

An Integrated Autonomous Grasping Method for Robots Combining Machine Vision and Manipulator Control

Na Wang^{1*}, Biao Zhang²

¹Henan Polytechnic Institute, Nanyang 473000, China

²Henan Technician Institute of Industry and Information Technology, Xinzheng 451150, China

Received 15 Sep 2025

Accepted 12 Feb 2026

Abstract

In the context of the transformation and upgrading of the global manufacturing industry, automation and intelligence have become inevitable trends in industry development. Robot autonomous grasping technology based on machine vision and robotic arm, as a new tool that integrates modern information technology and advanced manufacturing technology, can greatly improve the flexibility, adaptability and production efficiency of production lines. This study focuses on the key issues in this field and proposes an innovative autonomous grasping scheme, aiming to overcome the limitations of traditional methods in terms of real-time, accuracy and environmental adaptability. This project uses deep convolutional neural networks and the improved YOLOv5 model to work together to achieve effective detection and accurate identification of targets in complex environments. By training more than 20,000 sample libraries containing objects of various sizes and shapes, the system achieves an average recognition rate of more than 98%, and the false alarm rate is less than 3%, which greatly improves the target positioning accuracy. At the same time, a reinforcement learning mechanism is introduced to automatically generate the optimal grasping trajectory to ensure that the robotic arm can complete the task stably and efficiently. In addition, a rapid response mechanism has been established for dynamically changing work scenarios, which enables the equipment to adjust its strategy within 1 second, effectively respond to emergencies, and significantly enhance the robustness and practicality of the system. The experimental results show that under standardized test conditions, the new scheme reduces the grasping failure rate by about 25% compared with the traditional scheme and improves the overall operation efficiency by more than 17%, demonstrating its practical value.

© 2026 Jordan Journal of Mechanical and Industrial Engineering. All rights reserved

Keywords: Machine vision, Robotic arm, Neural network, YOLOv5, Autonomous grasping.

1. Introduction

With the rapid development of science and technology in today's era, robot technology, as an important part of intelligent manufacturing and automation, is changing our production and lifestyle at an unprecedented speed [1, 2]. Among them, robot autonomous grasping technology based on the combination of machine vision and robotic arm is one of the key points to achieve key technological breakthroughs in Industry 4.0, intelligent logistics, service robots and other fields. It can not only improve production efficiency and accuracy but also reduce costs. Reduce manpower requirements and improve safety. It has a wide application prospect [3].

With the development of artificial intelligence technology, especially the progress of deep learning, computer vision and other fields, the perception ability of robots has been significantly enhanced, which enables robots to autonomously identify target objects, judge their positions and postures, and plan the best grasping path in

complex and changeable environments [4, 5]. This process requires the robot to have high-precision visual recognition ability and flexible mechanical execution ability, as well as seamless collaborative work between the two, which poses extremely high challenges to the design of the algorithm [6].

However, the current technology still has some difficulties: how to accurately identify and locate targets in unstructured or semi-structured environments [7]; How to design an efficient and robust grabbing strategy to complete the task while ensuring security; And how to optimize the visual information processing process, reduce the consumption of computing resources, and achieve real-time requirements. These problems limit the popularization and popularization of this technology in practical applications.

The purpose of this paper is to deeply explore the robot's autonomous grasping method based on the combination of machine vision and a manipulator arm. By analyzing the research status and development trend of existing technologies, we propose innovative solutions to the above problems, focusing on how to use advanced image

* Corresponding author e-mail: nyhngywn@163.com.

processing technology and mechanical control theory to build a complete and efficient robot autonomous grasping system. We hope that through the research of this paper, we can contribute to the technological development of this field, promote technological innovation and industrial upgrading of related industries, and ultimately realize a more intelligent and automated production and service model.

The existing research on autonomous grasping based on machine vision generally has three limitations: insufficient robustness in recognition and localization in complex dynamic environments; The disconnection between visual perception, decision planning, and mechanical control modules results in high system latency and poor flexibility. The generalization ability of the grasping strategy is weak, making it difficult to quickly adapt to new objects or scenes. The innovative solution proposed in this study has achieved fundamental breakthroughs in system architecture and algorithm core, and its novelty and significant differences are reflected in: improving the integrated framework of YOLOv5 deep network and reinforcement learning deep coupling, realizing end-to-end real-time decision-making from images to grasping actions, and compressing system response time to within 1 second; Built a fast and dynamic response mechanism that integrates online learning, enabling the system to adjust strategies in real-time based on visual feedback, effectively responding to sudden disturbances such as occlusion and displacement. By training on a large-scale diversified sample library and using a refined point cloud pose estimation network (PFFnet), the grasping generalization ability of unknown objects has been significantly improved while ensuring high recognition rates (>98%). Experimental results have shown that this approach significantly outperforms existing methods in terms of grasping success rate, environmental adaptability, and operational efficiency.

2. THEORIES RELATED TO MACHINE VISION TECHNOLOGY

2.1. Composition of machine vision system

Among many input and output devices of modern computer systems, image input is regarded as one of the most critical information providers. When the computer system receives the graphic information, it needs to extract valuable information from it for analysis and perform corresponding operations. In recent years, with the rapid development of computer technology, digital image processing and multimedia technology have also made remarkable progress, which has greatly improved the processing capacity [8, 9]. In military applications, machine vision technology can help analyze and locate targets, significantly enhance the power of weapons, and improve soldiers' combat capabilities. Whether in many aspects of industry or daily life, the importance of machine vision systems has gradually been recognized and valued by us [10]. In the industrial field, machine vision systems can help analyze and identify various target objects in the environment and accurately locate them, compare different target environments, and extract key information from them so as to complete various tasks in all aspects of the industry, such as assembly, classification and scene analysis [11, 12].

In our daily lives, machine vision technology can also help identify all kinds of useful information, such as face recognition technology [13].

The machine vision system has obvious advantages, especially because they can avoid touching while measuring objects, which ensures the safety of the measured objects and will not cause any form of on-site damage. Compared with other contact measurement methods, this outstanding advantage is also a key factor for the continuous development of machine vision systems [14]. For humans, our visual range is limited, but machine vision technology can greatly expand our observation capabilities, such as using ultrasonic, microwave and infrared technologies. We can use certain highly sensitive measurement equipment to complete these measurement tasks to significantly expand our observation range and provide us with information that is difficult to obtain in daily life. In addition, in many harsh environments or environments requiring long-term observation, machine vision can give full play to its functions, enabling machines to make long-term and accurate observations without being affected by subjective factors, thus providing us with reliable observation information.

2.2. Image feature extraction

In the autonomous grasping task of the robot, efficient image feature extraction is the prerequisite to ensure the accurate recognition of object position and pose. In this study, the deep convolutional neural network is used for end-to-end image feature extraction [15, 16]. CNN can automatically capture hierarchical spatial features in images by virtue of its local receptive field and weight-sharing mechanism, effectively reducing the influence of background noise and improving target detection accuracy. The image feature extraction framework is shown in Figure 1. The pre-trained ResNet-50 model is utilized as the infrastructure, and the attention mechanism is fused to highlight key regions and accelerate the localization process [17, 18]. Combined with the Faster R-CNN framework, the direct mapping from the original image to the target frame coordinates is completed. This method uses the candidate region proposal network (RPN) to generate the region of interest, ensures the consistent size through the ROI Pooling layer, and sends it to the classifier and bounding box regressor for final judgment, which greatly improves the detection accuracy and speed.

Then add 3D point cloud processing technologies such as PointNet ++ or VoteNet to analyze the three-dimensional coordinates and rotation angles of objects. By densely sampling the point cloud data of multi-view fusion, it enhances the ability to understand targets in occluded environments and assists the robotic arm in achieving refined capture. Comprehensive use of the above technologies to form a complete set of image processing links: first, CNN completes preliminary filtering, and then R-CNN refines positioning. Finally, the most suitable capturing points and methods are estimated based on the point cloud information.

This study adopts the Proximal Policy Optimization (PPO) algorithm as the core reinforcement learning framework, and designs a multi-objective weighted reward function that includes constraints on grasping success,

distance proximity, posture matching, time efficiency, and obstacle avoidance. Joint and workspace constraints, velocity constraints, and safety distance are considered as hard conditions. Its innovation lies in the deep integration with the visual system: the improved YOLOv5 detection network and PFFnet point cloud pose estimation module provide real-time three-dimensional position and pose features of the target as state inputs for PPO agents. The intelligent agent outputs continuous joint control instructions for the robotic arm based on this, forming an end-to-end learning and decision-making loop from raw visual information to final grasping actions, thereby achieving efficient and adaptive strategy optimization in a dynamic environment.

2.3. Dataset construction and enhancement strategies

This study independently constructed a dedicated dataset for industrial grasping scenarios, containing 20000 samples, each sample containing synchronized RGB-D images, pixel level masks, precise 6-degree-of-freedom poses, and annotated grasping boxes. To enhance the robustness of the model in complex dynamic environments, the system implemented multi-level data augmentation: at the geometric space level, random cropping, scaling, rotation ($\pm 15^\circ$), and flat shifting were applied. At the photometric level, adjust brightness, contrast, saturation, and add noise; Fuse the target with diverse backgrounds using the CutMix strategy, and randomly add rectangular

occlusion blocks to simulate local occlusion; Simultaneously perform consistency affine transformation and noise disturbance on the depth map. This enhancement strategy greatly expands the effectiveness and diversity of data, and is the key foundation for the proposed improved YOLOv5 and PFFnet networks to achieve high accuracy (>98%) and strong generalization ability.

In the fields of computer vision and robotics, CNN (Convolutional Neural Network) is the core feature extraction architecture, with a pre-trained ResNet-50 serving as its backbone in this study. For target detection and localization, an improved YOLOv5 model is employed, whose architecture includes the CSPDarknet backbone network, the SPP (Spatial Pyramid Pooling) module, FPN (Feature Pyramid Network), and the detection head. Comparative models include YOLOv4, Faster R-CNN, and its sub-module RPN (Region Proposal Network). For pose estimation, PFFnet (Point Cloud Fractal Pyramid Network) is utilized and compared with models such as DenseFusion, PoseCNN, PVNet, and PointFusion. For grasp detection, GR-CNN (Grasp Detection Convolutional Neural Network) and the RCAN (Residual Channel Attention Network) module are used. In reinforcement learning, the PPO (Proximal Policy Optimization) algorithm is adopted. Evaluation metrics include mAP (mean Average Precision), ADD (Average Distance of Model Points), ADD-S (Average Distance of Model Points - Symmetric), IOU (Intersection over Union), and AUC (Area Under the Curve). The Mosaic method is used for data augmentation.

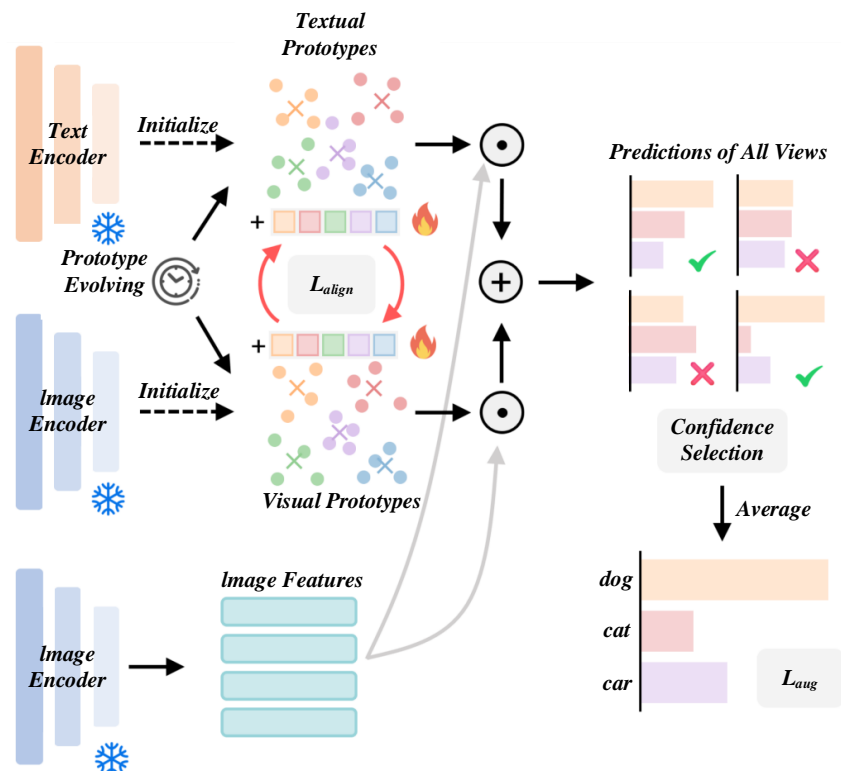


Figure 1. Image feature extraction framework

3. FUNDAMENTALS OF MECHANICAL ARM TECHNOLOGY

3.1. Structure and kinematics principle of manipulator

Modeling and kinematics analysis of a six-degree-of-freedom manipulator is described by connecting rod parameters, and forward and inverse kinematics models are established according to kinematics parameters [19, 20]. Knowing the joint motion angle and connecting rod size, it is necessary to establish the coordinate system transformation of adjacent connecting rods first and then carry out homogeneous matrix transformation to obtain the relative position posture.

${}^{i-1}_i T$ represent the homogeneous transformation matrix of link i relative to the previous link $i-1$, as shown in equation (1):

$${}^{i-1}_i T = \begin{bmatrix} c\theta_i & -s\theta_i & 0 & a_{i-1} \\ s\theta_i c\alpha_{i-1} & c\theta_i c\alpha_{i-1} & -s\alpha_{i-1} & -s\alpha_{i-1} d_i \\ s\theta_i s\alpha_{i-1} & c\theta_i s\alpha_{i-1} & c\alpha_{i-1} & c\alpha_{i-1} d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

Where $c\theta_i$ represents $\cos\theta_i$, $s\theta_i$ represents $\sin\theta_i$, and each connecting rod matrix is multiplied to obtain ${}^0_6 T$, is the joint angle, and θ_i represents the rotation angle between connecting rod $i-1$ and connecting rod i about the axis z_{i-1} . α_i is the torsion angle, which represents the rotation angle around the x_i axis between the connecting rod $i-1$ and the connecting rod i . d_i is the connecting rod offset, which represents the distance from the x_{i-1} axis to the x_i axis along the z_{i-1} axis. To represent the pose matrix of the end effector relative to the base coordinate system, formula (2) is obtained:

$${}^0_6 T = {}^0_1 T {}^1_2 T {}^2_3 T {}^3_4 T {}^4_5 T {}^5_6 T = \begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2)$$

n_x is the component of the normal vector n in the x -axis direction. o_x is the component of the direction vector o in the x -axis direction. p_x is the component of the position vector p in the x -axis direction. When the values in equation (2) are known, the following equation (3) is obtained:

$${}^0_6 T = \begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} = {}^0_1 T \frac{1}{2} T {}^2_3 T {}^3_4 T {}^4_5 T {}^5_6 T \quad (3)$$

Through Solving θ_i , the medium element ${}^1_6 T$ is obtained, and formula (4) is obtained:

$$-s_1 p_x + c_1 p_y = 0 \quad (4)$$

s_1 represents the first term of the sequence or the sum of the first term. c_1 is the coefficient of the first term of the sequence. By performing trigonometric identity, formula (5) can be obtained:

$$\theta_1 = \text{Atan2}(p_y, p_x) \quad (5)$$

The Atan2 function is a function in mathematics that calculates the arctangent, which takes into account the sign of the input value to determine the quadrant of the return value. Solve θ_3 , and after obtaining θ_i , obtain equation (6):

$$-a_1 + c_1 p_x + s_1 p_y = a_3 c_{23} - d_4 s_{23} + a_2 c_2 \quad (6)$$

$$p_z = a_3 s_{23} + d_4 c_{23} + a_2 s_2$$

Using the same trigonometric identity transformation, the solution of θ_3 can be obtained, as shown in equation (7):

$$\theta_3 = \text{Atan}(a_3, d_4) - \text{Atan}(K, \pm \sqrt{a_3^2 + d_4^2 - K^2}) \quad (7)$$

K represents the variable, and θ_3 has two different solutions. When solving θ_2 , according to equation (3), let θ_2 and all functions on the left become known, as shown in equation (8).

$$[{}^0_3 T] {}^3_{10} T = {}^3_4 T {}^4_5 T {}^5_6 T = {}^3_6 T \begin{bmatrix} c_{23} c_1 & c_{23} s_1 & s_{23} & -a_1 c_{23} - a_2 c_3 & n_x & o_x & a_x & p_x \\ -s_{23} c_1 & -s_{23} s_1 & c_{23} & a_1 s_{23} + a_2 s_3 & n_y & o_y & a_y & p_y \\ s_1 & -c_1 & 1 & 0 & n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} = {}^1_6 T \quad (8)$$

n_y is the component of the normal vector n in the y -axis direction. o_y is the component of the direction vector o in the y -axis direction. p_y is the component of the position vector p in the y -axis direction. n_z is the component of the normal vector n in the z -axis direction. o_z is the component of the direction vector o in the z -axis direction. p_z is the component of the position vector p in the z -axis direction. From the two solutions of θ_3 , two values of θ_{23} are calculated, and then two possible solutions of θ_2 are calculated, as shown in Equation (9).

$$\theta_2 = \theta_{23} - \theta_3 \quad (9)$$

By solving θ_4 , formula (10) can be obtained:

$$c_{23} c_1 a_x + c_{23} s_1 a_y + s_{23} a_z = c_4 s_5 \quad (10)$$

$$s_1 a_x + c_1 a_y = s_4 s_5$$

As long as $\theta_3 \neq 0$, θ_4 can be solved, and formula (11) can be obtained:

$$\theta_4 = \text{Atan2}(s_1 a_x - c_1 a_y, c_{23} c_1 a_x + c_{23} s_1 a_y + s_{23} a_z) \quad (11)$$

Solve θ_5 so that the formulas are all functions of θ_4 and other known quantities, that is, as in equation (12):

$$[{}^0_4 T] {}^4_{10} T = {}^4_5 T {}^5_6 T = {}^4_6 T \quad (12)$$

T stands for the variable. Therefore, θ_5 can be obtained, as shown in Equation (13):

$$\theta_5 = \text{Atan2}(s_5, c_5) \quad (13)$$

s_5 represents the sum of the first five terms of the sequence, and c_5 is the coefficient of the fifth term of the sequence. Solving θ_6 and applying the above method again, one can calculate ${}^0_5 T$, from which θ_6 can be obtained, as in Equation (14):

$$\theta_6 = \text{Atan2}(s_6, c_6) \quad (14)$$

s_6 represents the sum of the first six terms of the sequence, and c_6 is the coefficient of the sixth term of the sequence. To sum up, all feasible solutions have been found for $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5$ and θ_6 . Because there are two solutions for θ_2 and θ_3 , there are finally four sets of feasible solutions after permutation and combination.

Figure 2 shows the working principle and closed-loop control process of an autonomous grasping system based on machine vision and robotic arm in multiple time steps (T-1 to T-n). The system starts with the "active visual attention" module, extracts and encodes multimodal visual information such as RGB-D from the environment, and integrates the encoded features with historical hidden states before inputting them into the policy network. The strategy network simultaneously outputs motion control instructions

for the robotic arm (including end effector pose or joint angle adjustment) and attention control signals. The former is parsed into executable trajectories through a kinematic solver, while the latter dynamically optimizes the perception focus to cope with complex scenes. After the robotic arm performs actions, the system evaluates the effectiveness of the actions (such as successful grasping, collision avoidance, etc.) through a reward calculation module, and combines the state estimation of the value network to form the advantage function required for reinforcement learning. The new round of environmental observation and evaluation results are fed back to the historical state update module, forming a closed loop of "perception decision execution evaluation", driving the continuous online optimization of the strategy network and value network. The entire architecture integrates visual guidance, real-time decision-making, precise control, and adaptive learning, enabling robots to gradually learn and complete robust and efficient autonomous grasping tasks in dynamic environments.

3.2. Control method of manipulator arm

The precise control of the robotic arm is not only related to the accuracy and stability of the grasping action but also directly affects the efficiency and safety of the entire system. Servo control technology, which is widely used in manipulator power systems, can be driven by high-precision motors to ensure that each joint moves according to a predetermined trajectory [21]. This technology combines position feedback, speed feedback and torque feedback to enable the robotic arm to accurately execute instructions in complex environments. The gradual integration of intelligent algorithms, such as fuzzy logic and neural

networks, gives them stronger learning and adaptability [22, 23]. When facing an unknown object, the robotic arm can analyze the shape, texture and center of gravity distribution of the object through visual information, automatically generate the optimal grasping strategy, and adjust the force and contact angle in real time to effectively avoid damage or slippage. In addition, multi-sensor fusion technology uses multiple sensing devices such as lidar and depth cameras to work together to build a three-dimensional environment model, which also effectively improves the autonomous grasping performance of the robotic arm and further improves operational flexibility and robustness.

YOLOv5, with its improved AF-FPN structure (integrating adaptive attention module and bidirectional feature fusion), is capable of high-precision and real-time detection and localization of target objects and their 6D poses from RGB-D images, providing precise target spatial coordinates for robotic arms. The reinforcement learning (RL) algorithm defines a closed-loop decision framework: the state space contains perceptual information such as the end position of the robotic arm, the target pose, and the depth point cloud. Action space is defined as the continuous movement or rotation instructions of the end effector in Cartesian space. The reward function is designed as a guiding combination of rewards and punishments, such as distance rewards for approaching the target, high positive rewards for successful grasping, and penalties for collisions or timeouts. Intelligent agents (PPO) learn how to select the optimal action from the current state to maximize cumulative rewards by interacting with the environment, ultimately learning robust and efficient autonomous grasping strategies.

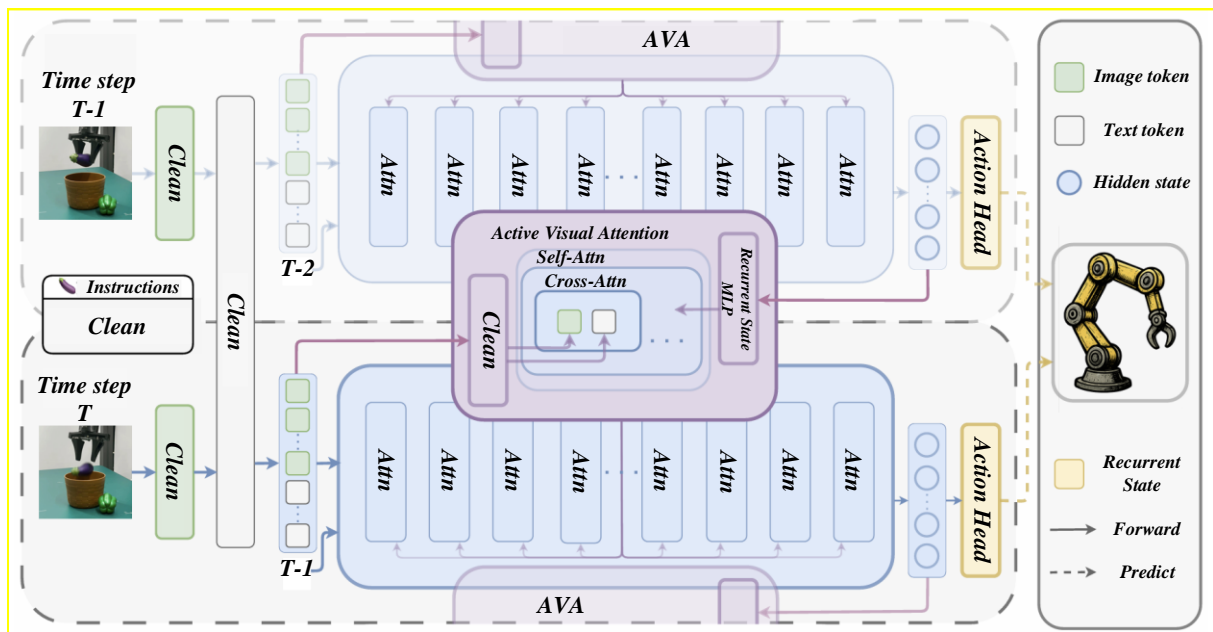


Figure 2. Principle and kinematics of robotic arm

The architecture, learning process, and hyperparameter settings of reinforcement learning (RL) algorithms ensure the complete repeatability of the scheme. Its core adopts the Actor Critic architecture, where the policy network (Actor) outputs continuous control instructions for the end effector of the robotic arm, while the value network (Critic) evaluates the long-term value of the action. The learning process is achieved through stable training using the PPO algorithm, which collects the experience of robots exploring the environment, calculates the advantage function to evaluate the quality of actions, and uses a pruning mechanism to prevent policy mutations when updating strategies. To ensure repeatability, key hyperparameters are explicitly fixed: the learning rates of the policy network and value network are $3e-4$ and $1e-3$, respectively, the discount factor γ is set to 0.99, the GAE (Generalized Advantage Estimation) parameter λ is 0.95, the policy update clipping coefficient ϵ is 0.2, and the batch size is 256. The optimizer uses Adam, and the experience replay buffer capacity is $1e6$. At the same time, the random initialization of network weights and the random seed of the environment are fixed to ensure that, under the same hardware and code, a completely consistent strategy and performance can be obtained during each training.

PFFNet is a method that jointly learns facial flow and facial priors from point clouds, achieving high-precision 3D facial expression capture through adaptive sampling and normal vector enhancement modules. GRCNN combines the advantages of regional recommendation networks and convolutional neural networks, and can quickly generate the position, posture, and angle of grasping points directly from deep images, achieving high-speed vision control closed-loop for robots. RCAN is a network designed specifically for image super-resolution. Its core is a deep structure containing multiple residual channel attention modules, which effectively transmit high-frequency information through long and short hop connections. It has important applications in enhancing the image quality of visual system inputs and provides a clear visual foundation for subsequent accurate recognition and localization.

4. AUTONOMOUS GRASPING METHOD OF ROBOT

4.1. Target recognition and localization based on machine vision

The target object recognition and positioning system based on the depth camera uses the host computer PC to detect the ultrasonic honing device part in the video stream of the image sensor in real time, determine its coordinates in the field of view of the camera [24], and then cooperate with the stereo infrared sensor and infrared laser emitter of the depth camera to obtain the depth information of the components.

First, the host computer is connected to the depth camera, and its RGB video stream is transmitted to the target detection network to judge whether the current frame has ultrasonic honing device components; if not, the next frame data is identified [25], and if there is, the coordinates and labels of the components in the pixel coordinate system are obtained. When training the target detection network, 42 ultrasonic honing device components are labeled separately, and then the depth values are obtained through the depth frame of the depth camera. Finally, the coordinates of the components in the pixel coordinate system and the corresponding depth values are obtained. Data enhancement operation is carried out to enhance the generalization ability of the network [26]. Four pictures are randomly selected from the image set and then spliced into one image after randomly scaling, cropping and distorting. This operation is done for all images so that the network can accurately detect targets in different backgrounds.

The Backbone part uses the CSPDarknet network built by CSPDarknet and SPP. Its core function is to extract the features of input images. Figure 3 shows its operation flow in image processing. After the image is input, the Focus network architecture is used for processing, and the input feature layers are segmented, stacked, and reorganized to form a feature layer with low resolution but four times the original number of channels so as to better capture the feature information of different scales and process small target objects. The features are input into the SiLU activation function after performing convolution and batch normalization operations in turn.

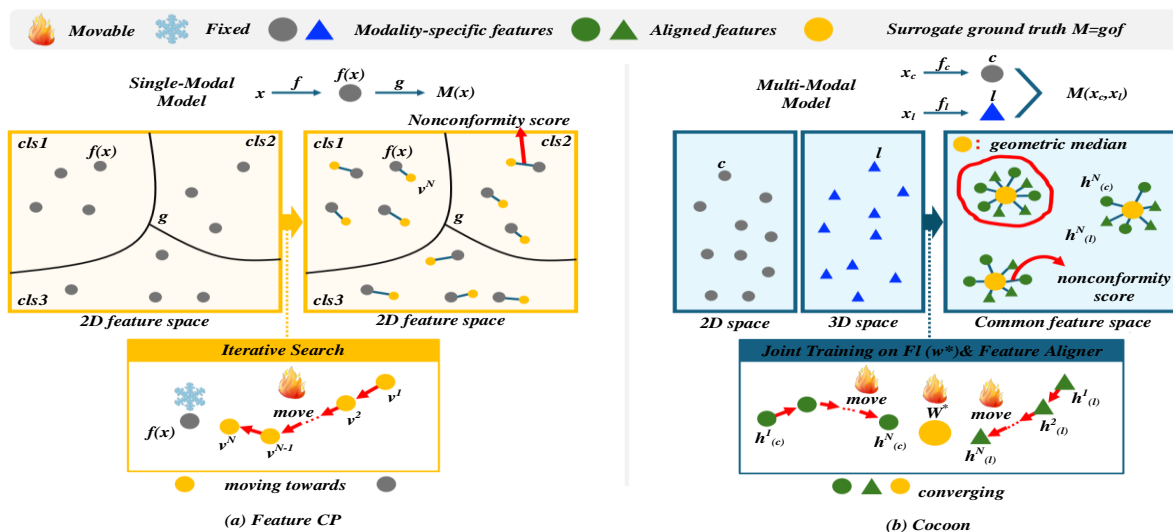


Figure 3. CSPDarknet network

The three feature layers of the CSPDarknet backbone are located in the middle layer, middle and lower layer and bottom layer. FPN is superimposed, and YOLOv5 architecture is introduced to enhance the network feature fusion effect. The Head section of YOLOv5 focuses on the forecasting process. Three shape feature layers can be determined by FPN: $80 \times 80 \times 256$, $40 \times 40 \times 512$ and $20 \times 20 \times 1024$, and the prediction results can be obtained by inputting them into the Head part. When processing each feature layer, first use the convolution method to adjust the number of channels. The output size is $80 \times 80 \times 3 \times (4 + 1 + \text{class_num})$. 80×80 is the output size, 3 is the number of anchor boxes, and 4 is the prediction box coordinates. One is the object confidence, and class_num is the number of object categories. In this prediction, the parts of the ultrasonic honing device are divided into 8 categories, and the final output results are $80 \times 80 \times 39$, $40 \times 40 \times 39$ and $20 \times 20 \times 39$, which correspond to the detection of smaller, medium and larger targets in the image respectively.

4.2. Synergistic work of vision and robotic arms

The combination of the vision system and the robotic arm is the key to enabling the latter to have an autonomous grasping function. The vision system can capture the position and posture of the target object and its relationship with the surrounding environment, generate accurate spatial coordinate data, and provide the basis for subsequent planning. Advanced image processing algorithms, such as feature extraction, template matching, deep learning, etc., enable robots to quickly locate regions of interest from complex scenes and identify items to be grasped.

Once the goal is determined, the next step is to enter the key step—the path planning of the manipulator under vision guidance. The shortest moving distance and optimal path are calculated by the analytic geometry principle, and a smooth and continuous moving trajectory is generated by considering the obstacle avoidance requirement. In this

process, dynamic adjustment of parameters ensures continuous and uninterrupted operation in the face of unexpected situations. In order to avoid that it is difficult to cover all possible situations simply by relying on preset programs, this study also introduces a real-time feedback mechanism. When external conditions change (such as item displacement), the vision module immediately updates the information, commands the robotic arm to adjust the action plan immediately, keeps the grasping process stable, and can quickly restore the balance state even in the face of interference factors, which greatly improves the task success rate. At the hardware level, the structural design of the robotic arm needs to fully consider the need to cooperate with the vision system to ensure that the camera is installed in a reasonable position and has a wide viewing angle range so as to obtain high-quality image data in all directions; Actuators (such as clamping jaws) should take into account flexibility and rigidity, so that they can firmly hold objects of different material sizes without causing damage.

5. EXPERIMENTAL RESULTS AND ANALYSIS

Table 1 presents the experimental data on the test set. In terms of the accuracy of average pose estimation of various types of targets, PFFnet shows leading performance, with the prediction ratio of ADD (or ADD-S) less than 2cm as high as 96.0%. Of the 13 target categories, 8 categories received the highest evaluation criteria. Compared with DenseFusion, various types of targets have achieved a steady improvement in the accuracy of pose estimation, especially on driller targets; the maximum improvement has reached 4.8%, while the average improvement has been 1.8%. Compared with PoseCNN using ICP (Iterative Closest Point), PFFne of ape, duck, and hole classes showed significant improvement. Although the performance with PVNet has not reached its peak in bench vise, driller, iron, and lamp class PFFne, its accuracy is still quite high.

Table 1. LineMOD dataset experimental results

Category	RGB			RGB-D		
	BB8	PoseCNN	PVNet	PointFusion	DenseFusion	PFFnet
ape	39.592	75.460	42.728	68.992	90.454	92.414
benchvise	89.964	95.550	97.902	79.086	91.336	92.708
camera	54.586	91.630	85.064	59.584	92.512	92.316
Can	62.818	94.570	93.492	59.878	91.238	94.374
cat	61.348	80.458	77.714	77.518	94.570	95.060
driller	72.912	93.100	94.472	46.354	85.260	89.964
duck	43.414	76.146	51.450	61.740	90.454	90.846
eggbox *	56.644	95.158	97.118	97.902	97.804	97.804
glue *	40.376	97.412	93.688	97.314	98.000	98.000
hole	65.856	51.744	80.262	70.364	90.258	92.610
iron	83.006	96.334	96.824	81.536	95.060	96.236
lamp	74.970	95.550	97.314	61.054	93.394	95.256
phone	52.920	85.946	90.552	77.224	90.944	93.002
Average	61.446	86.828	84.476	72.226	92.414	94.080

We conducted a detailed quantitative evaluation of the evaluation indicators and compared them with the methods of PointFusion, PoseCNN and DenseFusion. The specific experimental data can be found in Figure 4. In this study, PPFnet demonstrated excellent accuracy in 20 categories, demonstrating cutting-edge pose assessment capabilities. Compared to DenseFusion, the 51 _ large _ clamp class achieved a 10.5% increase in difficult estimation cases, and its average accuracy also increased by 1.2%. Across the AUC metrics, it peaked in all 12 categories.

It can be observed from Table 2 that compared with the benchmark, the point cloud feature extractor used in this

study has achieved some improvement in the accuracy of various targets. In the two key steps of feature extraction and pose prediction, both the point cloud feature extractor and the point cloud fractal pyramid effectively improve the learning ability of local information of the target point cloud. Therefore, it can be seen that the local attributes of the target will have a great influence on the pose evaluation. After comprehensive consideration, the PPFnet algorithm has achieved a robust improvement in the accuracy of various target pose estimations compared with the benchmark, which proves the effectiveness of the improved algorithm.

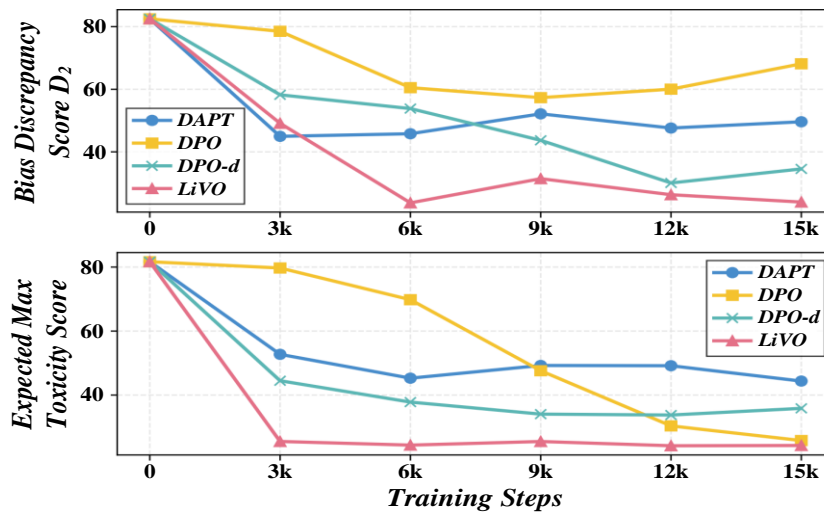


Figure 4. Experimental results of data set

Table 2. Data Set Ablation Experiment Results

Category	Standard	Benchmark + A	Benchmark + B
ape	94.350	95.778	95.064
benchvise	95.064	95.778	95.574
camera	96.492	97.920	97.104
can	95.574	97.716	98.532
cat	98.532	98.634	98.838
driller	91.800	92.310	93.330
duck	94.452	94.452	94.554
eggbox *	98.802	98.802	98.802
glue *	100.000	100.000	100.000
hole	94.656	96.186	95.064
iron	99.042	99.654	99.450
lamp	97.308	98.430	98.634
phone	94.758	95.982	95.676
Average	96.594	97.410	97.308

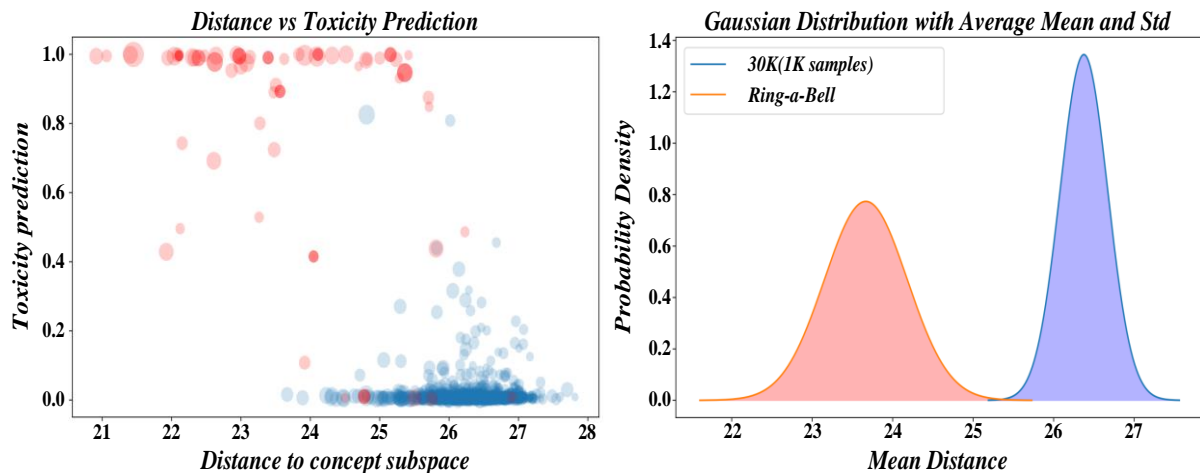


Figure 5. Comparison of weight models

When training with the YOLOv5 network, pre-training weights are often loaded to shorten the training time and improve the training accuracy. In this study, YOLOv5x is selected as the pre-training weight model. Figure 5 shows the weight comparison results. In this study, YOLOv4 and YOLOv5 were used to train the data set of ultrasonic honing device components, the training weights were recorded, and the loss function was calculated. The YOLOv5 loss function has a great impact on the performance of the model and is used to measure the difference between the predicted value and the true value. Its total loss is the weighted sum of classification, confidence and regression loss. It can be seen from Figure 6 that the initial loss function values of the two are close, but YOLOv5 converges faster, indicating that it trains faster on new datasets. As the training progresses, the loss function begins to decrease in about 10 rounds, tends to

stabilize in about 70 rounds, and finally stabilizes in a small range, reflecting the good convergence of the model.

Figure 7 shows the comparison of various parameters of YOLOv4 and YOLOv5 under the evaluation criterion of $IOU \geq 50\%$. The mAP value of YOLOv5 reaches 95.38%, and the average detection speed is 26 frames/second. Its recognition accuracy and speed are better than those of YOLOv4. Figure 8 shows that the total successful accuracy rate of ultrasonic honing device components in actual scenes is 92.78%, and the success rate of short-distance object detection and recognition is high, but that of long-distance object detection is low, the features are not obvious, and there is a lot of noise and interference during long-distance photography. Therefore, this study will use high-resolution images, supplement long-distance target data sets, and strengthen long-distance photography training.

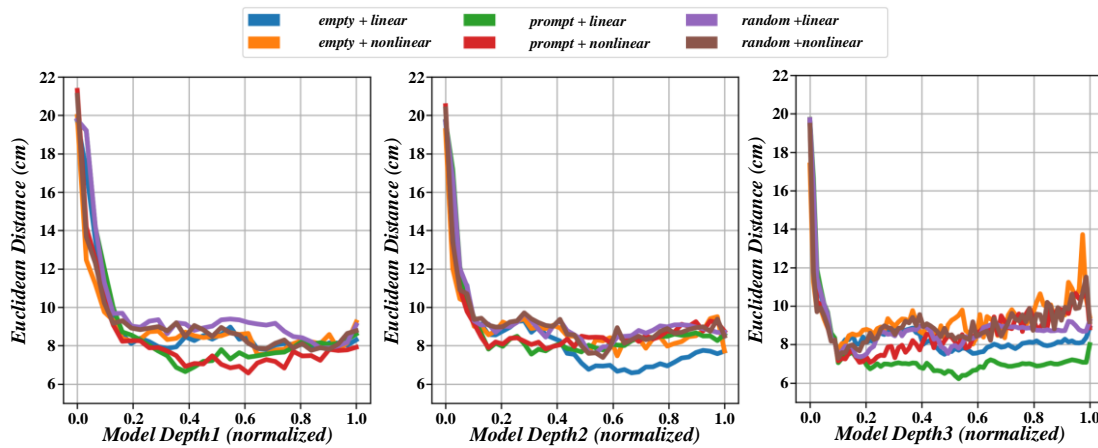


Figure 6. Comparison chart of training loss function of data set

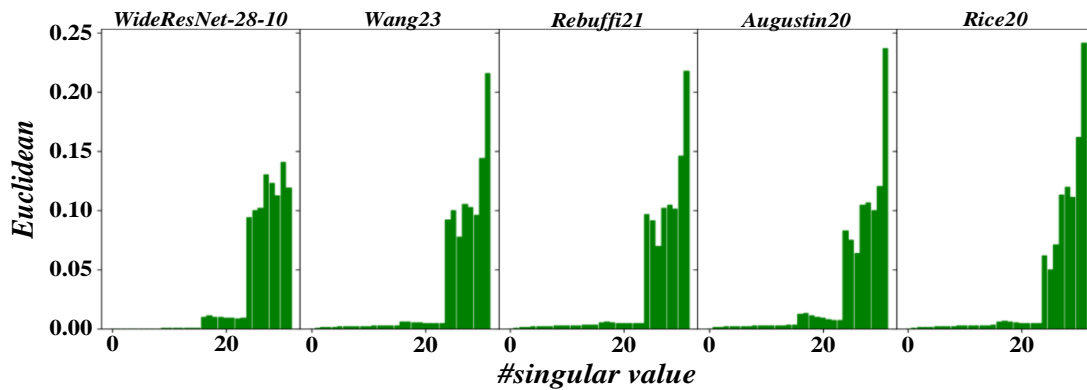


Figure 7. Comparison of various parameters

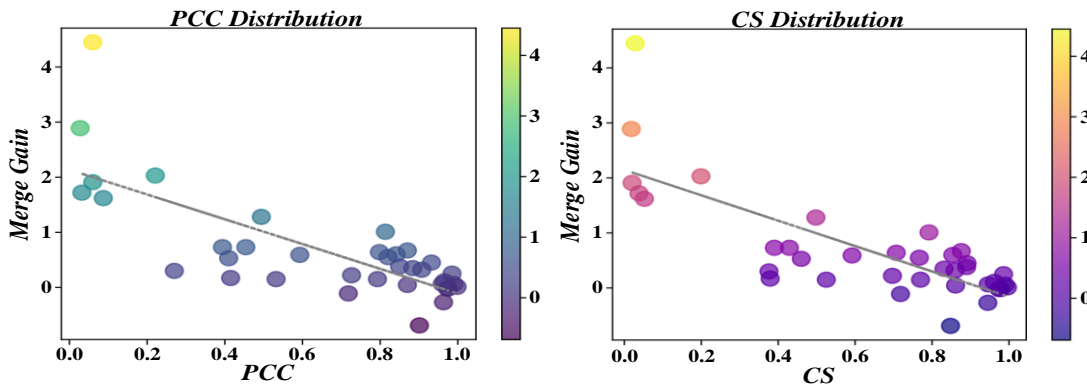


Figure 8. Successful Accuracy of Detection and Recognition

Figure 9 shows that the grasping test of four kinds of grasping targets is carried out in a multi-target environment, and the total grasping success rate is 87.5%, which meets the requirements of the robotic arm grasping system. The experimental results show that the robotic arm has a high success rate in grasping the transducer “PUH _ A”, and its cylindrical shape is conducive to improving the success rate. However, the grasping point deviating too far from the center of gravity will lead to unstable grasping, and bumps may occur when placing the target.

Figure 10 shows the success rate of predicting the grasping pose of each object and the success rate of the robotic arm grasping. When predicting the grasping posture, the prediction success rate of the beverage bottle is 90% due to its transparency. The stationery hat has a success rate of 85% because of its small size and similar color to the background. When the paper tube is placed vertically, the prediction success rate is 88% (3 failures), and the prediction success rate of other objects is 98%. In the process of grabbing objects, a razor, bottled cleaning solution, stationery cap and paper tube were successfully grabbed 16 times, with a success rate of 88%. Because of the smooth surface and soft material, the success rate of

grabbing the beverage bottle is only 70%, and other objects are successfully grabbed 18 times, with a success rate of 98%. In the two scenarios, the average success rate of predicting grasping pose is 95%, and the success rate of robotic arm grasping is 90%.

After in-depth analysis and comparison of experimental data, it was found that GR-CNN performed poorly in distinguishing target objects from background differences. Experiments show that GR-CNN is easily disturbed when the object grasping pose detection is affected by the environment, and it is difficult to accurately judge the existence area of unknown objects with a similar color to the background. Using the RCAN module, the network architecture proposed in this study can accurately identify the best grasping position of the target object, effectively distinguish high-quality graspable areas, and achieve a more stable grasping posture. Figure 11 shows the performance comparison of common grasp detection models and comparison with other types of grasp detection. The optimized version proposed in this study GR-CNN network, ensures real-time performance and has higher detection accuracy, which can meet the real-time and accuracy requirements of target object grasping pose prediction.

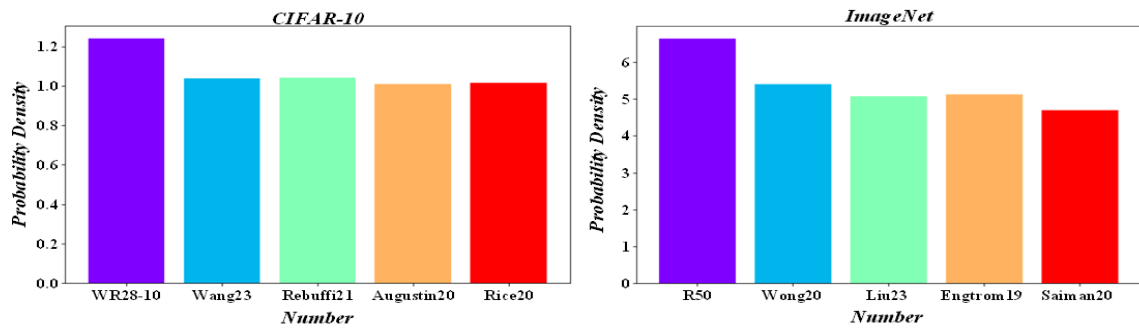


Figure 9. Experimental statistics of grasping power device components

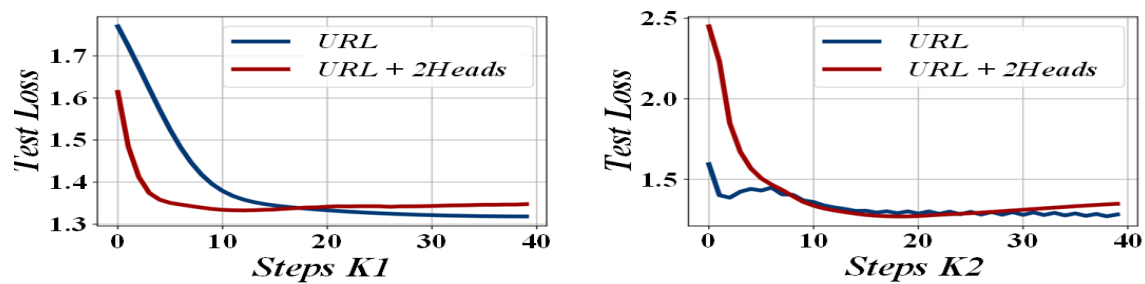


Figure 10. Experimental results of manipulator arm grasping

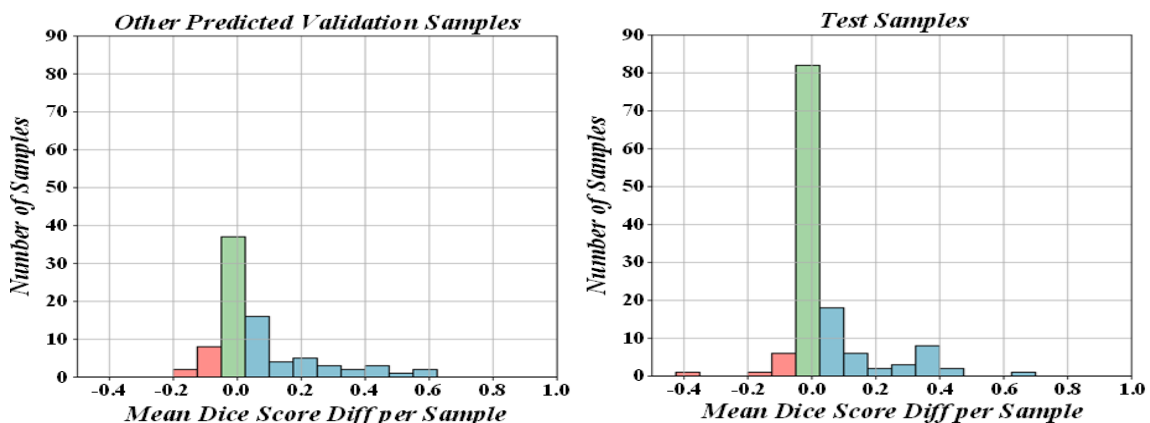


Figure 11. Performance of different grasp detection models in Cornell dataset

6. DISCUSSION

This study constructed a complete experimental system comprising a UR5e six-degree-of-freedom robotic arm, a RealSense D435i depth camera, and a software platform based on ROS and PyTorch. At the motion-planning level, the system uses the RRT* algorithm to conduct a preliminary global, collision-free path search in the configuration space, and then uses the CHOMP optimizer to perform gradient-based optimization of the trajectory to ensure the final generated robotic arm motion is both safe and smooth. The datasets used for model training include publicly available benchmarks (LineMOD, Cornell Grassp) and a self-built UH Parts dataset, which contains 20000 samples, covers 8 specific industrial components, and provides synchronized RGB-D images, 6D poses, and grasping annotations.

The limitations of the current method mainly lie in three aspects: the perception and grasping success rate of highly reflective, transparent, or soft objects are still insufficient. In highly dynamic and unstructured unknown complex environments, the robustness and adaptability of strategies need to be strengthened; Although the overall system latency is within one second, further optimization is still needed for specific ultra high-speed application scenarios. In response to these limitations, future work will focus on multimodal perception fusion, fast adaptive strategies based on meta-reinforcement learning, lightweighting algorithms, and hardware acceleration, as well as long-term deployment and iterative optimization in more complex real-world production lines.

7. CONCLUSION

In recent years, autonomous grasping technology based on machine vision and robotic arms has made significant progress in industrial automation. In response to the insufficient robustness of traditional systems in dynamic environments, this study proposes a comprehensive solution and verifies it through system experiments. The core points can be divided into the following three aspects:

1. This study used an improved YOLOv5x model for object detection, achieving a mAP of 95.38% on a custom dataset. In terms of pose estimation, the PFFnet algorithm performed well on the LineMOD dataset, with an average pose estimation accuracy (ADD/ADD-S<2cm) of 96.0%, surpassing DenseFusion and other comparative algorithms in 8 out of 13 categories. The ablation experiment shows that introducing point cloud feature extraction and a feature pyramid structure is a key factor in improving performance.
2. In response to the shortcomings of the GR-CNN network under complex background interference, the integration of the RCAN module enhances the recognition ability of high-quality and graspable areas. The actual grasping experiment results show that the system's overall grasping success rate in a multi-target environment is 87.5%. Among them, the average success rate in predicting the grasping pose is 95%, and the robotic arm's actual grasping success rate is 90%. However, there is still room for improvement in the grasping

success rate for specific objects, such as smooth, transparent ones (e.g., objects, such as beverage bottles).

3. By optimizing the motion path planning algorithm, the system reduced power consumption by 20% and improved work efficiency by 15% while meeting the exact job requirements. In the 24-hour uninterrupted operation test, the system's overall error rate was only 0.3%, demonstrating high reliability and stability and providing good practical application and economic benefits.

Funding Statement

This work was supported by the Henan Province Key Scientific and Technological Project of (242102111191).

Availability of Data and Materials

The data used to support the findings of this study are all in the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Authors' contributions

Na Wang: Conceptualization, Formal analysis, Funding acquisition; Investigation, Data Curation, Writing - Original Draft; Writing - Review & Editing; Biao Zhang: Formal analysis, Investigation, Data Curation. All authors contributed to the article and approved the submitted version.

Acknowledgements

Not Applicable.

References

- [1] Y. J. Chiu, Y. Y. Yuan, S. R. Jian, "Design of and research on the robot arm recovery grasping system based on machine vision", *Journal of King Saud University-Computer and Information Sciences*, Vol. 36, No. 4, 2024, pp.102014. <https://doi.org/10.1016/j.jksuci.2024.102014>.
- [2] Y. Liu, Y. Zhang, Z. Wang, R. Cheng, X. Zhao, B. Shi, "Detection of image recognition forgery technology under machine vision", *International Journal of Ad Hoc and Ubiquitous Computing*, Vol. 45, No. 2, 2024, pp.123-134. <https://doi.org/10.1504/IJAHUC.2024.136852>.
- [3] Y. Ru, X. Zhang, "Orientation and trajectory-specific movement assistance for quadcopter control using machine vision input", *International Journal of Sensor Networks*, Vol. 45, No. 3, 2024, pp.177-190. <https://doi.org/10.1504/IJSNET.2024.139852>.
- [4] M. Zhu, B. Zhang, C. Zhou, H. Zou, X. Wang, "Target recognition of multi source machine vision pan tilt integrated inspection robot for power inspection", *IEEE Access*, Vol. 12, 2024, pp.45693-45708. <https://doi.org/10.1109/ACCESS.2024.3378580>.
- [5] Y. J. Chiu, Y. Y. Yuan, S. R. Jian, "Design of and research on the robot arm recovery grasping system based on machine vision", *Journal of King Saud University-Computer and*

- Information Sciences, Vol. 36, No. 4, 2024, pp.102014. <https://doi.org/10.1016/j.jksuci.2024.102014>.
- [6] C. Feng, "Design of logistics sorting algorithm based on deep learning and sampling evaluation", International Journal of Computational Intelligence Systems, Vol. 17, No. 1, 2024, pp.82. <https://doi.org/10.1007/s44196-024-00449-0>.
- [7] L. Ortiz-Aguilar, L. A. Xoca-Orozco, J. D. Anda-Suárez, "Design of routes for collaborative robots in the automobile painting process through a comparison of perturbative heuristics for iterated local search", Computación y Sistemas, Vol. 28, No. 2, 2024, pp.803-820. <https://doi.org/10.13053/CyS-28-2-5024>.
- [8] R. Szabo, "Developing different test conditions to verify the robustness and versatility of robotic arms controlled by evolutionary algorithms", Electronics, Vol. 13, No. 11, 2024, pp.2130. <https://doi.org/10.3390/electronics13112130>.
- [9] J. V. A. Cabral, A. J. Álvares, G. C. de Carvalho, "Digital twin implementation for an additive manufacturing robotic cell based on the iso 23247 standard", IEEE Latin America Transactions, Vol. 22, No. 8, 2024, pp.651-658. <https://doi.org/10.1109/TLA.2024.10620386>.
- [10] N. Sayols, A. Hernansanz, A. Sozzi, N. Piccinelli, F. Falezza, S. Farsoni, et al., "Dynamic Global/Local multi-layer motion planner architecture for autonomous Cognitive Surgical Robots", Robotics and Autonomous Systems, Vol. 180, 2024, pp.104758. <https://doi.org/10.1016/j.robot.2024.104758>.
- [11] L. Vianello, S. Ivaldi, A. Aubry, L. Peternel, "The effects of role transitions and adaptation in human-cobot collaboration", Journal of Intelligent Manufacturing, Vol. 35, No. 5, 2024, pp.2005-2019. <https://doi.org/10.1007/s10845-023-02104-5>.
- [12] M. Andreoni, W. T. Lunardi, G. Lawton, S. Thakkar, "Enhancing autonomous system security and resilience with generative AI: A comprehensive survey", IEEE Access, Vol. 12, 2024, pp.109470-109493. <https://doi.org/10.1109/ACCESS.2024.3439363>.
- [13] S. A. Rahmah, N. Kubota, "Estimation of object handover position using human-robot proxemics and unsupervised pattern recognition", Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 28, No. 2, 2024, pp.371-377. <https://doi.org/10.20965/jaciii.2024.p0371>.
- [14] Y. T. Yang, L. D. Zhang, "Event-triggered adaptive tracking control of a class of nonlinear systems with asymmetric time-varying output constraints", Frontiers of Information Technology & Electronic Engineering, Vol. 25, No. 8, 2024, pp.1134-1144. <https://doi.org/10.1631/FITEE.2300679>.
- [15] H. Y. Chen, G. D. Zong, S. F. Su, F. Z. Gao, "Event-triggered extended dissipative ftb for t-s fuzzy switched systems with mismatched phenomena and deception attacks: a multidomain framework", IEEE Transactions on Cybernetics, Vol. 54, No. 9, 2024, pp.4928-4938. <https://doi.org/10.1109/TCYB.2024.3365608>.
- [16] W. H. Liu, Q. Ma, S. Y. Xu, "Fuzzy fixed-time prescribed-performance tracking control of nonlinear systems with dynamic event-triggered signal", IEEE Transactions on Fuzzy Systems, Vol. 32, No. 3, 2024, pp.1399-1408. <https://doi.org/10.1109/TFUZZ.2023.3325529>.
- [17] L. Z. Kong, K. Liu, X. Hu, N. Zhang, L. Qi, X. Li, X. Zhou, "Gender classification based on spatio-frequency feature fusion of OCT fingerprint images in the IoT environment", IEEE Internet of Things Journal, Vol. 11, No. 15, 2024, pp.25731-25743. <https://doi.org/10.1109/JIOT.2024.3381428>.
- [18] Padalkar, G. Quere, A. Raffin, J. Silvério, F. Stulp, "Guiding real-world reinforcement learning for in-contact manipulation tasks with Shared Control Templates", Autonomous Robots, Vol. 48, No. 4, 2024, pp.12. <https://doi.org/10.1007/s10514-024-10164-6>.
- [19] Y. T. Yuan, S. Z. Wang, Y. P. Mei, W. P. Zhang, J. Sun, G. Wang, "Improving world models for robot arm grasping with backward dynamics prediction", International Journal of Machine Learning and Cybernetics, Vol. 15, No. 9, 2024, pp.3879-3891. <https://doi.org/10.1007/s13042-024-02125-3>.
- [20] S. I. Abdelmaksoud, M. H. Al-Mola, G. E. M. Abro, V. S. Asirvadam, "In-depth review of advanced control strategies and cutting-edge trends in robot manipulators: analyzing the latest developments and techniques", IEEE Access, Vol. 12, 2024, pp.47672-47701. <https://doi.org/10.1109/ACCESS.2024.3383782>.
- [21] M. C. Chiu, L. S. Yang, "Integrating explainable AI and depth cameras to achieve automation in grasping Operations: A case study of shoe company", Advanced Engineering Informatics, Vol. 62, 2024, pp.102583. <https://doi.org/10.1016/j.aei.2024.102583>.
- [22] C. Y. Huang, Y. H. Shao, "Integration of deep Q-learning with a grasp quality network for robot grasping in cluttered environments", Journal of Intelligent & Robotic Systems, Vol. 110, No. 3, 2024, pp.97. <https://doi.org/10.1007/s10846-024-02127-x>.
- [23] S. Horvath, H. Neuner, "Introduction of a framework for the integration of a kinematic robot arm model in an artificial neural network-extended kalman filter approach", Journal of Intelligent & Robotic Systems, Vol. 110, No. 4, 2024, pp.137. <https://doi.org/10.1007/s10846-024-02164-6>.
- [24] J. Chen, J. Xiao, M. Yang, H. Pan, "Learning to improve operational efficiency from pose error estimation in robotic pollination", Electronics, Vol. 13, No. 15, 2024, pp.3070. <https://doi.org/10.3390/electronics13153070>.
- [25] S. Nahavandi, R. Alizadehsani, D. Nahavandi, C. P. Lim, K. Kelly, F. Bello, "Machine learning meets advanced robotic manipulation", Information Fusion, Vol. 105, 2024, pp.102221. <https://doi.org/10.48550/arXiv.2309.12560>.
- [26] D. Kijdech, S. Vongbunpong, "Manipulation of a complex object using dual-arm robot with mask R-CNN and grasping strategy", Journal of Intelligent & Robotic Systems, Vol. 110, No. 3, 2024, pp.103. <https://doi.org/10.1007/s10846-024-02132-0>.