Mech. Sci., 16, 445–455, 2025
https://doi.org/10.5194/ms-16-445-2025

Mechanical Sciences
Open Access

Trajectory planning for manipulator grasping with obstacle avoidance and visual occlusion in a complex environment

Jue Wang[1], **Xiaoxiang Sun**[1], **and Wei Wang**[1,2]

[1]School of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China
[2]State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China

Correspondence: Wei Wang (wangw@zstu.edu.cn)

Received: 29 December 2024 – Revised: 6 May 2025 – Accepted: 19 June 2025 – Published: 10 September 2025

Abstract. Manipulators often face challenges of reliable trajectory planning under limited visibility. This paper proposes an obstacle avoidance grasping path planning method based on monocular vision. A multi-layer neural network model is to recognize target features, enabling rapid environmental perception. The environment topology is segmented using convex hulls, and multi-objective optimization is applied to achieve agile obstacle avoidance for the robot. The main contributions include the following: (1) an improved mask region convolutional neural network architecture is constructed to predict features under limited information, and (2) an innovative path planning strategy that combines Gaussian sampling with Hopfield neural networks is developed to improve the non-dominated sorting genetic algorithm. The algorithm achieves an identification accuracy of 99.50 % in complex scenarios, and the optimized trajectory shows improvements in smoothness and motion efficiency by approximately 60 % and 10 %, respectively. Through simulations and experiments, the significant effects on enhancing the robot's operability in complex environments have been validated.

1 Introduction

Vision servo-based robotic arm grasping systems are widely used in many industrial and agricultural fields, including logistics sorting, object grasping and fruit picking (Konstantinidis et al., 2023). Obstacles in randomly occluded scenes, such as naturally growing branches and leaves, and randomly placed packages bring challenges to the accurate grasping of target objects (Othman et al., 2018).

In visual recognition, Shang et al. (2014) proposed a monocular method for measuring the position of a translationally one-dimensional object containing at least two known feature points only. Zhang et al. (2023) developed a weighted voting integrated regression algorithm for predicting the dimensional deviation of the workpiece imaging. Pan et al. (2020) implemented object height measurement using a depth camera based on the Sobel operator and a low-pass filter, but it requires high hardware cost and computational overhead. The above studies mainly focus on single object geometric feature recognition. To further improve the recognition ability in complex scenes, Li et al. (2024) presented a canopy labeling method for U-Net and lightweight segmentation networks. The method introduces a lightweight backbone network. It greatly reduced the computational complexity required for large-scale canopy segmentation. For this purpose, Yoon and Han (2023) pointed out a box object detection method, which can realize depth information extraction in occluded scenes but with lower accuracy. Zhao et al. (2023) developed a 3D novel end-to-end parcel picking model designed to segment stacked objects hierarchically. Yu et al. (2019) constructed a strawberry fruit auto-detection model based on Mask R-CNN, demonstrating better generality and robustness when handling overlapping and hidden fruits. Ge et al. (2023) designed a visual strategy for grasping occluded objects, fusing color information and recoded depth information to improve accuracy in object segmentation.

In vision servo grasping and planning, Auh et al. (2024) applied a sequence planning method based on A* search al-

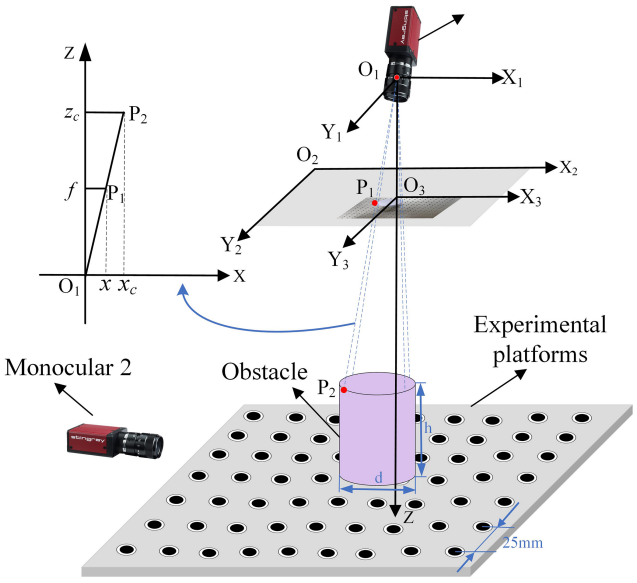


Figure 1. Monocular imaging schematic.

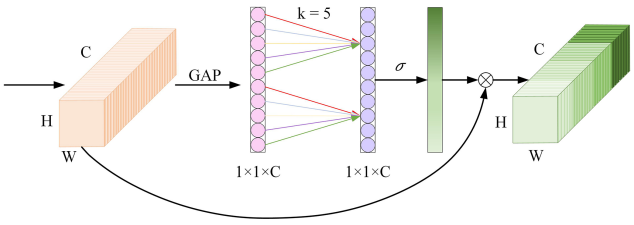


Figure 2. ECA module structure diagram.

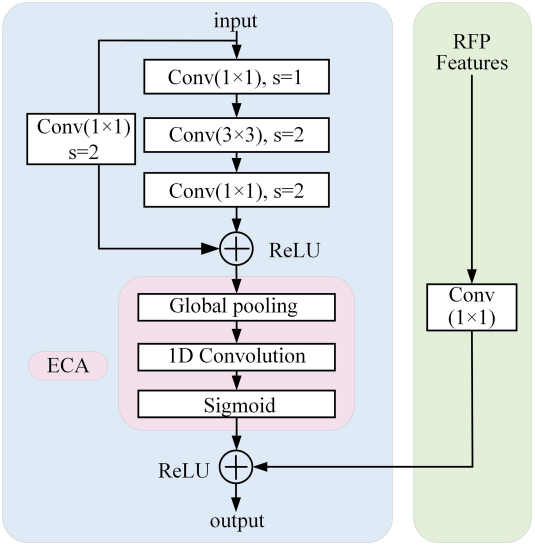


Figure 3. RFP add feature structure diagram.

gorithm applied to unload randomly stacked different boxes from a logistics container. Fan et al. (2022) established a PF-RRT* algorithm based on target bias strategy for UAV trajectory planning in cluttered environments. In Cong et al. (2020), based on the second generation of the non-dominated genetic algorithm (NSGA-II), the new population obtained from taboo search was added to the elite retention strategy to achieve the optimization of the solution efficiency. As the scenes of robotic applications become increasingly complex and unstructured, the demand for simultaneous optimization of multiple objectives (≥ 3) is increasing, while algorithms such as A* and RRT* are currently mostly used for single-objective solving, so path planning methods based on optimization strategies such as ant colony (Saenphon et al., 2014), genetic (Elhoseny et al., 2018), or particle swarm (Tavana et al., 2016) have been widely used.

In these methods, the third-generation non-dominated genetic algorithm (NSGA-III) is based on reference point selection by optimizing the selection and update mechanism of the population. The convergence efficiency of the algorithm and the global optimization solution search ability are improved. In the field of agriculture, Li et al. (2023) solved the path optimization problem based on the NSGA-III algorithm to achieve the goal of harvesting kiwi fruit. In the industrial field, Wang et al. (2021) employed the RRT* algorithm to search for collision-free paths and optimized the problem using the NSGA-III algorithm to solve a bi-objective path planning problem for arc welding. Xiao et al. (2020) combined the NSGA-III algorithm with a penalty function to solve a multi-objective optimization problem with constraints.

In conclusion, complex scene grasping under visual occlusion requires the integrated use of image recognition and motion planning. A deep neural network model is constructed for feature prediction with visual occlusion to realize accurate target localization. The Hopfield neural network and Gaussian sampling strategy are designed to optimize and improve the Non-dominated Sorting Genetic Algorithm (NSGA-III), enhancing the efficiency of path planning and the global optimization solving ability, which can realize autonomous target recognition and intelligent grasping in complex scenes.

2 Complex scene perception based on machine vision

Two monocular industrial cameras were used with an Improved Mask R-CNN to identify the geometric features of obstacles. The goal is to enhance machine vision's ability to perceive complex scenes.

2.1 Recognize and extract geometric features of obstacles

2.1.1 Extracting geometric features of obstacles

Features of obstacles are extracted using the holistically nested edge detection (HED) (Xie and Tu (2015)) instead of the traditional Canny edge detection algorithm due to the

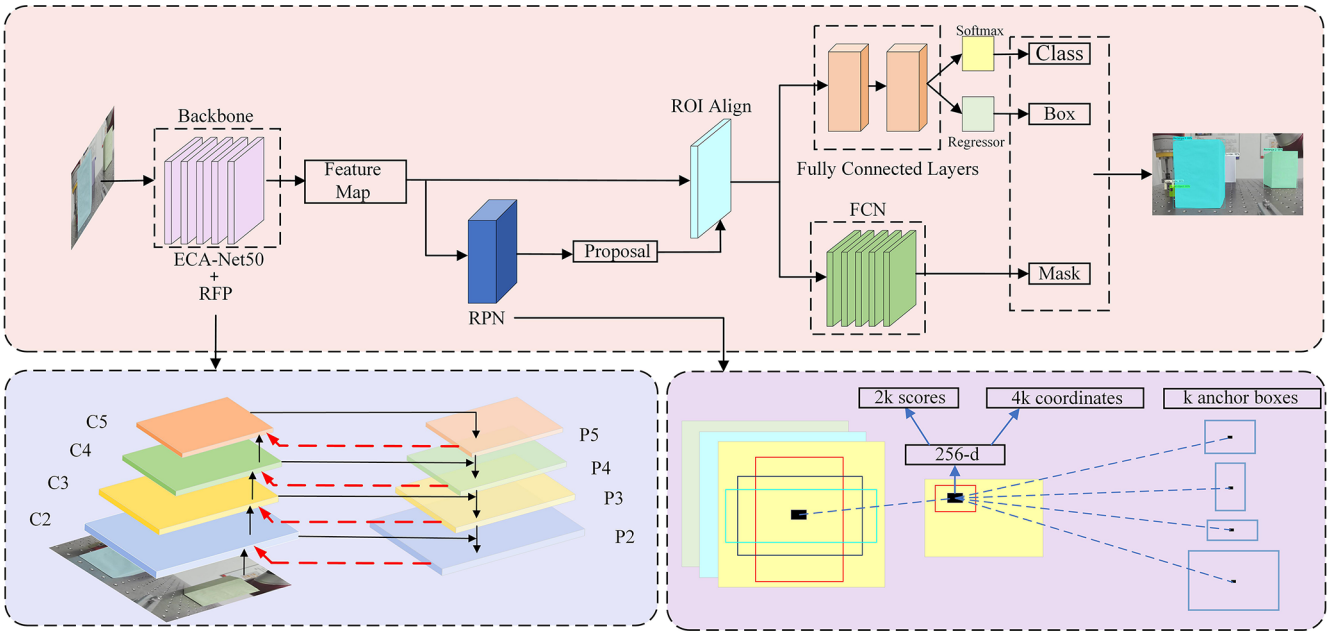


Figure 4. Improved Mask R-CNN framework.

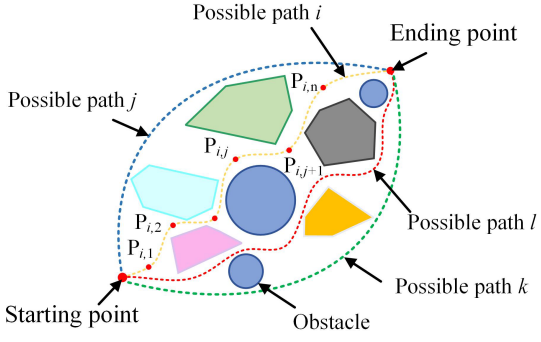


Figure 5. Schematic diagram of robotic arm running path.

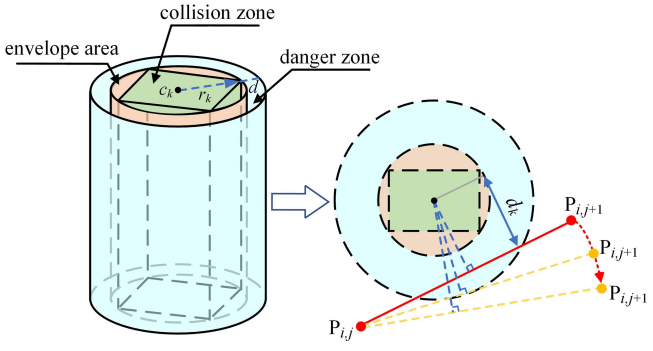


Figure 6. Collision risk diagram of a robotic arm.

better efficiency and accuracy. The contour of a rectangular body consists of scattered discrete points, and Hough transform is also used. The circle on the upper surface of the cylinder is generally elliptical in the image plane of the monocular camera. Ellipse fitting consists of two processes: one is the detection of the ellipse's edge points, and the other is ellipse fitting to the edge points to determine the center of the ellipse.

2.1.2 Monocular imaging principle

The geometric relationship of monocular industrial camera imaging is shown in Fig. 1, assuming that there exists any P_2 in the space and its projected position in the image is P_1. According to the similarity theorem of the right triangle, the following formula can be obtained:

$$x = \frac{x_c}{z_c} f \ , \quad y = \frac{y_c}{z_c} f, \tag{1}$$

where f is the focal length, x and y are the image coordinates of P_1, and x_c, y_c, and z_c are the coordinates of space P_2 under the coordinate system of the industrial video camera.

According to the camera imaging model, the projection transformation from the world coordinate system to the image coordinate system can be expressed by Eq. (2), where the image coordinates $[u, v, 1]^T$ are obtained by linear transformation of the inner reference matrix $\mathbf{A}$ and the outer reference matrix $[\mathbf{R}\ \boldsymbol{t}]$. The input is the $[X, Y, Z, 1]^T$ of the world coordinates and s is the scale factor.

$$s \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \mathbf{A}[\mathbf{R}\ \boldsymbol{t}] \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{2}$$

Table 1. Geometric information of the cuboid.

Cuboid	Actual (cm)	Measured (cm)	Error (%)
Length	10.00	10.01	0.1 %
Width	8.50	8.38	1.3 %
Height	12.03	11.97	0.4 %

Table 2. Geometric information of the cylinder.

Cylinder	Actual (cm)	Measured (cm)	Error (%)
Height	8.20	8.36	1.9 %
Diameter	12.10	11.88	1.8 %

The matrix **A** is then calibrated according to Zhengyou (1999).

2.2 Recognition of object visual occlusion

Mask R-CNN (He et al., 2017) is an instance segmentation algorithm. In this study, we proposed an Improved Mask R-CNN based on ECA-Net50 (Wang et al., 2020) and the Recursive Feature Pyramid (RFP) (Qiao et al., 2021), as shown in Fig. 4.

2.2.1 Backbone network ECA-Net50

ECA-Net50 uses ResNet50 as the backbone network framework and adds the Efficient Channel Attention (ECA) module after each residual block, as shown in Fig. 2.

Aggregated features are obtained by global average pooling (GAP). Channel weights are generated using a one-dimensional convolution operation with kernel size k, and weights are mapped to 0–1 intervals by the sigmoid function. The calculation can be represented in Eq. (3).

$$k = |t|_{\text{odd}} = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{\text{odd}}, \tag{3}$$

where $|t|_{\text{odd}}$ denotes the closest odd number to t, γ and b are set to 2 and 1, and C represents the number of channels of input features.

2.2.2 Recursive feature pyramid

RFP adds feedback connections that can be better adapted to the detection or segmentation tasks. The RFP adds the transformed features to the first block of each stage of ECA-Net50, as shown in Fig. 3.

The calculation process can be formulated as

$$f_i = F_i(f_{i+1}, x_i)x_i = B_i(x_{i-1}, R_i(f_i)), \tag{4}$$

where f_i is the feature map output from the current stage i, F_i denotes the top-down FPN operation, B_i represents the stage i of the bottom-up backbone network, and $R_i(f_i)$ indicates the feedback processing.

Table 3. Improved Mask R-CNN, U-Net and YOLOv8 comparison.

Algorithms	Precision	Recall	F1 score
Improved Mask R-CNN	99.50 %	97.00 %	98.23 %
U-Net	97.34 %	96.63 %	96.98 %
YOLOv8	89.73 %	85.48 %	87.55 %

3 NSGA-III algorithm based on Hopfield neural network with Gaussian sampling improvement (HG-NSGAIII)

This paper proposes an improved algorithm, HG-NSGAIII, based on the NSGA-III. The algorithm combines Hopfield neural network and Gaussian sampling to achieve global path planning and multi-objective optimization with more than three objectives, which improves the efficiency and accuracy of manipulator trajectory planning.

3.1 Construction of the optimization objective function

The safe operation of manipulators needs to take into account the mechanical performance, environmental characteristics and application requirements, but the increase of constraints and their mutual conflicts will increase the computational difficulty. For this reason, it is necessary to transform the environmental information and operation goals into executable objective functions to construct the optimization space. In this paper, we focus on three constraints: path length, obstacle threat and path smoothness.

3.1.1 Path length

The path length is a key factor that affects the indexes of robotic arm operation efficiency and energy consumption. The jth point on the ith path is denoted as $P_{i,j} = (x_{i,j}, y_{i,j}, z_{i,j})$, as shown in Fig. 5.

The sum of the lengths of all discrete path points is defined as the path cost function:

$$f_1 = \sum_{j=0}^{n-1} \| P_{i,j} \to P_{i,j+1} \|, \tag{5}$$

where $\| P_{i,j} \to P_{i,j+1} \|$ denotes the Euclidean distance between two neighboring path points $P_{i,j}$ and $P_{i,j+1}$, calculated as

$$\| P_{ij} \to P_{ij+1} \| = \sqrt{(x_{i,j+1} - x_{i,j})^2 + (y_{i,j+1} - y_{i,j})^2 + (z_{i,j+1} - z_{i,j})^2}. \tag{6}$$

3.1.2 Threats of obstacles

To measure the risk of collision, the obstacle is enclosed in a circle using the envelope method, with the center and radius of the circle denoted as c_k and r_k, as shown in Fig. 6.

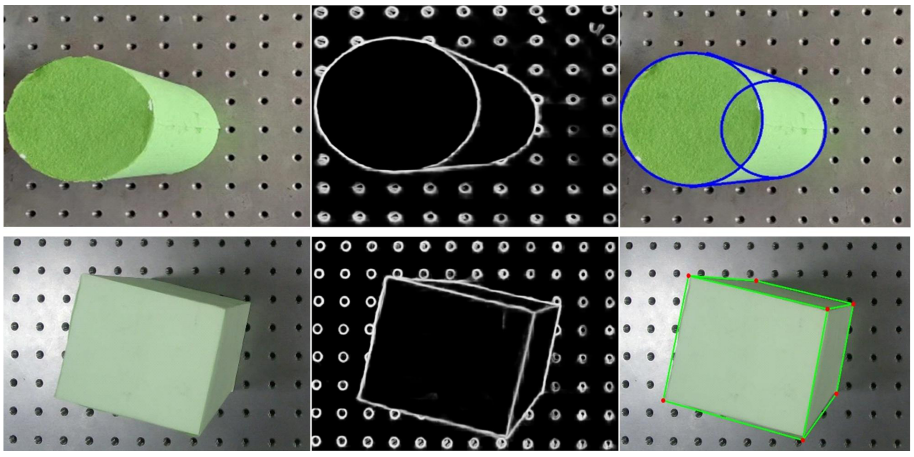


Figure 7. Feature extraction of the cylinder and cuboid.

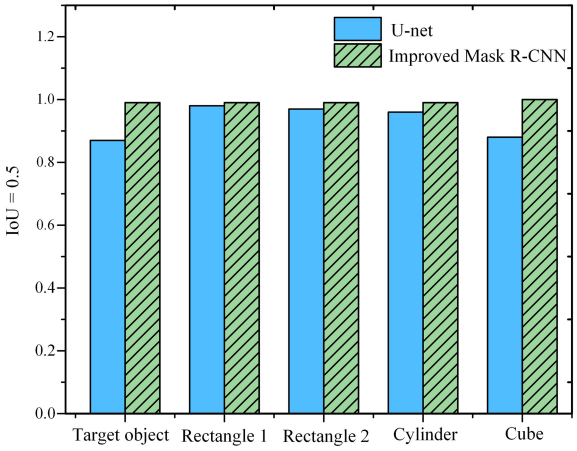


Figure 8. Comparison of Improved Mask R-CNN and U-Net.

The region with a radius of $r_k + d$ is defined as the dangerous region. When the trajectory of the manipulator moving from $P_{i,j}$ to the next path point $P_{i,j+1}$ traverses through the hazardous region, the manipulator is in possible danger, so obstacle threat cost can be described as

$$f_2 = \sum_{j=0}^{n-1} \sum_{k=1}^{m} T_k(P_{i,j} \rightarrow P_{i,j+1}), \tag{7}$$

where k is the number of obstacles and T_k is a segmented function on the distance d_k between the obstacles and local path $P_{i,j} \rightarrow P_{i,j+1}$ with the following expression:

$$T_k = \begin{cases} 0 & (d_k > d + r_k) \\ \gamma_c((d + r_k) - d_k) & (r_k < d_k \leq d + r_k) \\ +\infty & (d_k \leq r_k), \end{cases} \tag{8}$$

where $\gamma_c = 25$ represents the penalty coefficient.

3.1.3 Path smoothness

The degree of smoothing of the robotic arm path affects the operating efficiency of the robotic arm and the stability of the robotic arm. The combined force calculation function is

$$\boldsymbol{F}_{a+1,a} = \frac{P_{i,a+1} - P_{i,a}}{\| P_{i,a+1} - P_{i,a} \|} \tag{9}$$

$$f_3 = \sum_{j=1}^{n-1} (\boldsymbol{F}_{j+1,j} + \boldsymbol{F}_{j-1,j}), \tag{10}$$

where $\boldsymbol{F}_{a+1,a}$ is the result of normalizing the forces at the path point $P_{i,a}$ to the direction of the path point $P_{i,a+1}$.

3.2 HG-NSGAIII algorithm

This paper adopts Gaussian sampling strategy to generate initial paths. The sampling points are concentrated near the expansion points to reduce uncertainty and avoid local optimization. This optimizes the initial population of NSGA-III and improves efficiency of the algorithm.

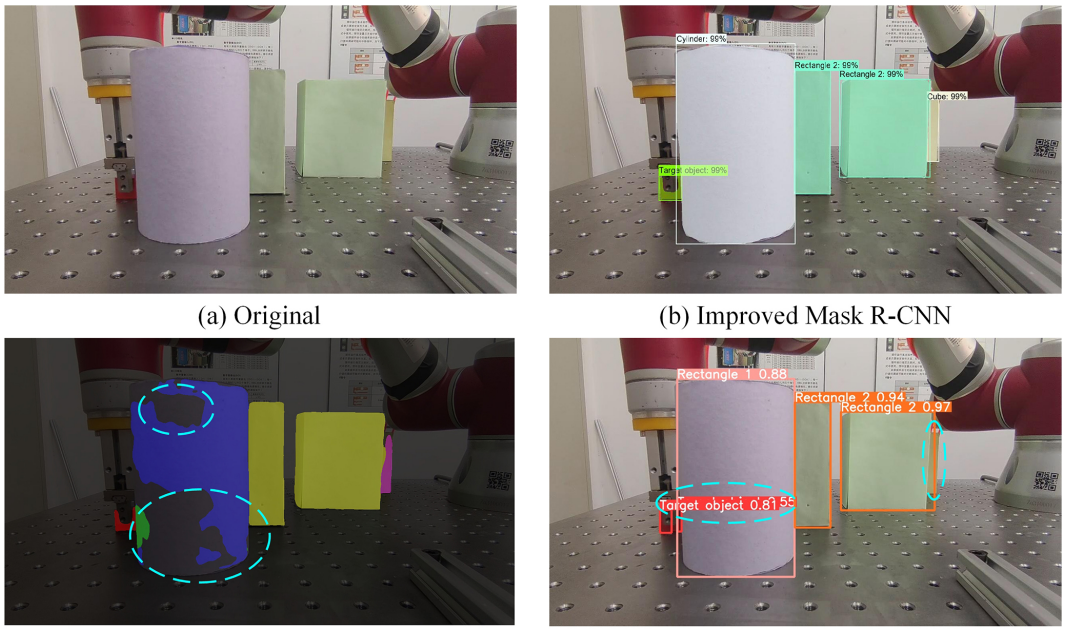
3.2.1 Specific flow of the HG-NSGA-III algorithm

Step 1: A monocular industrial camera is used to extract geometric features of obstacles to build a map of the working environment of the robotic arm.

Step 2: Establish the starting position of the robotic arm as the QS and the end as QE.

Step 3: A path planning algorithm based on Hopfield neural network algorithm is used to obtain a collision-free initial path P_{init}.

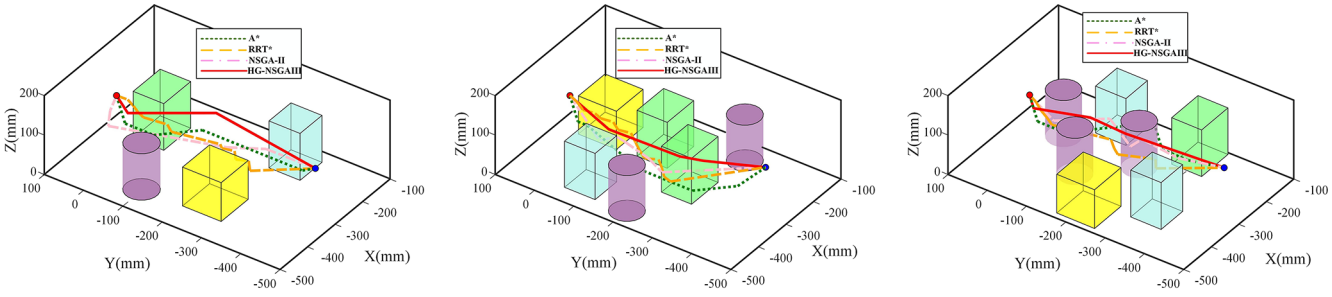
Step 4: The initial path is equally divided into N path points based on equal step length T and path length L_{init}.



(a) Original (b) Improved Mask R-CNN

(c) U-net (d) Yolov8

Figure 9. Image prediction with complex occlusion.



(a) Scene 1 (b) Scene 2 (c) Scene 3

Figure 10. 3D view of robotic arm trajectory planning in different scenes.

Step 5: A Gaussian sampling algorithm is used to generate $N \times$ num candidate points near N path points to form a candidate region.

Step 6: Randomly select the path points from each candidate region to generate an initial population P_t containing dim paths.

Step 7: The NSGA-III is used to perform multi-objective optimization on the initial population P_t in Step 6.

3.2.2 Nonlinear interpolation based on five-times B-splines

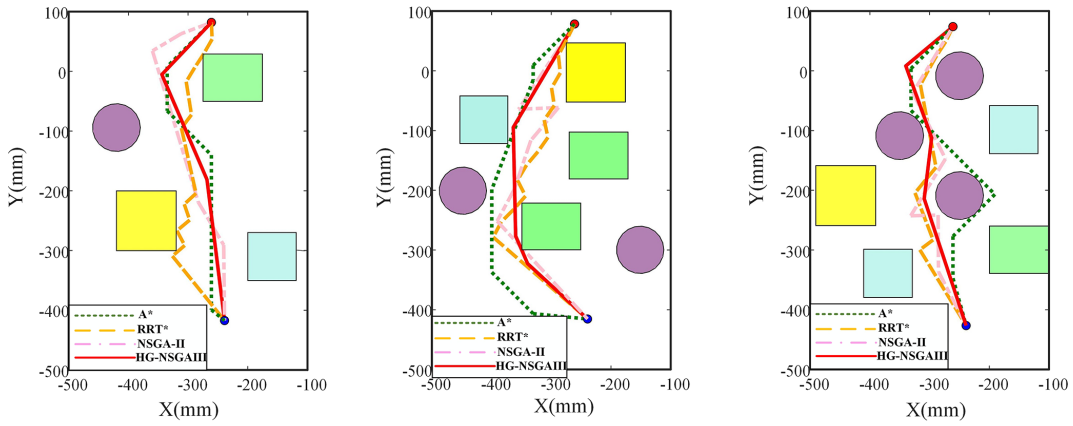
With the advantages of a high degree of freedom and good local modifiability, B-splines are widely used in robot path planning. The initial trajectory P_{optimize} is interpolated by fifth-order B-splines. Given $n+1$ control points $CP_0, CP_1, \cdots, CP_n$ and node sequence $u = \{u_0, u_1, \cdots, u_m\}$, the five-times B-spline function for the ith path can be constructed as follows:

$$P_{i,j} = \sum_{j=0}^{n} N_{j,5}(u) CP_j \quad u \in [0,1], \tag{11}$$

where $P_{i,j}$ is the path point and $N_{j,5}(u)$ is the B-spline basis function, which can be obtained by the Cox–de Boor recursive method.

At the start and end points, the velocity and acceleration are all 0, $v_{t_0} = 0$ and $v_{t_n} = 0$. The velocity is the first-order derivative of the position, and according to the B-spline theory, the first-order derivatives of the p-times B-spline curves are still p-1-times B-spline curves:

$$v_{i,j} = \frac{\mathrm{d}P_{i,j}}{\mathrm{d}u} = \sum_{j=0}^{n-1} N_{j+1,4}(u) D_j, \tag{12}$$



(a) Scene 1 (b) Scene 2 (c) Scene 3

Figure 11. Top view of robotic arm trajectory planning in different scenes.

Table 4. Solving path length and time and number of nodes.

Scenes	Methods	L_{t_i}/(mm)	φ_l	R_{t_i}/(s)	φ_t	N_{t_i}	φ_n
Scene 1	A*	570.73	3.18 %	10.5251	8.13 %	4	60 %
	RRT*	589.50	benchmark	5.9759	47.84 %	10	benchmark
	NSGA-II	584.50	0.84 %	11.4573	benchmark	4	60 %
	HG-NSGAIII	565.92	4.00 %	4.8901	57.32 %	2	80 %
Scene 2	A*	653.22	6.81 %	11.2631	benchmark	5	44.4 %
	RRT*	631.25	9.94 %	8.1350	27.77 %	9	benchmark
	NSGA-II	700.92	benchmark	8.9645	20.41 %	4	55.6 %
	HG-NSGAIII	585.40	16.48 %	5.6526	49.81 %	3	66.7 %
Scene 3	A*	617.99	benchmark	15.5786	benchmark	5	benchmark
	RRT*	568.39	8.03 %	7.1214	54.29 %	5	0 %
	NSGA-II	609.58	1.36 %	9.9561	36.09 %	5	0 %
	HG-NSGAIII	566.40	8.35 %	6.2613	59.81 %	3	40 %
Average	HG-NSGAIII	572.57	9.61 %	5.60	55.65 %	2.67	62.23 %

where D_j denotes the new control point, and the expression can be described as

$$D_j = \frac{p}{u_{j+6} - u_{j+1}}(CP_{j+1} - CP_j). \tag{13}$$

Similarly, its acceleration $\gamma_{i,j}$ can be further derived.

4 Simulation and experiment

4.1 Gripping simulation and emulation

4.1.1 Extraction of spatial features of obstacles

The contour features of the object are extracted, and the results are shown in Fig. 7. The geometric features of the object obtained by monocular measurement and vernier calipers are listed in Tables 1 and 2, respectively.

According to Tables 1 and 2, the error between the actual measurement value and the monocular measurement value is less than 2 %.

4.1.2 Training and prediction simulation under occlusion

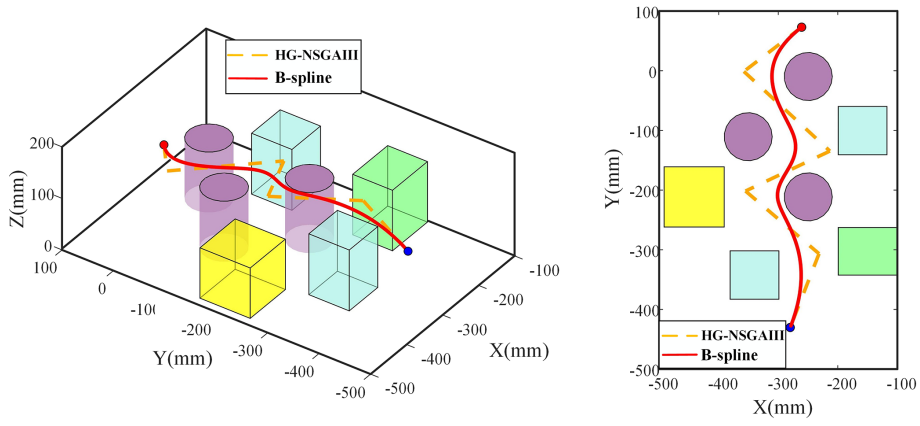
The data in this experiment contain 874 images, divided into training and validation sets according to 8 : 2. The image extension dataset is expanded by adding noise, increasing or decreasing brightness, rotating the image, etc. We use AutoDL, PyTorch 1.10 and Python 3.7 for training. After 300 rounds, the initial learning rate is 0.008, decreasing to 0.0008 and 0.00008 at the 60th and 200th rounds.

To validate the accuracy of the Improved Mask R-CNN, we compared it with U-Net and YOLOv8. Precision, recall and F1 score were used as evaluation metrics.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

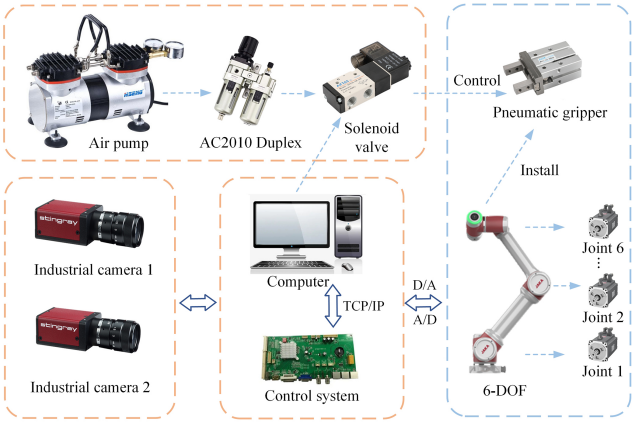
$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{16}$$

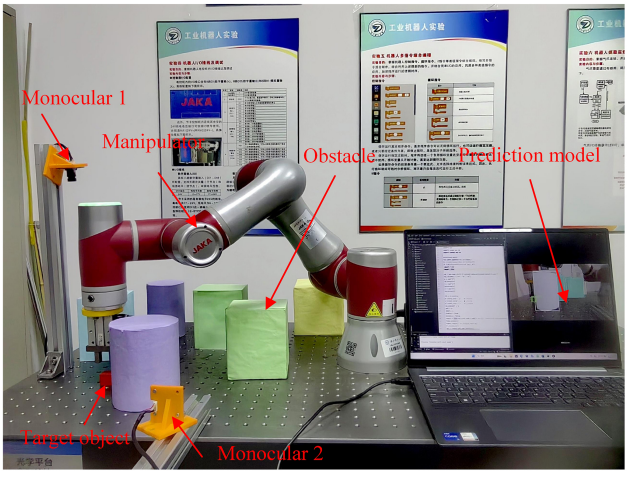


(a) 3D view (b) Top view

Figure 12. Further optimized trajectory.



(a) Diagram of communication and connection module



(b) Hardware system

Figure 13. Trajectory planning experiment platform.

where TP, FP and FN denote the sample size of true positive, false positive and false negative cases. The F1 score of the Improved Mask R-CNN is 1.25 % higher than U-Net and 10.68 % higher than YOLOv8, as listed in Table 3.

In the area of computer vision, IoU is an important evaluation metric commonly used in the task of image segmentation to measure the degree of overlap between the predicted results and the actual annotations.

$$\mathrm{IoU} = \frac{\mathrm{area}(U_\mathrm{p} \cap U_\mathrm{gt})}{\mathrm{area}(U_\mathrm{p} \cup U_\mathrm{gt})}, \tag{17}$$

where $\mathrm{area}(U_\mathrm{p} \cap U_\mathrm{gt})$ denotes the area of the intersection of the predicted mask and the real mask, and $\mathrm{area}(U_\mathrm{p} \cup U_\mathrm{gt})$ denotes the region of intersection. The results are shown in Fig. 8.

Three different scenarios are simulated. Taking scene 2 as an example, the predictions are compared using three visual recognition methods, and the results are shown in Fig. 9. Compared with U-Net and YOLOv8, the Improved Mask R-CNN has better adaptation to occlusion and higher accuracy with complete occlusion.

4.1.3 Autonomous generation of obstacle avoidance paths

The HG-NSGAIII algorithm is comprehensively compared with the A*, RRT* and NSGA-II to validate the algorithm's convergence efficiency and accuracy. Three different scenes are simulated, as shown in Figs. 10 and 11. In order to further evaluate the performance of each algorithm in terms of computational efficiency and accuracy, the optimization percentages of iteration time, path length and number of nodes are computed, respectively.

$$\begin{cases} \varphi_l = \dfrac{L_{\mathrm{t\_max}} - L_{t_i}}{L_{\mathrm{t\_max}}} \times 100\,\% \\ \varphi_t = \dfrac{R_{\mathrm{t\_max}} - R_{t_i}}{R_{\mathrm{t\_max}}} \times 100\,\% \qquad (i = 1, 2, \ldots, 4) \\ \varphi_n = \dfrac{N_{\mathrm{t\_max}} - N_{t_i}}{N_{\mathrm{t\_max}}} \times 100\,\%, \end{cases} \tag{18}$$

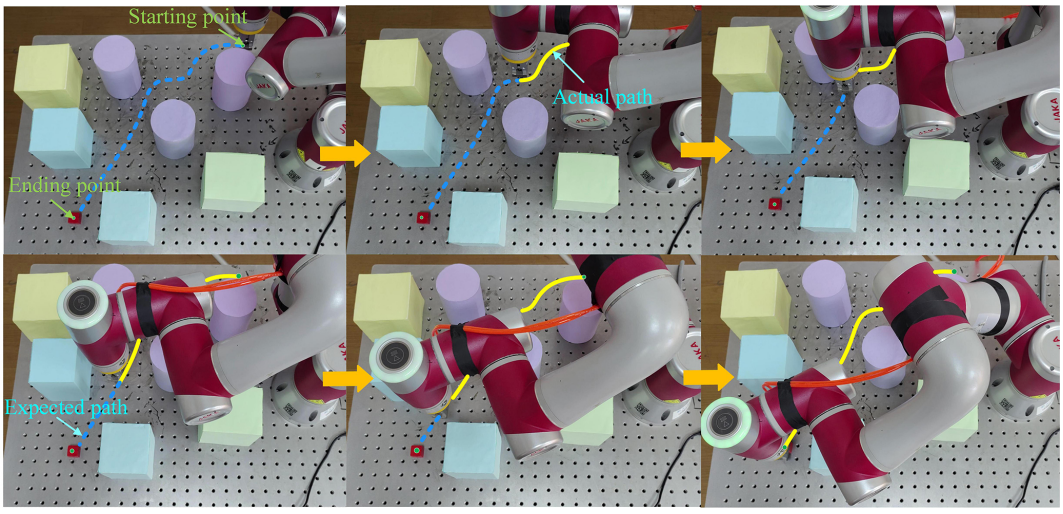


Figure 14. Obstacle avoidance trajectory planning.

where L_{t_i}, R_{t_i} and N_{t_i} represent path length, iteration time and the number of nodes optimized by the algorithms of this category. L_{t_max}, R_{t_max} and N_{t_max} symbolize the maximum value (benchmark of comparison) in which φ_l, φ_t and φ_n represent the percentage-related comparison data optimized by different algorithms, respectively, as listed in Table 4.

It can be seen that in the process of finding the global optimal solution, the HG-NSGAIII algorithm shortens the path length by 10 % on average compared with other algorithms and improves the computational efficiency by 55.65 %. The minimum number of inflection points is generated, which indicates that the HG-NSGAIII algorithm generates smoother and more continuous paths.

The initial optimal path generated under the constraints is characterized by poor geometric properties, and the trajectory often has sharp points that make it difficult to satisfy the C3 continuity. To improve the smoothness, the initial optimized path based on HG-NSGAIII is nonlinearly fitted with five-times B-splines, and the trajectory through the secondary optimization is shown in Fig. 12.

4.2 Experiment analysis

4.2.1 Construction of the experimental platform

In this paper, experiments were conducted based on a 6 DoF manipulator. The algorithm was programmed and operated on a computer, and the control systems of the manipulator arm and the computer were connected through the Internet. The air pump, AC2010 duplex and solenoid valve are connected to the pneumatic hand gripper. The solenoid valve is controlled by the computer to open and close the pneumatic hand gripper. The information is transmitted between the industrial camera and the computer, and the experimental platform is shown in Fig. 13a and b.

4.2.2 Autonomous capture of complex scenes

The experimental process of autonomous grasping by a robotic arm in a complex scene is shown in Fig. 14. The blue dashed line indicates the path not traveled by the robotic arm, and the yellow solid line indicates the path traveled by the robotic arm.

5 Conclusions

For target grasping scenarios with multiple occlusions, this paper proposes a trajectory planning method that combines the Improved Mask R-CNN and the HG-NSGAIII, and the main conclusions are as follows:

1. The monocular industrial camera extracts features of objects with different shapes. The error in the measured length is within 2 %, meeting the production requirements.

2. With the Improved Mask R-CNN algorithm, the accuracy of recognizing occluded objects in complex scenes is as high as 99.5 %, and the average accuracy of segmentation of various types of objects is 99 % (IoU = 0.5), which shows good accuracy and precision.

3. HG-NSGAIII has a strong global optimization solution search capability in complex scenes. Compared with RRT*, A* and NSGA-II, the NSGA-III algorithm based on the Gaussian sampling-improved Hopfield neural network is able to guarantee global convergence and obtain better optimized solutions. The motion efficiency and trajectory smoothness can be improved by about 9.61 % and 62.23 %, respectively.

Code and data availability. All data and codes used in this paper can be obtained on request from the corresponding author.

Supplement. The supplement related to this article is available online at https://doi.org/10.5194/ms-16-445-2025-supplement.

Author contributions. JW: writing (original draft), methodology. XS: conceptualization, methodology and software. WW: supervision, writing (review, editing) and funding acquisition.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. The authors thank the editors and reviewers for their efforts.

Financial support. This research has been supported by the National Natural Science Foundation of China (grant no. 52405041), the Key Research and Development Program of Zhejiang Province (grant no. 2023C01180), the National College Students Innovation and Entrepreneurship Training Program (grant no. 202410338048), the Zhejiang Provincial Xinmiao Talents Program (grant no. 2024R406A025), and the National Natural Science Foundation of China–Zhejiang Joint Fund for the Integration of Industrialization and Informatization (grant no. U23A20615).

Review statement. This paper was edited by Zi Bin and reviewed by two anonymous referees.

References

Auh, E., Kim, J., Joo, Y., Park, J., Lee, G., Oh, I., Pico, N., and Moon, H.: Unloading sequence planning for autonomous robotic container-unloading system using A-star search algorithm, Corros. Eng. Sci. Techn., 50, 101–610, https://doi.org/10.1016/j.jestch.2023.101610, 2024.

Cong, R., Qi, J., Wu, C., Wang, M., and Guo, J.: Multi-UAVs Cooperative Detection Based on Improved NSGA-II Algorithm, in: 39th Chinese Control Conference, 27–29 July 2020, Shenyang, China, 1524–1529 https://doi.org/10.23919/CCC50068.2020.9188354, 2020.

Elhoseny, M., Tharwat, A., and Hassanien, A. E.: Bezier Curve Based Path Planning in a Dynamic Field using Modified Genetic Algorithm, J. Comput. Sci.-Neth., 25, 339–350, https://doi.org/10.1016/j.jocs.2017.08.004, 2018.

Fan, J., Chen, X., Wang, Y., and Chen, X.: UAV trajectory planning in cluttered environments based on PF-RRT* algorithm with goal-biased strategy, Eng. Appl. Artif. Intel., 114, 105182, https://doi.org/10.1016/j.engappai.2022.105182, 2022.

Ge, J., Mao, L., Shi, J., and Jiang, Y.: Fusion-Mask-RCNN: Visual robotic grasping in cluttered scenes, Multimedia Tools and Applications, Multimed. Tools. Appl., 83, 20953–20973, https://doi.org/10.1007/s11042-023-16365-y, 2023.

He, K., Gkioxari, G., Dollár, P., and Girshick, R.: Mask R-CNN, in: IEEE International Conference on Computer Vision, 22–29 October 2017, Venice, Italy, 2980–2988, https://doi.org/10.1109/ICCV.2017.322, 2017.

Konstantinidis, F. K., Sifnaios, S., Arvanitakis, G., Tsimiklis, G., Mouroutsos, S. G., Amditis, A., and Gasteratos, A.: Multi-modal sorting in plastic and wood waste streams, Resour. Conserv. Recy., 199, 107244, https://doi.org/10.1016/j.resconrec.2023.107244, 2023.

Li, X., Lv, H., Zeng, D., and Zhang, Q.: An Improved Multi-Objective Trajectory Planning Algorithm for Kiwifruit Harvesting Manipulator, IEEE Access, 11, 65689–65699, https://doi.org/10.1109/access.2023.3289207, 2023.

Li, Z., Deng, X., Lan, Y., Liu, C., and Qing, J.: Fruit tree canopy segmentation from UAV orthophoto maps based on a lightweight improved U-net, Comput. Electron. Agr., 217, 108538, https://doi.org/10.1016/j.compag.2023.108538, 2024.

Othman, N. A., Salur, M. U., Karakose, M., and Aydin, I.: An Embedded Real-Time Object Detection and Measurement of its Size, in: 2018 International Conference on Artificial Intelligence and Data Processing, 28–30 September 2018, Malatya, Turkey 1–4, https://doi.org/10.1109/IDAP.2018.8620812, 2018.

Pan, Z., Jia, Z., Jing, K., Ding, Y., and Liang, Q.: Manipulator Package Sorting and Placing System Based on Computer Vision, in: Chinese Control And Decision Conference, 22–24 August 2020, Hefei, China, 409–414, https://doi.org/10.1109/CCDC49329.2020.9164071, 2020.

Qiao, S., Chen, L. C., and Yuille, A.: DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 20–25 June 2021, Nashville, TN, USA, 10208–10219, https://doi.org/10.1109/CVPR46437.2021.01008, 2021.

Saenphon, T., Phimoltares, S., and Lursinsap, C.: Combining new Fast Opposite Gradient Search with Ant Colony Optimization for solving travelling salesman problem, Eng. Appl. Artif. Intel., 35, 324–334, https://doi.org/10.1016/j.engappai.2014.06.026, 2014.

Shang, Y., Liu, J., Xie, T., and Yao, L.: A monocular pose mea- surement method of a translation-only one-dimensional objectwithout scene information, Optik, 125, 4051–4056, https://doi.org/10.1016/j.ijleo.2014.01.115, 2014.

Tavana, M., Li, Z., Mobin, M., Komaki, M., and Teymourian, E.: Multi-objective control chart design optimization using NSGA-III and MOPSO enhanced with DEA and TOPSIS, Expert. Syst. Appl., 50, 13–97, https://doi.org/10.1016/j.eswa.2015.11.007, 2016.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q.: ECA-Net: Efficient Channel Attention for Deep Convolu- tional Neural Networks, in: IEEE/CVF Conference on Computer Vision and Pat-

tern Recognition, 13–19 June 2020, Seattle, WA, USA, 11531–11539, https://doi.org/10.1109/CVPR42600.2020.01155, 2020.

Wang, X., Wei, J., Zhou, X., Xia, Z., and Gu, X.: Dual-Objective Collision-Free Path Optimization of Arc Welding Robot, IEEE Robot. Autom. Let., 6, 6353–6360, https://doi.org/10.1109/LRA.2021.3092267, 2021.

Xiao, J., Zhu, Z., Hu, X., Zhang, G., and Liu, L.: Research on payload distribution of UAV formation with constraints, in: 15th IEEE Conference on Industrial Electronics and Applications, 9–13 November 2020, Kristiansand, Norway, 1837–1842, https://doi.org/10.1109/ICIEA48937.2020.9248214, 2020.

Xie, S. and Tu, Z.: Holistically-Nested Edge Detection, in: IEEE International Conference on Computer Vision, 7–13 December 2015, Santiago, Chile, 1395–1403, https://doi.org/10.1109/ICCV.2015.164, 2015.

Yoon, J., Han, J., and Nguyen, T. P.: Logistics box recognition in robotic industrial de-palletising procedure with systematic RGB-D image processing supported by multiple deep learning methods, Eng. Appl. Artif. Intel., 123, 106311, https://doi.org/10.1016/j.engappai.2023.106311, 2023.

Yu, Y., Zhang, K., Yang, L., and Zhang, D.: Fruit detection for strawberry harvesting robot in non-structural environme- nt based on Mask-RCNN, Comput. Electron. Agr., 163, 104846, https://doi.org/10.1016/j.compag.2019.06.001, 2019.

Zhang, T., Zheng, J., and Zou, Y.: Weighted voting ensemble method for predicting workpiece imaging dimensional deviation based on monocular vision systems, Opt. Laser Technol., 159, 109012, https://doi.org/10.1016/j.optlastec.2022.109012, 2023.

Zhao, Y., Yang, J., Wang, S., and Li, X.: Towards One Shot & Pick All: 3D-OAS, an end-to-end framework for vision guided top-down parcel bin-picking using 3D-overlapping – aware instance segmentation and GNN, Robot. Auton. Syst., 167, 104491, https://doi.org/10.1016/j.robot.2023.104491, 2023.

Zhengyou, Z.: Flexible camera calibration by viewing a plane from unknown orientations, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, 20–27 September 1999, Kerkyra, Greece, 1, 666–673, https://doi.org/10.1109/ICCV.1999.791289, 1999.