Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 1500428, 9 pages
https://doi.org/10.1155/2022/1500428

Research Article

Improved SSD Model for Pedestrian Detection in Natural Scene

Feng Hong, Chang hua Lu, Wang Tao, and Weiwei Jiang

School of Computer and Information, Hefei University of Technology, Hefei, Anhui Province, China

Correspondence should be addressed to Feng Hong; hongfeng@czu.edu.cn

Received 27 July 2022; Revised 7 September 2022; Accepted 22 September 2022; Published 25 November 2022

Academic Editor: Kuruva Lakshmanna

The indagation improves the SSD network to improve the target detection performance in unmanned driving at night. In target detection, the goal is to identify and locate the different types of objects present in an image. The first-level target detection method pulls categorization information and target position information for use by the second-level target detection algorithm using the featured MAP created by the deep network. The depth characteristics, on the other hand, are processed using long-distance convolution and downsampling. Given the lack of geographical information, research alludes to the concept of semantic segmentation and proposes a method for improving the first-level target identification algorithm SSD by mixing shallow characteristics from the backbone network with deep features obtained through sampling. In addition, this research enhances the generation method and loss function of the preselection box by taking into account the peculiarities of pedestrian detection data. Undertakings to experiments on the two data sets provided by INRIA and Caltech show that the improved model USSD reported in this paper improves both the efficiency of detection and speed of retrieval.

1. Introduction

At present, most of the research on environment perception of unmanned driving focuses on daytime scenes, while the research on night scenes is relatively small, which makes the application of unmanned driving at night very limited. Unmanned driving at night can use infrared cameras to perceive temperature-sensing imaging of surrounding environmental objects, but the formed images are relatively common, and the images have shortcomings such as less texture information, more noise, and blurred images [1], so the night target detection is more difficult. Effective night target detection can reduce the occurrence of traffic accidents and has high application value. The popular camera terminal provides the possibility to obtain a large number of video image data, but at the same time, the huge amount of data makes the inefficient manual browsing become an obstacle to obtain information [2]. With the improvement of hardware performance, the rapid development of artificial intelligence technology began to replace manual browsing to obtain information from video images.

Deep learning-based target detection algorithms have now exceeded traditional detection methods and have become the mainstream of contemporary target detection algorithms, which mostly include single-stage (one-stage) and two-stage (two-stage) target detection algorithms. The two-stage algorithm is based on the Faster RCN (Faster Region) network [2] series of target detection methods, which cannot meet the requirements of real time, and it cannot meet the requirements of the candidate area. The speed is relatively slow. The one-stage algorithm is based on SSD (single-shot multibox detector) network [3] and YOLO (you only look once) network [4–6]. The idea of the SSD network and YOLO network is to transform the target detection task from a classification problem into a regression problem and complete the target location and classification at one time. The SSD network is based on the anchor point mechanism in the Faster RCN network, and a similar prior box method is proposed. The SSD network adds a feature pyramid-based detection method, that is, predicting targets on feature maps of different receptive fields.

Target detection includes two tasks: classification and location. However, these two tasks conflict with each other. The classification task requires the model to be invariant to the spatial position and pose transformation of the target, but the positioning task requires the model to be sensitive to it. In essence, the semantic segmentation task of pixel by pixel classification has the same problem [7]. Currently, popular regression-based target identification algorithms produce and categorise candidate frames of various sizes on the original image (as shown in Figure 1), then sort a large number of candidate frames using nonmaximum suppression, and lastly determine the target position using border regression. In this paper, we take the feature MAP pixels to generate candidate frames on the original image and classify them as a coarse-grained semantic segmentation task. Referring to the idea of considering classification and localization in semantic segmentation, the multiscale feature acquisition process is applied to the backbone network, and the method of combining deep features with shallow features is used for target detection. In order to improve the speed of target detection, this paper selects the network model DarkNet [4] which has an excellent performance in speed and precision as the backbone network, simplifies convolution, and proposes the USSD model. Finally, according to the characteristics of pedestrian detection data, this paper improves the generation method and loss function of the preselected border.

The paper is organized as follows: the related work is presented in Section 2; Section 3 consists of the method section. Section 4 discusses the experimentation design. Finally, in Section 5, the research work is concluded.

2. Related Work

Like other computer vision tasks, before the emergence of convolution neural networks, traditional target detection methods rely on simple visual features such as texture, color, edge, and hand-designed features. Then, the position, size, and category of the target are obtained by using a regression model and classifier, respectively. The commonly used manually designed features include the scale-invariant feature transform (SIFT) feature, histogram of oriented gradient (HOG) feature, and local binary pattern (LBP) feature [8]. The commonly used classifiers are support vector machines (SVM) and AdaBoost [9]. Literature [8–13] is a typical example of this kind of method. The artificial features used by them are limited by the designer's experience, the limitations of complex targets are fully exposed in the detection task, and the accuracy and robustness of the algorithm are difficult to meet the requirements of practical application.

Convolution neural network (CNN) has an excellent performance in the field of computer vision. The performance of R-CNN [14], the first target detection algorithm based on CNN, is far superior to the traditional algorithm based on artificial features at that time [15]. Since then, deep learning algorithm has been widely used in the field of target detection. The series models represented by R-CNN follow the idea of traditional methods: firstly, candidate frames of samples are generated; then, the features of candidate border regions are extracted and classified; and finally, the boundaries of candidate borders are adjusted. This kind of method is called a two-stage algorithm. The two-stage algorithm can usually achieve excellent detection results, but the cumbersome training steps, a large number of calculations, and complex candidate region extraction algorithm have become the bottleneck of its performance improvement. Fast R-CNN [16], Fast R-CNN [17], and mask R-CNN [18] have addressed the issues of computational redundancy and time-consuming candidate region extraction, but it is difficult to reduce a large number of parameters in the two stages of feature extraction and reclassification of candidate regions, and the improved algorithm still struggles to meet real-time requirements. To solve this problem, a one-stage algorithm, represented by YOLO [19], which only needs to extract features once, is proposed. In a one-stage algorithm, the target detection is regarded as a regression problem, and the image features are divided into grids to predict the target border. The speed of generating a preselected border on feature MAP is much faster than that of feature RE Extraction of the candidate border region. However, due to the lack of calculation of candidate region features, the target location accuracy and classification accuracy of the one-stage algorithm are inferior to those of the two-stage algorithm. In order to improve the speed and accuracy of target detection at the same time, SSD [20] uses the multiscale concept of traditional machine learning algorithm on the basis of YOLO [20, 21] prediction regression box and predicts the regression box on the feature MAP of different scales, which makes up for the shortage of regional feature calculation in YOLO algorithm and obtains the detection speed of YOLO [5, 6] and the detection accuracy of Fast R-CNN.

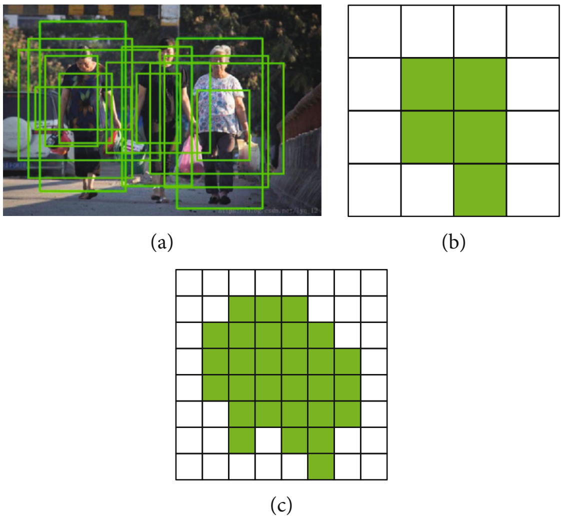
YOLO V3 [4] model put forward its own backbone network DarkNet53. The basic module of DarkNet53 is shown in Figure 2. The convolution of 1 × 1 is used for dimension reduction, which reduces the calculation amount. The overall structure of the model is shown in Table 1. A large number of basic modules are used. In order to avoid the negative effect of the gradient caused by pooling, the convolution with step size 2 is used instead of pooling to realize downsampling. The performance of this network is very good. Its classification performance is close to that of ResNet152, and its speed is nearly twice that of ResNet152. It has become one of the important backbone networks of the target detection model in recent years. The USSD model proposed in this paper conducts comparison experiments by replacing the backbone network, and finally, DarkNet53 is selected as the backbone network.

3. Method

In this section, we defined SSD, USSD, USSD's preselected border generation method, and loss function in detail.

3.1. SSD.

SSD is an additional structure added to the end of the underlying network, as depicted in Figure 3. SSD selects the commonly used depth model as the basic network to extract the overall features of the image, then transforms the extracted features to different resolution scales through a simple convolution layer and pooling layer, and then generates a fixed size and number of preselected framesets and the confidence degree of object categories in the frame through the features of different scales. Finally, a large number of preselected edges are processed by nonmaximum


(a)

(b)

(c)

Figure 1: (a) Partial preselected borders generated by the model. (b) Classification results on the 4 × 4 feature MAP. (c) Classification results on the 8 × 8 feature MAP.

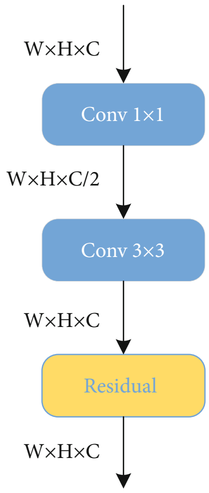


Figure 2: DarkNet53 basic block.

Table 1: DarkNet53.

Type	Filters	Size	Output
Convolution	32	3 × 3	256 × 256
Convolution	64	3 × 3/2	128 × 128
1 × basic block			128 × 128
Convolution	128	3 × 3/2	64 × 64
2 × basic block			64 × 64
Convolution	256	3 × 3/2	32 × 32
8 × basic block			32 × 32
Convolution	512	3 × 3/2	16 × 16
8 × basic block			16 × 16
Convolution	1024	3 × 3/2	8 × 8
4 × basic block			8 × 8

suppression operation; finally, the final positioning results are obtained.

The way SSD generates preselected borders is similar to the anchor mechanism in Fast R-CNN. As shown in Figure 4, an anchor border is generated by taking each point on the featured MAP as the center, and a series of concentric preselected borders are generated according to different border length and width height ratios. Suppose that m ($m = 6$ in SSD300 model, $m = 4$ in USSD model) is used to generate preselected borders. The border length $S_{\min}$ of the bottom feature MAP and the border length $S_{\max}$ of the top feature MAP are set. The border length of other feature graphs is calculated by the following formula:

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m - 1}(k - 1), \quad k \in [1, m]. \tag{1}$$

There are five values for the width height ratio α. The width and high pass formula (2) of the preselected border are calculated according to the width height ratio and the frame length:

$$w_k^a = s_k \sqrt{a_r}, h_k^a = \frac{s_k}{\sqrt{a_r}}, \tag{2}$$

where s_k is the anchor border length of the featured MAP of the layer, that is, the size when the height-width ratio of the preselected border is 1. In addition, a square preselected border of $s_k' = \sqrt{s_k \times s_{k+1}}$ will be set.

3.2. USSD. The image feature within the preselected frame range of the associated original image is represented by each element of the feature graph, which may be thought of as a coarse-grained semantic segmentation. Semantic segmentation models often use a continuous upsampling structure to obtain multiscale features that are easy to classify. In this paper, U-Net [22] is combined with SSD to ensure the classification accuracy of each pixel in the featured MAP (as

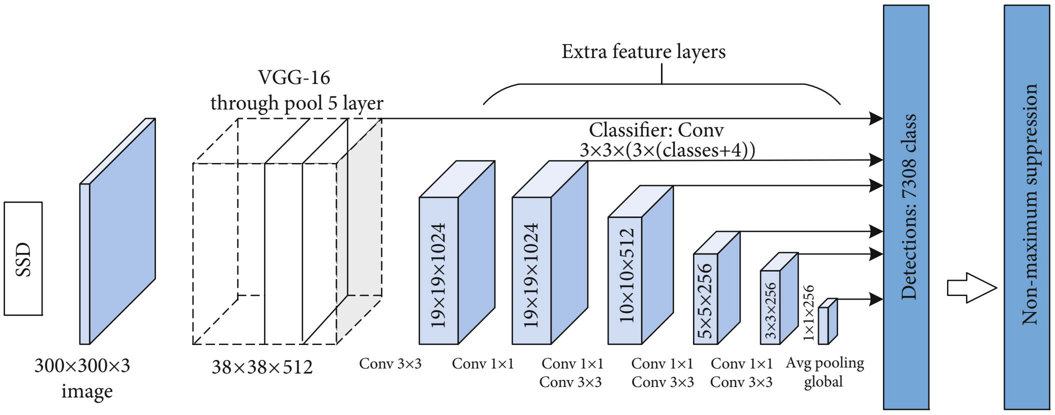


Figure 3: SSD structure diagram.

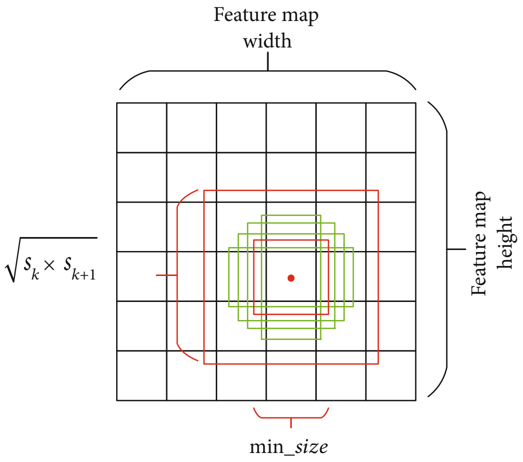


Figure 4: Schematic diagram of SSD precheck box generation.

shown in Figure 5). The difference between pedestrian detection and general target detection is that the number of pedestrian targets is large and dense, and spatial information is particularly important. Therefore, global pooling and full connection layer without local information and 1 × 1 resolution feature MAP without spatial information are no longer used in backbone network.

Classification and location are two contradictory tasks, and USSD considers the distinction between the two. The deep network model fully extracts features from images to obtain feature MAPs with rich classification information and converts classified information into feature MAPs with different resolutions through continuous upsampling. The shallow network model's categorization information is limited, and most of them focus on a small set of visual features, but shallow features have a lot of spatial position information. USSD combines the deep features with the shallow features to obtain the final multiscale features, giving consideration to both classification and positioning. In this paper, DarkNet53 was used as the backbone network. The classification accuracy of DarkNet53 is close to that of ResNet152, and the running time is only about 1/2 of that of ResNet152. The calculation of the semantic segmentation network is dense. In order to speed up the calculation and not lose the performance of the model, as shown in Figure 6, this paper USES more concise $1 \times k$ convolution and $k \times 1$ convolution instead of $k \times k$ convolution and continuous convolution [7].

3.3. USSD's Preselected Border Generation Method. Based on the anchor point idea in R-CNN, the SSD generates preselected borders by setting different aspect ratios and basic scales on multiple feature MAPs with different resolutions. Each preselected border needs to learn to predict the confidence of the categories and the bias about the coordinate position and width and height values. The former is the classification task, and the latter is the regression task. The purpose is to identify the target category as far as possible and get close to the marked border from the position. The generation effect of preselected borders is very important to the quality of detection results. The parameters of SSD-generated preselected borders were determined according to the clustering of VOC2007 and VOC2012 data sets [23]. The targets in the data set were of many categories and different shapes. Therefore, the generation method of width and height ratio symmetry in Figure 7(a) was adopted to evenly deal with various targets. However, this balanced generation method is not suitable for pedestrian detection. Pedestrian data is relatively regular. There is no case when the aspect ratio is larger than 1, and the height value is bigger than the width value. Pedestrian detection mostly considers walkers who are standing. The change of background and attitude and the distance between pedestrians and the camera do not affect the aspect ratio of pedestrians, which is within a relatively fixed range [24]. According to this characteristic, two improved generation methods of preselected border are designed in this paper.

$$h_i = h_{\min} + \frac{h_{\max} - h_{\min}}{k - 1} \times (i - 1), \quad i \in [1, 5], \tag{3}$$

$$w_i = h_i \times \text{ratio}. \tag{4}$$

Figure 7(c) is a revision of Figure 7(b). The ratio of width to height of pedestrian is within a certain range, so the way of taking fixed value is not suitable for the object of height

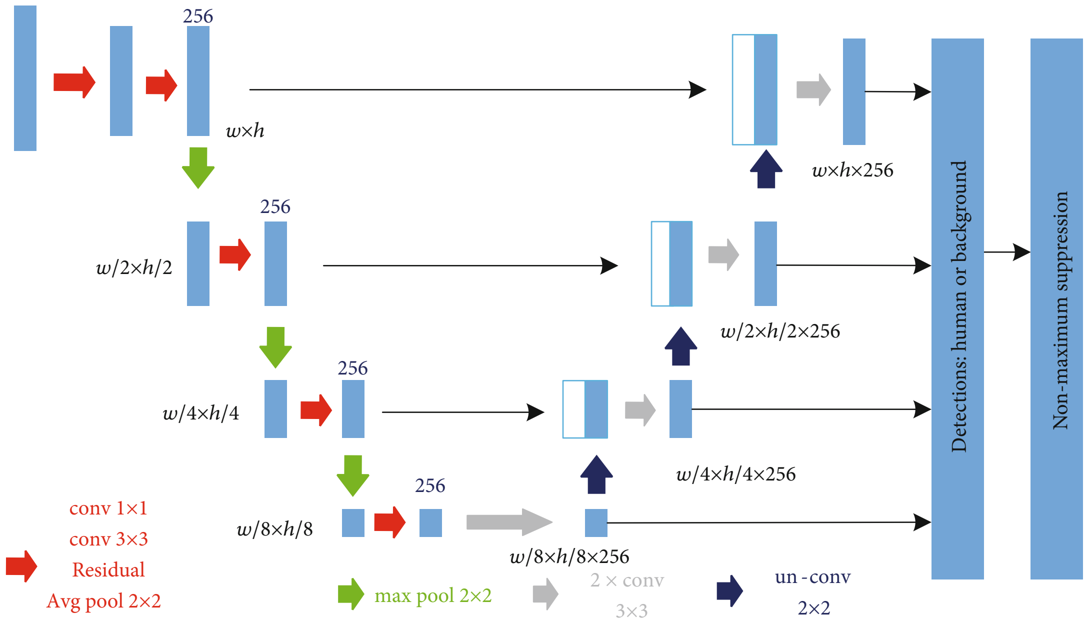


Figure 5: Schematic diagram of USSD.

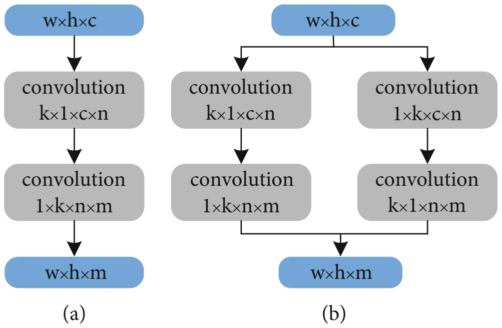


Figure 6: Two simplifications of the convolution.

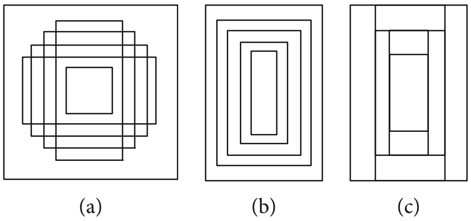


Figure 7: (a) SSD precheck box generation method. (b) Improved precheck box generation method 1. (c) Improved precheck box generation method 2. Figure 7(b) is a generation method of preselecting boxes of different heights with fixed aspect ratio. In the process of generating candidate borders, 5 candidate boxes are generated on each layer of different feature MAPs. A maximum height and minimum height should be set for each layer of feature MAP, and the generation process is shown in Equations (3) and (4).

to width comparison. Considering that too many preselected borders will affect the detection speed, the correction method is to generate three preselected borders in the way of Figure 7(b) to increase the height and reduce the width. Based on the second preselected border, two preselected borders with abnormal width/height ratio are generated.

3.4. Loss Function. As shown in Equation (5), the loss function is divided into two parts: confidence loss and positioning loss. In this paper, the confidence loss is improved.

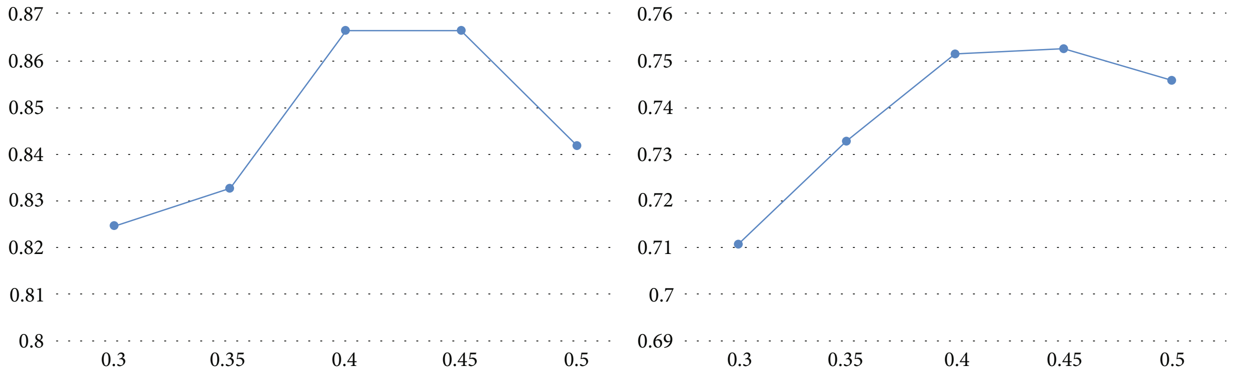
$$L(x, c, l, g) = \frac{1}{N}\left(L_{\text{conf}}(x, c) + \alpha L_{\text{loc}}(x, l, g)\right), \tag{5}$$

where N is the number of matching to the real border, l represents the predicted border, and g represents the real border. Furthermore, because the confidence loss and positioning loss are prone to have a big numerical difference, the default value is 1 to avoid biased introduction of and adjust the proportion between them.

Confidence loss is calculated based on the model's classification certainty for the target category in the bounding box, while the SSD model uses the crossentropy loss in the common classification task. Equation (6) is divided into two terms: the former represents the matching degree

Table 2: Data set statistics.

—	—	Pedestrian	Positive sample	Negative sample
INRIA	Train set	1280	614	1218
	Test set	566	288	453
Caltech	Train set	192K	67K	61K
	Test set	155K	65K	56K



(a)

(b)

Figure 8: The model set AR (a) on the test set with different aspect ratios MAP (b).

between the predicted border and the real border on the category; the latter represents the probability that there is no entity within the predicted border to judge it as the background; the greater the prediction probability, the smaller the loss will be.

$$L_{\text{conf}}(x, c) = -\sum_{i\in \text{Pos}}^{N} x_{ij}^{P} \log\left(\widehat{c}_i^{P}\right) - \sum_{i\in \text{Neg}} \log\left(\widehat{c}_i^{0}\right), \qquad (6)$$

$$\widehat{c}_i^{P} = \frac{\exp\left(c_i^{P}\right)}{\sum_{P} \exp\left(c_i^{P}\right)}, \qquad (7)$$

where I represents the bounding box ordinal number of model output, J represents the real bounding box ordinal number, P represents the category ordinal number, $P = 0$ refers to the background, $x_{ij}^{P} \in \{0, 1\}$ represents whether the predicted border matches the real border, and the judgment of whether the border matches depends on the IOU of the two, with the threshold value of 0.5. Probability $\widehat{c}^{P}$ is processed by Softmax.

Crossentropy loss takes a fair approach to data, that is, to treat both positive and negative samples equally, as well as simple samples and difficult samples equally. There are a large number of background samples in the pedestrian detection data set that are easy to be classified. These samples will account for most of the loss in the training process and have a great impact on the model gradient, which will degrade the model to a certain extent. Although the method of introducing weight factor can partially solve the problem of negative and positive sample imbalance, it cannot balance the simple and difficult samples. Focus loss [25] can reduce the weight of simple samples, focus the training of the model on difficult samples, and further solve the problem of sample imbalance.

Table 3: Experimental results of a method for generating preselected boxes with fixed aspect ratio.

Ratio	0.3	0.35	0.4	0.45	0.5
AR	82.47	83.26	86.66	86.65	84.18
MAP	71.08	73.26	75.12	75.23	74.55

$$L_{\text{conf}}(x, c) = -\sum_{i\in \text{Pos}}^{N} \left(1 - \widehat{c}_i^{P}\right)^{\gamma} x_{ij}^{P} \log\left(\widehat{c}_i^{P}\right) - \sum_{i\in \text{Neg}} \left(1 - \widehat{c}_i^{0}\right)^{\gamma} \log\left(\widehat{c}_i^{0}\right), \qquad (8)$$

where $(1 - \widehat{c}_i^{P})^{\gamma}$ is the weight factor and the difficulty degree of the sample is judged according to the probability $\widehat{c}_i^{P}$. A large prediction probability means that the sample is a simple sample that is easy to be classified and USES a small weight; otherwise, it means that the difficult sample uses a large weight. $\gamma \geq 0$ is an adjustable super parameter, known as a focusing parameter, and a criterion for judging the difficulty of controlling samples.

The positioning loss calculates the deviation value between the prediction box and the real box, including the center point coordinate deviation and the size deviation of the bounding box, using the loss of $\text{smooth}_{\text{L1}}$. Unlike the confidence loss, the positioning loss is only calculated for the positive class samples with solid inside the frame, and the negative class samples without solid inside the frame are ignored.

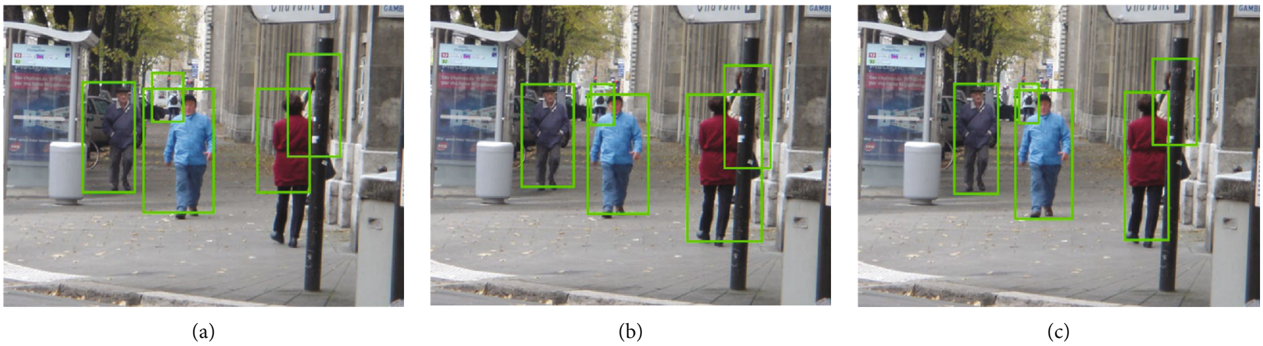


FIGURE 9: Pedestrian detection and comparison of three preselection box generation methods.

$$L_{\mathrm{loc}}(x,l,g)=\sum_{i\in\mathrm{Pos}}^{N}\sum_{m\in\{cx,cy,w,h\}}x_{ij}^{k}\mathrm{smooth}_{\mathrm{L1}}\left(l_i^m-\widehat{g}_j^m\right),$$

$$\mathrm{Smooth}_{\mathrm{L1}}(a)=\begin{cases}0.5a^2, & \text{if}|a|<1,\\ |a|-0.5, & \text{otherwise},\end{cases} \tag{9}$$

where $\{cx,cy,w,h\}$ represents the abscissa and ordinate of the center point and the width and height of the border, respectively. Border regression means that the prior border obtained by nonmaximal suppression is adjusted by gradient descent to obtain the final predicted border.

4. Experimental Design

4.1. Data Set. Common pedestrian detection data sets include MIT, KITTI, INRIA, Caltech, and CityPersons,. In this paper, INRIA and Caltech data sets are used for experiments. INRIA pedestrian detection data set includes images of different scenes of various sizes, mainly from long-term network collection, with a small amount of data [26]. Caltech pedestrian detection data set mainly includes about 10 hours of video, with a video resolution of 640 × 480 and a frame rate of 30 Hz. The video is taken from vehicle dashcam or other camera equipment. The content of the video is about roads and pedestrians in the city. The tagged video contains about 250K frames, including 2300 pedestrians and 350K frames, and the tagged information includes border information and shielding information. The image of pedestrian in the data set is a positive sample, while the image of nonpedestrian is a negative sample. Table 2 shows the statistics for the two data sets.

4.2. Experimental Result. In order to verify the effectiveness of the algorithm, specific experiments are carried out in this paper. Faster R-CNN, SSD, and YOLO V3 were selected for the effect comparison on the test set. Common target detection evaluation criteria were adopted in the experiment: mean average precision (MAP), average recall (AR), and frame per second (FPS) [24].

In order to determine the appropriate primary frame aspect ratio, this paper carried out a set of comparative experiments, from 0.3 to 0.5 for 0.05 step; set up the primary frame aspect ratio of the experiment; the result is shown in Figure 8; the use of extreme aspect ratio (0.3) of the experimental results is poorer; an aspect ratio from 0.4 to 0.45 can obtain good effect; the subsequent experiments of USSD use a wide high percentage which is 0.45.

In this paper, according to the characteristics of pedestrian data, the generation mode of preselected border is improved, and two improved generation modes of preselected border are proposed. In order to verify the effectiveness of the method, a second group of comparative experiments is conducted. The experimental results are shown in Table 3. The method of producing preselected borders with precise dimensions based on pedestrian characteristics is beneficial, as is the method of reserving preselected borders with anomalous width/height ratios while maintaining the width/height ratio of the preselected border. Figure 9 shows the effect comparison of pedestrian detection using three methods of generating the edge of the preselected box. On the basis that the pedestrian can also be accurately located, the predicted border generated by the improved method proposed in this paper is more in line with the

TABLE 4: Experimental results of the generation mode of precheck box.

Model	AR	MAP
USSD+Figure 6	83.96	72.53
USSD+Figure 6	85.07	74.26
USSD+Figure 6	86.65	75.23

TABLE 5: Final experiment result.

Model	INRIA		Caltech		
	AR	MAP	AR	MAP	FPS
Faster R-CNN	82.75	73.4	80.26	67.22	55
VGG16+SSD	82.47	71.62	79.32	65.07	69
YOLO v3	85.94	74.57	81.14	69.10	81
VGG16+USSD	86.65	75.23	83.05	69.35	70
DarkNet53+USSD	86.45	74.60	82.61	69.02	86

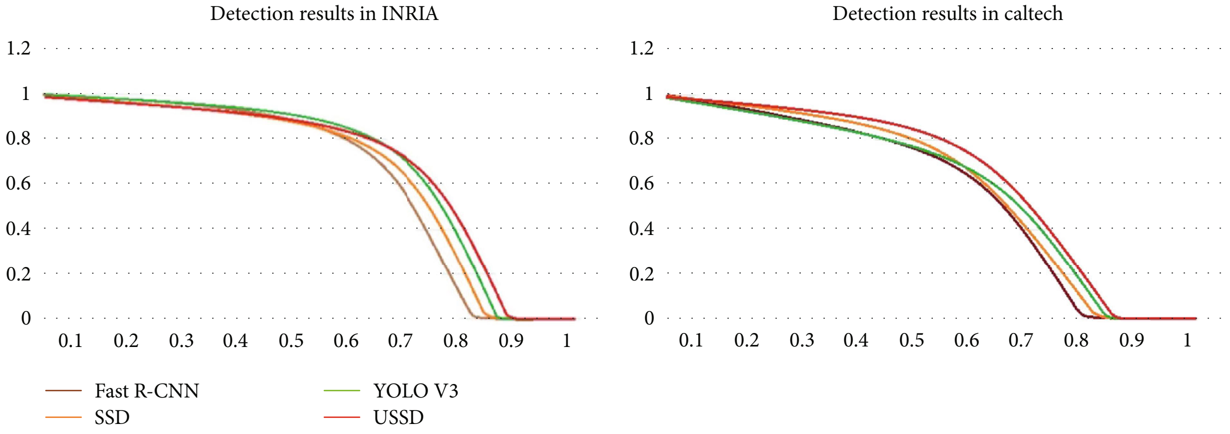


Figure 10: PR comparison chart of the model on the two data sets.

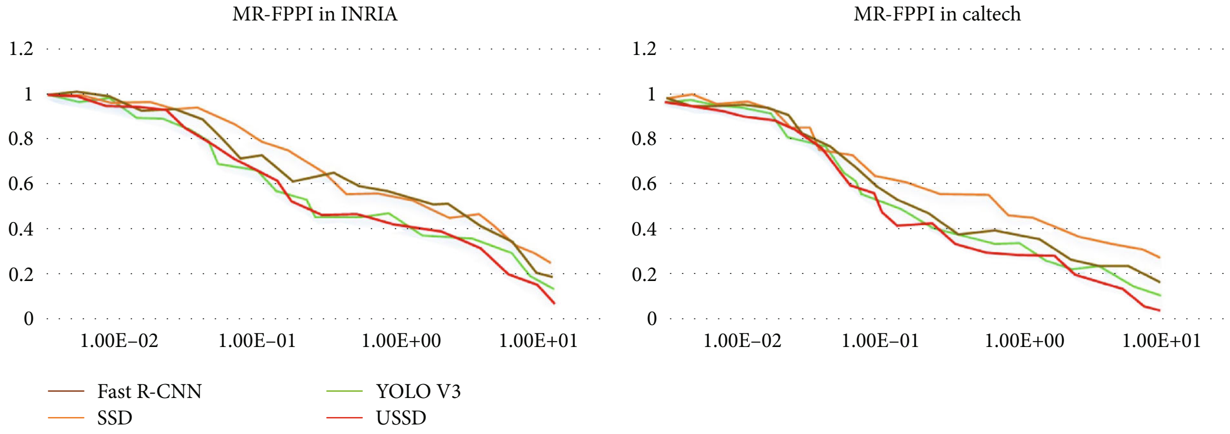


Figure 11: MR-FPPI comparison chart of the model on the two data sets.

pedestrian target. Table 4 shows experimental results of the generation mode of precheck box.

The final comparison experimental results are shown in Table 5. Figure 8 shows the precision-recall curve of the model on the two data sets. Compared with the detection effect of the SSD model and USSD model using VGG16 as the backbone network on the data set, the improved model USSD in this paper has better performance in AR and MAP indicators [27]. In order to improve the detection speed, USSD replaced the backbone network with DarkNet53 and used the two methods in Figure 5 to replace the general convolution and continuous convolution, and the detection speed and effect were good [26, 28]. The convolution calculation is a loss for the sparse certain classification performance at the expense of methods for calculating speed, but the pedestrian detection is a binary classification problem; difficulty is far less than the general classification of the target detection task; the simplified model also can adapt to the demand, so the complete Faster R-CNN and YOLO V3 model failed to make adequate classification advantage; and the detection speed is also weaker than for convolution model of USSD simplified in this paper. Finally, Figures 10 and 11 show the comparison of the model on the indicators of Miss Rate (MR) and False Positive per Image (FPPI). USSD has better performance.

5. Conclusions

The paper analyzes the method of unmanned target detection in complex scenes. In view of unmanned driving, pedestrian detection is the crucial procedure, and the target detection algorithm at this stage is not effective in detecting small and medium targets. This paper, SSD network model is applied to pedestrian detection, and the USSD model is proposed to improve the model according to data characteristics. In the process of generating precheck box, in order to obtain more spatial information from the feature MAP and refer to the multiscale feature combination method that combines classification and positioning in semantic segmentation, USSD transforms the deep features into the feature MAP with multiple resolutions through upsampling and combines with the shallow features extracted from the backbone network. Furthermore, this research purposely presets the dimensions and aspect ratio of preselected boundaries and recommends two presets, based on the features of pedestrians with similar aspect ratios. Finally, experiments were carried out on INRIA and Caltech pedestrian detection data sets, and the results were improved. Moreover, the simplification of convolution did not cause performance degradation but improved the detection speed, which could better meet the practical application requirements. The main difficulty of the previous research is to efficiently identify and detect small targets in a large background under uncertain conditions in harsh environments (occlusion, highly similar background and foreground, natural weather, etc.).

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no competing interest.

Acknowledgments

This manuscript was supported by the National Fund Cultivation Project (No. CZ2021GP08) and Chizhou University Natural Key Project (No. CZ2021ZRZ10).

References

[1] D. M. Martínez, F. Soriguera, and I. Pérez, "Autonomous driving: a bird's eye view," *IET Intelligent Transport Systems*, vol. 13, no. 4, pp. 563–579, 2019.

[2] M. Lu, Y. Wang, and G. Pan, "Generating fluent tubes in video synopsis[C]," in *2013 IEEE International Conference on IEEE, Acoustics Speech and Signal Processing (ICASSP)*, pp. 2292–2296, Vancouver, BC, Canada, 2013.

[3] T. W. Yan and H. Garciamolina, *SIFT: a tool for wide-area information dissemination[C]*, Usenix annual technical conference, New Orleans, 1995.

[4] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement[J]," https://arxiv.org/abs/1804.02767.

[5] J. Redmon and A. Farhadi, "YOLO9000: better faster stronger," 2016, https://arxiv.org/abs/1612.08242.

[6] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," 2018, https://arxiv.Org/abs/1804.02767.

[7] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, *Large kernel matters–improve semantic segmentation by global convolutional network[C]*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[8] I. Laptev and B. Caputo, *Recognizing human actions: a local SVM approach[C]*, International Conference on Pattern Recognition, UK, 2004.

[9] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class Ada Boost[J]," *Statistics & Its Interface*, vol. 2, no. 3, pp. 349–360, 2006.

[10] N. Dalal and B. Triggs, *Histograms of oriented gradients for human detection[C]*, International Conference on computer vision & Pattern Recognition. IEEE Computer Society, United States, 2005.

[11] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features[C]," *Computer Vision and Pattern Recognition*, vol. 1, 2001.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models[J]," *IEEE Transactions on Software Engineering*, vol. 32, no. 9, pp. 1627–1645, 2010.

[13] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

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

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[16] R. Girshick, "Fast R-CNN[C]," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[17] S. Ren, K. He, R. Girshick, and J. Sun, *Faster R-CNN: towards real-time object detection with region proposal networks[C]*, Advances in neural information processing systems, 2015.

[18] K. He, G. Gkioxari, P. Dollar et al., *Mask R-CNN[C]*, Proceedings of the IEEE international conference on computer vision, 2017.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, *You only look once: unified, real-time object detection[C]*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[20] W. Liu, D. Anguelov, D. Erhan et al., *SSD: single shot multibox detector*, Springer, Cham, 2016.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, *You only look once: unified real-time object detection*, Proc. IEEE Conf. Comput. Vis. Pattern Recogn., 2016.

[22] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention*, Springer, Cham, 2015.

[23] L. Q. Zuo, H. M. Sun, Q. C. Mao, R. Qi, and R. S. Jia, "Natural scene text recognition based on encoder-decoder framework," *IEEE Access*, vol. 7, pp. 62616–62623, 2019.

[24] Y. Freund and R. Schapire, *Experiments with a new boosting algorithm*, Proc. 13th Int'l Conf. Machine Learning, 1996.

[25] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, *Focal loss for dense object detection[C]*, Proceedings of the IEEE international conference on computer vision, 2017.

[26] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1907–1915, 2017.

[27] T. Y. Lin, M. Maire, S. Belongie et al., *Microsoft COCO: common objects in context*, Proc. Eur. Conf. Comput. Vis., 2014.

[28] H. Zhang, Y. Fu, L. Feng, Y. Zhang, and R. Hua, "Implementation of hybrid alignment algorithm for protein database search on the SW26010 many-core processor," *IEEE Access*, vol. 7, pp. 128054–128063, 2019.