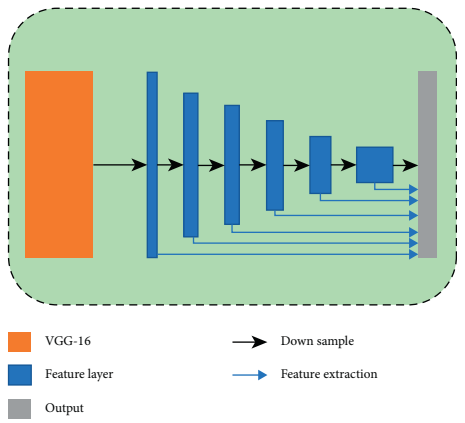
	1	2	3	4	5	6	7	8	X
1	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	1	0
X	0	0	0	0	0	0	0	0	0

(e)

Figure 3: Frame comparison. From: TUD-Crossing-test (frame: 1, 13, and 26). (a) Frame 1. (b) Frame 13. (c) Frame 26. (d) Matrix with dummy boxes (1 and 13). (e) Matrix with dummy boxes (13 and 26).



(a)



(b)

Figure 4: Continued.

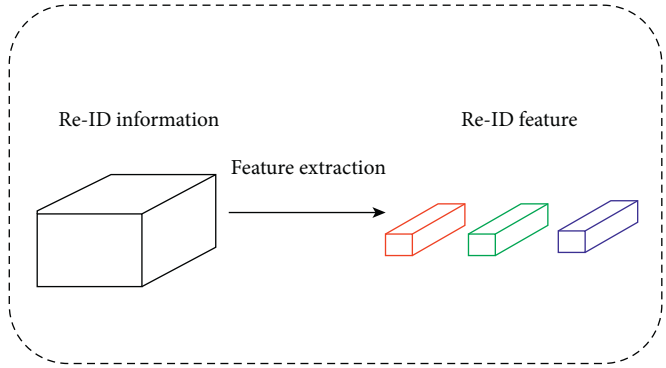


Figure 4: Model architecture. (a) Design concept. (b) Object detector. (c) Association extractor.

coordinates of all the objects in all frames. Considering the time gap, we can calculate the velocity and acceleration of them.

4.2. Object Detector

4.2.1. Backbone of Detection Network. The backbone network of our method is VGG-16 based. Its network architecture is shown in Figure 5. As shown, the architecture of the VGG-16 mainly consists of three core components: the convolution layer, the pooling layer, and the full connection layer [14, 33]. The convolution layer extracts feature from the image through convolution core to form a feature map. The pooling layer reduces the size of the model, improves the calculation speed and robustness of the extracted features. And the full connection layer is used for regression and classification. There are different architectures between the pooling layer and the convolution layer. Among them, the VGG-16 can be understood as the connection between the multilayer convolution layer and the pooling layer, with different architectures, and the rules of information transmission in the network are also different.

As stated before, the difference between SSD-based and other target tracking algorithms lies in the multiscale feature map strategy. The following three points introduce the core concept of multiobject tracking:

(1) The multiscale feature map refers to that after the fifth convolution layer of VGG-16, the image changes the full connection layer to the convolution layer. In general, the feature map in front of the CNN network is large, and, after passing through multiple pooling layers, the size of the feature map will be reduced. By setting different sizes of prior boxes, the larger feature map can be used to detect smaller objects, while the smaller feature map can detect larger objects' body.

(2) Our framework uses the provided SSD detectors directly to extract detection results from different feature images, which reduces parameter training to a certain extent.

(3) Our framework uses the anchor mechanism of Faster R-CNN for reference, sets a prior box of different scale size for each cell of the feature map, and the predicted bounding boxes are based on the prior box. For each prior box of each cell, a set of independent detection values corresponding to a bounding box is output.

As shown in Figure 6, the input picture of the SSD network is 300* 300, so the model is also called SSD-300, and the output characteristic picture size through the backbone network VGG-16 is 38*38*512. Then, the information flow direction on the feature map is divided into two directions. On the one hand, the information is transferred to the regression classification layer through the fifth convolution layer to predict directly, so as to train the loss function, where $[4^*(c+4)]$ represents c classification numbers, and each unit has 4 prior boxes and 4 position parameters. On the other hand, the information is transferred to the next layer. The full connection layer of the VGG-16 network is changed to the convolution layer. As the information is pooled before input, some information on the feature map will be lost. In order to increase the receptive field and not reduce the image size, the concept of Atrous conv (dilated convolution) is introduced. The convolution kernel is expanded by division rate-1 and filled with 0. The expanded convolution kernel is convoluted with the characteristic image. The superparameter division rate = 6. In the subsequent information transfer process, the heavy information is output in two directions, where s represents the step length of convolution kernel. In order to reduce the low edge recognition rate, the parameter p (padding) is introduced to expand the edge of the feature map to 0. Considering the weight sharing of the feature map to the convolution kernel, the relevant parameters of the whole network are shown in Table 1.

The convolution kernel size in Table 1 is as follows: 3*3*1024. Considering that the characteristic graph shares the weight on convolution kernel, it means that the convolution kernel size is 3*3, a total of 1024 convolution kernels. The number of priori boxes output from each layer to the classification regression layer can be seen from the

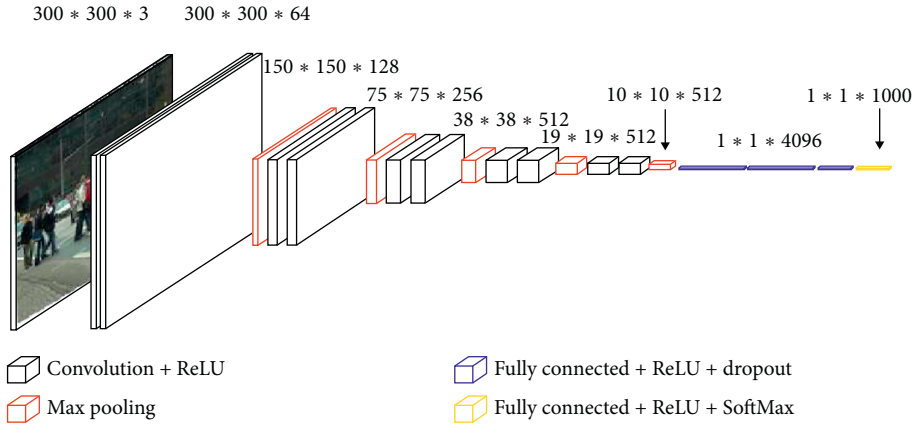


Figure 5: VGG-16 architecture.

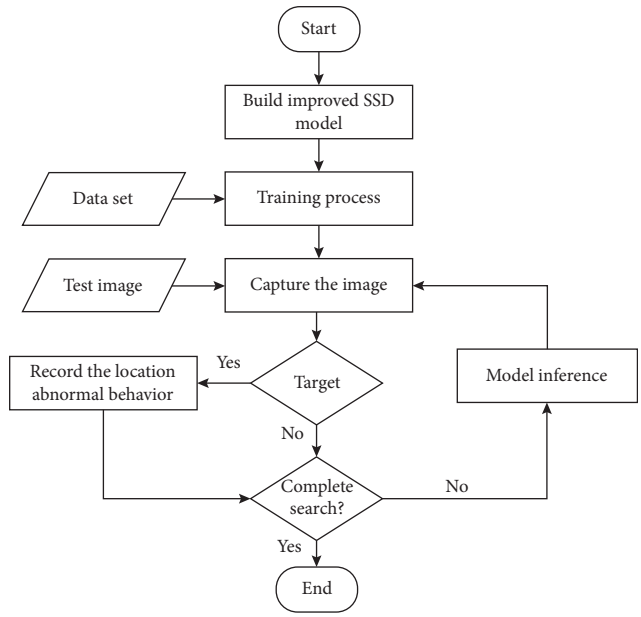


Figure 6: Model framework.

Table 1: Related parameters in SSD.

Layers name	Inputs	Size	Steps	Padding	Number of prior boxes	Outputs
VGG-16	300 * 300 * 3	——	——	——	38 * 38 * 4	38 * 38 * 512
Conv6	38 * 38 * 512	3 * 3 * 1024	——	——	——	19 * 19 * 1024
Conv7	19 * 19 * 1024	1 * 1 * 1024	1	0	19 * 19 * 6	19 * 19 * 1024
Conv8_1	19 * 19 * 1024	1 * 1 * 256	1	0	——	19 * 19 * 256
Conv8_2	19 * 19 * 256	3 * 3 * 512	2	1	10 * 10 * 6	10 * 10 * 512
Conv9_1	10 * 10 * 512	1 * 1 * 128	1	0	——	10 * 10 * 128
Conv9_2	10 * 10 * 128	3 * 3 * 256	2	1	5 * 5 * 6	5 * 5 * 256
Conv10_1	5 * 5 * 256	1 * 1 * 128	1	0	——	5 * 5 * 128
Conv10_2	5 * 5 * 128	3 * 3 * 256	1	0	3 * 3 * 4	3 * 3 * 256
Conv11_1	3 * 3 * 256	1 * 1 * 128	1	0	——	3 * 3 * 128
Conv11_2	3 * 3 * 128	3 * 3 * 256	1	0	1 * 1 * 4	1 * 1 * 256

above table that there are 8732 prior boxes, so SSD is essentially a new dense sampling method, which is also one of the key points to ensure that the target can be identified in an emergency.

4.2.2. Multiscale Fusion. On the basis of VGG-16, multiscale feature map detection is introduced to improve the detection accuracy. Its network structure is shown in Figure 7.

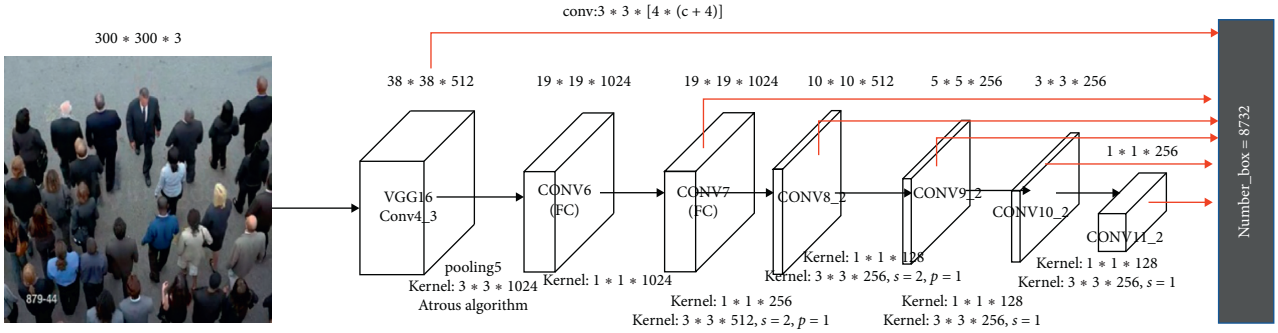


Figure 7: SSD network structure.

The feature maps of multiscales are extracted from Conv4_3, Conv7, Conv8_2, Conv9_2, Conv10_2, and Conv11_2, six layers. These six feature maps are in sizes of (38, 38), (19, 19), (10, 10), (5, 5), (3, 3), and (1, 1) and then upsampled into the same size of (38, 38). Through the fusion of six feature maps, the model can have the advantages of both large-scale and small-scale feature maps and effectively improve the detection accuracy.

4.2.3. Training Process. The SSD model has been successfully applied to existing data sets [14, 23–25, 34, 35], and it is of great application value to apply our SSD-based model to traffic hubs, large stadiums, squares, and places with dense crowds, especially to detect abnormal crowds in violent emergencies. We follow almost the same training policy as SSD in common traffic environment. The training process mainly includes the matching of prior frame and real target and loss function.

(i) Matching the Prior Box. The matching of SSD prior frame and real target is mainly divided into two parts. First, for each real target in the picture, the prior frame with the largest Intersection-over-Union (*IoU*) [36] is found in equation (1). Its equation represents the ratio of the intersection and union of prior frame and real target boundary frame area. Therefore, the higher the intersection and union ratio, the more matching the prior frame and real target. In the second part, considering that in the first part, there are fewer positive samples matching the real target, and there are more prior boxes matching as the background, which are called negative samples, which will lead to the imbalance of positive and negative samples in the model. Generally, *IoU* is set to be greater than a certain threshold (generally 0.5, strictly 0.6). In this way, there may be multiple prior boxes to match the real target, so as to ensure that the proportion of positive and negative samples is close to 1 : 3:

$$IoU = \frac{\text{area}(k) \cap \text{area}(G)}{\text{area}(k) \cup \text{area}(G)}, \tag{1}$$

where area(k) is the area of the prediction box and the area(G) is the area of the real target (grounding truth) box.

(ii) Loss Function. Because target detection involves two parts (target location and classification), SSD is defined as the weighted sum of location error and confidence error, which is different from the traditional mean square error loss function and cross entropy loss function in equation (2):

$$L(x, c, l, g) = \frac{1}{N}\left(L_{\text{conf}}(x, c) + \alpha L_{\text{loc}}(x, l, g)\right), \tag{2}$$

where N represents the number of samples of matched default boxes, c is the number of classifications, l is the predicted value of the position of the corresponding boundary box of the prior box and is the encoded value, the g is the position parameter of the real target and is the encoded value, and x is the {0,1} indicator. $L_{\text{conf}}(.)$ is the loss function of confidence, $L_{\text{loc}}(.)$ is the loss function of position, α is the weight coefficient, and the general value is 1.

The SoftMax loss function is used as the confidence loss function in the following equation:

$$L_{\text{conf}}(x, c) = -\sum_{i \in \text{Pos}}^{N} x_{ij}^{p} \log\left(\hat{c}_i^p\right) - \sum_{i \in \text{Neg}} \log\left(\hat{c}_i^0\right), \tag{3}$$

where x_{ij}^p indicates that the ith prior box matches the jth real target, and the real target category is p. The $\hat{c}_i^p$ is the category probability of prediction output, and the $\hat{c}_i^0$ is the probability of prediction as background, corresponding to negative samples.

The position loss function is

$$L_{\text{loc}}(x, l, g) = \sum_{i \in \text{Pos}}^{N} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^{k} g\left(\hat{l}_j^m - \hat{g}_j^m\right), \tag{4}$$

where cx and cy represent the center coordinate of the target and w and h represent the width and height of the target. x_{ij}^k indicates that ith prior box can match the jth real target, and the real target category is k. In order to ensure the robustness and stability of the solution, $g(.)$ adopts smooth L1 loss function in the following equation:

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if} |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{5}$$

4.3. Association Extractor. The first phase of the framework is called a functional extractor. The main function is to extract the crowd characteristics in the video. The object detector is used to detect and calibrate the video frame and

target center. And the extracted information can be used in the following steps [37].

First, the tracking is in 8-dimensional state space $(u, v, \gamma, h, \dot{x}, \dot{y}, \dot{\gamma}, \dot{h})$ that contains the bounding box center position (u, v), aspect ratio γ, and height h, and $(\dot{x}, \dot{y}, \dot{\gamma}, \dot{h})$ represent their velocities in image coordinates. Through the Kalman filter, the motion state and trajectory in next frame can be estimated.

Then, the ID-matching can be divided into 2 parts: the motion matching and appearance matching.

The motion matching is calculated by Mahalanobis distance based on the position information by the Kalman filter. The matching process is

$$d^{(1)}(i, j) = \left(d_j - y_i\right)^T S_i^{-1}\left(d_j - y_i\right), \tag{6}$$

where d_j indicates the state (u, v, γ, h) of the jth bounding box and y_i indicates the position of the ith tracker step. S_i is the covariance matrix estimated by the Kalman filter.

Besides the motion matching, the appearance matching is introduced to improve the matching accuracy. Appearance matching uses feature vectors, which are extracted by the detector. The feature map of the bounding box will be extracted and then stored in the list through the ReID (reidentification) network module. By calculating the minimum cosine distance of all feature maps in the list, we can get the appearance matching value, as is shown in

$$d^{(2)}(i, j) = \min\left\{1 - r_j^T r_k^{(i)} | r_k^{(i)} \in R_i\right\}, \tag{7}$$

where r_j represents the appearance descriptor of the jth bounding box and $||r_j|| = 1$, and R^i is the feature information for the ith track in the last kth associated appearance descriptors for each track.

We can calculate the final result of ID-matching measurement by the two values, as is shown in equation (8). λ is the weight of the value:

$$C_{i,j} = \lambda d^{(1)}(i, j) + (1 - \lambda) d^{(2)}(i, j). \tag{8}$$

With the Hungarian algorithm [38], the information of the matching detection frame and the tracking frame is correlated, and the ID-match result can be obtained.

5. Results and Evaluation

The computation is based on TensorFlow version 1.8 deep learning framework in Anaconda environment. Its configuration is as follows: Intel Core i5-8300H, CPU 2.30 GHz, 8 GB memory, 1 TB hard disk, and the GPU is NVIDIA GeForce GTX 1050.

The hyperparameters are as follows.

The learning rate and the learning rate decay are set to 0.001 and 0.94. Batch size is set to 50. The epoch is 3000 times. For the optimizer, we choose Adam which has low memory requirements. We set *IoU* threshold to 0.1 to make a better effect of complex environment.

5.1. Evaluating Indicator. The classification problem in target detection is multiple classification, and the proportion of positive and negative samples of each class is uneven when facing this kind of skew problem. The steps are shown as follows:

(1) For category c, the prediction boxes of all categories c output by the algorithm are first sorted by confidence.

(2) Select the top k prediction boxes, calculate false positive (FP) and true positive (TP), and make recall equal to 1.

(3) Calculate precision.

(4) Repeat Step 2 and select different K to make recall equal to 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0, respectively.

(5) If recall is greater than each threshold, the corresponding 11 maximum precisions will be averaged; i.e., average precision (AP) will be obtained; AP is for a single category, and mAP is to sum and average all categories of AP.

The formula of accuracy rate and recall rate is shown as equations (9) and (10), where TP is true positive, TN is true negative, FP is false positive, and FN is false negative:

$$P = \frac{TP}{(TP + FP)}, \tag{9}$$

$$\text{recall} = \frac{TP}{(TP + FN)}. \tag{10}$$

5.2. Mean of Average Precision (mAP). This experiment is to discuss the generalization performance of the training model, the normal data with little or no occlusion, and the abnormal data with more occluded and edge objects are used in the test set. We will discuss them separately in this section and make P-R curve to calculate mAP.

For the normal dataset, it mainly includes four categories of cars, buses, bicycles, and pedestrians. The corresponding P-R curve is shown in Figure 8.

According to the properties of P-R curve, the more convex the curve is, the better the effect is. The value of AP was calculated, respectively, such as $AP(car) = 94.94\%$, $AP(bus) = 96.97\%$, $AP(bike) = 99.24$, and $AP(person) = 97.42\%$. The mAP is the average of each class of AP. The result shows that the mAP of the model is 97.14% in the normal test set, and some of the test results are shown in Figure 9, where the left value represents the category label and the number on the right represents the confidence score of the category the target is. For the category label number, 2 represents bicycle, 7 represents car, 14 represents motorbike, and 15 represents person.

For abnormal data, *tt* mainly includes three types of objectives, and they are pedestrians, cattle, and cars. The corresponding P-R curve is shown in Figure 10. Calculating the AP value of each class, the values are as follows: $AP(person) = 98.73\%$, $AP(cow) = 93.9\%$, and $AP(car) = 94.85\%$. The recognition rate of cattle is relatively low, because the training data set of cattle accounts for about 2% of

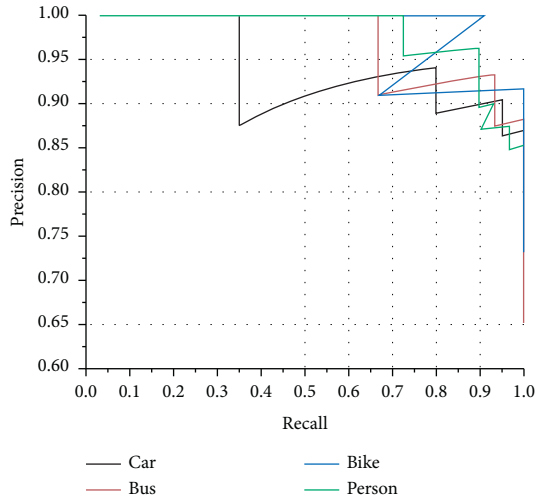


Figure 8: Normal video classification P-R curve.

Figure 9: Test results of some normal data.

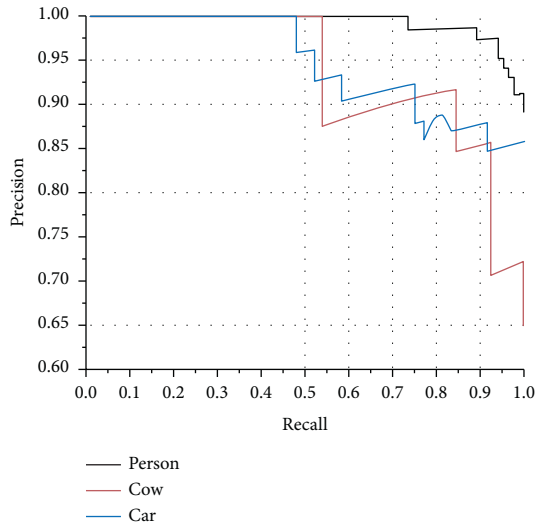


Figure 10: Abnormal video classification P-R curve.

the total data set. Therefore, there may be underfitting of the model for cattle recognition. Then, the mAP value under abnormal condition is 95.83%. Some of the test results are shown in Figure 11.

By testing the data sets in the above two states, it is concluded that the SSD model has about 95% and more recognition accuracy for objects of different scales. Also, it has a certain detection accuracy when there are many edge targets in the abnormal state. This is due to SSD using similar anchor mechanism combined with multiscale feature map detection, which can identify different size targets and edge targets.

Figure 11: Test results of some abnormal data.

Table 2: Comprehensive test results.

Algorithm		Normal state	Abnormal state
SSD	mAP	97.14%	95.83%
	FPS	11	8
YOLO	mAP	90.56%	88.97%
	FPS	23	15

5.3. Detection Speed. The target detection algorithm not only has certain requirements for accuracy but also pays attention to detection speed. We run the SSD detection model in the same data sets and the same test environment, compared to YOLO which also represents of the one-stage algorithm. The corresponding speed and accuracy are shown in Table 2.

It can be seen in Table 2 that the detection speed in normal state is slightly faster than that in abnormal state, but the accuracy of the former is 1.31% higher than that of the latter. Compared to YOLO, the detection accuracy of YOLO is significantly lower than that of SSD, but its detection speed is higher than that of SSD. This is because SSD inputs multilayer feature maps into the prediction layer, while YOLO only performs one convolution and single feature map prediction. Therefore, SSD is more suitable for target detection in emergency.

6. Conclusion

This work introduces the research achievements of target detection and its application in the field of transportation in recent years. Based on the core algorithm of SSD target detection, an improved shot multibox detector is proposed, for more effective object detection in sudden emergency or violent environment. The main conclusions are as follows:

(1) The SSD model is able to outperform the YOLO framework on abnormal behavior objects or specific unexpected objects.

(2) These test data for normal and abnormal states are produced. The detection accuracy and speed are improved. The detection speed for abnormal states is higher than normal states; on the contrary, the detection accuracy for normal is higher than it for abnormal.

For future work, a technique for weakly supervised image segmentation will be integrated with the improved SSD framework. We also plan to improve the detection speed and accuracy using more matching strategies and appropriate data set for sudden emergency or violent environment. The next step is to analyze the movement characteristics of pedestrians in the emergency state. In addition, it is necessary to predict the pedestrian trajectory in the emergency state, which provided means support for traffic safety management.

Data Availability

The data can be found by contacting the e-mail lizhihong@bucea.edu.cn.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

Acknowledgments

This paper was supported by the Beijing Social Science Fund Project (2020GLB020).

References

[1] S. Shumin, Y. Zhaosheng, Y. Wei, and Z. Maolei, "Research on multi-objective decision model based on location-allocation analysis during emergency traffic evacuation," in *Proceedings of the 2010 IEEE International Conference on Emergency Management and Management Sciences*, Beijing, China, August 2010.

[2] H. Meiyan, "Study on emergency management of traffic jam and improvement of emergency mechanism during ice disaster in Hunan," in *Proceedings of the 2010 IEEE International Conference on Emergency Management and Management Sciences*, Beijing, China, August 2010.

[3] W. Chuanmei and T. Hengqing, "Structural analysis of factors affecting urban traffic safety," in *Proceedings of the 2010 International Conference on Mechanic Automation and Control Engineering*, Wuhan, China, June 2010.

[4] Z. Li and W. A. Xu, "Pedestrian evacuation within limited-space buildings based on different exit design schemes," *Safety Science*, vol. 124, Article ID 104575, 2020.

[5] M. Hassaballah, M. A. Kenk, K. Muhammad, and S. Minaee, "Vehicle detection and tracking in adverse weather using a deep learning framework," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, 2020.

[6] W. Sun, M. Sun, X. Zhang, and M. Li, "Moving vehicle detection and tracking based on optical flow method and immune particle filter under complex transportation environments," *Complexity*, vol. 2020, 2020.

[7] X. Chen, L. Qi, Y. Yang et al., "Video-based detection infrastructure enhancement for automated ship recognition and behavior analysis," *Journal of Advanced Transportation*, vol. 2020, 2020.

[8] X. Chen, S. Wang, C. Shi, H. Wu, J. Zhao, and J. Fu, "Robust ship tracking via multi-view learning and sparse representation," *Journal of Navigation*, vol. 72, no. 1, pp. 176–192, 2019.

[9] J. Kampars and J. Grabis, "Near real-time big-data processing for data driven applications," in *2017 International Conference on Big Data Innovations and Applications (Innovate-Data)*, Prague, Czech Republic, August 2017.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (Voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[11] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft Coco: common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Springer International Publishing, New York, NY, USA, 2014.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[13] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, Las Vegas, NV, USA, June 2015.

[14] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in *Computer Vision–ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Springer International Publishing, New York, NY, USA, 2016.

[15] R. Girshick, "Fast R-CNN," in *The IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, Santiago, Chile, December 2015.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR, Kauai, HI, USA, December 2001.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, June 2005.

[18] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, June 2008.

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2013, https://arxiv.org/abs/1311.2524.

[20] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "Densebox: unifying landmark localization with end to end object detection," 2015, https://arxiv.org/abs/1509.04874.

[21] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: an advanced object detection network," 2016, https://arxiv.org/abs/1608.01471.

[22] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," 2016, https://arxiv.org/abs/1612.08242.

[23] Z. Li and F. Zhou, "FSSD: feature fusion single shot multibox detector," 2017, https://arxiv.org/abs/1712.00960.

[24] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: deconvolutional single shot detector," 2017, https://arxiv.org/abs/1701.06659.

[25] J. Yi, P. Wu, and D. N. Metaxas, "ASSD: attentive single shot multibox detector," *Computer Vision and Image Understanding*, vol. 189, Article ID 102827, 2019.

[26] W. Wu, Y. Yin, X. Wang, and D. Xu, "Face detection with different scales based on faster R-CNN," *IEEE Transactions on Cybernetics*, vol. 49, no. 11, pp. 4017–4028, 2019.

[27] S. Roy and M. S. Rahman, "Emergency vehicle detection on heavy traffic road from cctv footage using deep convolutional neural network," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox'sBazar, Bangladesh, February 2019.

[28] O. Maksymiv, T. Rak, and O. Menshikova, "Deep convolutional network for detecting probable emergency situations," in *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, Ukraine, August 2016.

[29] C. Wang, "Research and application of traffic sign detection and recognition based on deep learning," in *Proceedings of the 2018 International Conference on Robots & Intelligent System (ICRIS)*, Changsha, China, February 2018.

[30] L. Du, W. Chen, S. Fu, H. Kong, C. Li, and Z. Pei, "Real-time detection of vehicle and traffic light for intelligent and connected vehicles based on Yolov3 network," in *Proceedings of the 2019 5th International Conference on Transportation Information and Safety (ICTIS)*, Liverpool, UK, July 2019.

[31] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: detecting pedestrians in a crowd," 2018, https://arxiv.org/abs/1807.08407.

[32] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: detecting pedestrians in a crowd," 2017, https://arxiv.org/abs/1711.07752.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[34] B. Leibe, N. Sebe, and M. Welling, "SSD: single shot multibox detector," *Computer Vision – ECCV*, vol. 2016, p. 9905, 2016.

[35] B. Schroeder, A. Merchant, and R. Lagisetty, "Reliability of nand-based SSDS: what field studies tell us," *Proceedings of the IEEE*, vol. 105, no. 9, pp. 1751–1769, 2017.

[36] R. Hamid, T. Nathan, G. JunYoung, S. Amir, R. Ian, and S. Silvio, "Generalized intersection over union: a metric and a loss for bounding box regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR, Long Beach, CA, USA, June 2019.

[37] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, Beijing, China, September 2017.

[38] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, Phoenix, AZ, USA, September 2016.