A Monocular vision positioning and tracking system based on deep neural network

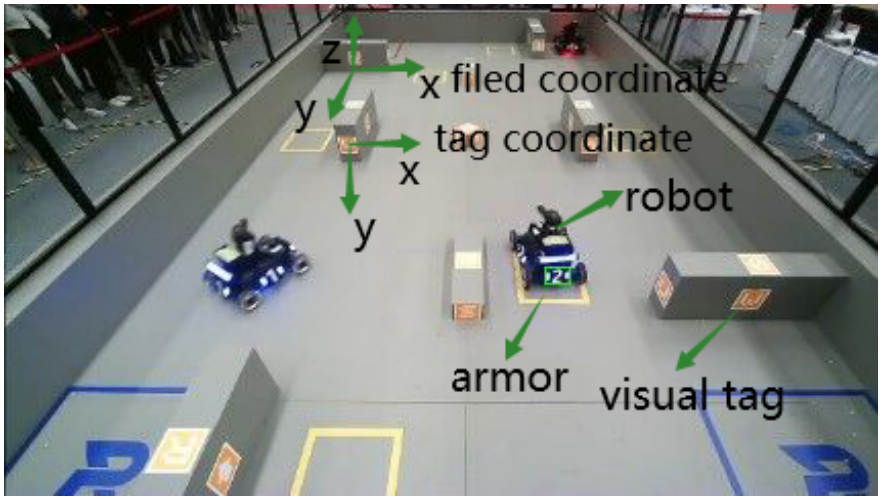
Huijun Li[1], Yu Zhang[1], Bin Ye[1], and Hailong Zhao[1]

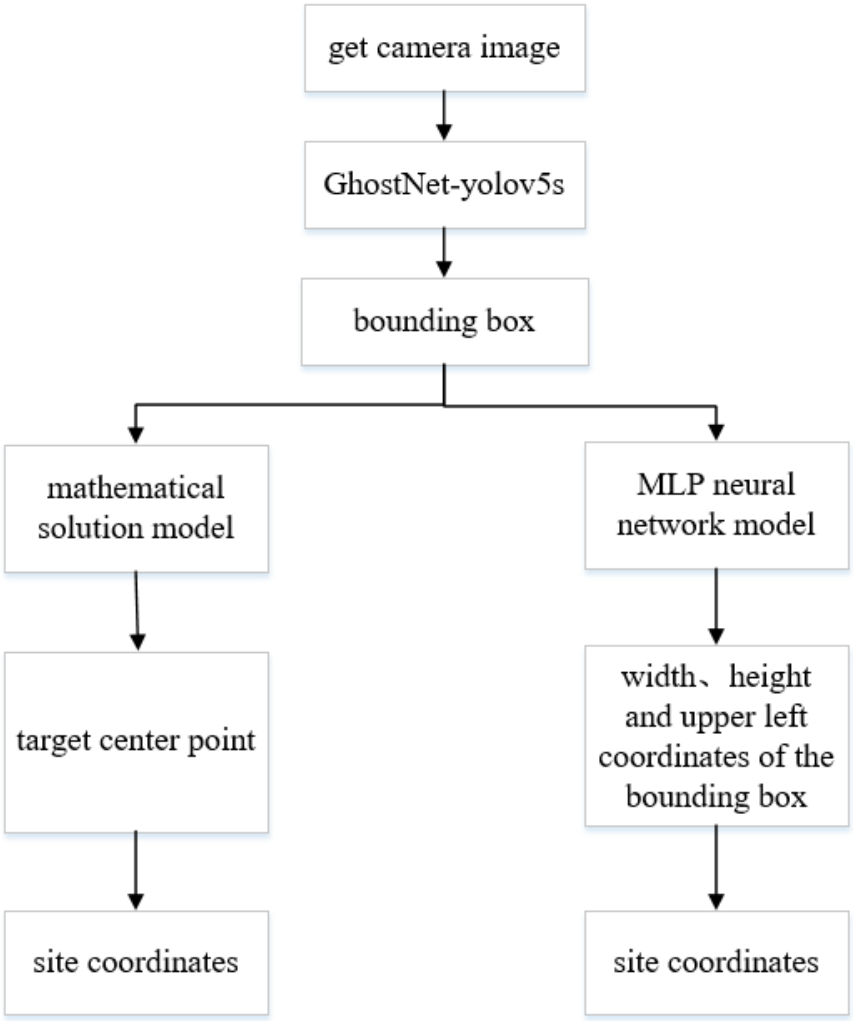
[1]China University of Mining and Technology

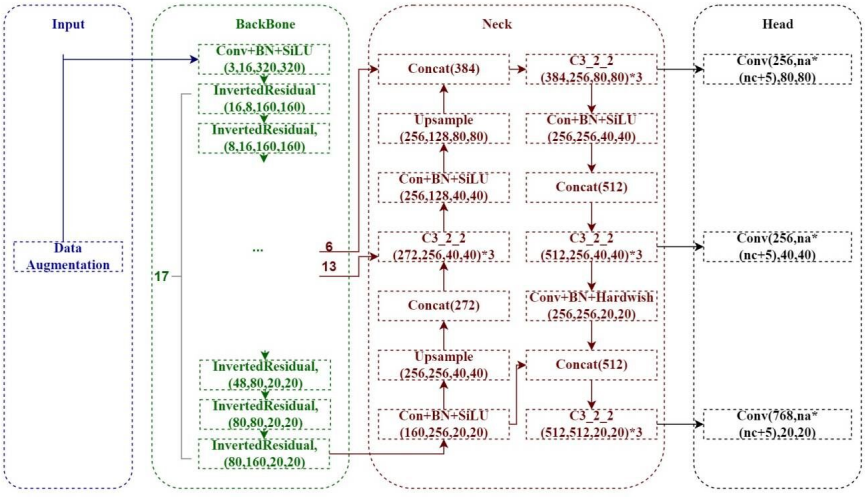
November 4, 2022

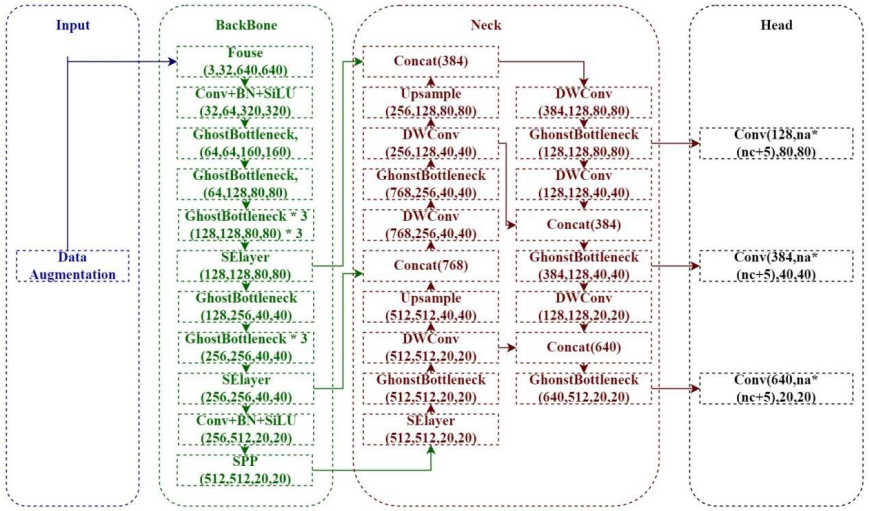
Abstract

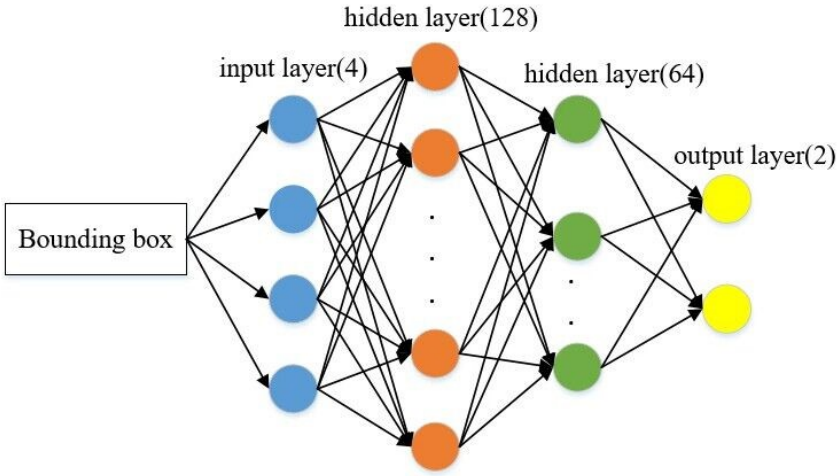
In order to locate the mobile robots in three-dimensional indoor environment, mostly global navigation satellite system-denied space, a monocular visual space positioning algorithm based on deep neural network is proposed. First, we employ the lightweight YOLOv5 algorithm for target detection, and the LibTorch deep learning framework is used for model deployment to improve the inference speed. Moreover, a multi-layer perceptron (MLP) neural network with four inputs and two outputs is constructed, which regress the coordinates of the robot in the field coordinate system to complete the target localization, and this method is compared with the mathematical model solving algorithm to reflect the accuracy and superiority of positioning algorithm based on deep neural network. The proposed positioning and tracking system has been successfully applied to ICRA robot competition, and results show that the positioning error estimated by our method is within 10cm whilst having good real-time performance.

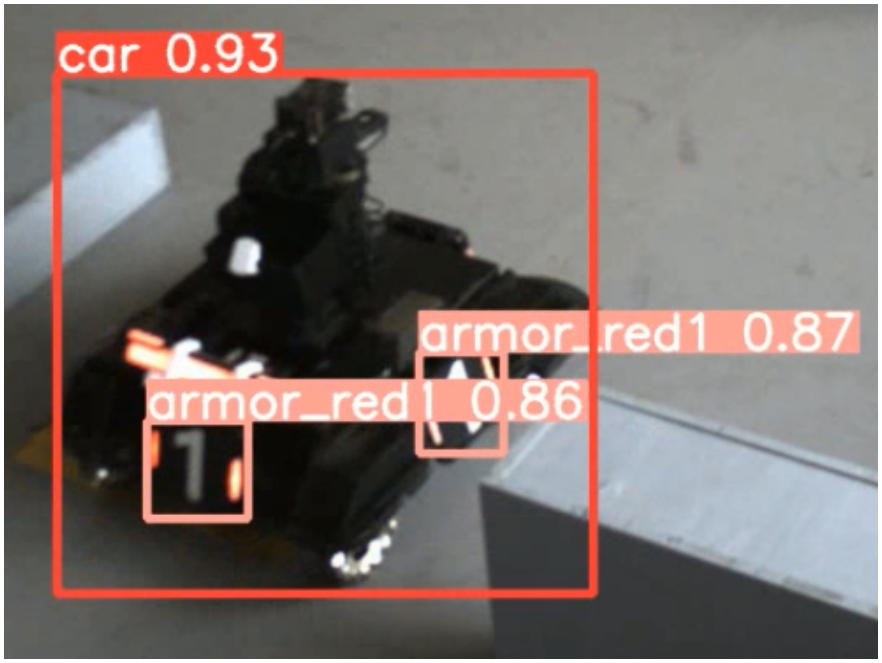


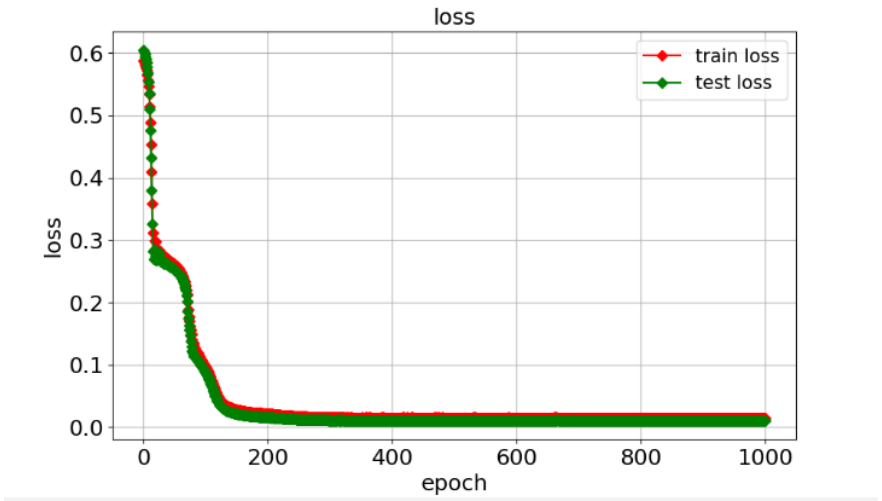


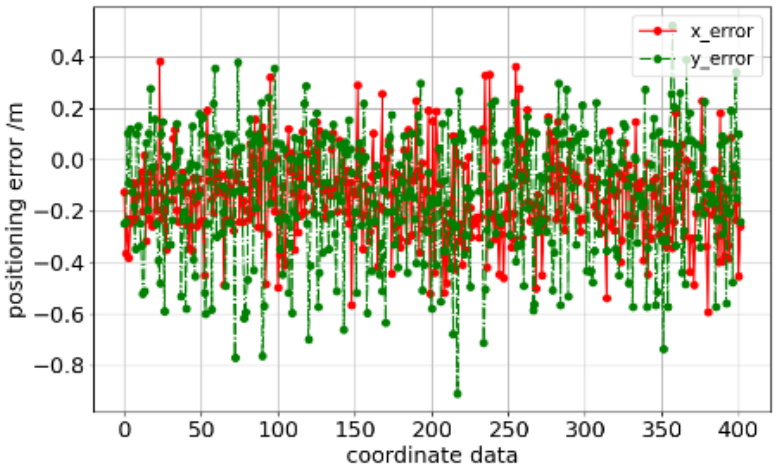


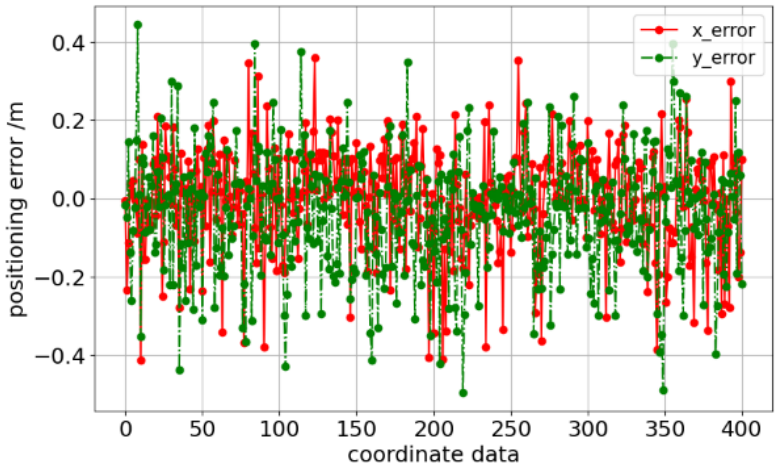


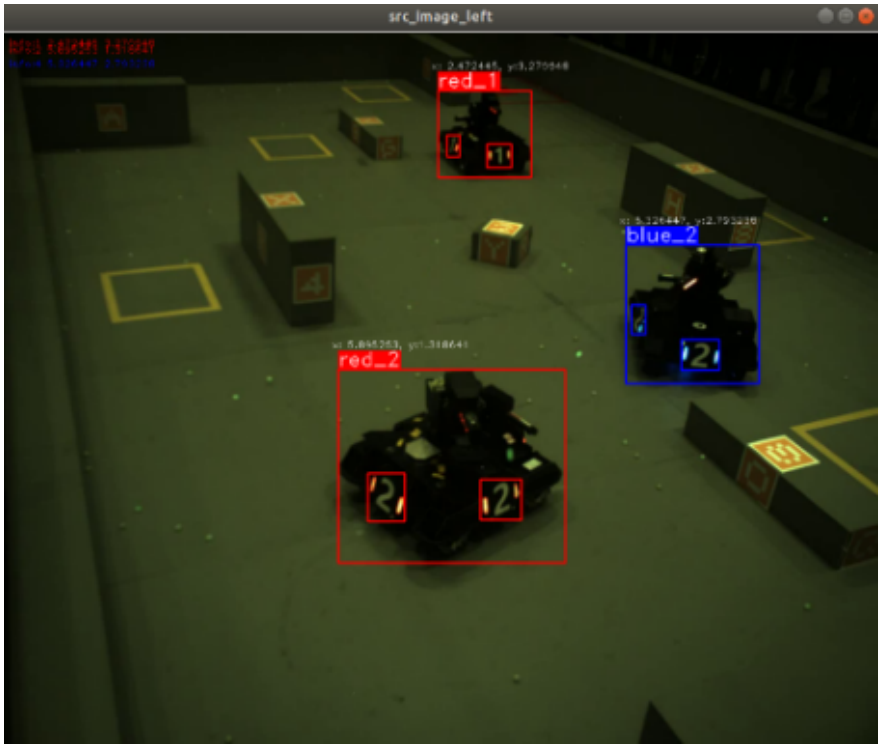


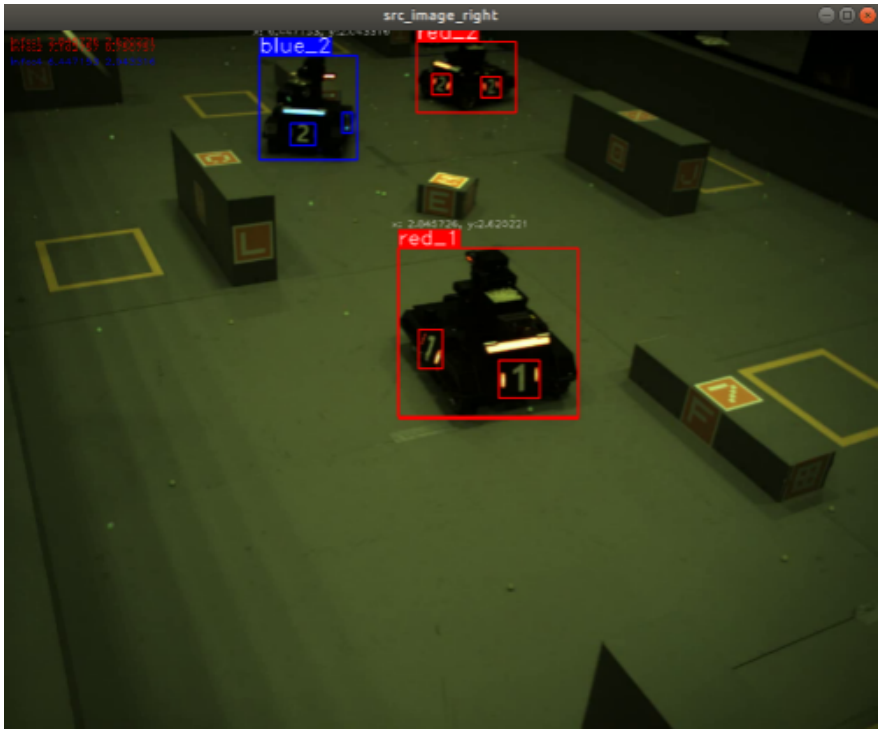


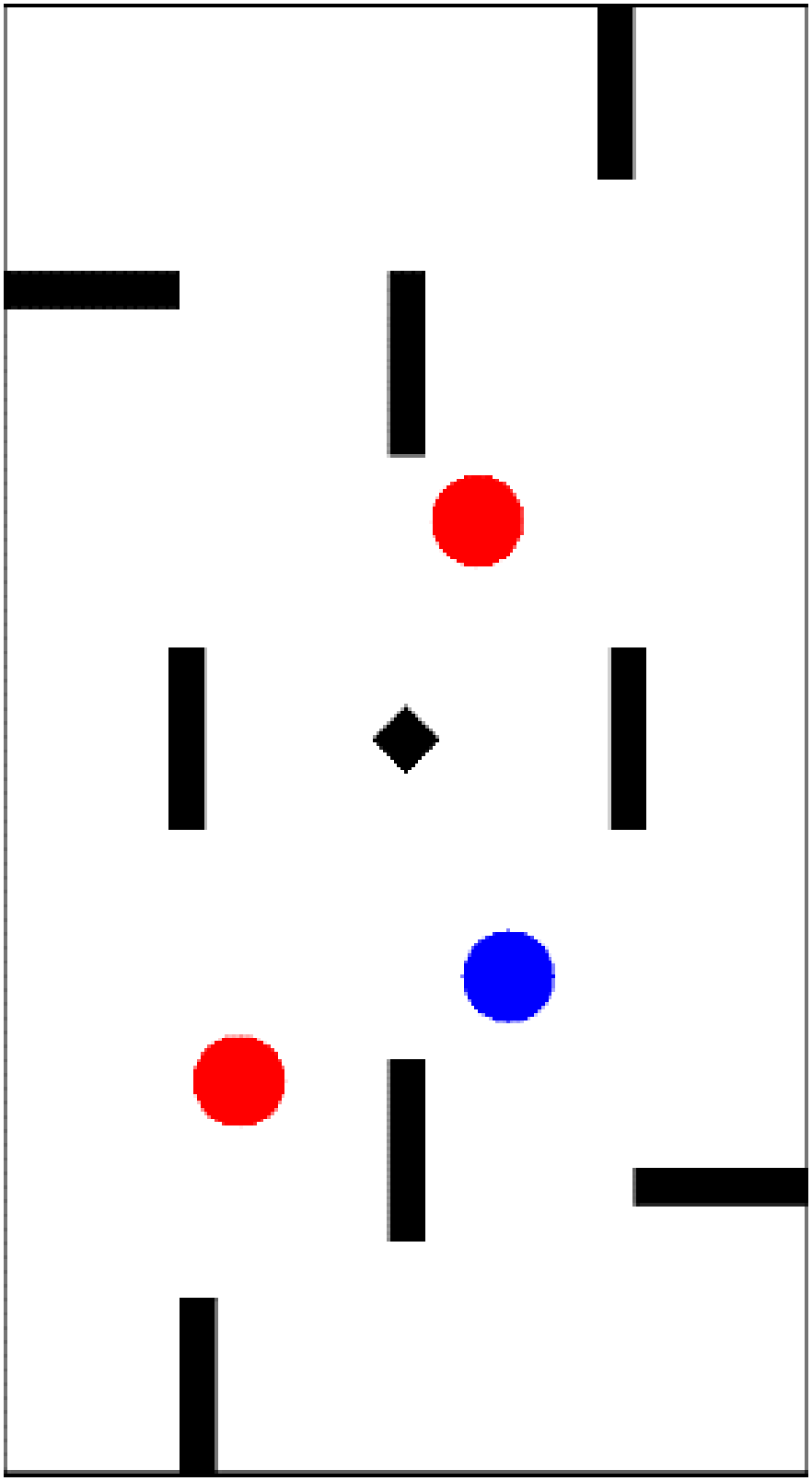












A Monocular vision positioning and tracking system based on deep neural network

Huijun Li, Yu Zhang, Bin Ye, Hailong Zhao
School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, 221116, China

Corresponding author: Yu Zhang (e-mail: TS20060105A31@ cumt.edu).

ABSTRACT In order to locate the mobile robots in three-dimensional indoor environment, mostly global navigation satellite system-denied space, a monocular visual space positioning algorithm based on deep neural network is proposed. First, we employ the lightweight YOLOv5 algorithm for target detection, and the LibTorch deep learning framework is used for model deployment to improve the inference speed. Moreover, a multi-layer perceptron (MLP) neural network with four inputs and two outputs is constructed, which regress the coordinates of the robot in the field coordinate system to complete the target localization, and this method is compared with the mathematical model solving algorithm to reflect the accuracy and superiority of positioning algorithm based on deep neural network. The proposed positioning and tracking system has been successfully applied to ICRA robot competition, and results show that the positioning error estimated by our method is within 10cm whilst having good real-time performance.

Keywords: YOLOv5; Neural networks; LibTorch; Target localization; Tracking

1. INTRODUCTION

Spatial positioning technology is one of the important research hotspots in the field of computer vision, which has been widely applied in autonomous driving, mobile robots, aerospace and other fields. Among the vision-based positioning methods, monocular visual positioning technology can only use one camera to complete the positioning work. Compared with multi-ocular visual positioning, it has the advantages of simple structure, high stability and good real-time performance [1]. Therefore, it is of practical significance and application value to carry out research on localization technology based on monocular vision.

Most of the traditional monocular vision positioning methods are based on the geometric principle. By establishing an appropriate model and combining the imaging process of the camera, the object's position and attitude can be solved. The most classic approach is the n-point Perspective solution (Perspective-n-Point, PnP) [2]. It used the projection relationship between n feature points in the image and their corresponding spatial points to determine the pose and position of the target relative to the camera, which is widely used in the field of computer vision. However, PnP solution has some disadvantages such as low accuracy and poor stability. Zhi et al. [3] proposed a real-time image registration and the target localization algorithm. This algorithm using improved ORB (Oriented FAST and Rotated BRIEF) to extract features, and use RANSAC algorithm to achieve precise matching and the transformation of the model parameters, CUDA is used to improve the real-time performance of localization, but the algorithm is highly dependent on the matching accuracy. Monocular visual localization technology is also widely used in SLAM (Simultaneous Localization and Mapping). For example, MonoSLAM algorithm, LSD-SLAM algorithm, ORB-SLAM algorithm and their improved algorithms have achieved good accuracy and effect in monocular visual positioning system. Liu et al. [4] proposed an algorithm combining object detection with ORB-SLAM2. It applies YOLOv4 target detection to the global mapping process, and uses ORB-SLAM2 for global mapping to determine the position and pose of the object in the world coordinate system, which provides more effective information for the positioning process. However, the internal parameters of the model established by this method are relatively large. For traditional methods, they require accurate mathematical models, which are relatively complex to solve. When the scenes are changing, the parameters need to be adjusted to meet new demands. At the same time, the adaptability and generalization ability of the traditional methods will be weak in the environment with many targets and complex background [5].

With the rapid development of deep learning, many localization algorithms based on neural networks have gradually emerged [6]. Taira et al. [7] proposed the InLoc method. It used multi-scale dense CNN features to achieve dense matching and performed pose verification through virtual view synthesis, which surpassing the highest level of indoor positioning accuracy at that time. However, this method still lacks enough information on pose selection.

Kendall et al. [8] proposed an end-to-end network PoseNet. A single picture could directly output the pose information after passing through this network, which could control the positioning error of indoor scenes within 0.5m. However, this algorithm could only predict the pose between frames, and its generalization ability and robustness were poor. In order to make full use of the information of sequence frames, Wang et al. [9] proposed the ESP-VO network model, which used deep recursive convolutional neural networks (RCNNs) to train and configure in an end-to-end manner. It could directly calculate the pose from a series of original RGB images, thus improving the positioning effect. However, the accuracy and robustness of this model need to be further improved. In addition, there are also localization methods that use neural networks for coordinate regression. In addition, Brachmann et al. [10] proposed a visual localization method based on neural network regression of scene coordinates DSAC++, which solved the visual localization problem and achieved high localization accuracy by learning the regression part of image scene coordinates and using local linearization to effectively optimize the poses, but the image information that can be utilized is limited. Sarlin et al. [11] proposed a scenario agnostic network called PixLoc. It learned data prior information from pixel to pose by end-to-end training, and optimized pose by separating model parameters and scene geometry. It could be used for accurate positioning of multiple scenes, but the algorithm needs good initialization and is susceptible to changes in perspective. Compared with traditional positioning methods, positioning algorithm based on neural network can obtain more accurate positioning and better generalization performance through end-to-end learning without establishing complex geometric models, which is a research hotspot of current visual positioning algorithms [12]. The above localization algorithms based on neural networks have achieved good localization results, but there are still some problems, such as it is difficult to achieve high positioning accuracy for dynamic objects, and the process of obtaining or labeling training data is cumbersome.

Based on the requirements of easy installation, high accuracy and fast speed of the sentry system in ICRA Artificial Intelligence Challenge, this paper proposes a monocular visual positioning and tracking system based on neural network. It uses the information obtained from target detection to establish MLP regression model to obtain the coordinates of the robot in the field, which can achieve better accuracy and speed for robot localization in the process of movement. The effectiveness and accuracy are verified in the robot competitions.

2. APPLICATION SCENARIOS

The ICRA Artificial Intelligence Challenge adopts the form of automatic shooting confrontation between the red and blue robots. Our robots need to sense the battlefield environment, find and hit the enemy robots' armor to win. In order to accurately obtain the enemy's position, the sentry system is required to provide the robot with the coordinate information of the opponent in the field.

The competition field is shown in Figure 1. The robot, armor area and visual tags are all marked in Figure 1. The sentry system consisting of cameras is fixedly placed on the brackets on both sides of the diagonal of the field to ensure that the whole field of vision can be captured.

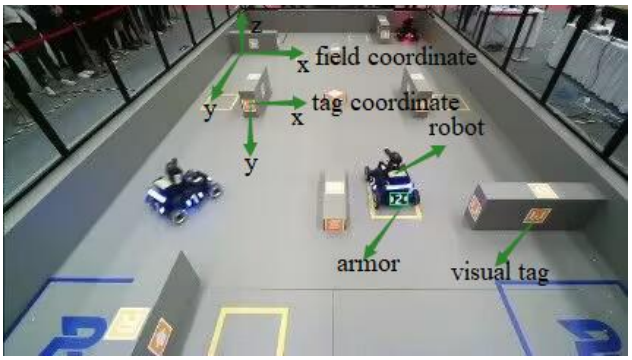


FIGURE 1. ICRA robot competition field

In order to ensure the real-time and accuracy of localization, the lightweight GhostNet-yolov5 algorithm is used in the target detection part, and the traditional mathematical solution method and MLP neural network algorithm are respectively used in the spatial localization part for two groups of comparison experiments. The overall architecture of the positioning system is shown in Figure 2.

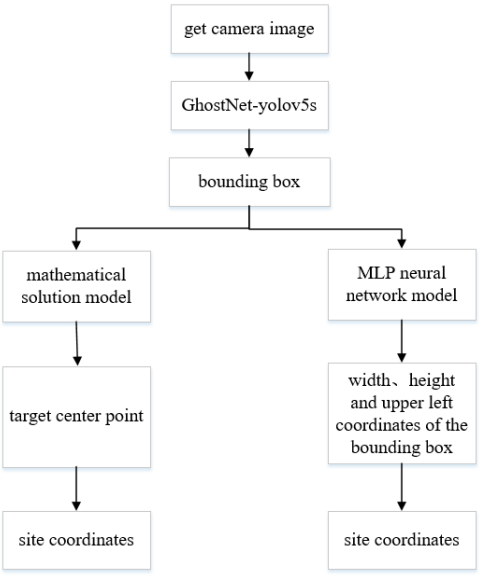


FIGURE 2. The overall architecture of positioning system

The image is predicted by Ghostnet-yolov5 model and the detection information of the robots can be obtained. The center point coordinates of the bounding box are used to solve the site coordinates of the target by mathematical solution algorithm. On the other hand, the coordinate value of the upper left corner, width and height of the bounding box are used as input of MLP neural network to regression the target's site coordinates. Through the comparison of the two methods, the superiority of the neural network algorithm

in positioning is verified and the coordinates of the robot in the field are determined.

3. YOLO v5 OBJECT DETECTION ALGORITHM

Object detection is a prerequisite for accurate spatial localization. Commonly used deep learning object detection algorithms include two-stage object detection algorithm (such as Faster R-CNN) and one-stage object detection algorithm (such as YOLO series). Compared with the two-stage algorithm, YOLO series can directly regression the category, confidence degree and location of the target through the network, with fast detection speed and strong advantages in model deployment. Since YOLOv1 was proposed in 2016, YOLO series has been continuously updated and improved [13-15], and has been widely used in intelligent transportation, defect detection, face recognition and other scenes. Therefore, this paper adopts the latest YOLOv5 for object detection, providing algorithm support for subsequent spatial localization.

The YOLOv5 algorithm inputs the entire image into the network, and directly returns the position and category of the bounding box at the output layer by meshing the image. The network structure of YOLOv5 is divided into four parts: input layer, Backbone, Neck network and prediction layer. The input layer uses Mosaic data enhancement, adaptive anchor calculation and adaptive scaling of the image to improve the speed and robustness of target detection. The Backbone part uses a series of convolutional neural networks, including normal convolution, Focus, BottleneckCSP and SPP structures to extract the features of the image and increase the acceptance range of the backbone features. The Neck network uses the FPN + PAN network structure to aggregate features, which strengthens the feature information of the network and pass it to the prediction layer [16]. The CIOU loss function and weighted NMS non-maximum suppression are used at the prediction layer to generate the output results and improve the detection accuracy.

In this paper, the structure of Yolov5s network is optimized, and two lightweight networks MobileNetV2 [17] and GhostNet [18] are respectively used as backbone networks to replace the original backbone for training and effect testing. MobileNetV2 follows the deepwise separable convolutions from MobileNetV1 network and introduces the inverted residual structure and linear bottleneck module. The inverted residual structure firstly uses 1×1 convolution to increase the dimension, then uses 3×3 deep convolution to extract features, and finally uses 1×1 point-by-point convolution to reduce the dimension, which strengthens the feature extraction ability of the network. The linear bottleneck layer uses linear convolution to replace the combination of original convolution and ReLU (Rectified Linear Unit) function [19], which helps to retain information. The MobileNetV2 network greatly reduces the computational cost while improving the accuracy. The core of GhostNet is Ghost module, which firstly uses ordinary convolution to obtain a small number of feature maps, then performs cheap operation to generate redundant feature maps, and finally uses concat operation to obtain feature maps of the same size as the original feature maps generated by these two steps. The SE attention module [20] is also added to GhostNet for better feature extraction. Compared with ordinary convolutional neural networks, the design of this module reduces the computation cost of network parameters.

Figure 3 shows the structure of the yolov5s-MobileNetV2 network. The original backbone network is replaced by 17 inverse residual modules from MobileNetV2. The original Focus structure is replaced by the standard convolutional module, which consists of a convolutional BN layer and a SiLU activation function. Figure 4 shows the structure of yolov5s-GhostNet network. The backbone network consists of Focus structure, GhostBottleNeck structure, SE attention module and SPP structure. The backbone network uses Ghost module and SE attention module (named GhostBottleNeck) instead of the original C3 module.

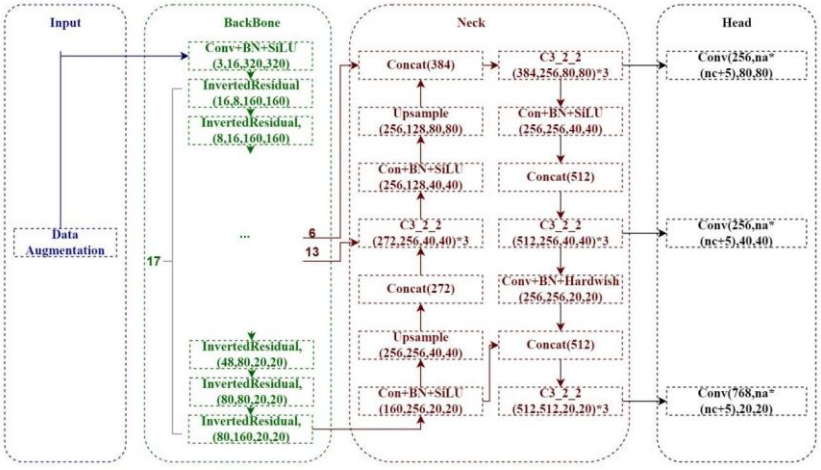


FIGURE 3. The structure of the yolov5s-MobileNetV2 network

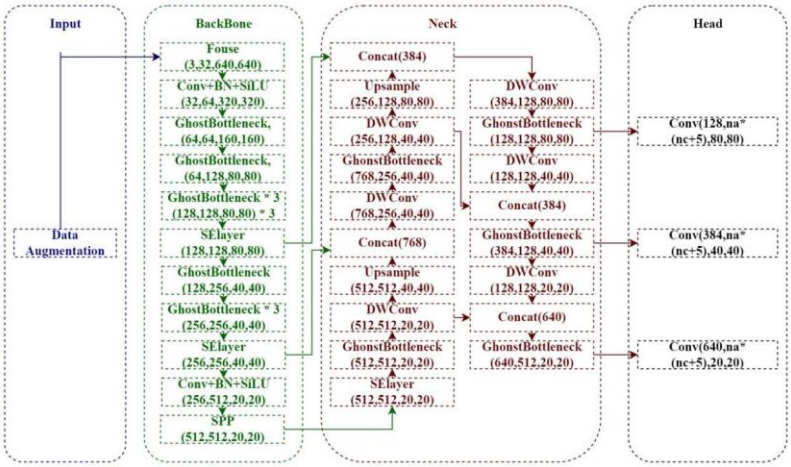


FIGURE 4. **The structure of the yolov5-GhostNet network**

4. OBJECT SPATIAL POSITION MODEL BASED ON MATHEMATICAL SOLUTION

Based on the information of visual labels in the competition field, this paper designs a mathematical solution method. In the competition, the center point of the bounding box obtained by the target detection is used as the known pixel coordinates, the four corner points of the visual tags in the field are used as the feature points, then according to the transformation relationship between the pixel coordinate system, the camera coordinate system, the tag coordinate system and the world coordinate system, the absolute pose relationship between the target and the camera can be calculated. The coordinate system conversion equations in this process are as follows:

1) Transformation of the pixel coordinate system to the camera coordinate system: let K be the camera internal matrix, $(u,v,1)$ is the pixel coordinate, f_x is using pixels to describe the length of the focal length along the x-axis, f_y is using pixels to describe the length of the focal length along the y-axis, c_x, c_y is the principal point coordinates, and (x_c, y_c, z_c) is the camera coordinate, by the formula (1):

$$z\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}f_x & 0 & c_x\\0 & f_y & c_y\\0 & 0 & 1\end{bmatrix}\begin{bmatrix}x_c\\y_c\\z_c\end{bmatrix}=K\begin{bmatrix}x_c\\y_c\\z_c\end{bmatrix} \quad (1)$$

From formula (1), we can obtain the formula (2):

$$\begin{bmatrix}x_c\\y_c\\z_c\end{bmatrix}=K^{-1}z\begin{bmatrix}u\\v\\1\end{bmatrix} \quad (2)$$

2) Let $(X_{tag}, Y_{tag}, Z_{tag})$ be the tag coordinates (the direction of the tag coordinate system is shown in Figure 1), and R_{tag}^{w}, t_{tag}^{w} be the rotation and translation from the label coordinate system to the camera coordinate system by the formula (3):

$$\begin{bmatrix}x_c\\y_c\\z_c\end{bmatrix}=R_{tag}^{c}\begin{bmatrix}X_{tag}\\Y_{tag}\\Z_{tag}\end{bmatrix}+t_{tag}^{c} \quad (3)$$

Next the formula (4) can be obtained:

$$\begin{bmatrix}X_{tag}\\Y_{tag}\\Z_{tag}\end{bmatrix}={R_{tag}^{c}}^{-1}\begin{bmatrix}x_c\\y_c\\z_c\end{bmatrix}-{R_{tag}^{c}}^{-1}t_{tag}^{c} \quad (4)$$

3) Transformation of the tag coordinate system to the world coordinate system: The rotation and translation of the visual tags in the field transformed to the site coordinate system are represented by R_{tag}^{w}, t_{tag}^{w} respectively, (X_w, Y_w, Z_w) is the world coordinate (the direction of the world coordinate system is shown in Figure 1), so we have the formula (5):

$$\begin{bmatrix}X_w\\Y_w\\Z_w\end{bmatrix}=R_{tag}^{w}\begin{bmatrix}X_{tag}\\Y_{tag}\\Z_{tag}\end{bmatrix}+t_{tag}^{w} \quad (5)$$

By substituting formula (4) and formula (2) into formula (5), we can obtain the formula (6):

$$\begin{bmatrix}X_w\\Y_w\\Z_w\end{bmatrix}=R_{tag}^{w}{R_{tag}^{c}}^{-1}K^{-1}z\begin{bmatrix}u\\v\\1\end{bmatrix}-R_{tag}^{w}{R_{tag}^{c}}^{-1}t_{tag}^{c}+t_{tag}^{w} \quad (6)$$

2) Calculate the unknown parameters z in formula (6): let $M=R_{tag}^{w}{R_{tag}^{c}}^{-1}K^{-1}\begin{bmatrix}u\\v\\1\end{bmatrix}$, $N=-R_{tag}^{w}{R_{tag}^{c}}^{-1}t_{tag}^{c}+t_{tag}^{w}$, denote the elements in the third row and first column of M、N and the world coordinates (X_w, Y_w, Z_w) as $M(3,1)$, $N(3,1)$ and z_w, then the formula (7) can be obtained by deformation:

$$z=(z_w+N(3,1))/M(3,1) \quad (7)$$

Considering that all robots in the competition field can be regarded to be moving on the same horizontal plane, so z_w can be treated as 0 in formula (7). Through the above mathematical solution method, the value of world coordinates can be obtained according to the transformation of each coordinate system.

5. SPATIAL LOCATION MODEL BASED ON MLP NEURAL NETWORK

In Section 4, we discussed how to use mathematical solution method to obtain the coordinates of the robot in the field. In this section, we designed an MLP neural network to achieve the location of the robot. The MLP neural network uses the information acquired after object detection to train the regression model, then obtains the site coordinates of the robot in an end-to-end way.

The MLP neural network is mainly composed of input layer, hidden layer and output layer. There can be multiple neurons in each hidden layer, and the different layers are fully connected. Its learning process consists of forward propagation of signal and backward propagation of error [21]. Weights, biases and activation functions are the three important elements of the network. The weights represent the strength and importance of the connections between neurons, the bias is to control the activation state of neurons, and the activation function is to make a nonlinear mapping between input and output. MLP neural network has strong nonlinear fitting ability and adaptive learning ability, which can deal with complex multi-input and multi-output nonlinear systems.

In neural networks, the number of nodes and layers of the hidden layer can fully regulate the neural network, and the use of activation function also affects the nonlinear fitting ability of the network, so it is very important to set the parameters of the hidden layer. According to the competition requirements, this experiment independently built a MLP neural network with four inputs and two outputs to train the regression model. The four parameters in the input layer are the upper left coordinates, width and height of the bounding box obtained after object detection, and the two parameters in the output layer are the coordinates of the robot in the field. In order to make the MLP network have good generalization performance, it is necessary to adjust the number of nodes and layers in the hidden layer during the design. Theoretically, the increase of the number of hidden layers can bring more complex computing power, but it may also cause the increase of computing time and the problem of overfitting. Therefore, the network chooses two hidden layers. The number of neurons in the hidden layer is usually adjusted by trial-and-error method, that is, the number of nodes in each layer is gradually increased from less to more until the accuracy of the model can't be improved any more. In general, a layer with more nodes followed by a layer with fewer nodes will have better performance.

According to the above principles, the final structure of the hidden layer is as follows: A total of two hidden layers were designed, with the number of 128, 64 respectively. In order to ensure the learning efficiency, ReLU activation function was used in both hidden layers, and Dropout was used to randomly drop some proportion of neuron nodes to prevent overfitting. The output layer uses Sigmoid activation function, and the specific MLP neural network structure is shown in the figure 5.

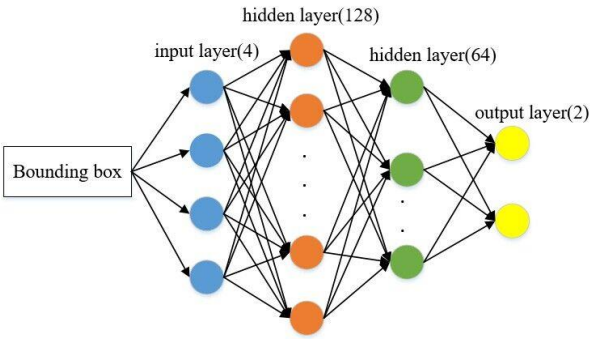


FIGURE 5. MLP neural network structure diagram

6. EXPERIMENTAL VERIFICATION AND ANALUSIS

6.1. EXPERIMENTAL ENVIRONMENT

The specific experimental environment is shown in Table 1.

TABLE 1. Experimental environment

Description	Parameters
Operating system	Ubuntu18.04
CPU	Intel Core i7-10875H
GPU	NVIDIA GeForce RTX 2060
Memory	32GB
Camera	Dahua Industrial Camera A5131CU210(Focal length: 6mm)
Deep learning framework	Pytorch1.10、CUDA10.2、cuDNN8.0
Python	Anaconda3

6.2. OBJECT DETECTION EXPERIMNET

6.2.1. DATA AND PROCESSING

In the target detection experiment, DJI COCO open source dataset is used to mix the locally collected dataset with a total of 5661 images, named as the CUMT-GUIDE dataset. The dataset is divided into six categories: survival robot, Red No.1 robot, Red No.2 robot, Blue No.1 robot, Blue No.2 robot and death robot.

The labels in PASCAL VOC format in DJI COCO open source dataset are converted to YOLO format, and the locally collected dataset is directly labeled to YOLO format by LabelImg annotation software. After integration, all data are divided into training set and test set in a ratio of 8:2.

6.2.2. EXPERIMENT AND ANALYSIS

In the experiment, the total training epochs is 1000, the batch size is 32, the image size is 640×640, and the learning rate is 0.001. Yolov5s is selected as the training model, and the backbone network structure of Yolov5s is replaced by MobileNetV2 and GhostNet respectively. Three sets of experiments were designed and compared. At the same time, in order to make full use of resources and speed up the inference speed, the three best models after training were deployed on the C++ project using the LibTorch framework,

and the frame rate is used to reflect the inference speed. The obtained experimental results are shown in Table 2.

TABLE 2. Experimental results of target detection

Model	mAP@0.5	mAP@0.5:0.95	P	R	Parameters	FPS
Yolov5s	0.873	0.628	0.897	0.773	14.8M	43
MobileNet-yolov5s	0.848	0.571	0.861	0.758	3.4M	47
GhostNet-yolov5s	0.865	0.607	0.876	0.716	6.6M	50

From Table 2, we can see that all three object detection models have obtained good detection results. Among them, after MobileNetV2 and GhostNet are introduced as the network's backbone, the model maintains high accuracy while the number of network parameters are reduced, and the complexity of the network is also decreased, which has the effect of lightweighting. Since the GhostNet-yolov5s model among the three has the advantages of higher detection accuracy, smaller parameter size, and faster inference, this model is finally chosen as the base model for target detection. Figure 6 shows the effect of real-time target detection.

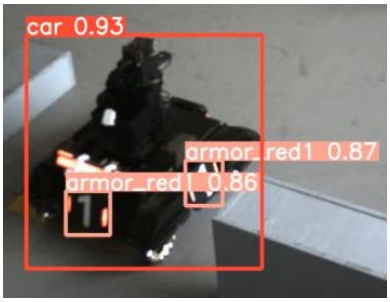


FIGURE 6. The result of object detection

6.3. SPATIAL POSITIONING EXPERIMNET

6.3.1. DATA AND PRE-PROCESSING

The bounding box information obtained after target detection is saved as the input data for the MLP regression model, and the coordinates obtained by the robot using the AMCL localization algorithm are used as labels. 3000 sets of data are collected and divided into 2600 sets of training data and 400 sets of test data.

Before training the regression model, good data pre-processing can play an important role in the training process and the performance of the neural network. For the training data in this experiment, since the input features and labels have different sources and units of measure, the input data and labels in the dataset are normalized to the same value interval [0,1]. For the input data, the two coordinate values of the bounding box as well as its width and height are divided by the image resolution of vertical and horizontal pixel values respectively; for the label data, divide the two actual coordinates by the width and height of the field respectively. This normalization operation reduces the amount of computation and accelerates the convergence of the model, which can achieve ideal results.

6.2.2. EXPERIMENT AND ANALYSIS

After data preprocessing, MLP regression model training can be carried out based on Pytorch1.10 deep learning framework. Epochs is set to 1000 and Adam optimizer is used, the learning rate is 0.001 and the loss function is SmoothL1 Loss. The convergence process of loss function is visualized as shown in Figure 7.

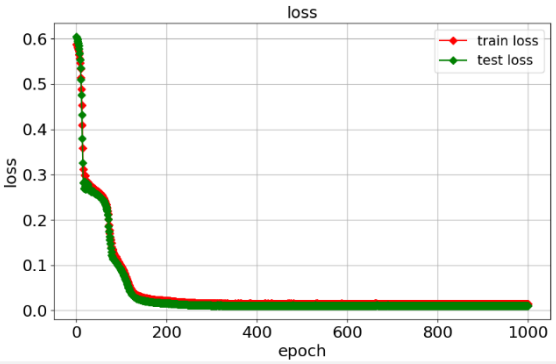
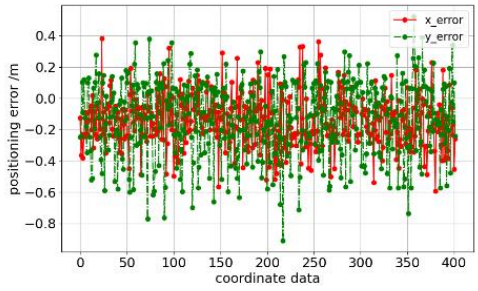
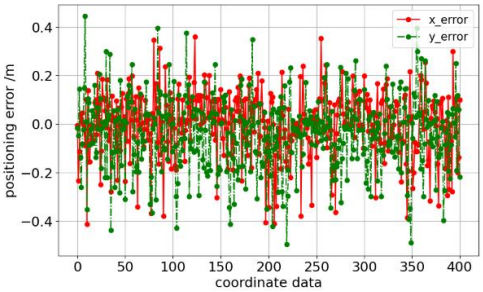


FIGURE 7. The change process of the loss function

Next, the data in the test set are tested and the errors in the x-direction and y-direction between the predicted and true values in the test set were visualized. In order to compare two methods, 400 sets of data are also collected for testing using the mathematical solution algorithm and the errors of the coordinate data are visualized. The final error visualization results of the two methods are shown in Figure 8.



(a) Mathematical solution error



(b) Neural network solution error

FIGURE 8. The error of the two models

From the above results, it can be seen that the error peak and error range of the neural network model are significantly smaller than those of the mathematical solution model, which initially reflects the superiority and stability of the neural network algorithm. On this basis, the mean error values of the test data in x and y directions are further compared, and the results are shown in Table 3.

TABLE 3. Mean error

Model	Error (X-direction)	Error (Y-direction)
Mathematical solution model	13.51cm	11.16cm
MLP neural network model	7.68cm	3.82cm

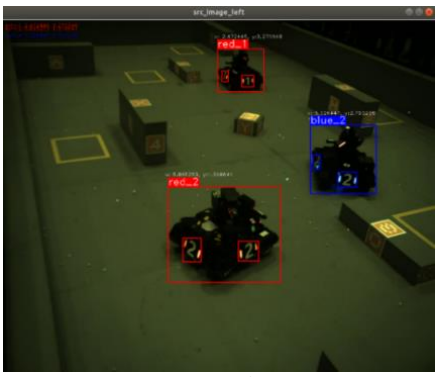
The analysis of the above data shows that the error of coordinate values obtained by the mathematical solution model is much larger than that obtained by the MLP regression model, which proves that the prediction performance of the neural network model is better. For the neural network model, the MLP reflects an end-to-end mapping relationship, so the source of error is mainly the input data obtained after the target detection. For the mathematical solution model, the solution process involves monocular camera calibration, PnP pose estimation, selection of visual tag's corners and other steps, so the sources of errors are more extensive, and the corresponding effect will become worse. Among them, in view of the error caused by the selection of visual tag's corners, the coordinates of four corners of five groups of visual tags are manually selected and mathematically solved, and the experimental results are shown in Table 4.

TABLE 4. The error of selecting different tag corner points

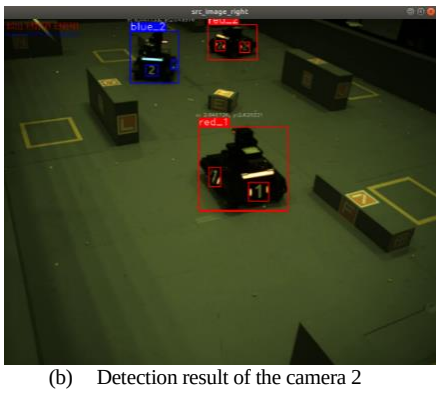
Corner coordinates of visual tag	Error(X-direction)	Error(X-direction)
[958, 341]、[1002, 348]、[993, 389]、[954, 380]	13.51cm	11.16cm
[960, 339]、[1001, 347]、[996, 390]、[955, 380]	12.32cm	11.2cm
[959, 338]、[1001, 347]、[996, 390]、[955, 380]	11.51cm	10.14cm
[960, 339]、[1001, 349]、[996, 389]、[955, 379]	12.45cm	12.16cm
[958, 340]、[1002, 348]、[993, 389]、[954, 378]	14.67cm	11.45cm

From Table 4, it can be seen that the selection of different tag corner points does bring some error to the mathematical solution algorithm, which affects the accuracy of this method. This error is larger than that of the neural network model, which also proves the accuracy and superiority of the neural network model for robot positioning in this experiment.

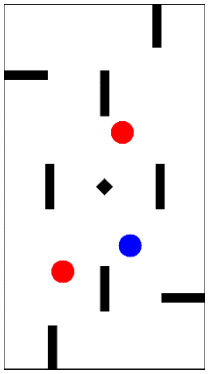
Finally, the MLP neural network model is deployed by LibTorch, and the coordinates predicted by the model are back-normalized to the original order of magnitude. The detection results of two monocular cameras located on both sides of the diagonal of the field are fused to obtain the coordinates of the robot. The results are shown in Figure 9. In the perspective of Figure (a), the origin of the site coordinate system is located at the upper left corner of the field. For the Red No.1 robot, the coordinates detected by the camera in the lower left corner of the field (denoted as camera 1, the image is shown in Figure (a)) are (2.47m, 3.07m), and the coordinates detected by the camera in the upper right corner of the field (denoted as camera 2, the image is shown in Figure (b)) are (5.57m, 2.42m). Then the detection results of the two cameras are fused, and the fused coordinates (2.49 m, 2.85 m) are displayed in the simulation map. It can be seen that the position of the red No.1 robot in the simulated field is almost the same as the actual field, as well as the positions of the other two robots, thus verifying the accuracy of the positioning and tracking system based on neural network. In the real-time tracking process of the robot, the positioning can also get a good effect, which provides a good help for the robot in the competition.



(a) Detection result of the camera 1



(b) Detection result of the camera 2



(c) Site simulation map

FIGURE 9. The results of fusion target detection

7. CONCLUSION

A monocular visual positioning and tracking system based on neural network is proposed. Firstly, the light-weight network structure GhostNet is used to replace backbone of YOLOv5, and then target detection is carried out based on the improved GhostNet-Yolov5 algorithm. Next, MLP neural network was used to establish a regression model to predict the coordinates of the robot in the site according to the information of the bounding box. The results of ICRA AI Challenge in 2021 show that the proposed method is reliable and practical.

REFERENCES

[1] Liang Y, He Y, Yang J, et al. An Efficient Vehicle Localization Method by Using Monocular Vision[J]. Electronics, 2021, 10(24): 3092.

[2] Lu X X. A review of solutions for perspective-n-point problem in camera pose estimation[C]. Journal of Physics: Conference Series. IOP Publishing, 2018, 1087(5): 052009.

[3] Zhi X, Yan J, Hang Y, et al. Realization of CUDA-based real-time registration and target localization for high-resolution video images[J]. Journal of Real-Time Image Processing, 2019, 16(4): 1025-1036.

[4] Liu Y, Xu M, Jiang G, et al. Target localization in local dense mapping using RGBD SLAM and object detection[J]. Concurrency and Computation: Practice and Experience, 2022, 34(4): e6655.

[5] Kim M, Kim J, Jung M, et al. Towards monocular vision-based autonomous flight through deep reinforcement learning[J]. Expert Systems with Applications, 2022, 198: 116742.

[6] Wang R, Wan W, Di K, et al. A high-accuracy indoor-positioning method with auto mated RGB-D image database construction[J]. Remote Sensing, 2019, 11(21): 2572.

[7] Taira H, Okutomi M, Sattler T, et al. InLoc: Indoor visual localization with dense matching and view synthesis[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7199-7209.

[8] Kendall A, Grimes M, Cipolla R. Posenet: A convolutional network for real-time 6-dof camera relocalization[C]. Proceedings of the IEEE international conference on computer vision. 2015: 2938-2946.

[9] Wang S, Clark R, Wen H, et al. End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks[J]. The International Journal of Robotics Research, 2018, 37(4-5): 513-542.

[10] Brachmann E, Rother C. Learning less is more-6d camera localization via 3d surface regression[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4654-4662.

[11] Sarlin P E, Unagar A, Larsson M, et al. Back to the feature: Learning robust camera localization from pixels to pose[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 3247-3257.

[12] Esfahani M A, Wu K, Yuan S, et al. From local understanding to global regression in monocular visual odometry[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2020, 34(01): 2055002.

[13] Sharif M I, Li J P, Amin J, et al. An improved framework for brain tumor analysis using MRI based on YOLOv2 and convolutional neural network[J]. Complex & Intelligent Systems, 2021, 7(4): 2023-2036.

[14] Huang Y Q, Zheng J C, Sun S D, et al. Optimized YOLOv3 algorithm and its application in traffic flow detections[J]. Applied Sciences, 2020, 10(9): 3079.

[15] Kumari N, Ruf V, Mukhametov S, et al. Mobile Eye-Tracking Data Analysis Using Object Detection via YOLO v4[J]. Sensors, 2021, 21(22): 7668.

[16] Yao J, Qi J, Zhang J, et al. A real-time detection algorithm for Kiwifruit defects based on YOLOv5[J]. Electronics, 2021, 10(14): 1711.

[17] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

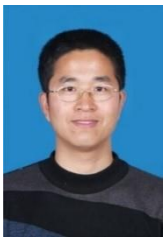
[18] Han K, Wang Y, Tian Q, et al. Ghostnet: More features from cheap operations[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1580-1589.

[19] Schmidt-Hieber J. Nonparametric regression using deep neural networks with ReLU activation function[J]. The Annals of Statistics, 2020, 48(4): 1875-1897.

[20] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[21] Rezaeipanah A, Ahmadi G. Breast cancer diagnosis using multi-stage weight adjustment in the MLP neural network[J]. The Computer Journal, 2022, 65(4): 788-804.

AUTHOR BIOGRAPHY



Huijun Li received the Ph.D. degree in engineering from the Institute of Automation, Chinese Academy of Sciences, 2008. He is an associate professor and has been working in the School of Information and Control Engineering, China University of Mining and Technology. He is mainly engaged in the research of intelligent robots, machine vision and its applications. He has published more than 10 academic papers in professional journals. He has applied for 4 invention patents, 2 utility model patents, and 4 software Copyrights. He has won 1 Excellent Instructor Award of RoboMaster2016 National College Student Robot Competition. Guided students to win the champion of the eastern division of RoboMaster2016 national college student robotics competition and the first prize of the national finals. He participated in one 863 sub-project, two National Natural Science Foundation projects, and undertook more than 10 projects entrusted by enterprises.

Yu Zhang received the B.S. degree in Measurement and control technology and instruments from Yantai University, Shandong, China, in 2016. He is currently a Master student from China University of Mining and Technology and majors in Control science and Engineering. His research interests covers computer vision and autonomous driving perception.

Bin Ye received his Ph.D. in control theory and Control Engineering from Jiangnan University in 2008. He is now an associate professor at the School of Information and Control Engineering, China University of Mining and Technology. His current research interests include autonomous mobile robots, intelligent environment sensing, and quantum computing. He has published 1 monograph and 1 textbook, and obtained 5 national invention patents and 2 software Copyrights. He has published more than 30 academic papers, of which 15 were retrieved by SCI (9 as the first author) and 10 were retrieved by EI. As the project leader, He has completed 1 National Natural Science Foundation of China Youth Fund, 1 special research fund for Doctoral Program of Ministry of Education, and 2 Youth research funds of China University of Mining and Technology. At present, HE is a communication evaluation expert of the National Natural Science Foundation of China and a reviewer of publications such as Communication in Nonlinear Science and Numerical Simulation and Control Theory and Application.

Hailong Zhao received the B.S. degree in Jiangsu University of Science and Technology, Jiangsu, China, in 2018. He is currently a Master student from China University of Mining and Technology and majors in Control science and Engineering. His research interests include visual SLAM and autonomous driving positioning.