An Autonomous Sorting System for Intelligent Warehouse Robots Based on Mask R-CNN

Ying Zhou[1†], Xu Ji[1*†], Yifei Guo[1†], Jing Tang[1†], Long jun Wu[1†]
{2806654757@qq.com, 493591532@qq.com, 3092307213@qq.com, 531013872@qq.com, 5905754@qq.com}
[1]Chengdu Technological University, No. 1, Section 2, Zhongxin Avenue, Chengdu, 611730, China
∗ Corresponding author.
† These authors contributed equally.

Abstract. A visual-guidance solution based on Mask R-CNN is proposed for the automated sorting of robots in intelligent warehouses. This approach solves the challenges of identifying and localising multi-category, stacked packages in complex environments. Built a custom dataset with multi-category package labels. Transfer learning was employed to train the Mask R-CNN model, enabling the simultaneous output of the target parcel category, bounding box, and a high-precision pixel-level mask. This mask provides a precise spatial contour and pose for robotic arm grasp planning, effectively preventing misgrabs. The average accuracy (mAP) of the system was 95.7% in the test set and 91.2% in the combination (mIoU). The actual sorting success rate reaches 98.5%, significantly outperforming traditional methods. This validates the solution's effectiveness and robustness in complex warehouse environments, providing a reliable path to enhance sorting automation.

Keywords: deep learning, intelligent warehouse, sorting system, Mask R-CNN, instance segmentation, robotics

1 Introduction

Deep learning is a methodology rooted in artificial neural networks, which solves complex problems by abstracting and learning features from visual data through multi-layered neural networks. From the early stages of simple neural networks to the introduction of the backpropagation algorithm, and subsequently to convolutional neural networks (CNN), and Recurrent Neural Networks (RNN)[1], the development of deep learning has seen its widespread application across various fields, including computer vision, speech recognition, natural language processing, and predictive analytics. For instance, in the field of computer vision, techniques such as R-CNN have been employed. A series of models, including those for image classification and object localisation applications. Building upon this foundation, Mask R-CNN emerges as a pivotal advanced technique.

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365267

By incorporating a mask prediction branch that operates in parallel with bounding box regression, it achieves pixel-level instance segmentation. With the rapid advancement of intelligent warehousing systems, the demand for sorting accuracy and automation continues to rise. The stacking, occlusion, and label diversity of parcels within warehouse environments present significant challenges for visual recognition systems. Deep learning, particularly object detection models like Mask R-CNN with instance segmentation capabilities, has become the key technology for achieving precise grasping. Leveraging its ability to simultaneously output object bounding boxes and pixel-level masks, Mask R-CNN [1]demonstrates core value across multiple stages in intelligent warehouses, including parcel localisation, pose estimation, and robotic arm obstacle avoidance during sorting. However, despite Mask R-CNN's significant advances in detection accuracy, its high computational complexity and relatively slow processing speed present a pressing challenge. When deployed on modern logistics sorting lines demanding extreme real-time responsiveness, enhancing processing efficiency while maintaining accuracy remains an urgent issue requiring resolution.

2 Background

Mask R-CNN is an efficient object instance segmentation framework capable of detecting objects within images and generating high-quality segmentation masks. It serves as an extension of Faster R-CNN[2], incorporating a branch for predicting segmentation masks while maintaining rapid training and inference speeds. The introduction of Mask R-CNN occurred against the backdrop of a fully developed two-stage object detection framework. Its evolutionary trajectory began with R-CNN, which laid the groundwork for deep learning applications in object detection by introducing candidate regions and convolutional neural network feature extraction. Subsequent developments like Fast R-CNN significantly enhanced efficiency through shared convolutional computations; Faster R-CNN further advanced this by integrating region proposal networks to achieve end-to-end training, thereby establishing a mature and efficient object detection framework. However, these models were constrained to output bounding boxes around objects, failing to provide precise contour information.[3] To address this challenge and meet the demands of instance segmentation—the higher-level task of pixel-level recognition for each object instance in an image—He et al. proposed Mask R-CNN in 2017. Building upon Faster R-CNN's parallel branch architecture, this model innovatively introduced a mask prediction branch, as illustrated in Figure 1.

2.1 Technical Foundations of a Benchmark Model

This achieves integrated output for object classification, bounding box regression, and pixel-level segmentation. Crucially, addressing the high spatial sensitivity of mask generation, the model replaces the original ROIPooling operation with an ROIAlign layer [4], effectively eliminating spatial misalignment between feature maps and the original image [5], thereby significantly enhancing segmentation mask accuracy. By inheriting the efficiency and accuracy of existing detection frameworks while achieving precise instance segmentation, Mask R-CNN rapidly established itself as the benchmark model in this domain. It provides robust technical support for visual tasks requiring precise object pose perception, such as intelligent warehouse sorting.

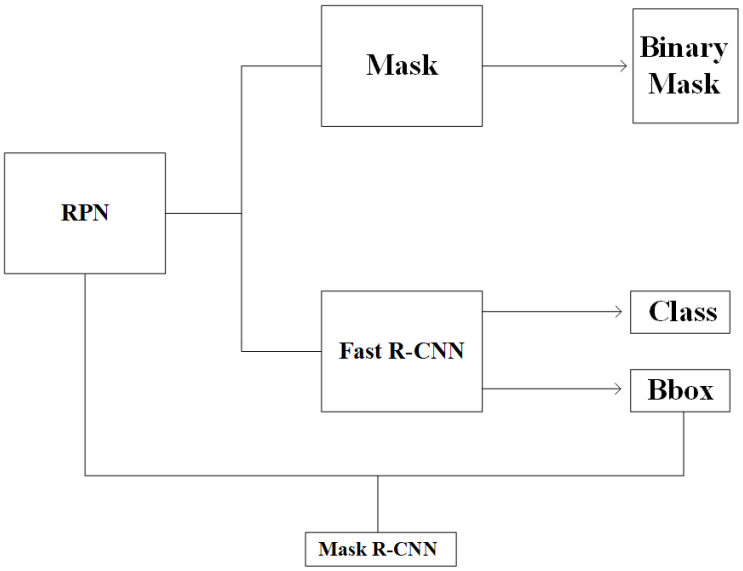

Fig. 1. Mask R-CNN Process Flow

2.2 The Mask R-CNN Architecture Overview

The model architecture of Mask R-CNN comprises four core components, representing key extensions to Faster R-CNN as illustrated in Figure 1. Firstly, the backbone network handles fundamental feature extraction, typically employing architectures such as ResNet or ResNeXt combined with a feature pyramid network to efficiently capture and fuse features across different semantic levels within an image. Secondly, the Region Proposal Network performs sliding-window scans across the feature maps generated by the backbone network to produce a series of candidate regions potentially containing targets. The third critical component is the ROIAlign layer. This layer employs a sophisticated bilinear interpolation algorithm to precisely map candidate regions of varying sizes onto feature map patches of a fixed scale. This effectively resolves the pixel misalignment issues inherent in the previous ROIPooling operation[6], laying a robust foundation for subsequent pixel-level predictions.

3 Application Of MASK R-CNN

3.1 Authors and Affiliations

Enhancing the accuracy of Mask R-CNN for instance segmentation can be achieved through the integration of attention mechanisms[7], which have become one of the most commonly employed modules [8]. By integrating attention mechanisms, Mask R-CNN's ability to focus on critical regions is enhanced, optimising feature extraction during segmentation while suppressing irrelevant back-

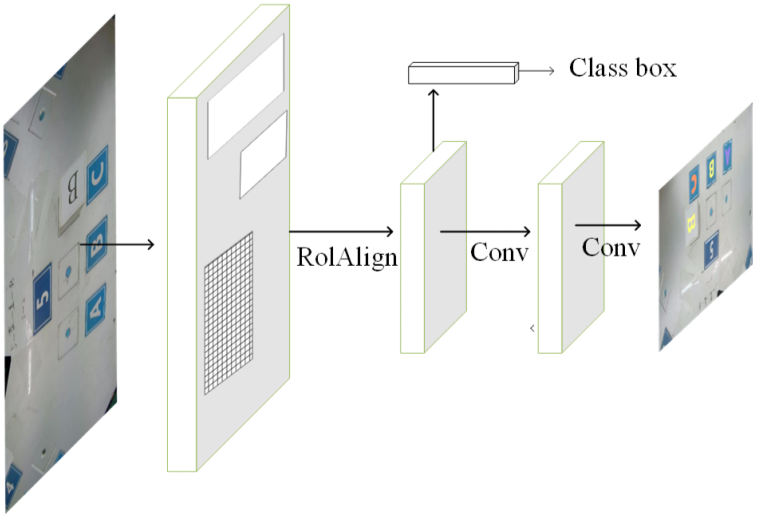

Fig. 2. Mask R-CNN Framework for Instance Segmentation

ground features. Wang et al. further improved segmentation accuracy by introducing a non-local attention module. This module aims to capture long-range dependencies within complex scenes, thereby more precisely delineating the contours of occluded or boundary-blurred objects.

3.2 Identify the Headings

The Feature Pyramid Network is a widely employed network architecture in the field of instance segmentation [9], enhancing multi-scale feature fusion through the introduction of top-down pathways. Its application within Mask R-CNN has also yielded positive results. Specifically, FPN enhances the model's ability to process objects at different scales by integrating deep semantic information with shallow detail information. Following the introduction of FPN, Mask R-CNN achieved a significant improvement in mAP on the COCO dataset, with a particularly notable increase in the APS metric for small object segmentation. This demonstrates a marked enhancement in the model's robustness when segmenting objects of various scales. Kirillov et al [10]. systematically explored the role of FPN within the Mask R-CNN framework. By integrating feature information across multiple hierarchical levels, they enhanced the representation of instances with markedly differing sizes within images, thereby achieving superior segmentation results.

4 Sorting system experimental platform

The sorting system experimental platform, as shown in Figure 4, comprises a KUKA KR 3 R540 industrial robot, an industrial camera, a GRM533Y-CLK remote module, and a SIEMENS

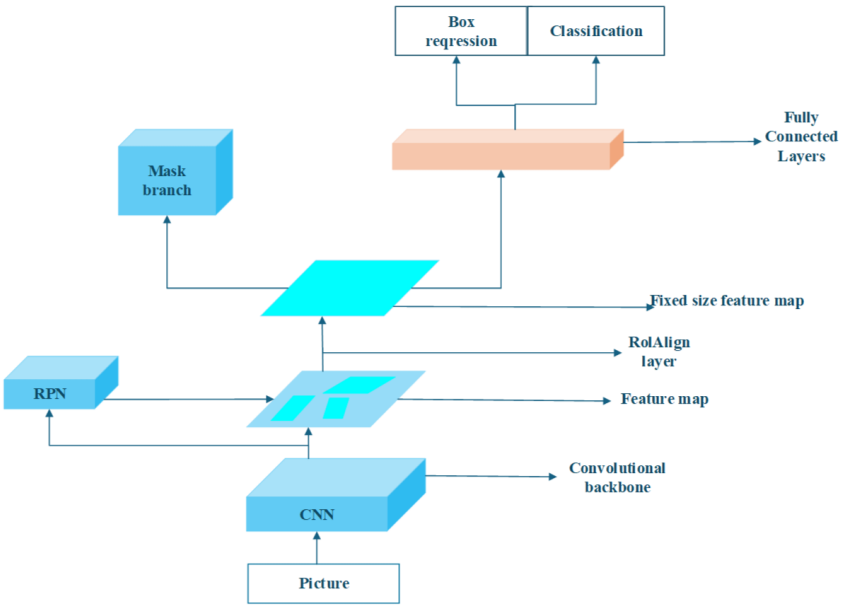

Fig. 3. Mask R-CNN Process Flow

S7-1511-1 PN controller.

4.1 Image Acquisition and Window Display Operations

The industrial camera is first connected to the host computer via USB to enable image acquisition and analysis through the host's image capture software. Acquired images often contain significant noise; within machine vision systems, unnecessary or interfering image data is termed noise [12]. To enhance post-acquisition image quality and reduce material recognition errors caused by noise, this system must process the captured images. First, the acquired image information must be displayed within a window. The `read_image` operator is employed to capture the image, while the `get_image_size` operator retrieves its dimensions. Subsequently, the `dev_open_window_fit_image` and `dev_display` functions are utilised to present the acquired image within a suitable window.

4.2 Robotic Sorting System Software Operation

The software operation process of the robotic sorting system is illustrated in Figure 6. Firstly, the experimenter establishes communication between both ends by running the server programme and client programme, respectively. Subsequently, the server sequentially executes the image acquisition thread, object detection thread, and instance segmentation thread[10]. The image acquisition

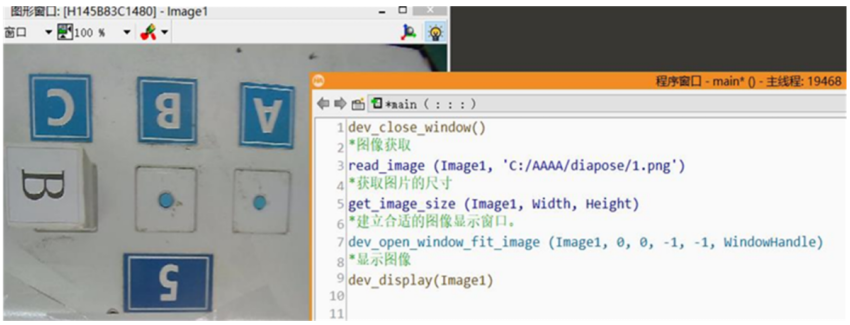

Fig. 4. Image Display Window

thread employs a USB industrial camera to capture colour and depth images from the scene, storing the associated image metadata. The object detection thread utilises the ResNet-50 network to analyse the acquired colour images, predicting the target's category, positional coordinates, and rotational angle within the scene. The client's data reception and processing thread receives and processes the target object information transmitted by the server, subsequently storing this data.

5 Object detection network construction

ResNet-50 comprises 49 convolutional layers and one fully connected layer. 'ID BLOCK ×2' denotes two identity residual blocks that preserve the feature map dimensions, while 'CONVBLOCK' represents convolutional residual blocks that increase the feature map dimensions. Each residual block contains three convolutional layers, resulting in a total of convolutional layers. For the entire training set, each channel is subtracted by the channel mean across the training set[11]. Following successive convolutional operations within the residual blocks, the number of channels within the image pixel matrix progressively deepens. Subsequently, a Flatten layer adjusts the image pixel matrix dimensions to batch_size × 2048. Finally, the image pixel Feature Alignment Module: Resolves misalignment of leaf-lattice features caused by object deformation or rotation by reconstructing and sampling corrections at feature points, thereby enhancing subsequent modules' perception of object boundaries.

a)To enhance the model's recognition robustness and generalization capability in practical sorting tasks, the selected target object categories should encompass diverse shapes, dimensions, and surface textures. This dataset employs three representative workpiece types—A, B, and C—as grasping subjects. Their morphology and grasping characteristics are representative, and all can be stably grasped by existing robotic grippers.

b)Image acquisition equipment aligns with the actual sorting system, employing a USB industrial camera with a resolution of 1920×1080. This ensures consistency in input data distribution and prevents model performance degradation due to resolution discrepancies.

c)During the specific data collection process, two to three target objects were randomly placed on the experimental platform each time. Various scenarios simulating conditions likely to occur dur-

Fig. 5. Sorting Experiment Platform

ing actual sorting operations—such as stacking, occlusion, and dispersed distribution—were replicated. This ensured the dataset encompassed isolated objects, multi-object scattered arrangements, and partial stacking configurations, thereby enhancing the model's adaptability to complex layouts matrix is input to the fully connected layer FC[12], with corresponding category probabilities output by the softmax classifier. The softmax function used in the classification tasks is expressed by Equation (1) below.

$$y_t = \frac{\exp(a_t)}{\sum_{i=1}^{m} \exp(a_i)} \tag{1}$$

Overall Network Architecture: Backbone Network: ResNet combined with FPN[13] Feature Pyramid Network serves as the backbone network, extracting multi-scale feature maps to accommodate both semantic information and spatial details of objects of varying sizes[14].

6 Attitude estimation and state analysis

Instance segmentation networks can isolate objects from scenes, thereby obtaining pixel coordinates on the object's surface. Using the intrinsic and extrinsic parameter matrices derived from camera calibration, these pixel coordinates can be transformed into a point cloud within the world coordinate system. The world coordinates corresponding to each pixel are: A point cloud is a collection of n points, denoted as Object surface normal estimation based on PCA is achieved by projecting the point cloud onto a plane and analyzing its principal direction of variation[14], as illustrated in Figure 7.Principal Component Analysis (PCA) theory: To eliminate the influence of coordinate

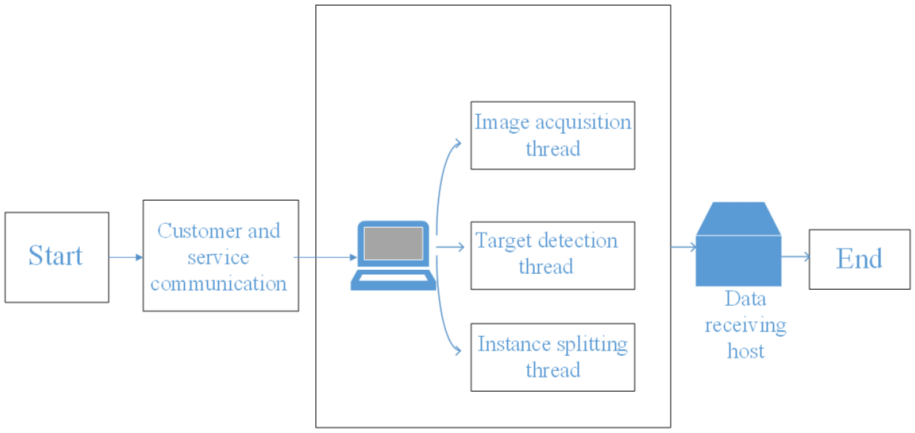

Fig. 6. Software Flowchart of the Robotic Sorting System

magnitude on subsequent analysis, all coordinates must undergo mean-shifting processing. First, calculate the mean position of the point cloud using the formula:

$$\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{2}$$

Redefining normalized zero-mean point clouds, among which . Based on the zero-mean point cloud, construct the covariance matrix [15].M to describe the statistical correlation of the point cloud in the three-dimensional space, expressed by the formula:

$$M = \frac{1}{n}\sum_{i=1}^{n} \tilde{p}_i \tilde{p}_i^T \tag{3}$$

Non-diagonal elements (such as M_{12}, M_{13}, M_{23}) respectively denote the correlations between point cloud coordinates x and y, x and z, and y and z. The specific elements may be expanded as follows:

$$\begin{aligned}
M_{11} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{p}_x)^2,\ M_{12} = M_{21} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{p}_x)(y_i - \bar{p}_y) \\
M_{22} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{p}_y)^2,\ M_{13} = M_{31} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{p}_x)(z_i - \bar{p}_z) \\
M_{33} &= \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{p}_z)^2,\ M_{23} = M_{32} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{p}_y)(z_i - \bar{p}_z)
\end{aligned} \tag{4}$$

Physically speaking, the size of an eigenvalue reflects the extent of a point cloud's dispersion along its respective eigenvector. A lesser eigenvalue indicates a reduced correlation along that particular path. Given the minimal correlation between surface points in the normal direction, the eigenvector linked to the least eigenvalue offers a rough estimate of the surface normal

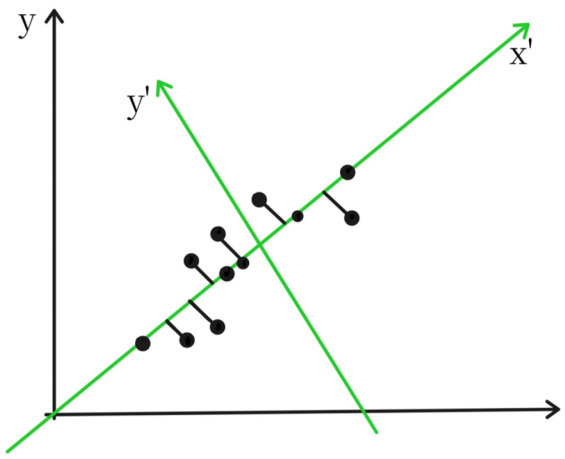

Fig. 7. Surface Analysis Method Projection

7 Applications of intelligent sorting

During robot operation, on-site materials can only be detected via sensors to confirm presence or absence, without directly identifying the current material type. To address this, a vision module has been incorporated. This enables real-time detection of material orientation, type, and position, rendering the entire control process more automated and intelligent[16]. Conventional vision systems employ industrial cameras for sampling, processing images through dedicated software. Such software typically handles only basic operations like shape matching and pattern recognition. At this stage, we incorporate an innovative mask R-CNN for image recognition.

7.1 Real-Time Material Recognition with Mask R-CNN

During the preliminary phase, we trained the model using a labelled dataset of sampled items, enabling it to autonomously recognise and classify the current image. During actual operation, when the sensor detects material arriving on the conveyor belt, it triggers the industrial camera to capture an image [17]. The captured image is fed in real-time to the deployed Mask R-CNN model for inference. This model not only outputs the confidence level for the material category but also generates precise pixel-level segmentation masks corresponding to each material instance, as illustrated in Figure 4. Based on the results of the recognition and segmentation output from Mask R-CNN, the host computer can precisely calculate the position of the center (X, Y coordinates) of the material within the robot coordinate system and its rotational angle[18]. This high-precision pose data is transmitted in real time via the communication interface to the robot controller, thereby guiding the robotic arm to execute precise grasping actions, as illustrated in Figure 8 . Owing to Mask R-CNN's robust instance segmentation capabilities, the system effectively distinguishes

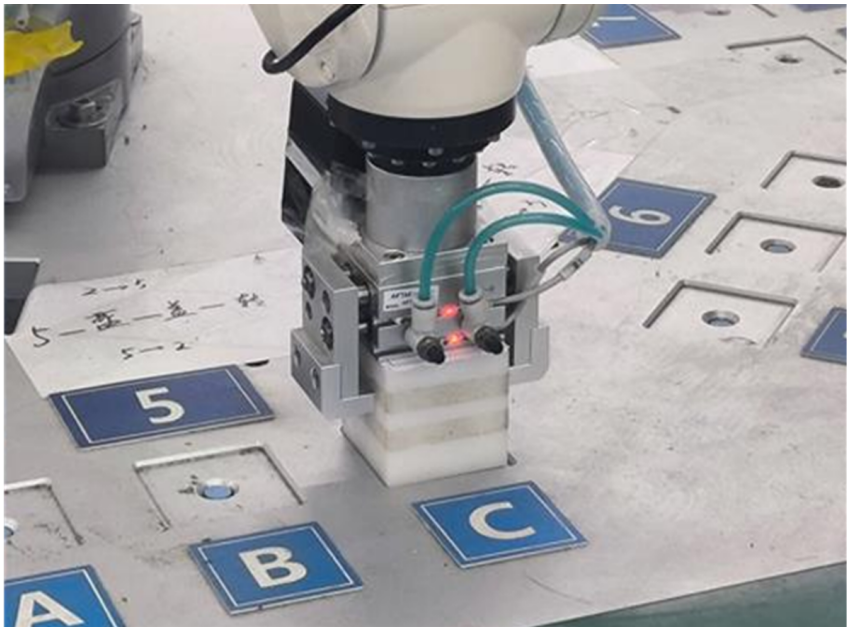

Fig. 8. Material Grasping

materials in contact or partially obscured, substantially reducing the risk of misgrasping or missed grasping.

7.2 Success Criteria and Experimental Design for Multi-Object Sorting

In multi-object sorting experiments, the criterion for success is defined as follows: all target objects must be grasped sequentially in a logical order, and no other objects must be dislodged or displaced during the operation for the experiment to be deemed successful [19]. The core results of the multi-object classification and sorting experiments conducted in this study are summarised in Tables 1 and 2, respectively, with experimental subjects comprising three distinct objects: a, b, and c. In the experimental data capture record sheet, different numbers of repeated experiments were set for different combinations of experimental subjects. This design is primarily based on two considerations: firstly, to ensure that all experimental subjects maintain a consistent average number of captures across all experiments, thereby eliminating potential data bias arising from uneven sample sizes; secondly, to ensure that the total number of experiments remains consistent across different subject quantity combinations, thus strictly controlling experimental variables and enhancing the comparability and rigour of the results.

Table 1: Results of the extraction

Objects Numbers	**1**			**2**			**3**		
Class	a	b	c	a	b	c	a	b	c
Experimental	80	80	80	120	120	120	160	160	160
Number of failures	0	0	0	0	0	0	6	2	2

Table 2: Recognition and Segmentation Accuracy

	Objects Numbers	**a**	**b**	**c**
Method of this paper	Experiential times	200	200	200
	Number of successes	200	195	197
	Rate of success	100%	97.5%	98.5%

7.3 A Qualitative Leap in Sorting: From Simple Matching to Mask R-CNN Precision

For a robot to be deemed 'successful capture' as per the table, it needs to maintain a steady hold on the target object and precisely position it at the intended spot, avoiding any irregular events like unintended collisions or tipping during the entire process. Failure to meet any of the above conditions is deemed a grasping failure. Experimental data indicate that the grasping success rates for all objects remain at a high level: object a achieved 99.67%, object b reached 100%, and object c attained 99.22%. This demonstrates that the constructed grasping system exhibits excellent stability and reliability. Experimental results demonstrate that the system achieves an average recognition and segmentation accuracy (mAP) of 98.2% across multiple material types. Within actual production settings, the extensive success rate in sorting consistently surpasses 99%, confirming the vast practical utility and capabilities of sophisticated deep learning models in industrial vision evaluation and robotic management.

8 Conclusion

This paper designs and implements a robot vision guidance system based on mask r-cnn instance segmentation model, which effectively solves the urgent needs of insufficient sorting accuracy and low intelligent level in the intelligent storage system. This research deeply integrates advanced in-depth learning technology into the traditional sorting control process, completely replacing the traditional image processing scheme that can only achieve basic shape matching, significantly improving the accuracy and intelligence of sorting operations, and providing reliable technical support for improving the efficiency of intelligent warehousing sorting.

Although the current system has achieved the expected goal and achieved satisfactory application results, there is still a large space for optimization. In the future, in-depth research can be focused on the following directions. First, explore lighter network structure or efficient model com-

pression technology to further improve the real-time response speed of the system, so as to adapt to higher production throughput requirements and meet the efficient operation requirements of large-scale warehousing and sorting. Secondly, the research and development of adaptive online learning mechanism can quickly identify new product categories without retraining the model, reduce the cost of system adaptation, and enhance its scene adaptability, which has important practical application value. Finally, promote the deep integration of the system and the warehouse management system, build a fully intelligent end-to-end warehousing logistics solution, and realize the automation and intelligence of the whole process of warehousing, sorting and scheduling, which is also the core development goal of the future intelligent warehousing field.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proc. IEEE Int. Conf. Comput. Vis.; 2017. p. 2961-9.

[2] Zhang H, Liang H, Ni T, Wang S, Li Y. Research on a multi-object sorting system based on deep learning. Sensors. 2021;21(18):6238.

[3] Tang P, Liu Y, Wei H, Wang J. Automatic recognition algorithm for digital instrument readings at offshore step-up substations based on Mask-RCNN. Infrared Laser Eng. 2021;50(S2):163-70.

[4] Huang X, Zhang Y, Wang L, Deng Y. Infrared Image Segmentation and Temperature Reading of Composite Insulator Strings Based on Mask-RCNN Algorithm. High Volt App. 2021;57(09):87-94.

[5] Liu Z, Fu H, Li Y, He C. Detection of Power Equipment in Infrared Images Based on Mask-RCNN Transfer Learning. Data Acquis Process. 2021;36(01):176-83.

[6] Zhong W, Liu X, Yang K, He Y. Segmentation and Recognition of Multiple Target Leaves in Complex Backgrounds Based on Mask-RCNN. J Zhejiang Agric. 2020;32(11):2059-66.

[7] Li H, Li M, Li K, Zhang Y. Application of Mask R-CNN Model in Road Surface Defect Detection. Sci Technol Innov. 2020;29:131-2.

[8] He D, Shi W, Lin Z, Lin J. Building Extraction from Remote Sensing Images Based on an Improved Mask-RCNN. Comput Syst Appl. 2020;29(09):156-63.

[9] Yang Z, He S, Feng W, Yang Y. Intelligent Identification and Application of Wear Particles Based on Mask R-CNN Network. J Tribol. 2021;41(01):105-14.

[10] Yue Y, Tian B, Wang H, Li J. Apple Detection in Complex Environments Based on an Improved Mask R-CNN. Chin J Agri Mach Chem. 2019;40(10):128-34.

[11] Ou P, Lu K, Zhang Z, Guo F. Object Recognition and Spatial Localisation Based on Mask R-CNN. Comput Meas Control. 2019;27(06):172-6.

[12] Li D, He W, Guo B, Wang C. Building an Object Detection Algorithm Based on Mask-RCNN. Sci Surv Mapp. 2019;44(10):172-80.

[13] Shi J, Zhou Y, Zhang Q. Service robot object recognition system based on improved Mask RCNN and Kinect. Chin J Instrum Meas. 2019;40(04):216-28.

[14] Xu X. Research on Key Technologies for Intelligent Warehouse Systems Based on Machine Vision; 2024. PhD thesis, Hefei University of Technology.

[15] You J. Road Crack Detection Based on an Improved Mask-RCNN. Video Eng. 2022;46(06):7-9+19.

[16] Liu F. Research on Key Technologies of Fruit Grading and Sorting System Based on Machine Vision. Mech Electr Inf. 2021;28:56-7+60.

[17] Song Z, Zhang J, Qin X, Chen W. A Quantification Method for Pigment Deposition in Cephalopod Beaks Based on Mask-RCNN Image Segmentation. Fishery Modern. 2021;48(05):70-8.

[18] Nie Z, Ren J, Lu J. Ship Traffic Detection in Foggy Background Based on Mask R-CNN. Trans Beijing Inst Technol. 2020;40(11):1223-9.

[19] Li H, Xia Q. Research on Robotic Fish Localization Technology Based on Mask R-CNN. Robot Technol Appl. 2019;05:36-9.