Cross-Layer Feature Fusion and Decentration Aberration Correction of Circular Points for Automated Guided Vehicle Terminal Positioning

Guiyang Zhang, Lanyu Yang, Kunkang Cao, Zengguang Man, Jian Wu, and Boning Li

***Abstract*—To address the visual detection and positioning challenge of the Automated Guided Vehicle (AGV) terminal, this study proposed a high-precision target recognition and positioning strategy based on cross-layer feature fusion and eccentricity error correction. In the remote coarse positioning stage, small number of receiver domains underwent continuous multi-layer convolution, which broadened the receptive field of a single element in the feature map of a small circular target. Next, a cross-layer connection feature pyramid was constructed, and the deconvolution module was utilized to enlarge the feature map and fuse it with the shallow features, focusing more on the fine-grained image recognition and enhancing the applicability of circular marker detection. In the near-end fine positioning stage, the factors impacting the deviation between the projection point of the circular feature center and the fitting center were analyzed. Based on this analysis, the deviation of the iterative fitting center was corrected to approximate the real center projection sub-pixel coordinates. The experimental results demonstrated that the proposed method achieved dynamic positioning accuracy of better than 2.0 mm, static positioning accuracy of better than 1.5 mm, and a false recognition rate of less than 1.33% in the range of 3 m from AGV to the tray. Consequently, this method has significant application potential in enabling rapid and stable terminal vision positioning.**

***Index Terms*—Automated guided vehicles (AGV), target detection, feature fusion, decentration aberration correction, vision positioning.**

I. Introduction

AUTOMATED Guided Vehicles (AGV) are widely used in industrial applications, but their current routine track and preset mode limits their intelligence and scalability. These AGV lacks the abilities to autonomously perceive cross-scene environment, target state, and pose changes of terminal effectors, which hinders their capabilities [1], [2]. With the rapid advancements in technologies related to intelligent manufacturing, there is growing interest in integrating vision technology and robots to achieve real-time target recognition and precise positioning. Combining computer vision and robot technology enables mobile robots to perceive the surrounding environment and utilize the captured information to complete specific tasks. This integration has numerous advantages, including improved transportation efficiency, optimized resource allocation, and reduced production costs [3], [4]. As a result, this technology is being widely used in fields such as warehousing logistics and autonomous parking.

The generalization ability and positioning accuracy of the target recognition model are crucial indicators for evaluating AGV performance. The traditional convolutional networks typically follow a top-down approach. With the augmentation of network layers, there's a concurrent enlargement of the receptive field and enrichment of semantic information. This inherent top-down structure presents challenges for multi-scale object detection. Particularly, marker points situated at long distances may be diminutive, causing their features to diminish progressively with depth escalation, consequently undermining detection performance. To achieve different levels of feature information interaction, Zhong et al. [5] highlighted feature fusion technology as the key for detecting target behavior. Continuous down-sampling was used to produce a new feature pyramid by combining different scales to form a feature layer, this was then returned to the multi-box detector to predict the final detection result. Joseph Redmon utilized this idea to construct a feature pyramid by upsampling to obtain the YOLOv3 network model [6], which further improved upon the YOLO series methods. Liang et al. proposed the Feature Fusion and Scaling-based Single Shot Detector (FS-SSD) [7], which uses the average pooling operation to add supplementary scaling branches of the deconvolution module to form a feature pyramid and integrates context analysis to improve small vehicle detection accuracy. In addition, a novel Expansion-Squeeze-Excitation Fusion Network (ESE-FN) based on aggregating the discriminative information of actions and interactions from both RGB videos and skeleton sequences by attentively fusing multimodal features

Manuscript received 18 April 2023; revised 5 October 2023; accepted 17 October 2023. Date of publication 20 October 2023; date of current version 1 November 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61876024, in part by Higher Education Colleges in Jiangsu Province under Grants 23KJA460001 and 21KJA510003, and in part by Suzhou Municipal Science and Technology Plan Project under Grant SYG202129. *(Corresponding author: Lanyu Yang.)*

Guiyang Zhang is with the School of Mechanical Engineering, Changshu Institute of Technology, Suzhou 215500, China, and also with the Optical Measurement Center, TZTEK Technology Company, Ltd., Suzhou 215153, China (e-mail: dr_gyzhang@outlook.com).

Lanyu Yang, Zengguang Man, and Jian Wu are with the School of Mechanical Engineering, Changshu Institute of Technology, Suzhou 215500, China (e-mail: yang_cslg@163.com; yaotingm@163.com; wujian@cslg.edu.cn).

Kunkang Cao is with the Optical Measurement Center, TZTEK Technology Company, Ltd., Suzhou 215153, China (e-mail: tztek_cv@foxmail.com).

Boning Li is with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA (e-mail: cvph2016@126.com).

Digital Object Identifier 10.1109/JPHOT.2023.3326305

was proposed [8], and the experimental results show that the ESE-FN performs better in terms of normal action recognition task. However, unfavorable factors such as long distance, low resolution, weak light source, and complex environment can still reduce target detection accuracy, leading to false alarms and missed actual detection in the industrial field. The target image captured during the visual positioning process is usually characterized by interframe continuity, thus two methods were derived: single frame detection and sequence frame detection. For single frame detection, Chen et al. [9] proposed the Local Contrast Method (LCM), which considers the huge difference between the target and the background in the image and combines the human visual attention mechanism. On this basis, improvement strategies such as Relative Local Contrast Measure (RLCM) [10], Multiscale Patch-based Contrast Measure (MPCM) [11], and Double-Neighborhood Gradient Method (DNGM) [12], have been derived. However, these methods are easily disturbed by clutter and noise, resulting in low detection accuracy. When detecting targets through sequence frame, the strong correlation between adjacent images is typically leveraged. Various methods have been proposed, such as Spatial Local Contrast (SLC), which uses spatial contrast to enhance the target, Temporal Local Contrast (TLC), which uses temporal contrast to enhance the target, and Spatial-temporal local contrast filter (STLCF) [13], which combines both spatial and temporal contrast to form a spatial-temporal contrast filter. Numerous studies have been conducted on small target recognition using deep learning. Lu et al. [14] used deep learning strategies to initially locate target information in space and then adopted graph matching and flow graph of sequence images to eliminate false alarms and enhance recognition credibility. However, their method did not make optimize missed detections. The CNN-based target detector was proposed and enhanced. The detector generated synthetic targets and divided the detection task into two steps [15], [16]: candidate target extraction and candidate recognition. Eliminating false alarms excessively can lead to missed detection, and strong feature learning can make up for missed detection, but it will lead to false alarms. Therefore, balancing missed detection and false alarms is crucial for improving the robustness and accuracy of target detection.

The technology for visual positioning of robots using visual sensors can be classified into three categories based on the system structure. The first category is monocular vision positioning, which establishes a mapping relationship between the two-dimensional pixel coordinates of the target point and the three-dimensional space coordinates using the feature information related to the target object, such as point and linear features [17], [18]. The second category is binocular vision positioning, which uses a binocular camera to capture left and right images of the same scene, calculates the pose information of the target object in the robot coordinate system using feature matching and triangulation principle, and completes the positioning task of the target [19]. The third category is omnidirectional vision positioning, which obtains more three-dimensional information by placing multiple cameras at different angles to solve the mismatch problem caused by ambiguity [20]. Cui et al. [21] proposed a method to recognize the cross-section features of an axisymmetric spacecraft based on contour and center constraints. After that, a simplified model of the spacecraft was obtained through three-dimensional reconstruction. Zhang et al. [22] utilized the perspective projection equation of the ellipse and constructing the binocular stereo-vision constraint to achieve pose measurement of the spatial circular target with a known radius. Besides, Klimchik et al. [23] studied the selection of optimal measurement pose points in the robot kinematics parameter identification process to improve the accuracy and stability of parameter identification. They compared the observable indexes of optimal selection of various measurement pose points. Liu et al. [24] proposed a method to estimate the pose of spatial circles using binocular stereo vision triangulation technology. They first constructed a special vector to calculate the normal direction of the spatial circle. Then, they estimated the pose by calculating the projection pixel coordinates of the center of the circle on the left and right camera images. Finally, the pose of the spatial circle was estimated using binocular stereo vision triangulation technology. However, this method can lead to large pose measurement errors for the measuring circular targets with small long-distance sizes and large projection errors. Therefore, precisely extracting the center coordinates of the circular mark points is the basis of high-precision positioning.

Circles possess desirable morphological characteristics, such as clear descriptions, measurable radii, and normal vectors, and normal vectors, making them ideal for visual morphological detection. Additionally, circles demonstrate strong resistance to noise and image blur, making them a common choice for landmarks preparation and navigation [25], [26]. To enhance the accuracy of moment-based image borders extraction, Tabatabai et al. [27] proposed an algorithm for sub-pixel edge localization utilizing the first three gray moments, This algorithm provides a closed-form edge localization solution without requiring interpolation or iteration. However, it is sensitive to image gray additive and multiplicative noise. Gu et al. [28] developed a spherical imaging error variation law and error correction model, which used simulation to analyze the spherical center imaging error by aligning the spherical target with the camera coordinate system's origin. Nonetheless, this model is an approximation of the ideal model, and the six spatial positions of the spherical target must be known in advance. Additionally, the ratio of the space ball radius to the object distance must be kept within a specific range to avoid model failure. Subsequently, an exact analytical expression for the point center imaging error was derived [29], with no theoretical error. However, the circle center imaging error can still be affected by ellipse fitting, imaging quality, and edge positioning accuracy. Furthermore, ensuring the collinearity between the imaging ellipse and the imaging plane center is challenging. Researchers established a distortion error model for the projection of spatial circles based on the perspective projection imaging process [30]. They then obtained the distortion law of spatial circle projection through simulation analysis. However, this method has several shortcomings, including a large amount of computation and the requirement of knowing the posture of the spatial circle in the camera coordinate system.

The aforementioned methods provided a closed analytical solution to the target positioning problem based on circular landmarks by establishing the spatial elliptical cone equation and extending the circular projection equation through circular feature back projection. However, there are more apparent interference factors in the running scenes of mobile robots, and the small circular target with long-distance size occupies fewer pixels in the feature area. Therefore, it is necessary to achieve effective feature fusion and target detection through more delicate features. Moreover, the inaccuracy of ellipse contour extraction significantly affects the accuracy of three- dimensional reconstruction of space targets, resulting in a considerable deviation between the target positioning result and the actual position. Consequently, it is crucial to deeply analyze the long-distance target detection and close-range precise positioning to provide accurate visual guidance information for AGV motion.

The proposed strategy in this paper aims to enhance the accuracy and reliability of AGV terminal vision detection and positioning, inspired by the observations mentioned earlier. The proposed small target detection algorithm utilizes the parallel-connected convolutional neural network PaRNet as the backbone network and constructed the feature pyramid through cross-layer connection feature fusion, enabling stable detection of small landmarks and Regions Of Interest (ROI) region selection. Additionally, the factors influencing the projection point of the circular feature center and the deviation between the fitting centers were analyzed. The coordinates of the circle center's projection point were determined using the principle of simple ratio invariance and straight line invariance of the projective transformation, providing essential data for the AGV's high-precision visual positioning.

The remaining sections of this study are outlined as follows: In Section II, the framework of terminal positioning and visual positioning mathematical model is demonstrated. In Section III, the cross-layer feature fusion and decentration aberration correction of circular points for target location is presented and assessed. In Section IV, the experiments are conducted to verify the feasibility and effectiveness of the proposed method. Finally, in Section V, the conclusion based on the results obtained from the experiments is drawn.

II. Preliminaries

A. Framework of Terminal Positioning

Typically, to enable AGV detection and location using vision technique, one or more cameras are installed on vehicle. These cameras capture real time scene images, which are used to determine the position of the target relative to AGV. This information provides crucial control data, enabling AGV to move precisely and efficiently.

The overall scheme presented in this paper aims to achieve both accuracy and timeliness in AGV positioning by using a binocular camera for tray recognition and positioning. The pallet recognition and positioning process is divided into two steps: far-end coarse positioning and near-end precise positioning. In the far-end coarse positioning process, AGV is at a significant distance from the tray position, making it difficult for the camera to distinguish the tiny circular marker information on the tray. Therefore, deep learning target detection is utilized to identify the approximate position of the tray, and AGV's steering is adjusted to roughly aligned AGV and the tray position. When the relative position of AGV and tray is within a certain threshold, mark point recognition technology is employed. To meet timeliness requirements, target detection technology is used for direct detection of the area where the mark points are located during near-end precise positioning. This allows for the precise acquisition of all ROI in the scene image captured by the camera, that is, the area where all the mark points are located. Due to the perspective projection of the camera, the spatial circle appears as an ellipse in the image coordinate system. Therefore, high-precision ellipse edge sub-pixel fitting is necessary in the ROI region to obtain the coordinates of the projected ellipse center point. By performing the three-dimensional reconstruction of the mark point center coordinates on the tray, the position coordinates of the tray relative to the camera on AGV can be calculated. Furthermore, utilizing the triangular geometric relationship, the angle between the tray and the horizontal line can be determined to aid in adjusting AGV's motion direction.

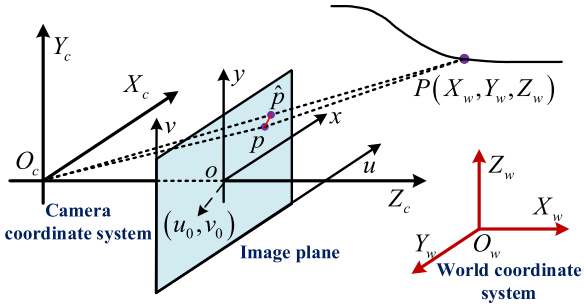


Fig. 1. Pinhole imaging model of camera.

B. Visual Positioning Mathematical Model

To achieve target positioning using visual mode, it is necessary to capture the target's image using a camera. Camera imaging not only serves as the information source of the vision system, but also helps obtain the internal and external parameters of the camera through imaging model, which enable pose measurement. The commonly used pinhole projection model is shown in Fig. 1, where $O_w - X_w Y_w Z_w$ is the world coordinate system. $O_c - X_c Y_c Z_c$ is the camera coordinate system. $o - uv$ is a two-dimensional image coordinate system.

Based on the ideal pinhole imaging model, $P(X_w, Y_w, Z_w)$ in the world coordinate system that can be converted to the two-dimensional image pixel coordinate system by the following equation:

$$s\begin{pmatrix}\mu\\ \nu\\ 1\end{pmatrix}=\begin{pmatrix}\alpha & \gamma & \mu_0 & 0\\ 0 & \beta & \nu_0 & 0\\ 0 & 0 & 1 & 0\end{pmatrix}\begin{pmatrix}\boldsymbol{R} & \boldsymbol{t}\\ 0 & 1\end{pmatrix}\begin{pmatrix}X_w\\ Y_w\\ Z_w\\ 1\end{pmatrix} \quad (1)$$

where $\boldsymbol{R}$ and $\boldsymbol{t}$ are the rotation matrix and translation vector between the world coordinate system and the camera coordinate

system, respectively. s is the scale factor. μ and ν are the pixel coordinate of point. P, α, and β are the equivalent focal lengths in the directions of the X_c and Y_c axes, respectively. μ_0 and ν_0 are the principal point coordinates of the optical axis and the image plane. γ is a non-vertical coefficient of the angle between the X_c and Y_c axes.

The precisely extraction of the center coordinates of marker points is crucial for achieving high-precision positioning. However, the moment of the actual continuous edge of the marker point is differs from that of the digital sampling edge after imaging, owing to the sampling effect of the imaging system and the impact of defocus, penumbra, uneven illumination distribution, and other factors. As a result, there are notable errors in the existing Zernike moment for sub-pixel edge positioning. In this study, the sub-pixel coordinates obtained based on the Zernike moment edge model are fitted, and the eccentricity error is subsequently corrected to obtain precise center coordinates. When the binocular camera captures two images containing marker points from different perspectives, the triangulation method can be used to obtain the three-dimensional space coordinates of the marker points in the world coordinate system, thereby completing the visual positioning function.

III. Cross-Layer Feature Fusion and Decentration Aberration Correction of Circular Points

A. Circular Point Remote Detection

When the tray is at a significant distance from AGV, an efficient detection model is required for circular marker points with small image sizes. To achieve this, the proposed small target detection algorithm constructs an appropriate using the parallel connected convolutional neural network PaRNet as the backbone network. The feature pyramid is constructed using the cross-layer connection feature fusion method, which aims to achieve stable detection of small landmarks and reduce interference from complex backgrounds in the marker point extraction process.

This study utilizes two model architectures, PaRNet-35 and PaRNet-51, which are shown in Fig. 2. These architectures consist of parallel residual blocks interspersed with continuous convolutions, resulting in a lightweight network with improved generalization performance and accuracy. When using "Bottleneck", PaRNet-35 is selected, and when using the "Bottleneck" base block, PaRNet-51 is selected. Both models are composed of four parallel residual blocks, but PaRNet-51 has a deeper network layer due its use of BottleNeck as the base block in building the residual block. The network employs continuous multi-layer convolution of a small number of receptive domains to enhance feature learning ability, resulting in a broad receptive field of a single element in the feature map of a small target. Then, the parallel-based residual blocks are stacked to fully extract features.

Based on the parallel residual block structure, as shown in Fig. 3, the input part is divided into two paths, which compresses the image size through the traditional residual block, and the branch path, which fully extracts image features through the 3 $\times$ 3 maximum pooling layer and 1 $\times$ 1 convolution layer. The

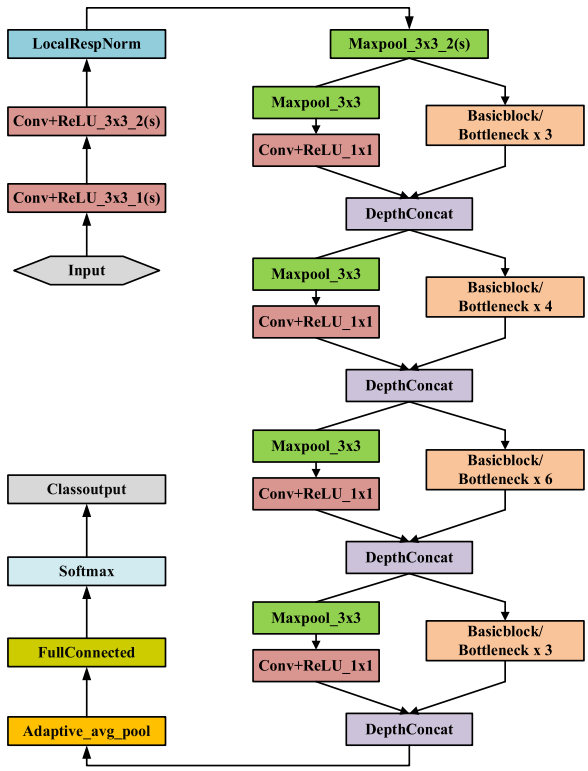


Fig. 2. Architecture of PaRNet-35.

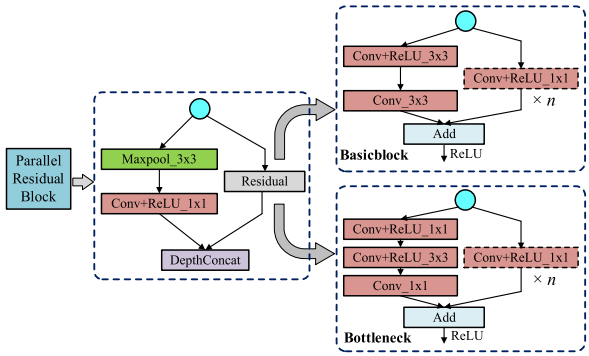


Fig. 3. Parallel residual block designing diagram.

channel fusion is achieved by increasing channel dimensions through the trunk and branch part. To obtain the PaRNet-35, the minimum number of residual blocks in series, where each module is connected in series by two residual blocks, is adopted, resulting in a 35-layer classification network model. The channel selection uses the incremental form of [64, 128, 256, 512, 1024].

PaRNet-35 is used as the backbone network for target detection and to construct a cross-layer connection feature pyramid. Operating scenes of mobile robots generally have more evident interference factors and weaker characteristics, requiring more delicate features for effective feature fusion and target detection. The output features of the 2nd, 3rd, and 4th parallel residual blocks are extracted and then fused by expanding the image size from bottom to top. To fully restore the detailed features of small targets, the traditional YOLOv3 feature fusion module

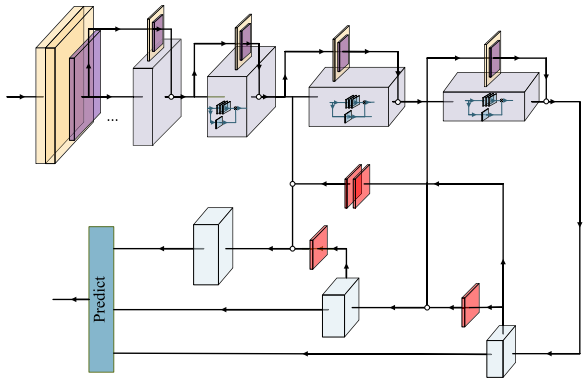


Fig. 4. Diagram of feature fusion by cross-layer connection.

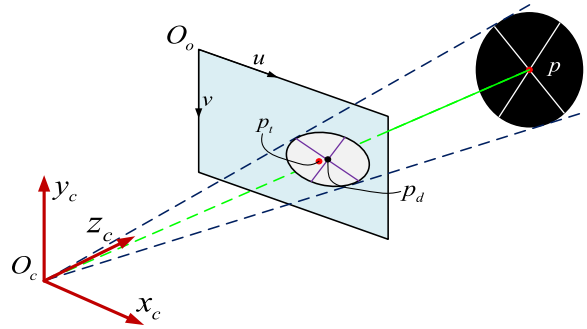


Fig. 5. Spatial circular projection model.

uses deconvolution instead of a simple up-sampling operation. Deconvolution is the inverse process of convolution that can restore the underlying image to its original or larger size through convolution operation and reflect the pixel-level features in the image. Compared to simple up-sampling, deconvolution can mine each pixel feature and restore more precise local features, making it more suitable for detecting small targets. However, deconvolution is prone to uneven overlap, especially when the kernel size cannot be divided by step size, resulting in a checkerboard effect. To alleviate this issue, the last layer of deconvolution has a step size of 1.

Following repeated convolution, the output features of the bottom layer with a channel number of 2048 are fused into each layer of the network. Multi-layer deconvolution and connected layers-by-layer and cross-layer were employed on the channel dimension. Additionally, the output features of the first parallel residual block are fused with other feature layers through the convolution layer and the maximum pooling layer to ensure the completeness and comprehensiveness of the features. The specific structure is illustrated in Fig. 4.

In this structure, the deconvolution module consists of two 2×2 deconvolution layers with a step size of 1, and each layer is activated by a rectified linear unit (ReLU) and then normalized by batch processing. Compared to the traditional bi-linear up-sampling method, the deconvolution layer focuses more on the fine-grained recognition of the image while improving the resolution of the feature layer, which contributes to improving the representativeness of the network features. A small change in input will lead to a large change in the loss function, making the gradient larger and alleviating the problem of gradient disappearance. The deconvolution module amplifies the feature map and fuses it with shallow features, and the final prediction layer is obtained by continuous convolution.

B. Center Coordinate Extraction Based on Decentration Aberration Correction

The small target detection network model mentioned earlier is capable of stable detection of far-end circular markers. Then the ROI region is selected and the edge sub-pixel coordinates of the circular marker points within it using Zernike moments. Once the edge pixel coordinates are obtained, they are fitted by an ellipse to obtain the elliptic equation:

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0 \tag{2}$$

where $W = [A, B, C, D, E, F]^T$, (2) can be expressed as a matrix form of $W^T X = 0$. The precise acquisition of the center coordinates of the ellipse can be formulated as an optimization problem that seeks the optimal solution of a certain function, which can be expressed as follows:

$$\min \left\| W^T A \right\|^2 \qquad s.t.\ \ W^T H W = 1 \tag{3}$$

where H is a quadratic coefficient matrix to satisfy the constraint $4AC - B^2 > 0$.

In general, there exists a certain angle between the normal vector on the surface of the circular feature and the optical axis of the camera. This perspective projection model, as shown in Fig. 5, can be used. The spatial circle is transformed by the model to form an ellipse on the image plane, which better reflects the actual situation.

The center point P of the space circle is transformed by the imaging model to obtain the projection point p_t, and the geometric center of the projection ellipse on the imaging plane is p_d. Based on the above analysis, it can be observed that p_t and p_d do not coincide. The pixel coordinates of the geometric center point p_d are:

$$\begin{cases} x_{p_d} = \dfrac{2\left(n^2+p^2-\eta_i^2 e^2\right)\left(2mw+2oq-2\eta_i^2 df\right)}{\left(2mn+2op-2de\eta_i^2\right)^2-4\left(m^2+o^2-\eta_i^2 d^2\right)\left(n^2+p^2-\eta_i^2 e^2\right)} \\ \quad -\dfrac{\left(2mn+2op-2de\eta_i^2\right)\left(2nw+2pq-2\eta_i^2 ef\right)}{\left(2mn+2op-2de\eta_i^2\right)^2-4\left(m^2+o^2-\eta_i^2 d^2\right)\left(n^2+p^2-\eta_i^2 e^2\right)} \\ y_{p_d} = \dfrac{2\left(m^2+o^2-\eta_i^2 d^2\right)\left(2mw+2pq-2\eta_i^2 ef\right)}{\left(2mn+2op-2de\eta_i^2\right)^2-4\left(m^2+o^2-\eta_i^2 d^2\right)\left(n^2+p^2-\eta_i^2 e^2\right)} \\ \quad -\dfrac{\left(2mn+2op-2de\eta_i^2\right)\left(2mw+2oq-2\eta_i^2 df\right)}{\left(2mn+2op-2de\eta_i^2\right)^2-4\left(m^2+o^2-\eta_i^2 d^2\right)\left(n^2+p^2-\eta_i^2 e^2\right)} \end{cases} \tag{4}$$

The pixel coordinates of the projection point p_t are:

$$\begin{cases} x_{p_t} = \dfrac{nq-wp}{mp-no} \\ y_{p_t} = \dfrac{ow-mq}{mp-no} \end{cases} \tag{5}$$

The variables used in (4) and (5) are as follows: (X_i, Y_i) is the coordinate value of the circular geometric edge point, η_i is the radius of the circle, $r_1, r_2, \ldots, r_9$ are the rotation matrix elements of the camera's external parameters, and t_φ, $(\varphi = x, y, z)$

is the translation component. The expressions of the remaining variables are given below:

$$\begin{cases} a=(r_8t_y - r_5t_z)\,,\ b=(r_2t_z - r_8t_x)\,,\ c=(r_5t_x - r_2t_y)\,, \\ d=(r_5r_7-r_4r_8)\,,\ e=(r_1r_8-r_2r_7)\,,\ f=(r_2r_4-r_1r_5)\,, \\ h=(r_7t_y - r_4t_z)\,,\ j=(r_1t_z - r_7t_x)\,,\ k=(r_4t_x - r_1t_y)\,, \\ o=(h+Y_id)\,,\ p=(j+Y_i)\,,\ q=(k+Y_if)\,, \\ m=(a-X_id)\,,\ n=(b-X_ie)\,,\ w=(c-X_if) \end{cases} \tag{6}$$

The offset between the two points was used as the distortion error of the circular feature projection, which leads to the following equation:

$$\Delta = \sqrt{(x_{p_t} - x_{p_d})^2 + (y_{p_t} - y_{p_d})^2} \tag{7}$$

The distortion error described in (7) is influenced by the distance between the spatial center and the image plane, as well as the angle between the spatial circular plane and the image plane. In cases of small inclination and close projection, the error can reach up to 5.2 um. At long distance and large angles, the deviation of the center coordinates is even greater. Therefore, accurately obtaining the fitting center coordinates of the geometric center of the projection ellipse is crucial.

After obtaining the fitting ellipse, the two points that are farthest from the edge coordinates of the ellipse are connected, and the midpoint coordinates of the long axis are calculated. The distance Δd between the midpoint of the long axis and the center of the fitting circle is then determined. If Δd satisfies the threshold, the midpoint of the long axis is set as the center coordinates. If Δd is too large, the projection point coordinates of the center of the circle are obtained by the intersecting the long axis and the projection ellipse at coordinates (u_1, v_1) and (u_2, v_2) respectively, using the principle of simple ratio invariance and straight line invariance of the projection transformation:

$$\begin{cases} \frac{u_o-u_1}{u_2-u_1} = \frac{\mathrm{R}}{2\mathrm{R}} \\ \frac{v_o-v_1}{v_2-v_1} = \frac{\mathrm{R}}{2\mathrm{R}} \end{cases} \Rightarrow \begin{cases} u_o = \frac{u_1+u_2}{2} \\ v_o = \frac{v_1+v_2}{2} \end{cases} \tag{8}$$

where (u_o, v_o) is the projection coordinates of the center of the circle in space, R is the radius of the circus in space.

The above steps further correct errors in the fitted ellipse's geometric center and obtain the projection point's coordinate value, which is highly close to the circle's real center. This algorithm is characterized by its simple calculation and high extraction accuracy.

Subsequently, the projection coordinates of the center $P(X_w, Y_w, Z_w)$, are determined as (u_l, v_l) and (u_r, v_r), and the left camera coordinate system is selected as the world coordinate system. Additionally, u_{0l}, v_{0l}, α_l, β_l represent the intrinsic parameters of the left camera. After completing the stereo registration of the homonymous points in the left and right camera images, the physical spatial coordinates of the speckle point in the world coordinate system can be obtained using (9).

$$\begin{cases} X_w = \frac{(u_l-u_{0l})Z_w}{\alpha_l} \\ Y_w = \frac{(v_l-v_{0l})Z_w}{\beta_l} \\ Z_w = \frac{\alpha_l\beta_l(t_zu_r-D)}{\Xi} \end{cases} \tag{9}$$

TABLE I
PERFORMANCE OF DIFFERENT NETWORKS ON CIFAR-10

	Params(10^6)	FLOPs(10^8)	TOP-1 (%)	TOP-5 (%)
GoogLeNet	6.6	3.9	13.01	0.72
ResNet-50	25.6	11	18.32	1.29
ResNet-101	44.5	20	16.71	1.04
PaRNet-35	22.8	4.4	14.19	0.67
PaRNet-51	31.7	10	13.38	0.64

where l and r are the left and right cameras, respectively. $\Xi = \beta_l(A - r_7u_r)(u_l - u_{0l}) + \alpha_l(B - r_8u_r)(v_l - v_{0l}) + \alpha_l\beta_l(C - r_9u_r)$. The coefficients are determined as follows:

$$A = \alpha_r r_1 + u_{0r}r_7, B = \alpha_r r_2 + u_{0r}r_8,$$
$$C = \alpha_r r_3 + u_{0r}r_9, D = \alpha_r t_x + u_{0r}t_z.$$

After obtaining the three-dimensional coordinates of the circular marker points on the tray in the world coordinate system, AGV can use the optical center of the camera as the origin and calculate the Euclidean distance between the target point and the camera optical center. Additionally, the angle information can be calculated through coordinate operations.

IV. EXPERIMENTAL RESULTS AND ANALYSES

A. Verification of Detection Accuracy

The performance of the backbone network is initially validated using the CIFAR-10 dataset. PaRNet is utilized for this purpose, and the small batch gradient descent method is employed for training. The training is performed for 1000 cycles using a learning rate of 0.1, momentum of 0.1, and a batch size of 16. The experimental setup includes GTX2070, Ubuntu 16.04, PyTorch 1.6.1, CUDA 10.1, CUDNN 7.6.5, and Python 3.6.9. Table I summarizes the model parameters (Params), FLOPs, TOP-1 error rate, and TOP-5 error rate of PaRNet-35 and PaRNet-51. Furthermore, a comparison is made with ResNet-50, ResNet-101, and GoogLeNet, and the results are presented as follows:

Based on the number of parameters and computational load, PaRNet has more parameters but less computation compared to networks with similar layers. This is due to its parallel advantage and sparsity similar to the GoogLeNet network. Moreover, as the number of network layers increases, the accuracy of ResNet improves, and the employed network benefits from this advantage. During the first 1000 training cycles, both PaRNet-35 and PaRNet-51 achieved more accurate classification results under a low network load.

The dynamic variation trend of the network's accuracy on CIFAR-10 over the iteration period is presented in Fig. 6. The results show that PaRNet outperforms traditional network models in terms of the rate of decline and oscillation in the early stages and the stability and accuracy in the later stages. The addition of parallel modules allows the network to be shunted, resulting in a large descent gradient in the early stage, which leads to a greater rate of decline in the classification error for

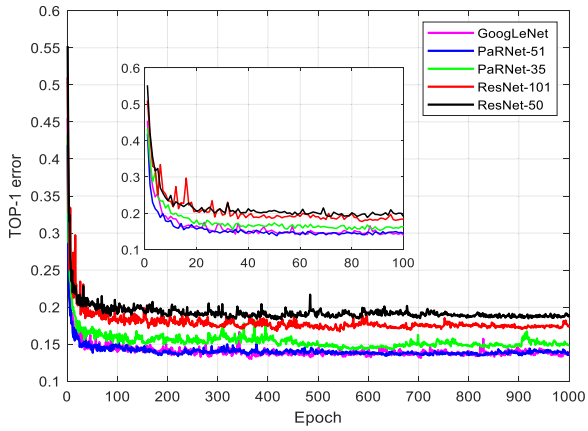


Fig. 6. Change of TOP-1 value in different networks with the training epochs.

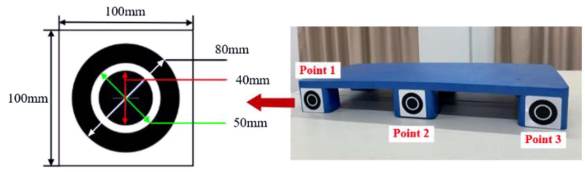


Fig. 7. Design and size of circular points.

PaRNet. Additionally, PaRNet network model is more stable and robust compared to other models, as evident from the smoother curve of PaRNet after the 500th iteration period.

Considering the distance of the camera from the tray and the low proportion of the mark points in the captured image, a white background and a black foreground, consisting of multiple concentric rings with two different colors, was designed. The size diagram of the circular points can be seen in Fig. 7.

The dataset collection and training process focused mainly on the tray's feet landmarks captured under varying conditions such as different heights, scenes, and light intensities. A total of 600 tray images were collected, with 450 used for training and 150 for verification. The optimization algorithm used was the Adaptive Moment Estimation, with an initial learning rate of 0.001 and the StepLR mechanism for dynamically reducing the learning rate. Freezing training was used to prevent the initial training weight values from being destroyed and speed up the process. This involved 50 freezing training generations followed by 100 non-freezing training generations.

To ensure accurate detection of circular markers on the tray and the complete detection of multiple markers in the image, this study introduced additional evaluation indicators, namely the False Alarm Rate (FAR) and the Missed Detection Rate (MDR), in addition to the traditional precision index, which can be determined via (10).

$$\begin{aligned} FAR &= AlarmPic/TotalPic \\ MDR &= MissedPic/TotalPic \end{aligned} \tag{10}$$

where $AlarmPic$ is the number of images with false alarms, $MissedPic$ is the number of images with missed detection, and $TotalPic$ is the total number of images.

TABLE II
THE EVALUATION RESULTS OF TARGET DETECTION ALGORITHMS

Method	Performance evaluation indicators			
	Precision	FAR	MDR	FPS
YOLOv3	92.45	3.33	11.33	33.50
YOLOv4	94.67	2.00	8.67	36.90
YOLOv5m	93.56	2.67	9.33	43.03
Proposed	95.78	1.33	6.67	33.10

The positive and negative detection results of the circular marker points were obtained using the proposed algorithm, as shown in Fig. 8. In this test, the trays were distributed at different positions in the field of view of stereoscopic camera.

It can be seen from Fig. 8(a) that when the tray is placed in different poses and heights, the constructed network model can accurately and stably detect the circular points imaged by the left and right cameras. By contrast, in Fig. 8(b), When extreme situations occur in sequence, (e.g., the angle between the normal vector of the tray front panel and the camera optical axis is too large, the exposure is too strong, the marker points exceed the field of view, and the scale is too small when the target distance is too far), the original image information is relatively missing, which results in missed detections. However, in practical applications, pre-detection of the entire tray can be used to avoid large distances and azimuth angles between AGV and target during the movement process.

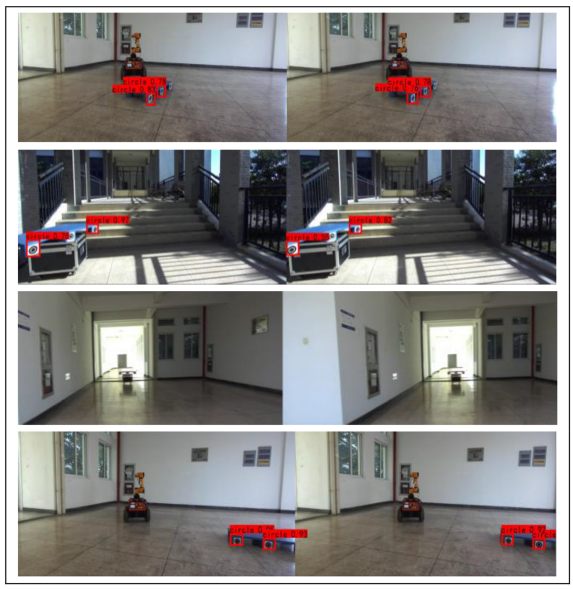
To further evaluate the accuracy and timeliness of the detection, the results obtained by the following typical representative methods were compared. Four indicators in different states were measured, including precision, false alarm rate, missed detection rate, and Frames Per Second (FPS), as shown in Table II.

The four detection methods achieved different level detection results under the same conditions, with a precision reaches of over 92%, a missed detection rate of no more than 11.33%, and a false alarm rate of less than 3.33%. The proposed method had better accuracy (Precision: 95.78%, FAR: 1.33%, MDR: 6.67%). YOLOv5m had a faster detection speed, with a FPS of up to 43.03, while YOLOv4 had a more balanced performance index. Accuracy and timeliness are critical factors affecting the algorithm's practical value in mobile robot positioning applications, and different detection methods must be chosen based on the actual situation. The key to ensuring the algorithm's utility is to improve detection accuracy as much as possible, provided that the detection speed meets the required standards.

B. Positioning Accuracy Verification Experiment

ZED stereo camera was utilized for image acquisition and positioning in this study. The camera has an imaging resolution 2208×1242 pixels, and the pixel size is $du \times dv = 2 \times 2\ \mu\text{m}^2$. The extraction accuracy of the center of the circle was evaluated and analyzed using simulation images. An image with a black circle and white background, with a size of 1680×1680 pixels, was generated, consisting of three concentric circles, upon which the center is extracted to verify, as illustrated in Fig. 9.

(a) Positive results of detection



(b) Negative results of detection

Fig. 8. Target detection results under different conditions.

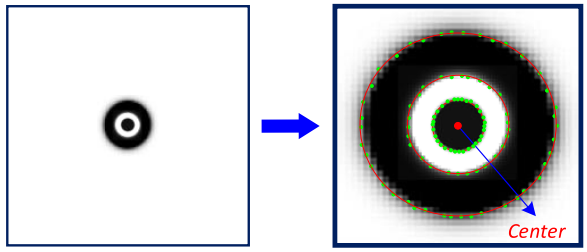


Fig. 9. Circular feature point simulation and center location.

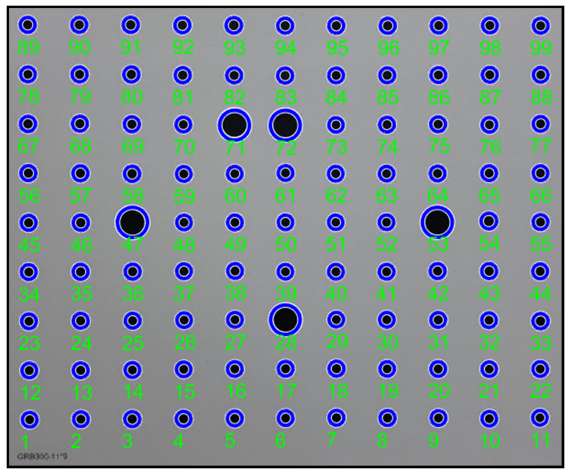


Fig. 10. Extraction and coding of planar calibration plate.

The center of the circle in Fig. 9 has the coordinate $(840, 840)$, and the outer circle has a radius of $r = 50$ pixels. The simulation image was further corrupted by adding Gaussian noise with a mean value of 0 and a variance of 00.002, and 0.004, respectively. The gray moment, fitting method, Zernike moment, and the proposed method were then employed to locate the center coordinates. The fitting center results for each method are shown in Table III.

As shown in Table III, the proposed method achieved the highest accuracy in center circle extracted. This can be attributed to the introduction of eccentric error correction theory, which yielded center coordinates closer to the true center of the circle. These results can provide a reliable data basis for high-precision camera calibration and spatial positioning.

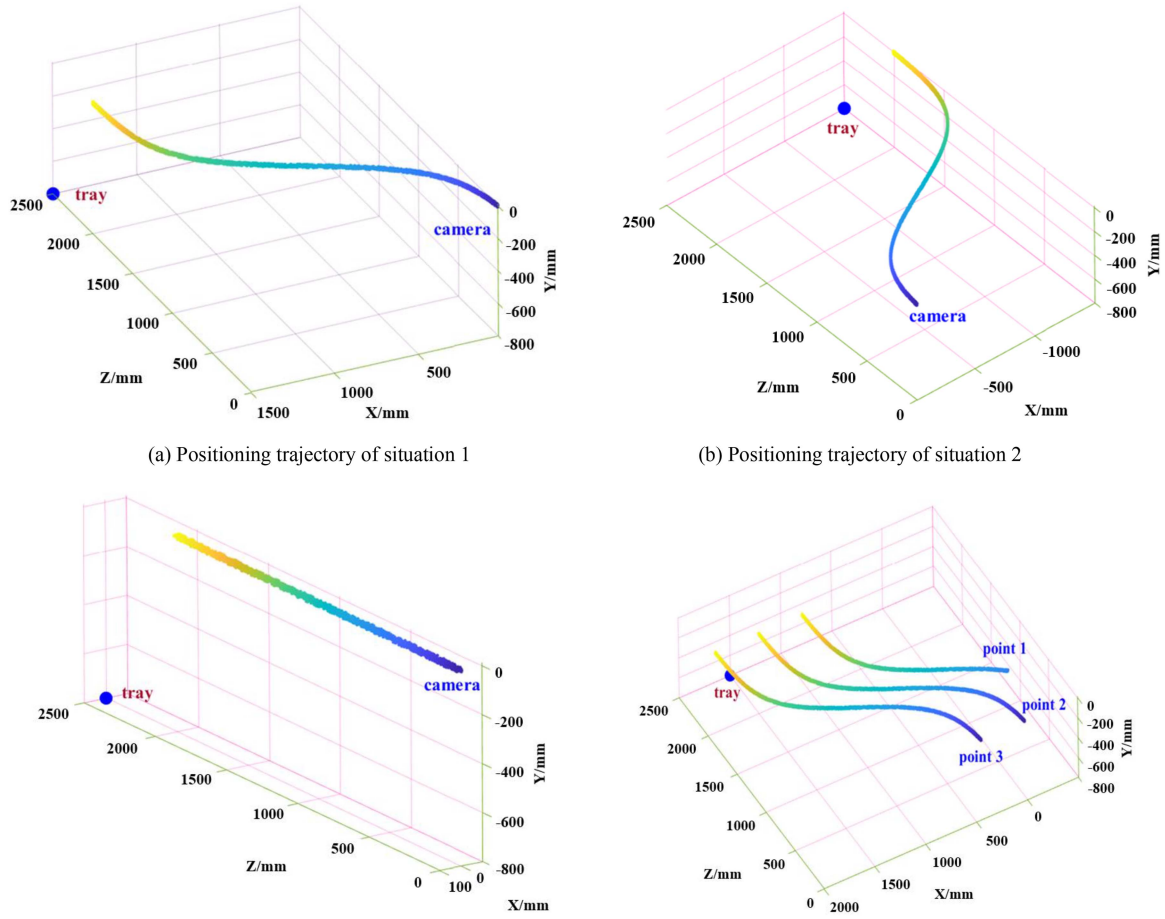
In order to achieve accurate target positioning, obtaining the internal and external parameters of the binocular camera beforehand is essential. Therefore, the next step involved camera calibrated. The calibration was performed in a space of approximately 3 m $\times$ 3 m $\times$ 1.5 m. The calibration plate used and feature point coding are illustrated in Fig. 10 .

The intrinsic parameters and pose relationship matrix of the binocular camera system were obtained after calibration, as presented in Table IV.

To simulate an actual working scenario, the ZED camera was mounted on a platform at a height of 0.8 m above the ground, while the tray was placed on a shelf at a height of

TABLE III
CENTER LOCATION ACCURACY OF SIMULATIVE IMAGE

Methods	Noise type					
	Noiseless		White noise (σ^2 =0.002)		White noise (σ^2 =0.004)	
	Fitting center	Error	Fitting center	Error	Fitting center	Error
Gray moment	(840.024,840.029)	0.0376	(840.026,840.054)	0.0599	(840.032,840.053)	0.0619
Quasi legality	(840.035,839.970)	0.0461	(840.065,839.870)	0.1453	(839.847,839.863)	0.2054
Zernike moment	(839.959,840.014)	0.0433	(839.963,840.018)	0.0411	(839.953,840.015)	0.0493
Proposed method	(840.014,839.991)	0.0166	(840.025,839.989)	0.0273	(840.027,839.986)	0.0304



(a) Positioning trajectory of situation 1

(b) Positioning trajectory of situation 2

(c) Positioning trajectory of situation 3

(d) Detail of trajectory of situation 1

Fig. 11. Trajectory graph of the four situations.

TABLE IV
STEREO CAMERA INTERNAL PARAMETERS AND POSE MATRIX

Camera internal parameters	Position attitude relationship between stereo cameras
$\left[f_u^r, f_v^r\right]=\left[1076.4775, 1130.1791\right]$	$\boldsymbol{R}=\begin{bmatrix}1.0000 & 0.0003 & -0.0004\\ -0.0003 & 1.0000 & -0.0037\\ 0.0003 & 0.0037 & 1.0000\end{bmatrix}$
$\left[u_0^r, v_0^r\right]=\left[1075.4949, 620.5816\right]$	
$\left[f_u^l, f_v^l\right]=\left[1074.4627, 1130.1092\right]$	$\boldsymbol{t}=\begin{bmatrix}-120.0852\\ -0.2140\\ 0.4508\end{bmatrix}^T$
$\left[u_0^l, v_0^l\right]=\left[1075.4949, 623.9368\right]$	

1.0 m above the ground. The accuracy of the marker point reconstruction determines the tray's pose information, and hence, its validity can be assessed based on the reconstruction accuracy. Since obtaining the true value of the actual three-dimensional space coordinates is difficult, the precision was verified through displacement distance measurement. The camera position was fixed and the initial tray's position was set approximately 5 m in front of the camera. The tray was then translate along the axis direction, and images were captured at intervals of 0.5 m.

TABLE V
DISPLACEMENT POSITIONING RESULTS AND ERRORS

NO.	Marker points	Initial value of 3D coordinates (mm)			Current value of 3D coordinates (mm)			Estimated distance (mm)	True value of distance (mm)	Error (mm)
		X	Y	Z	X	Y	Z			
1#	Point 1	-213.79	643.57	2296.46	289.00	649.04	2306.04	502.91	502.11	0.80
	Point 2	640.21	638.40	2230.08	1147.16	641.57	2239.58	507.05	506.13	0.92
	Point 3	640.21	638.40	2230.08	1147.16	641.57	2239.58	507.05	506.04	1.01
2#	Point 1	-213.79	643.57	2296.46	789.61	649.42	2298.70	1003.42	1004.56	1.14
	Point 2	640.21	638.40	2230.08	1648.35	640.84	2239.37	1008.19	1007.09	1.10
	Point 3	640.21	638.40	2230.08	1648.35	640.84	2239.37	1008.19	1009.37	1.18
3#	Point 1	-213.79	643.57	2296.46	1286.97	644.11	2301.77	1500.77	1502.18	1.41
	Point 2	640.21	638.40	2230.08	2147.96	646.57	2238.77	1507.80	1506.53	1.27
	Point 3	640.21	638.40	2230.08	2147.96	646.57	2238.77	1507.80	1509.16	1.36

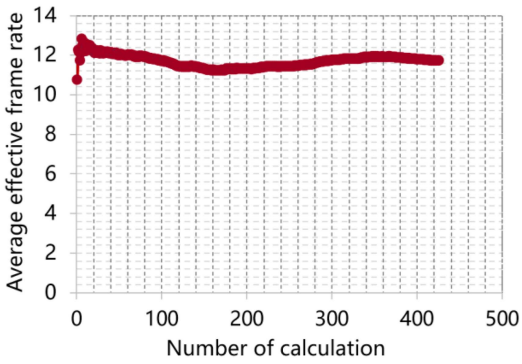


Fig. 12. Calculation efficiency of the proposed algorithm.

Additionally, OptiTrack dynamic vision measurement system was utilized to obtain precise measurements, which served as the ground true value. This system was used to collect images and locate landmarks, and displacement was calculated based on the previous and last two times of positioning. The performance of the proposed approach was then evaluated by comparing the results with the true value. The measured results are shown in Table V. The results in Table V illustrate an increase in measurement error as the translation distance increases, with a minimum error of 0.8 mm. Within a range of 5 m from the camera, the global error is less than 1.5 mm.

C. AGV Continuous Motion Location Experiment

To evaluate the detection and positioning performance of AGV during continuous motion, the serial movement of the tray was simulated by manually controlling the camera's motion. While maintaining a fixed camera height, the camera was moved towards the tray until it was close enough. Three different scenarios were tested: in Situation 1, AGV gradually approaches the target tray from the left side; in Situation 2, AGV gradually approaches the target tray from the right side; and in Situation 3, AGV gradually approaches the target tray from the front side.

The three-dimensional space points acquired through the camera's real-time image acquisition and positioning are typically coordinate values in the left camera coordinate system. To provide an intuitive representation of the data, the coordinate values of the initial position were added to all trajectory points to establish a spatial coordinate system. Additional, the coordinates of the initial position were used as the origin to establish a spatial coordinate system. The trajectories of the four situations are illustrated in Fig. 11, where (a), (b) and (c) exhibit good real-time positioning performance. Fig. 11(d) is the three trajectories of three mark points in (a). When the camera begins to move, mark point 1 fails to be detected due to illumination or distance, leading to a loss of the trajectory of mark point 1 in the initial segment (Fig. 11(d)). However, the proposed method is resilient and only requires detection of any two mark points, and thus this case can still achieve its final trajectory as depicted in Fig. 11(a), which proves the efficacy of the proposed method.

The detection and positioning speed of the algorithm in AGV continuous motion and positioning mode was by determining the effective number of frames processed per second using the sliding timing of the image sequence images as shown in Fig. 12. To accurately reflect the algorithm's processing efficiency, only the number of images that successfully calculated the relative position information of the pallet with respect to AGV forklift were counted, and those that were not were eliminated. As a result, an effective average frame rate was obtained in Fig. 12, indicating that the average effective frame rate per second exceeds 11 frames when AGV is in the continuous positioning state. The detection rate of YOLOV3 for small targets can theoretically be maintained at 30 frames per second, indicating that the proposed detection+positioning function has considerably fast processing speed in this study.

The dynamic positional accuracy in the X, Y and Z directions, as depicted in Fig. 13, is found to be better than 2.0 mm, with the Z-direction translation accuracy being slightly worse. The cameras employed in the measurement system primarily use an area array CCD and cylindrical mirror, which makes it easier to measure movement along the direction perpendicular to the cylindrical mirror's central axis with accuracy using the area array CCD. Therefore, measuring the component movement along the Z-axis direction is slightly less accurate than the other two translation vectors.

To summarize, the proposed algorithm provides highly accurate target parameters within a spatial range of 3 m $\times$ 3 m $\times$ 1.5 m. The results indicate that the proposed approach is effective and achieves superior comprehensive performance for target detection and localization.

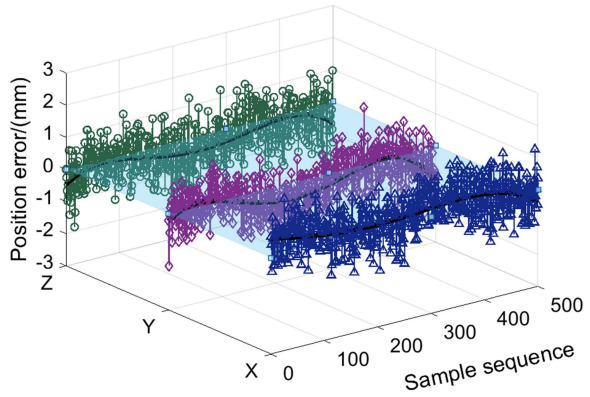


Fig. 13. Error of the target positioning.

V. Conclusion

The proposed approach in this study aims to achieve AGV terminal detection through cooperative target detection and position measurement. To achieve this, an enhanced deep learning algorithm with cross-layer connection is employed to detect the marker points' area of interest, followed by restoring more refined local features to improve the algorithm's adaptability to small targets. Moreover, a high-precision three-dimensional reconstruction technique is demonstrated using ellipse center coordinates to ensure the reliability of the targets and avoid the loss of landmarks due to environmental factors. The proposed algorithm is verified through both simulation and actual experiments, and the results show its effectiveness and superiority in achieving high positioning accuracy. The algorithm can meet the end positioning requirements of AGV and has practical engineering application value.

Future improvements to the algorithm's performance could involve incorporating context information into the network structure and training mode, as well as integrating timing information and convolution operations between adjacent frames. Furthermore, the algorithm's efficiency could be enhanced by implementing a multi-GPU parallel working mode, which would reduce the operation cycle of AGV robot.

V. Disclosures

The authors declare that they have no known competing financial interests or personal relationships that may have appeared to influence the work reported in this paper, and they declare no competing interests.

References

[1] G. H. Zhao et al., "Positioning error compensation for parallel mechanism with two kinematic calibration methods," *Chin. J. Aeronaut.*, vol. 33, no. 9, pp. 2472–2489, 2020.

[2] W. Guan, S. Chen, S. Wen, Z. Tan, H. Song, and W. Hou, "High-accuracy robot indoor localization scheme based on robot operating system using visible light positioning," *IEEE Photon. J.*, vol. 12, no. 2, Apr. 2020, Art. no. 7901716.

[3] Y. Zou et al., "Novel standardized representation methods for modular service robots," *Int. J. Social Robot.*, vol. 14, pp. 699–712, 2022.

[4] J. J. Zhu et al., "Underwater object recognition using transformable template matching based on prior knowledge," *Math. Problems Eng.*, vol. 2019, 2019, Art. no. 2892975.

[5] S. Zhong et al., "Behavior prediction for unmanned driving based on dual fusions of feature and decision," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3687–3696, Jun. 2021.

[6] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767 [cs.CV]*.

[7] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian, "Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1758–1770, Jun. 2020.

[8] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-squeeze-excitation fusion network for elderly activity recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5281–5292, Aug. 2022.

[9] C. L. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581, Jan. 2014.

[10] J. Han, K. Liang, B. Zhou, X. Zhu, J. Zhao, and L. Zhao, "Infrared small target detection utilizing the multiscale relative local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 612–616, Apr. 2018.

[11] Y. Wei et al., "Multiscale patch-based contrast measure for small infrared target detection," *Pattern Recognit.*, vol. 58, pp. 216–226, 2016.

[12] L. Wu, Y. Ma, F. Fan, M. Wu, and J. Huang, "A double-neighborhood gradient method for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 8, pp. 1476–1480, Aug. 2021.

[13] L. Z. Deng et al., "Infrared moving point target detection based on spatial–temporal local contrast filter," *Infrared Phys. Technol.*, vol. 76, pp. 168–173, 2016.

[14] M. L. Lu et al., "A novel Gaussian ant colony algorithm for clustering cell tracking," *Discrete Dyn. Nature Soc.*, vol. 2021, 2021, Art. no. 9205604.

[15] J. M. Du et al., "CNN-based infrared dim small target detection algorithm using target-oriented shallow-deep features and effective small anchor," *IET Image Process.*, vol. 15, no. 1, pp. 1–15, 2021.

[16] Y. Xia, J. Li, and L. F. Zhou, "A two-stage visual tracking algorithm using dual-template," *Int. J. Adv. Robotic Syst.*, vol. 13, no. 5, 2016, Art. no. 1729881416666797.

[17] X. Tian et al., "Deep multi-view feature learning for EEG-based epileptic seizure detection," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 1962–1972, Oct. 2019.

[18] Y. Chang and W. Wang, "A deep learning-based weld defect classification method using radiographic images with a cylindrical projection," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5018911.

[19] R. Wang et al., "Three-dimensional reconstruction of shoe soles via binocular vision based on improved matching cost," *Mathematics*, vol. 10, no. 19, 2022, Art. no. 3548.

[20] G. Y. Zhang et al., "Parallel optimization of tridimensional deformation measurement based on correlation function constraints of a multi-camera network," *Appl. Opt.*, vol. 61, no. 31, pp. 9225–9237, 2022.

[21] J. S. Cui et al., "A measurement method of motion parameters in aircraft ground tests using computer vision," *Measurement*, vol. 174, 2021, Art. no. 108985.

[22] H. Zhang, K.-Y. K. Wong, and G. Zhang, "Camera calibration from images of spheres," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 499–502, Mar. 2007.

[23] R. A. Boby and A. Klimchik, "Combination of geometric and parametric approaches for kinematic identification of an industrial robot," *Robot. Comput.-Integr. Manuf.*, vol. 71, 2021, Art. no. 102142.

[24] S. G. Liu, X. X. Song, and Z. H. Han, "High-precision positioning of projected point of spherical target center," *Opt. Precis. Eng.*, vol. 24, no. 8, pp. 1861–1870, 2016.

[25] Y. H. Li et al., "Algorithm of locating the sphere center imaging point based on novel edge model and Zernike moments for vision measurement," *J. Modern Opt.*, vol. 66, no. 2, pp. 218–227, 2019.

[26] J. K. Nichols et al., "A kinect-based movement assessment system: Marker position comparison to Vicon," *Comput. Methods Biomech. Biomed. Eng.*, vol. 20, no. 12, pp. 1289–1298, 2017.

[27] A. J. Tabatabai and O. R. Mitchell, "Edge location to subpixel values in digital imagery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 2, pp. 188–201, Mar. 1984.

[28] F. F. Gu et al., "Analysis and correction of projection error of camera calibration ball," *Acta Optica Sinica*, vol. 32, no. 12, 2012, Art. no. 1215001.

[29] W. P. Cao, R. S. Che, and D. Ye, "Estimation of the center of rotation and 3D motion parameters from stereo sequence images and virtual validation using three-COMERO," *Pattern Recognit. Lett.*, vol. 28, no. 13, pp. 1593–1599, 2007.

[30] L. Yang, B. Wang, R. Zhang, H. Zhou, and R. Wang, "Analysis on location accuracy for the binocular stereo vision system," *IEEE Photon. J.*, vol. 10, no. 1, Feb. 2018, Art. no. 7800316.