Research on Blind Obstacle Ranging based on Improved YOLOv5

Yongquan Xia[1], Yiqing Li[2]*, Jianhua Dong[3], Shiyu Ma[4]
School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, Henan, China[1, 2, 3]
School of Economics and Management, Beijing Jiaotong University, Beijing, China[4]

***Abstract*—An improved model based on YOLOv5s is proposed for the problem that the YOLOv5 network model does not have high localization accuracy when detecting and identifying obstacles at different distances and sizes from the blind, which in turn leads to low accuracy in measuring distances. There are two main core ideas: firstly, a feature scale and a corresponding prediction head are added to YOLOv5 to improve the detection accuracy of small objects on blind paths. Secondly, SK attention mechanism is introduced in the feature fusion part. It can adaptively adjust the perceptual field for feature maps of different scales and more accurately extract objects of different distances and sizes on the blind path, which can improve the accuracy of detection and the accuracy of subsequent distance measurement. It was experimentally demonstrated that the improved YOLOv5 model improved the mAP by 6.29% compared to the original YOLOv5 model based on a small difference in time consumption. And for each category of AP values, the improvement ranged from 2.13% to 8.19%, respectively. The average accuracy of the measured distance from the obstacle at 1.5m to 3.5m from the camera is 98.20%. This shows that the improved YOLOv5 algorithm has good real-time performance and accuracy.**

Keywords—Binocular ranging; object detection; attention mechanism

I. Introduction

Blind corridors are one of the main guides for blind people to walk around. However, they are occupied by obstacles in almost every city. This causes inconvenience to the blind. Therefore, accurate identification and distance measurement of obstacles on blind paths is essential for the blind to travel safely.

In life and industrial scenarios, commonly used distance measurement methods include infrared distance measurement, stereo vision distance measurement, laser distance measurement and ultrasonic distance measurement. Among them, stereo vision distance measurement is a technology that allows computers to simulate human visual system to achieve distance measurement. The principle of the traditional stereo vision-based binocular ranging method is to use two cameras to acquire left and right images from the same viewpoint [1]. Then a stereo matching algorithm is used to find out the parallax. Finally, the distance from the target object to the camera is obtained by constructing similar triangles through the camera imaging principle. Stereo matching technology is an important part of binocular distance measurement. Before matching, it first extracts image features.

The mainstream stereo matching algorithms are GC (graph cuts) algorithm based on global feature matching [2], SGBM (semi-global block matching) algorithm based on semi-global feature matching [3], and SIFT(scale-invariant feature transform) algorithm [4], SURF(speeded up robust features) algorithm [5] and ORB(Oriented FAST and Rotated BRIEF) algorithm [6] based on local feature matching, etc. GC, SGBM algorithm can construct parallax images with high accuracy. However, the matching speed is slow, it is difficult to meet the requirements of high real-time, and the parallax image information is relatively simple. SIFT, SURF and ORB are three feature matching algorithms. Generally, feature points are used to match the left and right images from the same viewpoint, and then the distance is calculated according to the disparity of the feature points. The feature points that can be matched together are often those whose gray values change significantly and have similar trends. However, this often leads to the problem of mismatch, and the disadvantage of time-consuming detection of all feature points of the whole map (including uninterested areas).

The matching algorithms introduced above are always unsatisfactory in terms of speed or accuracy. Therefore, this paper uses a deep learning based method to detect the target from the left and right images of the binocular camera, and obtains parallax after getting the position information of the target. Finally, the distance from the target object to the camera is obtained by the binocular ranging principle.

In this research, the purpose is to improve the performance of target detection in blind channel images to solve the above problems. In addition, the speed of detection poses a great challenge to detection algorithm, because blind channel obstacle ranging is often real-time. The You Only Look Once (YOLO) neural network [7] is one-stage target detection algorithm that improves detection speed by making the target classification and localization set a one-level regression problem. YOLOv5 is the fifth version of YOLO, and its target detection performance is better than the first four versions, which is the most widely used version.

Although YOLOv5 can achieve multi-target detection quickly and accurately, it is difficult to apply directly to one specific image for target detection. Therefore, this paper proposes an improved model based on YOLOv5s for the detection and recognition of blind roadway pictures taken by binocular cameras. First of all, for the problem that the depth feature map after YOLOv5 feature fusion causes the loss of small object feature information due to excessive down-sampling, an up-sampling operation is added in the Neck part

*Corresponding Author.

to get a new feature scale for feature fusion with the shallow features of the Backbone network. And a corresponding prediction head is added in the head to improve the detection accuracy of small objects on blind corridors. Secondly, the SK attention mechanism is introduced after the last four feature fusions, which can adaptively adjust the sensing domain for the four different scales of feature maps. Thus, objects of different distances and sizes on the blind path can be extracted more accurately. The accurate recognition of blind obstacle and the accurate measurement of distance are achieved.

Author contributions can be summarized as follows:

- A new feature scale fusion layer is added in YOLOv5, which fuses shallow features in the network to maximize the retention of feature information and avoid the loss of small object features in the deep network.
- The attention mechanism Selective Kernel Networks (SKNet) is added to the feature fusion part in YOLOv5. It can adaptively adjust the perception field for feature maps of different scales, and it has different effects for objects of different distances and sizes on the blind path, so as to extract the target information more accurately.
- Mean Average Precision (mAP) of 0.843 is achieved with the model proposed in this paper on a homemade blind pathway dataset with complex objects. The improved YOLOv 5 model is used to detect and recognize binocular blind road pictures. After obtaining the position information of the same obstacle in the two pictures, the distance of the obstacle is obtained by using the principle of binocular ranging. At the obstacle distance from 1.5 m to 3.5 m from the camera, the average accuracy of measuring distance can reach 98.20%. It's very accurate.

The main part of this paper consists of five parts. Section I is the introduction, which introduces the background and significance of blind distance measurement for obstacle, as well as the research method of this paper and the reasons for adopting this method. Section II is the related work, which introduces the related research about target detection. Section III is the core part of the paper, which introduces the improved model of this paper, and how and why it was improved. In Section IV, the model proposed in the Section III was tested, and the experimental results and analysis were given. Section V is a summary of the overall research methods of this paper and the outlook for the future.

II. Related Work

A. Traditional Object Detection Algorithm

The traditional target detection algorithm [8] can be roughly divided into three steps: region selection, feature extraction and target classification. Region selection refers to first locating the target location. Then the images or image sequences are traversed by sliding windows of different scales and aspect ratios. Finally, all possible positions containing the detected target are framed by an exhaustive strategy. Feature extraction refers to the extraction of visual features of the target candidate regions. For example, SIFT (Scale Invariant Feature Transformation), HOG (Gradient Direction Histogram) [9], Haar [10] and LBP (Local Binary Pattern) [11] and other feature extraction operators are used for feature extraction. Object classification refers to the classification of the target by using the classifier through the extracted features. The commonly used classifiers are DPM (Deformed Part Model) [12], Adaboost [13], SVM (Support Vector Machine) [14] and so on.

Although traditional object detection methods have achieved certain results in the field of detection, they also have drawbacks. The high time complexity and many redundant windows in the region selection stage by sliding window selection candidate region strategy lead to the performance degradation of subsequent feature extraction and classification. Due to the object's own factors and environmental factors, the manual feature design method is poor in robustness, versatility and detection accuracy [15]. Traditional target detection methods have been difficult to meet people's pursuit of high performance.

B. Deep Learning based Object Detection Algorithms

In 2012, AlexNet [16] based on deep Convolutional Neural Network (CNN) won the ImageNet image recognition competition with a significant advantage. Since then, deep learning has been receiving widespread attention. Object detection is also gradually entered the era of deep learning.

The mainstream object detection algorithms [17] based on deep learning are classified into one-stage object detection algorithms and two-stage object detection algorithms according to whether exist a candidate frame generation stage. The two-stage target detection algorithm first extracts candidate frames for image targets, and then classifies the detection results based on the candidate frames. The representative algorithms have R-CNN [18], Fast R-CNN [19], Faster R-CNN [20] and R-FCN [21], etc. Single-stage target detection algorithms do not generate candidate frames and compute the detection results directly on the image. The classical single-stage target detection algorithms have SSD [22], Retina-Net [23] and YOLO series.

The YOLO algorithm was proposed by Redmon et al. in 2016, and there have been seven generations of iterations so far. YOLOv1 [7] is the first object detection algorithm based on regression analysis. It directly divides the image into regions and predicts the bounding box and probability of each region simultaneously. Its detection speed has been greatly increased. However, the detection accuracy is low, especially for small objects. YOLOv2 [24] built a new backbone network Darknet-19 on top of v1. Darknet-19 has fewer convolution layers, and its performance is higher. YOLOv2 maintains the advantage of fast inference while significantly improving prediction accuracy. However, the backbone network of YOLOv 2 is deeper and the recall rate of small target detection is low. And it is poor for dense group target detection. The underlying network of YOLOv3 [25] is Darknet-53, which borrows the residual structure of ResNet [26] for multiscale prediction. YOLOv3 has obviously improved the overall detection accuracy, detection accuracy for small objects, the inference speed and so on. YOLOv4 [27] uses CSPDarkNet53 [28] instead of DarkNet53 and SPP+PAN instead of FPN. The computation is further reduced and the detection accuracy is improved. In June 2020, Jocher proposed YOLOv5 [29]. Compared with YOLOv1~YOLOv4, YOLOv5 has small size and short inference time, and is suitable for edge devices with limited computational and storage resources. YOLOv5 contains 4 models, in descending order according to model depth and feature map width, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Among them, YOLOv5s was iterated to version 6.1 in March 2022.

III. Blind Obstacle Recognition based on Improved YOLOV5

A. *Improved YOLOV5*

The network model in this paper is improved based on YOLOv5s version 6.1.The improved part is shown in Fig. 1 at the red box line. Firstly, one up-sampling and one down-sampling operation were added in each Neck part. The up-sampling operation is continued after the second up-sampling to obtain a new feature scale for feature fusion with the shallow features of the Backbone network, and a corresponding prediction head is added to the head. The aim is to improve the detection accuracy of small objects on blind paths. Secondly, the SK attention mechanism is introduced after the last four feature fusions, which can adaptively adjust the perceptual field for the four different scales of feature maps. Thus, the objects of different distances and sizes on the blind corridor can be extracted more accurately.

B. *Improved Multi-scale Feature Detection*

The YOLOv5 network uses three resolutions of the output feature maps to detect objects of different sizes. When a blind person walks on a blind path, there are inevitably smaller obstacles on the road, so a feature scale to focus on smaller targets and a prediction head accordingly are added, as shown in the Head section in Fig. 1. After the second up-sampling in the Neck part, the up-sampling operation is continued to to obtain a feature map with a resolution of 160×160. This feature map is feature fused with the shallow high-resolution feature layer obtained in the backbone network after the first C3 structure. The obtained high-resolution feature information makes the added prediction head more sensitive to small objects.

Fig. 2 shows the prediction part of the model with four different scales of feature layers, and the grid-based prediction is performed on these four feature layers in turn. The prediction results of each feature layer are the center adjustment parameters t_x and t_y, the width and height adjustment parameters t_w and t_h, and the confidence score of the prior frame, respectively. The center (x, y) and width and height (w, h) of the predicted frame are obtained based on the first four adjustment parameters. Based on the ground truth of the object, the network establishes the loss between the predicted and true values and calculates the loss for each feature layer.

Through the feedback of the loss, the model gradually optimizes the performance and completes the training. The loss of each feature layer is calculated in the same way and is obtained by calculating the sum of the bounding box regression loss, category loss and confidence loss. As shown in (1).

$$\text{Loss} = \lambda_1 L_{\text{cls}} + \lambda_2 L_{obj} + \lambda_3 L_{\text{ciou}} \tag{1}$$

In equation (1), the rectangular box loss (L_{ciou}) is obtained using CIoU loss [30], and the classification loss (L_{cls}) and confidence loss (L_{obj}) are calculated by BCE (Binary Cross Entropy) loss. λ_1, λ_2 and λ_3 are balance coefficients.

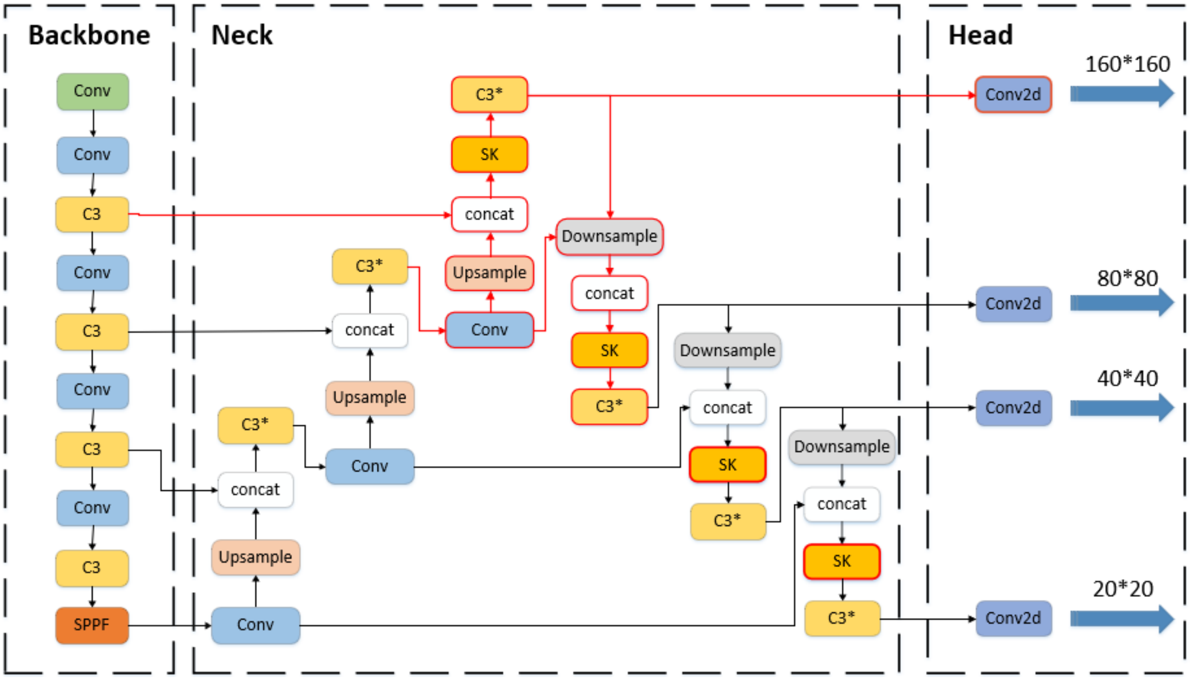

Fig. 1. Improved YOLOv5 Network Model.

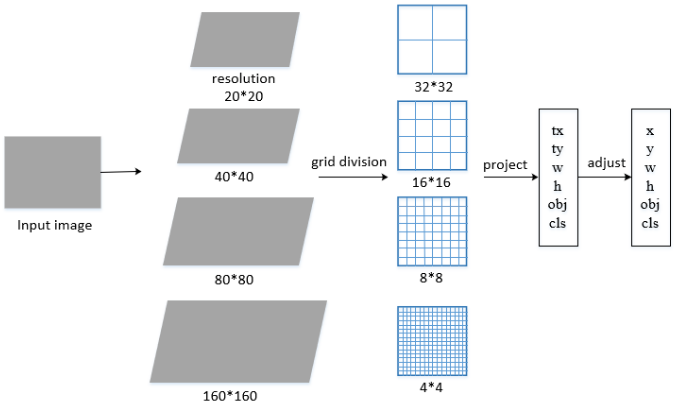

Fig. 2. Improved Model Prediction Section.

C. The SK Attention Mechanism Introduced

The size of the receptive fields of neurons in the visual cortex of the brain is adjusted according to the stimulus when people with normal vision look at objects of different distances and sizes. SKNet (Selective Kernel Networks) [31] is a kind of network with a similar function that adaptively adjusts the size of the receptive fields according to the multi-scales of the input information, so as to extract objects of different sizes and distances more accurately. Since the obstacles on the blind path are different in sizes and the distances, we added the SK attention mechanism to the four feature maps of different resolution sizes obtained after feature fusion in the Neck network. As shown in the Neck part in Fig. 2.

The structure of SKNet is shown in Fig. 3, it is composed of Split, Fuse, and Select. The part of Split obtains two feature maps $U1$ and $U2$ by convolution kernels of 3×3 and 5×5 respectively for the original feature map. Then the feature map U is obtained after the addition operation. The Fuse stage is to calculate each convolutional kernel weight. S is obtained using global average pooling(F_{gp}) for U with dimensionality$C \times 1$, and compact features Z is generated by fully connected layer(F_{fc}) with dimensionality compressed to be $d \times 1$. As in (2-4).

$$F_{gp}(U) = \frac{1}{W \times H}\sum_{i=1}^{H}\sum_{j=1}^{W} U(i,j) \tag{2}$$

$$s = F_{gp}(U) \tag{3}$$

$$z = F_{fc}(s) = \delta(BN(Ws)) \tag{4}$$

In (2), $W$and H are the width and height of the vector U, respectively. i and j are the i-th row and j-th column of U, respectively. In (4), δ is the ReLU activation function and BN denotes the batch normalization operation.

The Select part reclassifies the feature Z into two feature vectors a and b, which are the weights of the two convolution kernels, by a softmax operation (5).

$$a_c = \frac{e^{A_c^z}}{e^{A_c^z} + e^{B_c^z}}, \quad b_c = \frac{e^{B_c^z}}{e^{A_c^z} + e^{B_c^z}} \tag{5}$$

In equation (5), A_c,$B_c$are the cth row data of A and B. a_c,b_c are the cth elements of a,b respectively.

The final output feature map V is obtained by weighting the previous $U1$, $U2$ by the two weight matrices a and b. The following (6).

$$\mathbf{V}_c = a_c \cdot \mathrm{U1}_c + b_c \cdot \mathrm{U2}_c, \quad a_c + b_c = 1 \tag{6}$$

where $V = [V1, V2, \ldots, VC]$, Vc dimension is $H \times W$. Since the function values of a_c and b_c add up to 1, it is possible to set the weights for the feature maps in the branches.

Since the SKNet attention mechanism uses different convolutional kernel weights for different images, adding the SKNet network to the network to adaptively adjust the perceptual field for our four different resolution feature maps, which will have different effects for objects of different distances and sizes and improve the accuracy of detection.

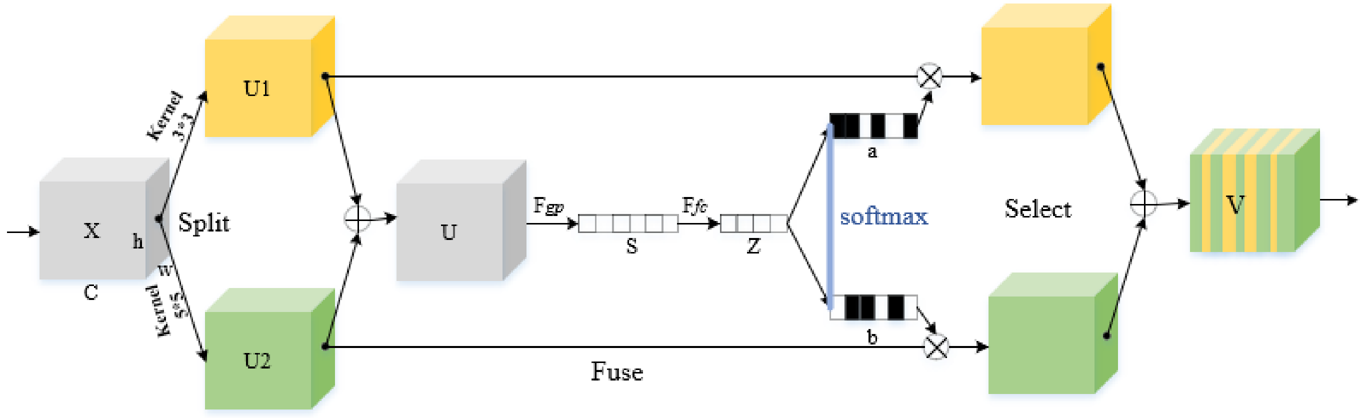

Fig. 3. SKNet Network Structure.

IV. Experiment and Analysis

A. Experimental Data

Since there is no authoritative dataset for blind roadway images, this research dataset consists of two main parts. One part is to use some images from The PASCAL Visual Object Classes Challenge 2007/2012 (VOC07+12) dataset, mainly selected images containing people, cats, dogs, and common vehicles. The other part is a manual field photography of the blind road surface and a browser search for keywords such as "blind road occupied" and "blind road obstruction" to get a web page with search results and download images. These two parts were filtered and aggregated to create a total dataset of 1245 images, each with a resolution between 1280×720 and 333×333. Some of the samples are shown in Fig. 4.

The dataset uses the PASCAL VOC format, and the targets in the images are selected by manual annotation of the frames. Eight types of common obstacles were selected as recognition objects, namely, people, bicycles, cars, motorcycles, cats, dogs, fire hydrants, and manhole covers. In addition, the dataset was randomly divided into training set, test set, and validation set according to the ratio of 6:2:2, where there were 747 images in the training set, 253 images in the test set, and 245 images in the validation set.

B. Model Training

Experimental environment configuration: The hardware environment is Intel(R) Core(TM) i7-6500 CPU with 2.50 GHz and 8 GB memory. The software environment is Windows 10 operating system, Python 3.6.15 development language, and PyTorch 1.2.0 deep learning framework.

Training parameters: For model training, the model is initialized using a pre-trained weight file on the COCO dataset, and the model input image size is 640×640. The training period is 100.Batch size is set to 16. The initial learning rate is 0.01. The minimum learning rate is 0.0001. The SGD optimizer is used for optimization. The momentum parameter is 0.937. The decay coefficient of the weights is 0.0005. The learning rate decreases by "cos". And the weights are saved every 10 The weights are saved every 10 epochs.

C. Performance Comparison Experiment

The improved network model predicts the images, and the results are shown in Fig. 5 and Fig. 6. Fig. 5 shows the image to be predicted, and Fig. 6 shows the prediction results. From the prediction results, it can be seen that the figure contains some objects that can be basically predicted accurately both in the strong light of day and in the weak light of street light at night. The rectangular box position accurately frames the outer edge of the target. For the partly obscured objects in the figure and some small objects at a distance can also be accurately detected.

To validate the performance of the proposed improved YOLOv5 network model, it is evaluated with homemade dataset. Precision, Recall, F1 Score and mean average precision (mAP) are used as evaluation metrics to assess the effectiveness of model training and prediction. The evaluation results are shown in Table I. From Table I, the Recall value, Precision value, and F1 score corresponding to each category of the improved YOLOv5s network model at Confidence of 0.5 are shown. The performance experiments were also compared with the original YOLOv5s network, the YOLO-FIRI network proposed in literature 32 [32] under homemade dataset. The accuracy and average elapsed time comparison results are shown in Table II.

The mAP values of the three algorithms and the AP values for each category can be seen in Table II. With Recall as the horizontal coordinate and Precision as the vertical coordinate, the curve plotted is the PR curve. The AP value is the area under the PR curve, which represents the average accuracy of a single category. It is used to measure how good the model is on a single category. The mAP is obtained by averaging the AP values of all categories, and is used to measure the goodness of the model in all categories. In terms of accuracy, this paper proposed YOLOv5s model improves the mAP by 6.29% compared to the original YOLOv5 model, and for each category of AP values, the improvement ranges from 2.13% to 8.19%, respectively. Compared to the YOLO-FIRI network, the mAP improved by 2.36%, and the AP values for each category ranged from 1.28% to 4.31%, respectively. This indicates that the improved network has better detection accuracy. The AP values for the three smaller object categories of cats, dogs and fire hydrants are 76.31%, 77.84% and 86.43%, which are 4.11%, 8.19% and 3.96% better than the original YOLOv5s model. This indicates that the improved YOLOv5 network model is more accurate in detecting small objects. In terms of inference time, the average time consumed by the improved model for each image is 12ms, which is not much different from the original YOLOv5s time consumed. Compared to the YOLO-FIRI model, the average elapsed time per image is 2ms faster. This shows that the improved model has a faster inference speed.

AS a whole, the improved YOLOv5s model has better accuracy compared to the original model for a single category versus all categories with comparable time consumption. This shows that the improved algorithm in this paper has good real-time performance and accuracy.

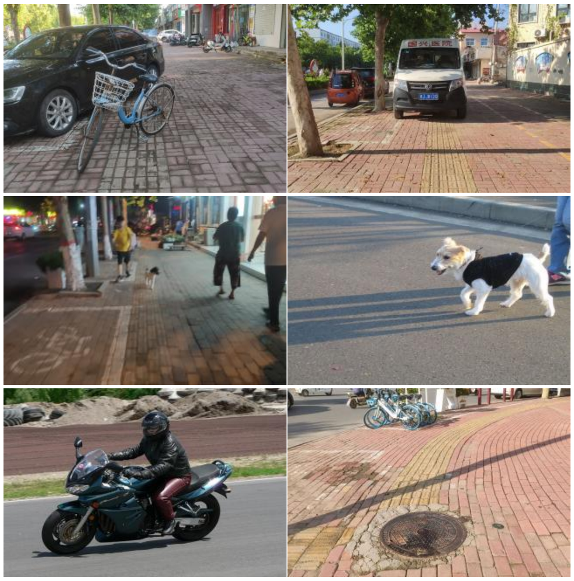

Fig. 4. Partial Sample.

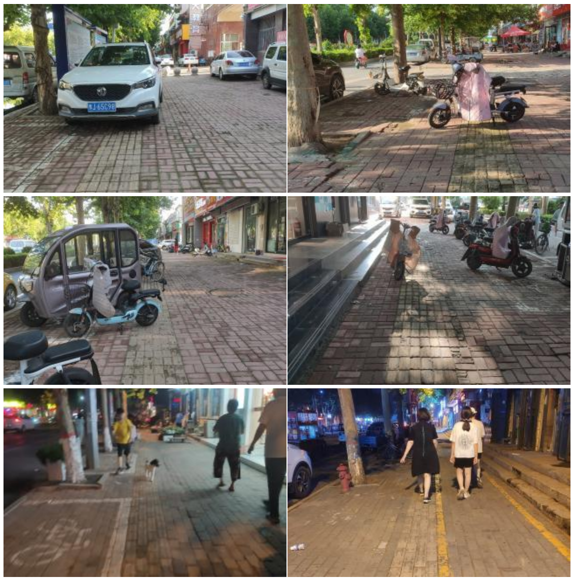

Fig. 5. Image to be Predicted.

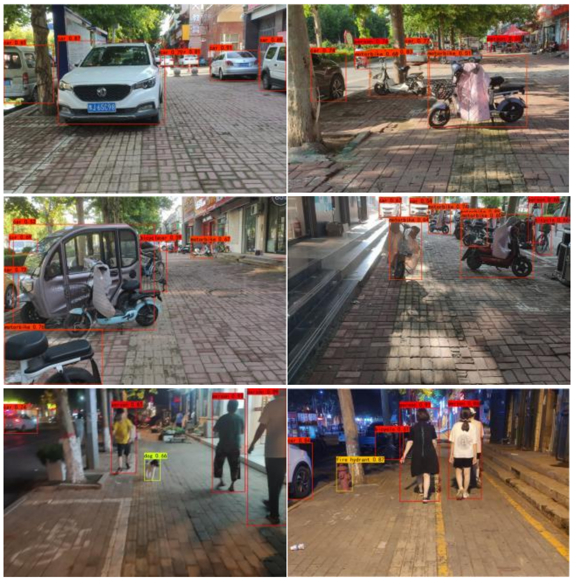

Fig. 6. Predicted Results.

TABLE I. EVALUATION RESULTS

Class	people	cat	dog	bicycle	fire hydrant	motorcycle	car	manhole cover
Precision/%	87.73	85.83	88.74	92.02	86.91	90.37	88.93	87.67
Recall/%	78.94	56.72	54.10	70.14	76.81	70.85	71.86	68.09
F1 Score	83.10	68.30	67.22	79.60	81.55	79.43	79.49	76.65

TABLE II. COMPARISON OF RESULTS

Algorithm	mAP/%	people	cat	dog	bicycle	Fire hydrant	motorcycle	car	Manhole cover	Time/ms
YOLOv5s	78.03	82.56	72.20	69.65	79.89	82.47	84.14	75.89	77.45	10.5
YOLO-FIRI	81.96	85.81	74.15	76.35	86.77	83.78	86.51	81.93	80.38	14
Ours	84.32	89.94	77.84	76.31	88.50	86.43	90.06	83.81	81.66	12

D. Distance Measurement Experiment

To verify the performance of the algorithm in ranging in this paper, this research used binocular cameras to capture pictures of obstacles at 1.5m, 2.0m, 2.5m, 3.0m and 3.5m at the same location. The binocular camera is calibrated with a known focal length of 2.13mm, a baseline length of 60.4mm, and an image element size of 3μm.

The prediction was performed after stereo correction of the left and right images at different distances (as shown in Fig. 7). The same nearest obstacle in the two pictures was screened by the center longitudinal coordinates of the outer rectangular box of the obstacle. And then the distance was obtained according to the difference of their center transverse coordinates using the principle of binocular ranging [33], as in (7-8). The experimental results are shown in Table III.

$$Z = \frac{fB}{d} \tag{7}$$

$$d = x_l - x_r = d_x \left(u_l - u_r \right) \tag{8}$$

In (8), u_l, u_r are the central pixel horizontal coordinates of the same object rectangular box on the left and right images. d_l, d_r are the central physical horizontal coordinates of the same object rectangular box on the left and right images. d_x is the image element size.

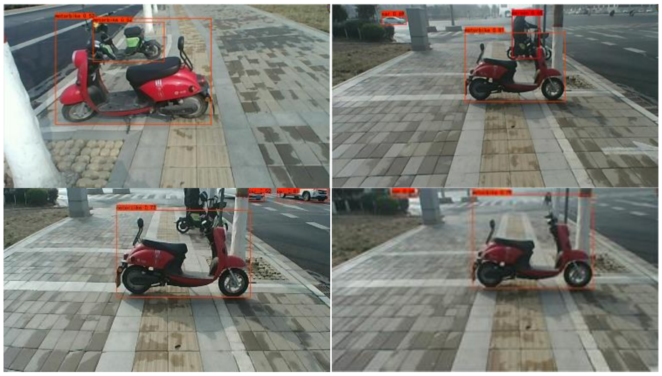

Fig. 7. Graph of Prediction Results for different Distances.

TABLE III. DISTANCE MEASUREMENT RESULTS

Actual distance /m	obstacle on the left/pixel	obstacle on the right /pixel	Parallax /pixel	Measuring distance /m	Accuracy /%
1.5	（475.5,159）	（447.5,158）	28	1.532	97.87
2.0	（542,228）	（521,226）	21	2.042	97.90
2.5	（721,255）	（705,255.5）	16	2.523	99.08
3.0	（790.5,209.5）	（776,208.5）	14.5	2.958	98.60
3.5	（716.5,246）	（704,246）	12.5	3.431	98.03

As can be seen from the data in Table III, the parallax, distance measurement results and accuracy of the nearest obstacle in the figure at a distance of 1.5m to 3.5m from the binocular camera. The measured distances obtained from parallax calculations were 1.532m, 2.042m, 2.523m, 2.958m, and 3.431m at actual distances from 1.5 m to 3.5m, respectively. The accuracy of the measured distance is above 97.87%, and the average accuracy is 98.20%. The accuracy at 2.0m and 2.5m reached 99.08% and 98.60%, respectively. The error was the smallest at the obstacle distance camera at 2.5 m, and the measured distance was the most accurate. This shows that the improved YOLOv5s algorithm has good accuracy and robustness in predicting the recognition of objects at different distances.

V. CONCLUSION

This paper proposes an improved model based on YOLOv5s for blind roadway picture detection. Firstly, a feature scale and corresponding prediction head are added in YOLOv5 to improve the detection accuracy of small objects on blind path. Secondly, SK attention mechanism is introduced in the feature fusion part to adaptively adjust the perceptual field for feature maps of different scales to more accurately extract objects of different distances and sizes on blind path. Finally, the improved network model is used to detect the left and right maps of the binocular camera, and the exact distance of the blind obstacle from the camera is obtained by the binocular ranging principle. The experimental results show that the improved YOLOv5s algorithm can achieve the accurate recognition of blind obstacle and the accurate distance measurement.

In the future, this research will achieve accurate blind path identification. Combining blind path identification with blind obstacle ranging proposed in this paper can form a complete navigation system for the blind. This will be of great help for blind person to travel.

ACKNOWLEDGMENT

This work was supported by the Key Scientific Research Project of Henan Universities. The project is entitled "Research on Stereo Vision Based Ranging Technology for Blind Navigation System". The project number is 22B510018.

REFERENCES

[1] Y. Zheng, P. Liu, L. Qian, S. Qin, X. Liu, Y. Ma, and G. Cheng, “Recognition and Depth Estimation of Ships Based on Binocular Stereo Vision,” *Journal of Marine Science and Engineering*, vol. 10, no. 8, pp. 1153, 2022.

[2] Y. Y. Boykov, and M. P. Jolly, “Interactive graph cuts for optimal boundary & region segmentation of objects in ND images,” *In Proceedings eighth IEEE international conference on computer vision. ICCV 2001 IEEE*, vol. 1, pp. 105-112, July 2001.

[3] H. Hirschmuller, “Stereo processing by semiglobal matching and mutual information,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328-341, 2007.

[4] W. Burger, and M. J. Burge, “Scale-invariant feature transform (SIFT),” *In Digital Image Processing Springer*, Cham, pp. 709-763, 2022.

[5] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features (SURF),” *Computer vision and image understanding*, vol. 110, no, 3, pp. 346-359, 2008.

[6] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” *In 2011 International conference on computer vision*, pp. 2564-2571, November 2011.

[7] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788, 2016.

[8] W. X. Dong, H. T. Liang, G. Z. Liu, Q. Q. Hu, and X. Yu, “A review of deep convolution applied to target detection algorithms,” *Computer Science and Exploration*, vol. 16, no. 5, pp. 1025, 2022.

[9] N. Dalal, and B. Triggs, “Histograms of oriented gradients for human detection,” *In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886-893, June 2005.

[10] R. Lienhart, and J. Maydt, “An extended set of haar-like features for rapid object detection,” *In Proceedings. international conference on image processing*, vol. 1, pp. I-I, September 2002.

[11] 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.

[12] 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.

[13] P. Wang, C. Shen, N. Barnes, and H. Zheng, “Fast and robust object detection using asymmetric totally corrective boosting,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 1, pp. 33-46, 2011.

[14] J. L. Balcazar, Y. Dai, and O. Watanabe, “Provably fast training algorithms for support vector machines,” *In Proceedings 2001 IEEE International Conference on Data Mining*, pp. 43-50, November 2001.

[15] D. C. Wang, C. S. Bai, and K. J. Wu, “A review of deep learning-based video target detection,” *Computer Science and Exploration*, vol 15, no. 9, pp. 1563-1577, 2021.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017.

[17] L. P. Fang, H. J. He, and G. M. Zhou, “A review of target detection algorithm research,” *Computer Engineering and Applications*, vol. 54, no. 13, pp. 11-18, 2018.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” *I2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587, 2014.

[19] R. Girshick, “Fast r-cnn,” *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440-1448, 2015.

[20] 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, pp. 1137-1149, 2017.

[21] J. Dai, Y. Li, K. He, and J. Sun, “R-fcn: Object detection via region-based fully convolutional networks,” *Advances in neural information processing systems*, vol. 0, pp. 379-387, 2016.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed and C. Y. Fu, et al. “SSD: Single shot multibox detector,” *In European conference on computer vision*, vol. 9905, pp. 21-37, 2016.

[23] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, “Focal Loss for Dense Object Detection,” *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327, 2020.

[24] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517-6525, 2017.

[25] J. Redmon, A. Farhadi, “YOLOv3: an incremental improvement,” unpulished, 2018.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.

[27] A. Bochkovskiy, C Y. Wang, H Y M. Liao,“YOLOv4: optimal speed and accuracy of object detection,” unpublished, 2020.

[28] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, “Scaled-yolov4: Scaling cross stage partial network,” *In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition,* pp. 13029-13038, 2021.

[29] G. Jocher, “YOLOv5[EB/OL],” unpulished, 2021.

[30] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, “Distance-IoU loss: Faster and better learning for bounding box regression,” *In Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, pp. 12993-13000, April, 2020.

[31] X. Li, W. Wang, X. Hu and J. Yang, “Selective Kernel Networks,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 510-519, 2019.

[32] S. Li, Y. Li, Y. Li, M. Li and X. Xu, “YOLO-FIRI: Improved YOLOv5 for Infrared Image Object Detection,” *in IEEE Access*, vol. 9, pp. 141861-141875, 2021.

[33] Z. Liu and T. Chen, “Distance Measurement System Based on Binocular Stereo Vision,” *2009 International Joint Conference on Artificial Intelligence*, pp. 456-459, 2009.