Research on mechanical part recognition method based on improved mask R-CNN instance segmentation

Kui Xiao[1], Lei Wang[2], Haoran Xu[3], Pengchao Zhang[4], Heng Zhang[5]
School of Mechanical Engineering, Shaanxi University of Technology, Hanzhong, Shaanxi, 723001, China
The Key Laboratory of Industrial Automation of Shaanxi Province, Shaanxi University of Technology, Hanzhong, 723001, China
[2]Corresponding author
E-mail: [1]*17734691980@163.com*, [2]*WangLei_sut@163.com*, [3]*357494247@qq.com*, [4]*8811202@qq.com*, [5]*3210001542@qq.com*

Received 31 August 2024; accepted 10 January 2025; published online 23 January 2025
DOI https://doi.org/10.21595/rsa.2025.24518

Abstract. Aiming at the problem of poor part recognition due to mutual occlusion between parts and the influence of different postures in the assembly scene, we propose an improved Mask R-CNN-based part recognition method for complex scenes. Firstly, the ResNet101 network is used to enhance the feature extraction capability of the network and improve the part recognition effect; secondly, the normalization layer of the backbone network is replaced to reduce the effect of batch size on the feature extraction of the model; lastly, the feature pyramid network structure is improved to enhance the transfer efficiency between the high and low layers of the network, and to enhance the capability of the feature capture; through the experiments on the homemade dataset, the average detection accuracy of this method is 4.7 % higher than that of the original Mask R-CNN. Through the experiments on the homemade dataset, it is found that compared with the original Mask R-CNN, the average detection accuracy of the method is improved by 4.7 %.The optimized network model proposed in this paper can improve the accuracy of part recognition, realize the accurate detection of parts in the complex environment such as stacking, occlusion and so on, and provide a solution for the recognition of parts in the complex environment.

Keywords: ResNet101, identification and localization, object detection, deep learning, instance segmentation.

1. Introduction

As an important part of mechanical production, parts are complex and numerous, and misidentification of parts can lead to reduced productivity and waste of resources, and even lead to accidents. Therefore, the accurate identification of mechanical parts is a hot issue in current research.

Object recognition is a popular task, Chen [1] Using a smartwatch for daily object recognition and Zhang [2] Monitoring the state of objects through vibration. It is also widely applied in the recognition of mechanical parts Traditional manual identification is inefficient and subjective; using a camera to obtain part images and using machine learning algorithms to perceive the target can overcome these problems, therefore, part identification and localization [3-4] has become an important part in the study of robot vision tasks. Currently, the mainstream machine vision-based part recognition methods are deep learning and template matching. The template matching based methods are to compare the feature vectors of the targets with the template vectors in the library, and the commonly used methods are supporting vector machine (SVM), edge detection, feature description methods and corner detection, etc. focusing on the description of the target features. For example, Peisi Zhong et al. [5] proposed a feature point matching based machine vision workpiece identification method for workpiece identification, which achieved accurate workpiece identification. Wang [6] et al. improved the matching accuracy of the algorithm by feature matching the stable points of descriptors under different affine variations as feature points. Sasikala [7] et al. identified the steering racks by calculating the number of maximal normalized

correlation components with a multi-objective and template model to achieve bogie recognition with less computational effort. The template matching method is highly depend on the feature operators and has stringent requirements on the detection conditions and new designs are needed for different tasks.

With the continuous improvement of computer performance and the development of deep learning technology, neural networks have achieved significant advantages in the field of part recognition. For example, Wang Yi-Yi et al. [8] recognized parts by cascading Faster R-CNN, and then designed a dichotomous rotation method to realize automated grasping of manipulators; Yuan Bin et al. [9] achieved high recognition accuracy and grasping success rate by extracting the region of interest from the target recognition results of YOLOv4 and feeding it into PSPnet semantic segmentation network, then performing sub-pixel level template matching on the region of interest. accuracy and success rate of crawling are achieved. Wang Xiangzhou [10] et al. proposed a system for detecting and locating main bolts in angle steel towers based on a lightweight YOLOv5-T neural network, which achieves accurate identification and localization of main bolts. It can be seen that deep learning can extract more feature information of parts through neural network with better recognition effect. However, the simple target detection algorithm cannot detect the edge features of the targets, and it cannot provide sufficient edge information for the subsequent grasping of mechanical parts. Therefore, the parts are detected by Instance Segmentation [11] (IS) and further information such as the exact position, edges and pose of the target is obtained [12], so as to obtain the complete information required for part grasping.

In this paper, an improved mask R-CNN [13] instance segmentation model is designed to solve the problem that the original model is inefficient in recognition in complex environments, which affects the accuracy of part recognition. Firstly, the feature extraction capability of the network is enhanced by using ResNet101 [14] as the backbone network. Secondly, the normalisation of the network is improved to reduce the impact of batch size on the recognition results. Then, the structure of the feature fusion module of the network is improved to increase the utilisation of semantic information by the network. Finally, the recognition effect of the model is evaluated by experiments on self-constructed datasets, and it is found that the improved network in this paper improves the detection accuracy by 4.7 % to 91.3 % compared with the original Mask R-CNN, which basically achieves the accurate recognition of mechanical parts in complex environments.

– We propose an improved Mask R-CNN model and it is very robust to occlusions between objects.

– We introduce a new feature fusion model and insert it into the Mask R-CNN model.

– We contribute a dataset for instance segmentation which includes many special conditions.

2. Mask R-CNN detection model

As shown in Fig. 1, Mask R-CNN as a two-stage instance segmentation network, it predicts the target instances by adding a branch on Faster R-CNN, which realizes the acquisition of information such as target species, coordinates, edges, etc., [15] and performs end-to-end learning. the overall structure of the network mainly consists of the Backbone, Region Proposal Network (RPN), Feature Pyramid Networks (FPN), pixel-level instance segmentation outputs and so on. Its main advantage is in Feature Pyramid Networks and Region of Interest Alignment (ROI Align).

2.1. Feature pyramid networks

The feature pyramid network in Mask R-CNN mainly includes top-down, bottom-up, horizontal join and convolutional fusion. By up-sampling the high-level feature map, connecting the high-level feature information with the low-level features side by side, and then using the feature pyramid as the base structure, predicting the feature maps of each level separately. it can utilize the features of the high-level and the low-level at the same time and predict them on different layers, outputting the features of different scales for the parts with different scales, using

the large feature maps to distinguish between the simple targets and the smaller feature maps to distinguish between the complex targets and guides the recognition of the shallow network, allowing the network to extract rich semantic information. This network structure, which simply changes the connection of the network, can trade-off the speed and accuracy of detection, obtain more robust target semantic information. It essentially no increase in calculations and substantially improve the network performance.

2.2. ROI align

Mask R-CNN uses the RoI Align layer to pool the region of interest into a fixed-size feature map, by traversing each candidate region, obtaining the image values of the pixels in the form of floating-point numbers and splitting the candidate region into a number of units, and then transforming the aggregation of the whole network into a continuous operation, eliminating the quantization operation of the network, and adopting the bilinear interpolation to compute the fixed coordinate positions , and then the maximum pooling operation is performed to calculate the position of the regression frame. Without quantizing the position information, the pixel information of small targets in the image distant view can be effectively retained, and the effective detection and recognition of small targets can be realized.

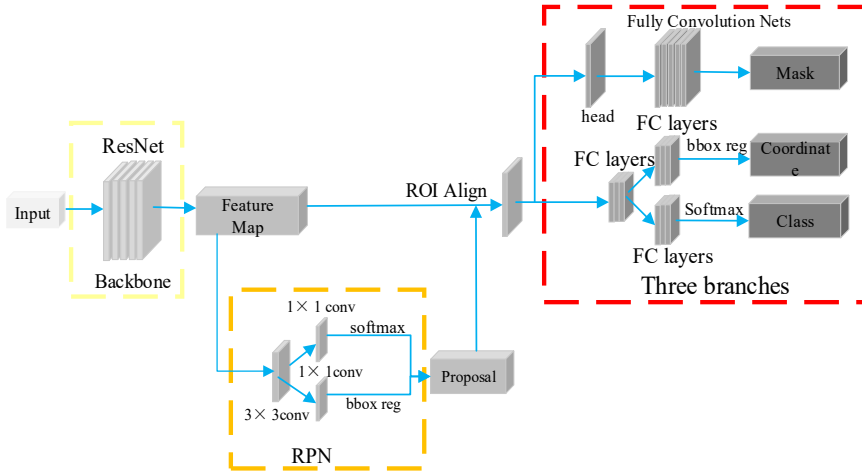


Fig. 1. Structure of the Mask R-CNN network

3. Improved mask R-CNN model

The stacking of parts leads to a serious loss of semantic information of the parts, so to improve the feature extraction ability of the network and the feature transfer ability of the network, so in this section, we propose to improve the backbone network and feature fusion network of the original model to realize the network's high efficiency of recognition in the case of part stacking.

3.1. Backbone network selection

The backbone networks of Mask R-CNN are generally ResNet50 and ResNet101, and the difference mainly lies in the depth of the network, the depth of the network determines the expression ability of the network, and the deeper the network is the stronger the learning ability. In order to enhance the feature extraction ability of the network, this paper adopts ResNet101 backbone network, and its specific network structure is shown in Fig. 2.

ResNet reduces the number of parameters and increases the learning ability of the network for features by introducing residual units and adding directly connected channels in the network,

which has the characteristics of few parameters and excellent recognition effect, etc. ResNet101 consists of a total of 101 layers, which includes the first 7×7 convolutional layer, four stages, the last layer is a global average pooling and a layer of fully connected layers. Each of these stages contains several residual blocks.

Each residual block contains two 3×3 convolutional layers followed by a batch normalization layer and an activation function, and there are also batch normalization layers and activation functions between the residual blocks.ResNet101 allows the network to focus directly on the residuals of the inputs and the outputs during the learning process, which enhances the feature extraction capability of the network and improves the recognition efficiency of the network.

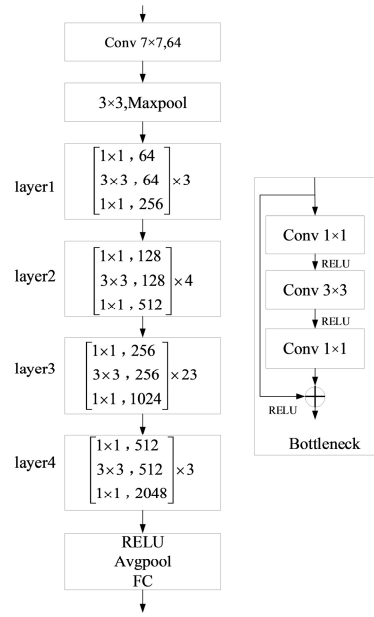


Fig. 2. ResNet101 network structure

3.2. Feature pyramid network improvement

Before the feature pyramid network (FPN [16]) fuses the features, the different stages of the backbone network first carry out the feature extraction of the targets through convolution, and then the feature pyramid network up-samples the feature maps of the upper layer and down-samples the feature maps of the following layer to realize feature fusion, and then the fused feature maps are fused with the lower features through the same operation. Because different stages have different sizes of sensory fields and contain different semantic information, the higher convolutional layer has more semantic information and the lower convolutional layer has more positional information, so the direct feature summation operation will weaken the feature extraction ability of the network and affect the detection effect of the network. In order to better output the corresponding targets on top of different layers, improve the efficiency of the feature fusion of the network to accelerate the operation of the network, and improve the detection performance of the network, this paper proposes an optimized feature pyramid network.

Fig. 3 shows the structure of the improved feature pyramid network, the optimized network obtains more robust target semantic features by adding lateral connections, top-down and bottom-up paths. Features are fused with neighboring paths by 1×1 convolution on each feature map; the blue solid line is the lateral connection, where features are projected by convolution on each feature map and fused with neighboring paths. The green diagonal solid line is the pyramid channel, which is the bottom-up information flow, where high-level feature maps are up-sampled

by using nearest-neighbor interpolation and then fused with lower-level features by convolution. The red dashed curves are skip connections, which are connected by 1×1 convolution to simplify the training of the feature pyramid network. Compared with the traditional feature pyramid network, the improved network can fuse more semantic information, improve the recognition accuracy of small parts, and reduce the impact of the lack of semantic information caused by the stacking of small parts on the recognition of parts, thus reducing the probability of misrecognition of parts.

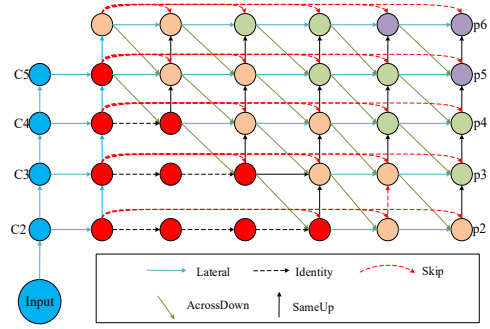
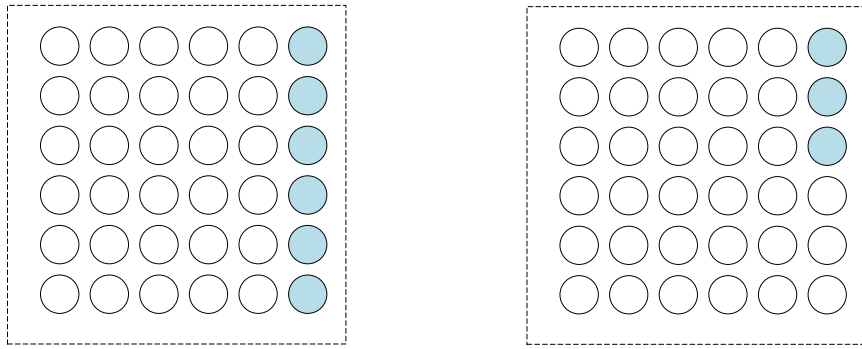


Fig. 3. Improved feature fusion network

3.3. Improvement of normalization method

The original Batch Normalization (BN) is only applicable to a larger number of batches, and the performance drops dramatically when the Batch Size is small. Group Normalization (GN) has greater stability at smaller Batch Size by dividing the input channels by groups and normalizing the computation for each group. A comparison of the two approaches is shown in Fig. 4, where the normalization is performed by the mean and variance obtained by aggregating the pixels in the green part.



a) Batchnormlization b) Groupnormalization

Fig. 4. Improvement of normalization method

4. Experimental validation

We implement our network using the publicly available Pytorch and the operating system is Ubuntu 20.04LTS, the computer processor is Intel (R) core (TM) i7-6800kCPU, while using NVIDIA TITAN Xp to accelerate the training, the memory is 32GB, and the dataset used in the experiment is the dataset of self-made parts. The image resolution is 640×640. The samples contain individual parts, part mixing and part stacking, etc. The total number of samples is 160 and it includes the condition of multiple part stacking and one part stacking, which is expanded to 1200 by data enhancement.

4.1. Experimental dataset

The experiments use self-made parts picture dataset, the samples cover the two main mechanical parts of bolts and nuts and mainly contain the picture dataset in the case of parts stacking. Bolts and nuts as the most common machine parts are more prone to the complex condition of their stacking. Therefore, we choose bolts and nuts as our objects. The dataset details are shown in Fig. 5, including individual parts, multiple parts and stacked parts.

a) Single b) Multie c) Stack

Fig. 5. Image datasets

In order to improve the generalization of the network and to simulate complex environments in production, we used various means of data enhancement such as changing the illumination, contrast and adding mosaics. [17] Not only the dataset be greatly expanded, but also retain image features, increase the diversity of sample training, and improve the generalization ability of the network. After enhancing the dataset, it is fused with the original image to make the robustness of the model more stable. The image resolution is 640×640.The dataset is labeled by Labelme labeling software to obtain a JSON format file.

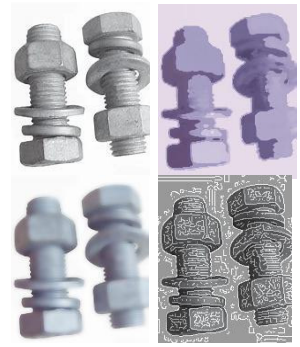


Fig. 6. Data enhancement effect diagram

4.2. Experimental results

The dataset is divided proportionally where 70 % is training set and 30 % is test set and the data format is coco dataset format. The loss variation curve of the improved network is shown in Fig. 6. The parameters for training the network are set, the training period is 60 and the initial learning rate is a training cycle is 1000 steps, the momentum factor is 0.9, in order to verify the model's detection and segmentation ability in different scenarios, the training weights of the last round are selected for testing.

Fig. 4 shows the loss change curve of the improved network, including the overall, Bbox, Class, and Mask loss change curve. Among them, Class and Bbox fluctuates greatly at the beginning of training, decreases rapidly in the first and middle stages, and then gradually stabilizes and converges in the middle and late stages. mask and total loss, which decreases rapidly at the beginning, stabilizes and converges in the middle and late stages.

When the network is trained, in order to obtain better network training weights, the model test results with iteration cycles of 20, 40, 60, and 80 are selected for evaluation [23], and the network training loss and detection accuracy are shown in Table 1, with the increase of the network training cycle and the change of the learning rate, the network loss is able to effectively decrease and

converge, indicating that the improvement of the detection accuracy of the network model is constantly improving. However, when the training period is greater than 60, the detection accuracy is almost unchanged after the number of training rounds increases.

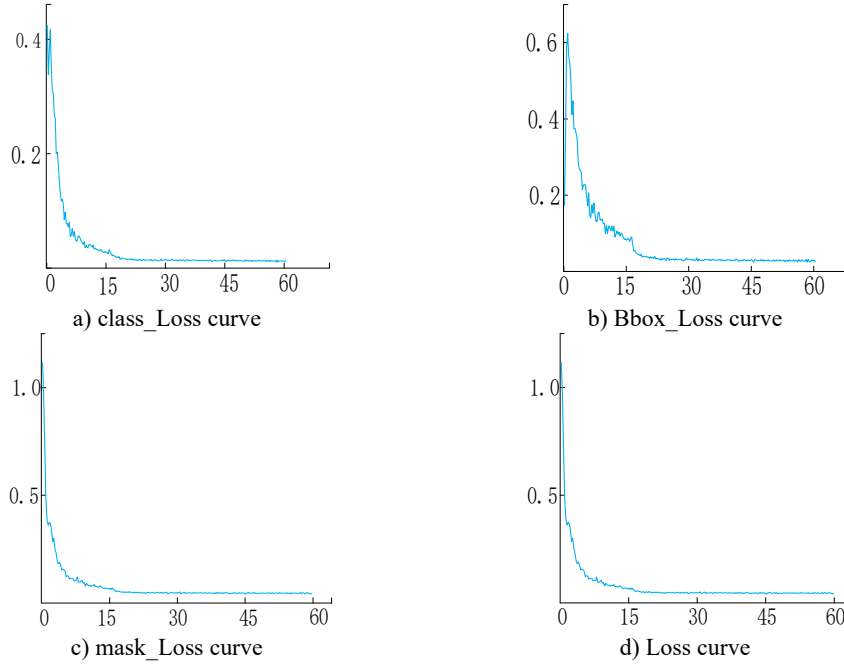


Fig. 7. Loss function change curve

Table 1. Network training loss and detection accuracy

Training cycle	Loss	Detection precision / %
20	0.8125	88.9
40	0.1139	90.3
60	0.1009	91.5
80	0.0919	91.6

Therefore, the network model trained with a cycle number of 60 is selected, and images are randomly selected from the test sample set to perform test experiments and calculate the detection accuracy of the network. The test results are measured using the detection precision Precision, recall rate Recall and average detection precision mAP indicators to visualize the performance of the improved network model, the formulas are shown in Eqs. (1-5):

$$precision = \frac{TP}{TP + FP}, \tag{1}$$

$$recall = \frac{TP}{TP + FN}, \tag{2}$$

$$F1 = 2 \times \frac{precision \cdot recall}{(precision + recall)}, \tag{3}$$

$$AP = \int_0^1 p(r)\, dr, \tag{4}$$

$$mAP = \sum_{i=1}^{C} \frac{Ap}{C}, \tag{5}$$

where: TP is the number of correctly identified and calibrated targets; FP is the number of incorrectly identified but calibrated targets; FN is the number of incorrectly identified and failed to calibrate targets. AP combines Precision and Recall to provide a comprehensive assessment. The F1 Score is the reconciled average of Precision and Recall and is a composite metric. MAP is an extension of AP, which is usually used to evaluate the performance of multi-category classification tasks or multiple queries. AP and MAP are mainly used to evaluate the performance of tasks such as information retrieval and target detection, focusing on the combination of Precision and Recall. F1 Score is a generic categorization performance metric suitable for unbalanced categorization tasks.

In order to verify the effectiveness of the part recognition model in this paper and analyze the role of the proposed improvements, the ablation experiments after the replacement of each module are designed [18-22], which are visually demonstrated by the accuracy and detection precision of the model, and the results are shown in Table 2. Firstly, in order to verify the effect of the depth of the backbone network on the detection accuracy, different backbone networks are used for the experiments respectively, and then the model accuracy is compared, and it can be found that with the improvement of the backbone network, the detection accuracy of the network is increased from 86.2 % to 89.2 %, which is an increase of 3 %, which proves that there is a significant improvement in the detection effect of the network with the increase of the depth of the network. Secondly, it was experimentally found that the detection accuracy of the network increased by 1.8 % to 88.4 % after the Group Normalizatin (GN) normalization, which proved the effectiveness of the module. Finally, the model after improving the feature fusion network is compared with the original model in the experiment, and it is found that the detection precision of the network is increased by 2.4 %, and the recall rate is increased from 89.2 % to 91.5 %, which proves that the improved feature fusion network is able to fuse the semantic information of more targets, and improve the detection efficiency of the network.

Table 2. Comparison of ablation experiment models

Model	Recall / %	Precision / %	F1 / %	Seg_mAP / %	Bbox_mAP / %	Size of the model / M	FPS / Gflops
Mask R-CNN	89.2	86.6	87.9	87.8	96.7	43.75	258.17
Mask R-CNN+r101	92.5	89.2	90.8	89.2	97.8	62.74	334.24
Mask R-CNN+GN	90.4	88.4	89.3	88.4	97.7	45.79	465.04
Mask R-CNN+M-FPN	91.5	89.0	90.2	88.9	97.6	47.05	356.32
Ours	94.2	91.3	92.7	92.1	98.0	66.03	480.34

In order to verify the reliability of this paper's algorithm relative to the current mainstream algorithms, this paper's algorithm is compared with the mainstream algorithms including Mask Cascade R-CNN, Mask Scoring R-CNN, as shown in Table 3, this paper's algorithm relative to other algorithms, in the case of about the same size, there is a significant improvement in the precision and recall. Therefore, the part recognition algorithm in this paper is effective.

Table 3. Comparative experiments with other common models

Model	Recall / %	Precision / %	F1 / %	Seg_mAP / %	Bbox_mAP / %	Size of the model / M	FPS / Gflops
Mask R-CNN	89.2	86.6	87.9	87.8	96.7	43.75	258.17
MS R-CNN [23]	91.3	89.1	90.3	88.9	97.8	60.01	458.17
Mask Cascade R-CNN [24]	92.3	90.4	91.3	88.4	97.7	62.79	465.04
Ours	94.2	91.3	92.7	92.1	98.0	66.03	480.34

Finally, in order to visualize the detection and segmentation effect of the model on the parts, the experimental needle extraction part of the picture for detection, Fig. 7 for the detection and segmentation results of the network in the case of parts stacking, as can be seen from the figure, the improved Mask R-CNN model in the parts of the parts of the stacking and other complex

situations not only can be effectively realized in the detection of the parts, but also the edge of the parts of the segmentation, has a good feasibility and robustness.

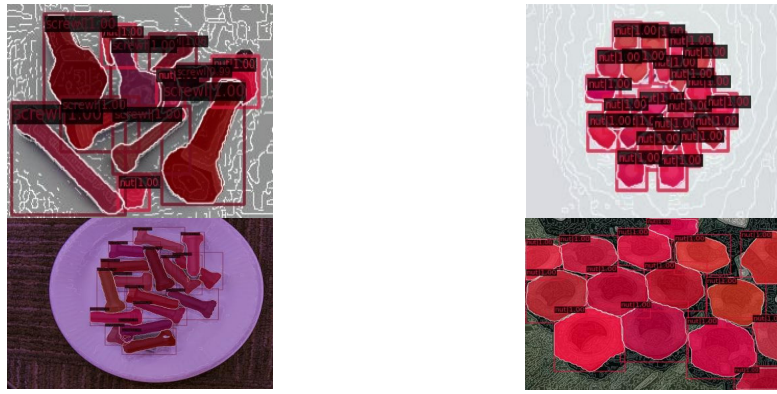


Fig. 8. Model detection and segmentation results

5. Conclusions

In this paper, we propose a improved Mask R-CNN, a better extraction and utilization of semantic and detailed features model for object detection. We have made it possible to carefully design its backbone network and feature fusion network so that it can better extract and transfer feature information, thus improving the recognition efficiency of the network, which makes it very promising for applications in the field of machine part recognition. As a result, the network in this paper is able to handle occluded objects in cluttered scenes. Our results are very encouraging as they show that our model can largely achieve part recognition in complex environments. We note that the recognition method in this paper still has deficiencies, the model is too large and need more time to work, the next step needs to be specialized training for part recognition in special conditions and to lighten the network and improve the recognition speed and increase the recognition speed.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62176146, the National Social Science Foundation of China under Grant 21XTY012, the National Education Science Foundation of China under Grant BCA200083, and Key Project of Shaanxi Provincial Natural Science Basic Research Program under Grant 2023-JC-ZD-34.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author contributions

Kui Xiao is primarily responsible for enhancing the algorithm, conducting experimental training of the model, planning the research content, writing the paper, and creating the illustrations. Lei Wang and Peng-Chao Zhang are mainly responsible for the research findings and revisions of the paper. Hao-ran Xu is responsible for organizing the experimental data. Heng Zhang is in charge of capturing the dataset and preprocessing the images.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] W. Chen et al., "ViObject: harness passive vibrations for daily object recognition with commodity smartwatches," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 8, No. 1, pp. 1–26, Mar. 2024, https://doi.org/10.1145/3643547

[2] D. Zhang, A. Polamarasetty, M. O. Shahid, B. Krishnaswamy, and C. Ma, "Metamaterial-based passive analog processor for wireless vibration sensing," *Communications Engineering*, Vol. 3, No. 1, Mar. 2024, https://doi.org/10.1038/s44172-024-00190-8

[3] Y. Wang et al., "Key technologies of robot perception and control and its intelligent manu-facturing applications," (in Chinese), *Acta Automatica Sinica*, Vol. 49, No. 3, pp. 494–513, 2023, https://doi.org/10.16383/j.aas.c220995

[4] Z. Tian et al., "Part recognition and assembly monitoring based on depth images," *Computer Integrated Manufacturing Systems*, Vol. 26, No. 2, pp. 300–311, 2020, https://doi.org/10.13196/j.cims.2020.02.003

[5] P. S. Zhong et al., "Workpiece recognition of assembly robot," (in Chinese), *Manufacturing Technology and MachineTool*, pp. 65–70, Mar. 2023, https://doi.org/10.19287/j.mtmt.1005-2402.2023.03.008

[6] Z. Wang, Z. Li, L. Cheng, and G. Yan, "An improved ORB feature extraction and matching algorithm based on affine transformation," in *Chinese Automation Congress (CAC)*, pp. 1511–1515, Nov. 2020, https://doi.org/10.1109/cac51589.2020.9327165

[7] N. Sasikala and P. V. V. Kishore, "Train bogie part recognition with multi-object multi-template matching adaptive algorithm," *Journal of King Saud University – Computer and Information Sciences*, Vol. 32, No. 5, pp. 608–617, Jun. 2020, https://doi.org/10.1016/j.jksuci.2017.10.001

[8] W. Yi, M. Zhengdong, and D. Guanglin, "Parts recognition method based on improved Faster RCNN," (in Chinese), *Journal of Applied Optics*, Vol. 43, No. 1, pp. 67–73, Jan. 2022, https://doi.org/10.5768/jao202243.0102003

[9] B. Yuan et al., "Research on part recognition and grasping methods of visual robots," (in Chinese), *Mechanical Design and Manufacturing*, 2024, https://doi.org/10.19356/j.cnki.1001-3997.20230824.045

[10] X. Z. Wang et al., "Bolt detection and positioning system based on YOLOv5s-T and RGB-D camera," (in Chinese), *Transactions of Beijing institute of Technology*, Vol. 42, No. 11, pp. 1159–1166, 2022, https://doi.org/10.15918/j.tbit1001-0645.2021.339

[11] X. Y. Liang et al., "Research on the progress of image instance segmentation based on deep learning," (in Chinese), *Acta Electonical Sinica*, Vol. 48, No. 12, pp. 2476–2486, 2020, https://doi.org/10.3969/j

[12] H. D. Zhang et al., "Research progress on visual based defect detection of automotive assembly parts," (in Chinese), *Chinese Journal of Scientific Instrument*, Vol. 44, No. 8, pp. 1–20, 2023, https://doi.org/10.19650/j.cnki.cjsi.j2311695

[13] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, Oct. 2017, https://doi.org/10.1109/iccv.2017.322

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Jun. 2016, https://doi.org/10.1109/cvpr.2016.90

[15] Z. Lin, S. Yoshikawa, M. Hamasaki, K. Kikuchi, and S. Hosoya, "Automated phenotyping empowered by deep learning for genomic prediction of body size in the tiger pufferfish, Takifugu rubripes," *Aquaculture*, Vol. 595, p. 741491, Jan. 2025, https://doi.org/10.1016/j.aquaculture.2024.741491

[16] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, Jul. 2017, https://doi.org/10.1109/cvpr.2017.106

[17] S. R. Yang et al., "Survey of lmage data augmentation techniques based on deep learning," (in Chinese), *Journal of Software*, p. 2024, Sep. 2024, https://doi.org/10.13328/j.cnki.jos.007263

[18] T.-K. Lau and K.-H. Huang, "A timely and accurate approach to nearshore oil spill monitoring using deep learning and GIS," *Science of The Total Environment*, Vol. 912, p. 169500, Feb. 2024, https://doi.org/10.1016/j.scitotenv.2023.169500

[19] Y. Zhang, Y. Ma, Y. Li, and L. Wen, "Intelligent analysis method of dam material gradation for asphalt-core rock-fill dam based on enhanced Cascade Mask R-CNN and GCNet," *Advanced Engineering Informatics*, Vol. 56, p. 102001, Apr. 2023, https://doi.org/10.1016/j.aei.2023.102001

[20] X. Qu, J. Wang, X. Wang, Y. Hu, T. Zeng, and T. Tan, "Gravelly soil uniformity identification based on the optimized Mask R-CNN model," *Expert Systems with Applications*, Vol. 212, p. 118837, Feb. 2023, https://doi.org/10.1016/j.eswa.2022.118837

[21] Z. Liu et al., "Automatic pixel-level detection of vertical cracks in asphalt pavement based on GPR investigation and improved mask R-CNN," *Automation in Construction*, Vol. 146, p. 104689, Feb. 2023, https://doi.org/10.1016/j.autcon.2022.104689

[22] X. R. Wu, T. T. Qiu, and Y. N. >Wang, "Multi-object detection and segmentation for traffic scene based on improved Mask R-CNN," (in Chinese), *Chinese Journal of Scientific Instrument*, Vol. 42, No. 7, pp. 242–249, 2021, https://doi.org/10.19650/j.cnki.cjsi.j2107749

[23] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6409–6418, Jun. 2019, https://doi.org/10.1109/cvpr.2019.00657

[24] Z. Cai and N. Vasconcelos, "Cascade R-CNN: high quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 5, pp. 1483–1498, May 2021, https://doi.org/10.1109/tpami.2019.2956516

Kui Xiao Master's degree student, Shaanxi University of Technology, mainly engaged in machine vision 3D reconstruction.

Lei Wang Professor, Shaanxi University of Technology, mainly engaged in intelligent computing, machine learning, pattern recognition, and big data analysis technologies.

Hao-Ran Xu Master's degree student, Shaanxi University of Technology, mainly engaged in computer vision target detection

Peng-Chao Zhang Professor, Shaanxi University of Technology, mainly engaged in robotics and control engineering technology research.

Heng Zhang Master's degree student, Shaanxi University of Technology, mainly engaged in computer vision image super-resolution.