Robot Arm Grasping and Automotive Unidirectional Device Defect Detection Based on Dual Modal Mask R-CNN

Chunsheng LI*, Jianlin CHEN

School of Intelligent Science and Technology, Geely University of China, No. 123, Section 2, Chengjian Avenue, Eastern New District, ChengDu, 641423, China
lichunsheng@bgu.edu.cn (**Corresponding author*), 369423772@qq.com, 1310531802@qq.com

Abstract: This study proposes an improved mask model based on a bimodal convolutional neural network for addressing the limitations of the existing detection methods with regard to the processing of multi-source information. The purpose is to improve the model`s recognition accuracy for complex targets in the context of robot arm grasping tasks, as well as the accuracy of defect detection for automotive unidirectional devices. Firstly, based on the mask model including a convolutional neural network, a bimodal input model is constructed by combining red-green-blue images and depth images in order to enhance the model's ability to perceive the three-dimensional structure of objects. Secondly, the model's feature extraction performance is enhanced by introducing a convolutional block attention module, and the final robot arm grasping system and the automobile unidirectional device defect detection system are constructed. The experimental results demonstrated that the proposed model achieved a detection accuracy of 0.97 for benchmark performance testing and a recognition time as low as 0.01 seconds for actual robot arm grasping tasks. In addition, the recognition accuracy of the proposed model in the context of five different robot arm grasping scenarios was as high as 98.84%, and the defect detection accuracy was as high as 98.25%. Therefore, this model can well optimize the efficiency of the current robotic arm grasping and defect detection tasks.

Keywords: Mask R-CNN, Defect detection, Mechanical arm, Deep learning, Unidirectional device, Target recognition.

1. Introduction

With the rapid development of industrial automation and intelligent manufacturing, robotic arms have come to be widely used in the production field. At the same time, due to an accurate and efficient target recognition and defect detection technology, robotic arms have become a key factor in improving production efficiency and product quality (Han et al., 2024; Yasir & Ahn, 2023). In complex industrial environments, robotic arms need to be able to accurately identify multiple target objects. Especially with regard to defect detection for automotive unidirectional devices, the traditional detection methods often struggle to detect subtle structural defects (Fişne, Kalay & Eken, 2024; Mani et al., 2023). At present, machine vision technology is being widely applied in this field. The Mask Region-based Convolutional Neural Network (Mask R-CNN) has shown great potential in industrial visual inspection due to its superior object detection and instance segmentation capabilities (Macías-Macías et al., 2023; Benmiloud, Kouzi & Ameur, 2023). However, the traditional single modal Mask R-CNN models mainly rely on Red-Green-Blue (RGB) images for object detection. This method has a limited ability to process 3D information in complex environments and it is difficult for it to cope with multi-dimensional object recognition tasks in different scenarios (Abdullah et al., 2024). Based on this, this study proposes a dual modal Mask R-CNN-based method for RAG and Automotive Unidirectional Device Defect Detection (AUDDD). The innovation brought about by this paper lies in the use of a bimodal data fusion method for optimizing the object detection capability of Mask R-CNN, and the combination of multi-level feature extraction and an attention mechanism for improving the feature extraction effect.

The remainder of this paper is structured as follows. Section 2 reviews the related work on Mask R-CNN–based industrial inspection, RGB depth fusion, and robotic grasping. Section 3 presents the proposed bimodal Mask R-CNN model, including the RGB depth fusion strategy (CBAM) and the two application pipelines – the robotic arm grasping and the automotive one-way device defect detection. Further on, Section 4 describes the employed datasets, the model implementation details, and evaluation protocols, while discussing the results of an ablation experiment and comparing the proposed model with the YOLOv5, Faster R-CNN, and HTC algorithms using MAE, MSE, the ROC curve, detection accuracy, and runtime. Finally, Section 5 summarizes the conclusions, and discusses the limitations of this paper, while also outlining possible future work directions.

2. Literature Review

Mask R-CNN, as an improved version of Faster R-CNN, can perform not only object detection but also instance segmentation. Kasinathan & Uyyala

https://doi.org/10.24846/v34i3y202501 ICI Bucharest © Copyright 2012-2025. All rights reserved

(2023) constructed a fall armyworm detection system on the basis of Mask R-CNN, which utilizes this model for instance segmentation in the case of fall armyworms to ensure crop quality and safety. The Average Precision (AP) of the Mask R-CNN model was found to reach 94.21%. To effectively detect COVID-19, Sahin et al. (2023) used Faster and Mask R-CNN models to classify novel coronavirus pneumonia and pneumonia infection types. The Mask R-CNN built on the ResNet-50 and ResNet-101 residual backbone networks obtained a mAP of 97.72% and 95.65%, respectively. Abdrakhmanov et al. (2023) proposed a detection method based on Mask R-CNN. The scalable architecture was used for segmenting lane markings and generating precise boundaries, and the multi-scale feature extraction technology was introduced for improving the detection effect in near and far fields of view.

An increasing number of researchers are investigating the potential of in-process monitoring and defect detection. This paper mainly focuses on enhancing the precision and efficacy of the proposed model`s segmentation and detection capabilities. O`Donovan et al. (2023) put forth a novel ladle pouring tracking method. This method employed adaptive histogram equalization with limited contrast for image enhancement, utilized the Mask R-CNN model for segmentation prediction, and introduced the Kalman filter to improve the accuracy of ladle position tracking. Based on the evaluated tracking accuracy, the average accuracy of this model reached 0.61. Diaz & Tittus (2023) proposed a fast wind turbine defect detection model based on an improved Mask R-CNN. To reduce the computational costs, a depthwise separable convolution was performed instead of the standard convolution, and image enhancement and transfer learning techniques were combined to improve model performance. The experimental results showed that the proposed model obtained a classification accuracy of up to 97.8%. Djavadifar et al. (2022) focused on defect detection tasks that conform to material manufacturing processes and evaluated the performance of four deep CNN models, including Mask R-CNN, in detecting fiber cutting defects. The results showed that the accuracy of the model for gripper and fabric detection reached 0.92 and 0.86, respectively.

In summary, although the Mask R-CNN model has made many breakthroughs with regard to the detection accuracy and speed, there are still some problems related to RAG tasks and AUDDD in complex industrial environments. For example, the traditional unimodal visual detection models cannot fully utilize multi-source information, which can affect the accuracy and robustness of the detection process. Based on this, this study proposes a RAG and AUDDD method based on the bimodal Mask R-CNN model (the proposed method), aiming to integrate bimodal data and promote the extraction of multidimensional features.

3. Research Methodology

3.1 Construction of the Dual Modal Mask R-CNN Detection Model

In RAG and AUDDD tasks, Mask R-CNN can be used to accurately identify and segment complex object shapes, enabling the robotic arm to accurately locate and grasp targets, while capturing small unidirectional defect areas (Gheisari et al., 2023; Moccia, Fiorentino & Frontoni, 2021). Figure 1 displays the Mask R-CNN framework (He et al., 2017).

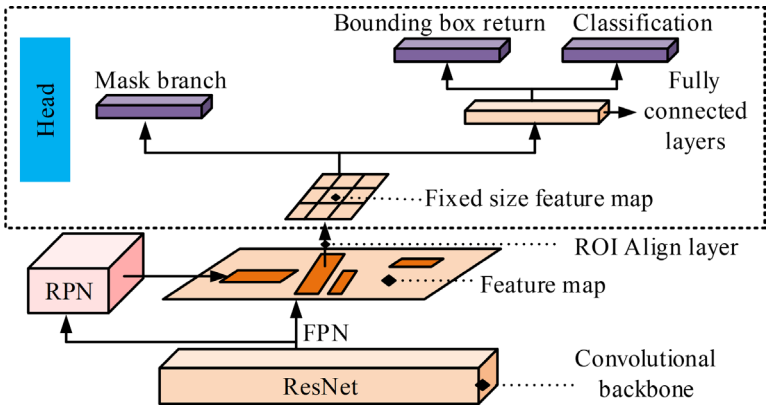


Figure 1. Structure of Mask R-CNN (adapted and redrawn based on the original architecture described in (He et al., 2017))

In Figure 1, Mask R-CNN first utilizes ResNet as the convolutional backbone network and extracts multi-level Feature Maps (Fmaps) from the input image. Secondly, the Feature Pyramid Network (FPN) fuses feature maps of different scales. Next, the Region Proposal Network (RPN) will extract the candidate regions from the Fmap generated by FPN. These candidate regions are precisely aligned through the Region of Interest Alignment (ROI Align) to avoid quantization errors in region cropping. Then, the aligned Fmaps are fed into the head section for object classification and bounding box regression, and precise localization and object detection tasks are completed. The function for Mask R-CNN is given by equation (1) (Goyal et al., 2023; Yang et al., 2022):

$$L = L_{cls} + L_{bbox} + L_{mask} \tag{1}$$

In equation (1), L_{cls}, L_{bbox}, and L_{mask} represent the loss for classification, bounding box regression, and mask. L is the total loss function of Mask R-CNN. To improve the segmentation effect, the model uses binary cross entropy loss to predict each pixel. By accurately calculating the mask loss of the target area, the model can better distinguish the shape and contour of the target area in complex scenes, as shown in equation (2):

$$L_{mask} = -\frac{1}{m^2}\sum_{i,j}\left[y_j \log\left(p_j\right) + \left(1-y_j\right)\log\left(1-p_j\right)\right] \tag{2}$$

In equation (2), m is the size of the predicted mask, while y_{ij} and p_{ij} are the true and predicted mask values of the pixel (i, j). Finally, feature fusion is performed at the FPN layer. The traditional Mask R-CNN relies on a single RGB image input and lacks the perception of object depth, position, and 3D structure. Therefore, the study introduces the bimodal Mask R-CNN, which fully utilizes the surface visual features and 3D structure information for objects by combining RGB images and depth images. To further enhance the use of depth information, the study introduces the HHA map to encode the depth image, extract features such as the horizontal disparity, object height, and normal vector, and improve the model's ability to understand the 3D structure of the object (Ding et al., 2022). The formula for horizontal disparity is shown in equation (3):

$$d = \arctan\left(\frac{x - c_x}{f_x}\right) \tag{3}$$

In equation (3), d is the horizontal disparity and c_x and f_x represent the horizontal coordinates and focal length of the camera, respectively. The calculation of the object height is shown in equation (4):

$$h = Z\sin\left(\theta\right) + Y\cos\left(\theta\right) \tag{4}$$

In equation (4), h is the height of the object. Z and Y represent the depth values and vertical coordinates of the pixels in the camera coordinate system, respectively and θ is the angle at which the camera rotates around the horizontal axis. The expression of the normal vector angle is shown in equation (5):

$$a = \arccos\left(\frac{n_z}{\sqrt{n_x^2 + n_y^2 + n_z^2}}\right) \tag{5}$$

In equation (5), a is the angle between the normal vectors, and (n_x, n_y, n_z) is the normal vector of the plane where the pixel is located. The processed HHA image and RGB image are separately input into the model to obtain an optimized bimodal Mask R-CNN structural framework, as shown in Figure 2.

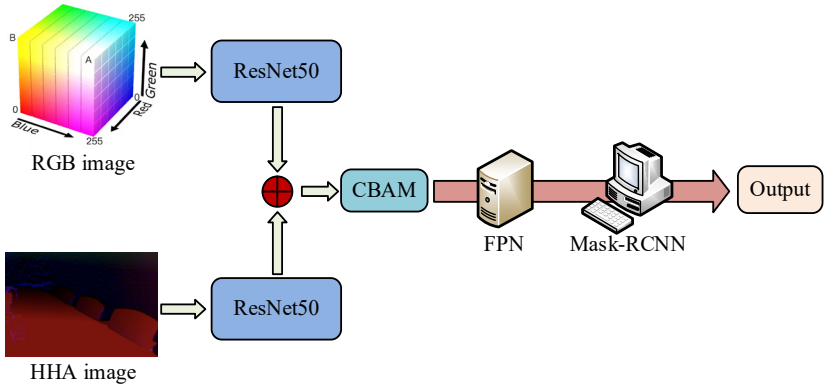


Figure 2. Bimodal Mask R-CNN structure diagram

ICI Bucharest © Copyright 2012-2025. All rights reserved

In Figure 2, the model uses ResNet50 as the backbone network and introduces two parallel branches to process the 2D information of RGB images and the 3D information of HHA images, respectively, and extract features at each layer from C2 to C5. Each feature layer also introduces the Convolutional Block Attention Module (CBAM) to enhance the expressive power of the features. Subsequently, the features of the two branches are fused at each level, and color, texture, and depth information are combined to further enhance the model's understanding of objects in complex scenes. The fused features are used to generate multi-scale Fmaps through the FPN. Finally, target classification, bounding box regression, and instance segmentation are performed on the Mask R-CNN head to complete the right detection and segmentation. Figure 3 shows the structure of CBAM (Woo et al., 2018).

The structure optimizes the Fmap through two steps. Firstly, the input Fmap is convolved to generate the Fmap F, and then the Channel Attention Module (CAM) extracts the important features F' in the channel dimension, enhancing the effectiveness of feature representation. Secondly, the Spatial Attention Module (SAM) enhances the spatial features of F' and generates the final optimized Fmap F''. The formula for CBAM is shown in equation (6) (Chen et al., 2023):

$$F' = M_c\left(F\right) \otimes F \tag{6}$$

In equation (6), F is the Fmap gained through the convolution operation for the previous layer. ⊗ is the symbol for multiplication between features. M_c represents CAM. F' is an important feature extracted by CAM. The expression for F'' is shown in equation (7) (Aydin et al., 2023):

$$F'' = M_s\left(F'\right) \otimes F' \tag{7}$$

In equation (7), M_s represents SAM. F'' is the attention weight value obtained by processing F' through SAM. The final attention weight value is weighted and summed with the original Fmap and then output as a new Fmap.

3.2 Robot Arm Grasping and AUDDD Framework Design Based on Bimodal Mask R-CNN

Based on the construction of the bimodal Mask R-CNN model in the previous subsection, in this study two application system frameworks were further implemented: the RAG system and the automotive unidirectional device defect detection system. The purpose is to achieve an accurate positioning, effective grasping, and efficient defect detection for the target. The first step was to build a RAG system, the overall architecture of which is shown in Figure 4 (Wang et al., 2022):

The RAG system in Figure 4 mainly consists of two parts: the 2D/3D Visual Processing Module

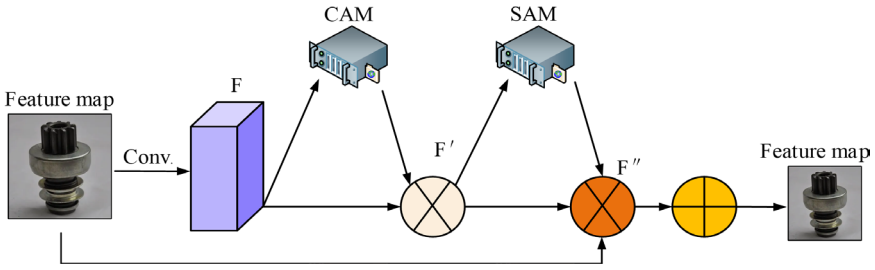


Figure 3. CBAM structure (adapted and redrawn based on the work of Woo et al. (2018))

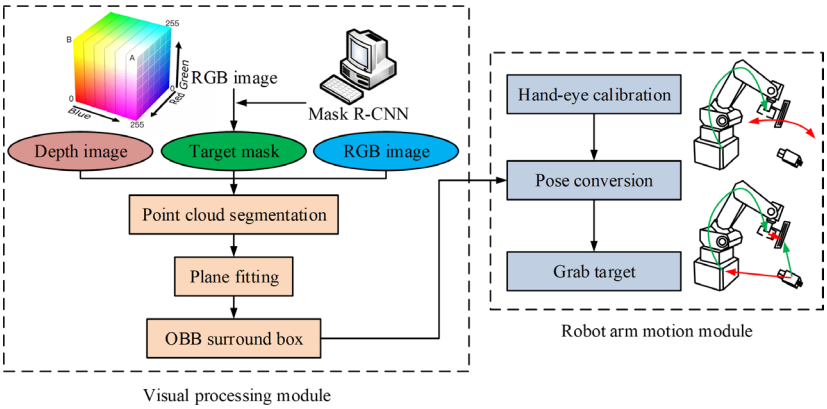


Figure 4. Framework structure diagram of the RAG system

(VPM) and the Robotic Arm Motion Module (RAMM). In the VPM, the system simultaneously obtains the RGB image and depth image of the target object through an RGB camera and a depth sensor. The RGB and depth images are input into a bimodal Mask R-CNN model for processing. The model extracts 2D and 3D features through parallel networks and generates precise segmentation masks for the target object (Hou et al., 2024). To achieve the effective fusion of bimodal features, a weighted summation method is employed for integrating the features of RGB images and depth images, as shown in equation (8):

$$F_{merge} = \alpha \cdot F_{RGB} + \beta \cdot F_{depth} \tag{8}$$

In equation (8), F_{merge} is the fused feature. F_{RGB} and F_{depth} are features for RGB images and depth images, respectively and α and β are weighting coefficients. Furthermore, after generating the target mask, the system performs point cloud segmentation on the depth image to extract the 3D point cloud data of the target. In the process of generating masks, the model predicts each pixel point through the sigmoid function to calculate the probability value of the target area, as shown in equation (9) (Zhou et al., 2023):

$$p_{i,j} = \frac{1}{1+e^{-(\omega \cdot x_{ij}+b)}} \tag{9}$$

In equation (9), $P_{i,j}$ is the predicted possibility of the pixel (i,j). ω and b represent the weights and biases and x_{ij} is the eigenvector. Subsequently, the plane fitting algorithm is employed to accurately calculate the plane information for the object, ensuring its accurate positioning in space. Next, the Orientation Bounding Box (OBB) is used for describing the boundaries of an object in space. In contrast to the conventional axis-aligned bounding box, the OBB is capable of adapting its position in accordance with the rotation angle of the object in question (Jotić et al., 2023). This makes it an optimal choice for describing objects with irregular shapes or rotation angles. The 6D pose matrix is shown in equation (10):

$$T = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \tag{10}$$

In equation (10), matrix T represents the pose of the object. R is the rotation matrix and t is the translation vector, representing the position of the object with respect to the robot base coordinate system. After obtaining accurate 6D pose information based on the 2D/3D VPM, the RAMM will perform target grasping tasks based on the calculated 6D pose information. Firstly, the system converts the position information in the camera coordinate system into the grasping coordinates required for the robotic arm operation through the hand-eye calibration step. After completing the pose conversion, the robotic arm performs precise grasping operations based on the planned path. At the same time, the system monitors the grasping process in real time, guaranteeing that the relative position between the robotic arm end-effector and the target object remains accurate and the grasping action is completed smoothly (Boyle et al., 2023). After presenting the RAG system framework, this study further presented the AUDDD system framework, as shown in Figure 5.

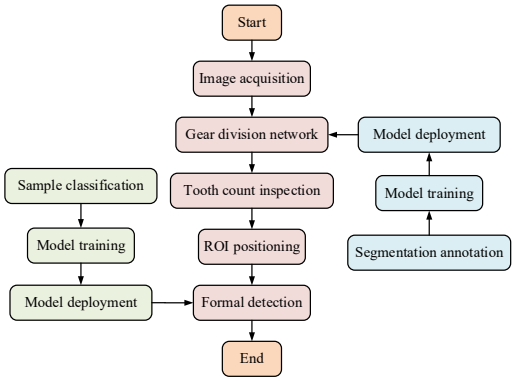


Figure 5. Framework structure diagram of the AUDDD system

The AUDDD system in Figure 5 consists of image acquisition, image processing, and classification modules. In the image acquisition module, the detection system obtains images of the car's unidirectional sensor through a high-definition camera. The collected images contain information on the appearance of key components such as gears. In the image processing module, the detection system uses the gear segmentation network to segment the one-way device image to extract the salient features of the gear area. Next, the system performs tooth count inspection to verify the structural integrity of the gear so as to ensure the accuracy of the subsequent defect detection. To optimize the alignment of ROI regions, the model uses binary cross entropy loss to improve the segmentation accuracy, as given by equation (11) (Tilki & Ölgün, 2023):

$$L_{ROI} = \sum_{i,j} (y_{ij} \log(p_{ij}) + (1+y_{ij}) \log(1-p_{ij})) \tag{11}$$

ICI Bucharest © Copyright 2012-2025. All rights reserved

In equation (11), L_{ROI} is the alignment loss, y_{ij} denotes the right label, and P_{ij} denotes the prediction possibility. After completing ROI localization, the detection system conducted a more in-depth detection for these ROIs. The specific inspection included details related to copper sleeves, chamfers, and semi-circular pieces.

After the detection was completed, the results were displayed on the system interface and entered the classification module. The classification module classified the images related to the unidirectional sensor into qualified and unqualified samples based on the detection results. For qualified samples, the system recorded and output a testing report. With regard to the unqualified samples, the system marked them and they entered the model training process. The trained model was redeployed to the detection process to ensure the continuous improvement and optimization of the system in actual production.

4. Results and Discussion

4.1 Performance Testing for the Bimodal Mask R-CNN Algorithm

To test the effectiveness of the proposed algorithm, this study implemented a self-made dataset for RAG and AUDDD, including RGB images and depth images from industrial environments. RGB images were captured by high-definition cameras, while depth images were obtained by depth sensors. The dataset contains 5,000 images, covering various robotic arm operation scenarios and different types of unidirectional device defects, and is segmented into training and testing sets in a 8:2 ratio.

In order to evaluate the performance of the bimodal Mask R-CNN model under different hyperparameter settings and find the optimal training configuration, this study employed the grid search method to perform hyperparameter combination tuning and a sensitivity analysis of the key hyperparameters. The parameter adjustment range includes the learning rate, batch size and dropout ratio. After performing a grid search over different combinations of key hyperparameters and testing the model performance on the validation set, the representative combination results were selected for display, and the final parameter configuration was determined accordingly. The results are shown in Table 1.

Table 1 shows the results for 9 typical hyperparameter combinations selected using the grid search strategy. The experiments carried out show that different parameter combinations have significant effects on model accuracy, the F1 Score value and training convergence speed. Combination 5 (a learning rate of 0.001, a Batch Size of 16, and a Dropout Ratio of 0.3) achieved the best accuracy and F1 value, namely 96.04% and 0.98 respectively, and the number of convergence rounds is relatively small, and the training stability is good. The performance of combinations 4 and 9 is relatively close, but there are slight overfitting or fluctuation problems. The combinations 3 and 8 show a certain underfitting or convergence instability. Therefore, the study finally uses combination 5 as the formal training parameters of the model, and the model's specific configuration is as follows: a learning rate of 0.001, a Batch Size of 16, a Dropout ratio of 0.3, the optimizer Adam, the Binary Cross-Entropy loss function, the number of training rounds was set at 300, and the Cosine Annealing strategy was used for learning rate decay. Further, the ablation experiment results are shown in Table 2.

Table 1. Model performance under different hyperparameter combinations (the Grid Search method)

No.	Learning Rate	Batch Size	Dropout ratio	Accuracy (%)	F1 Score	Epochs to Convergence
1	0.0005	8	0.1	93.12	0.91	268
2	0.0005	16	0.3	94.57	0.93	257
3	0.0005	32	0.3	92.48	0.89	281
4	0.001	8	0.3	95.63	0.96	254
5	0.001	16	0.3	96.04	0.98	243
6	0.001	32	0.1	93.87	0.92	266
7	0.005	8	0.3	94.20	0.94	273
8	0.005	16	0.1	92.33	0.90	279
9	0.005	32	0.3	94.27	0.94	288

As it can be seen in Table 2, Model 6 performs the best with regard to all the metrics involved in the ablation experiment, with accuracy values, recall values, and F1 values of 0.96, 0.97, and 0.98, respectively. The YOLOv5, Faster R-CNN, and Hybrid Task Cascade (HTC) algorithms were further selected for comparison purposes to evaluate the detection errors for two testing sets, as shown in Figure 6.

Figures 6(a) and 6(b) show the Mean Squared Error (MSE) and Mean Absolute Error (MAE) for YOLOv5, Faster R-CNN, HTC, and the proposed algorithm. As the sample size increases, the MSE and MAE of each algorithm gradually stabilize. Among them, the proposed algorithm obtained the lowest error for all sample sizes, with a stable MSE as low as 0.42 and a MAE as low as 0.26. Among the four algorithms, YOLOv5 obtained the highest error, with a MSE and MAE of 0.71 and 0.58, respectively. The MSE and MAE of Faster R-CNN were 0.60 and 0.54, respectively while HTC's MSE and MAE were 0.53 and 0.37, respectively. Figure 7 compares the Receiver Operating Characteristic (ROC) Curve for the four algorithms.

Table 2. Ablation results for six different models

Network structure	Accuracy	Recall	F1 value
VGG+FPN (Model 1)	0.85	0.84	0.83
ResNet50+FPN (Model 2)	0.87	0.85	0.87
ResNet50+ResNet50+FPN (Model 3)	0.88	0.89	0.89
VGG+FPN+CBAM (Model 4)	0.90	0.91	0.92
ResNet50+FPN+CBAM (Model 5)	0.92	0.93	0.93
ResNet50+ResNet50+FPN+CBAM (Model 6)	0.96	0.97	0.98

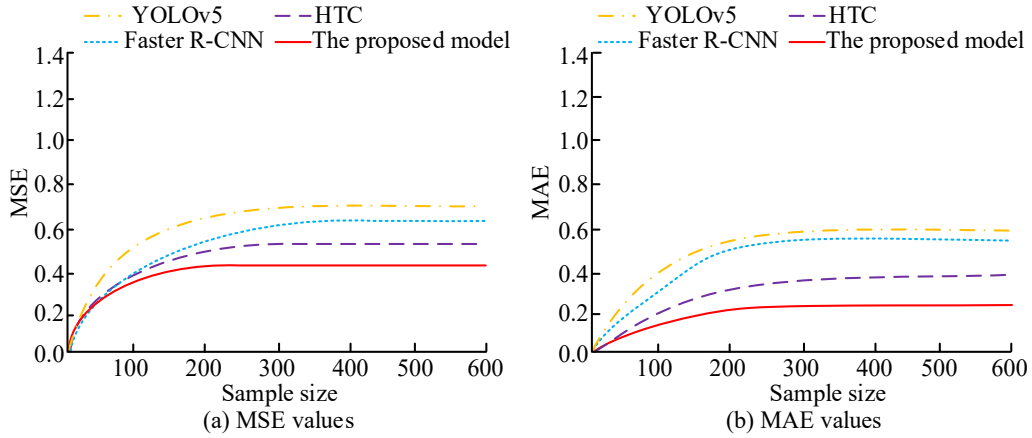


Figure 6. Error performance for the four algorithms

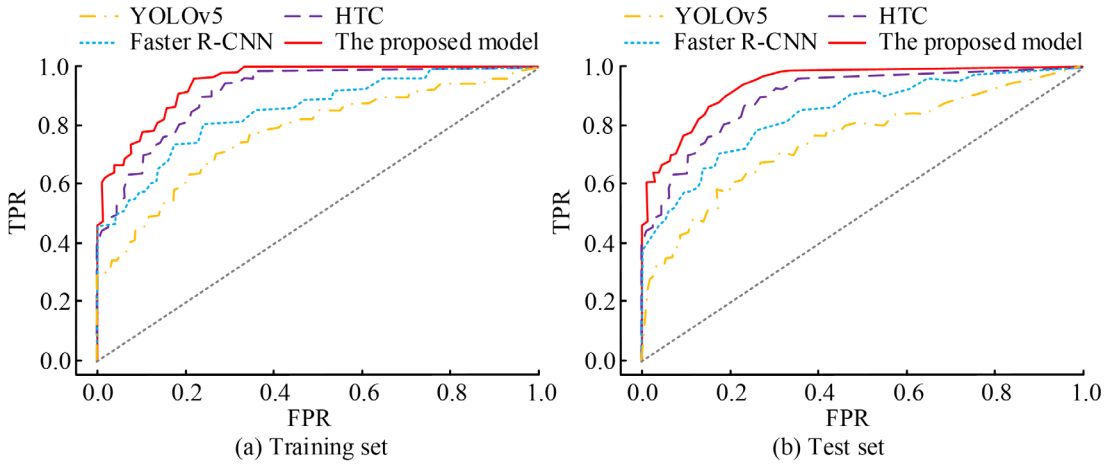


Figure 7. The ROC curves of the four algorithms for both datasets

ICI Bucharest © Copyright 2012-2025. All rights reserved

In Figures 7(a) and 7(b), the proposed algorithm exhibits optimal curves on both datasets, with curves close to the upper left corner, indicating a high True Positive Rate (TPR) and a low False Positive Rate (FPR) in classification tasks. In addition, HTC and Faster R-CNN perform well, and their ROC curves are also close to the ideal state, but slightly inferior to that of the proposed model.

4.2 Application Effect Analysis for the Dual Modal Mask R-CNN

In this study detection experiments were carried out for target objects in different robot arm grasping scenarios. Five common grasping scenarios are selected, namely Scenario 1 where objects are randomly stacked, Scenario 2 where objects are placed at different intervals, Scenario 3 where objects are occluded, Scenario 4 where objects with smooth reflective surfaces are grasped, and Scenario 5 where objects with complex shapes are grasped. Table 3 includes the comparative values for the recognition accuracy rate and recognition time for the four employed models in the five scenarios.

In Scenario 1, the recognition accuracy is as high as 97.52% for the proposed model, significantly higher than the value achieved by HTC, that is 95.48%, Faster R-CNN, namely 92.01%, and YOLOv5, namely 89.14%. In addition, the recognition time for the proposed model in Scenario 1 is only 0.05 seconds. Figure 8 compares the detection accuracy and detection time for the four algorithms in defect detection tasks.

As shown in Figures 8(a) and 8(b), the proposed algorithm maintains a high detection accuracy for all sample sizes, with a maximum of 98.25%, and the detection time is only 0.24 seconds. The HTC and Faster R-CNN algorithms feature a lower detection accuracy, reaching 93.48% and 90.59%, respectively with detection times as low as 1.02s and 1.78s, respectively. The detection accuracy of YOLOv5 is the lowest among all the employed models, that is only 85.14%, and the detection time is the highest at 4.08 seconds. HTC and Faster R-CNN only rely on single-modal RGB images and do not use depth information. Therefore, their recognition performance is slightly worse when dealing with complex surface structures. However, the feature extraction module of YOLOv5 cannot fully capture subtle defects, especially in multi-scale defect detection. Finally, the performance of the four models in actual detection cases is shown in Figure 9.

Table 3. Recognition accuracy and recognition time for the four employed models

Scene number	Network structure	Recognition accuracy rate /%	Recognition time /%
Scenario 1	YOLOv5	89.14	0.45
	Faster R-CNN	92.01	0.24
	HTC	95.48	0.17
	The proposed model	97.52	0.05
Scenario 2	YOLOv5	88.23	0.73
	Faster R-CNN	90.18	0.52
	HTC	93.24	0.29
	The proposed model	96.36	0.09
Scenario 3	YOLOv5	88.95	0.68
	Faster R-CNN	90.71	0.46
	HTC	91.38	0.18
	The proposed model	97.04	0.07
Scenario 4	YOLOv5	90.05	0.41
	Faster R-CNN	92.34	0.27
	HTC	95.98	0.19
	The proposed model	98.84	0.01
Scenario 5	YOLOv5	89.93	0.55
	Faster R-CNN	91.85	0.30
	HTC	96.04	0.22
	The proposed model	98.12	0.02

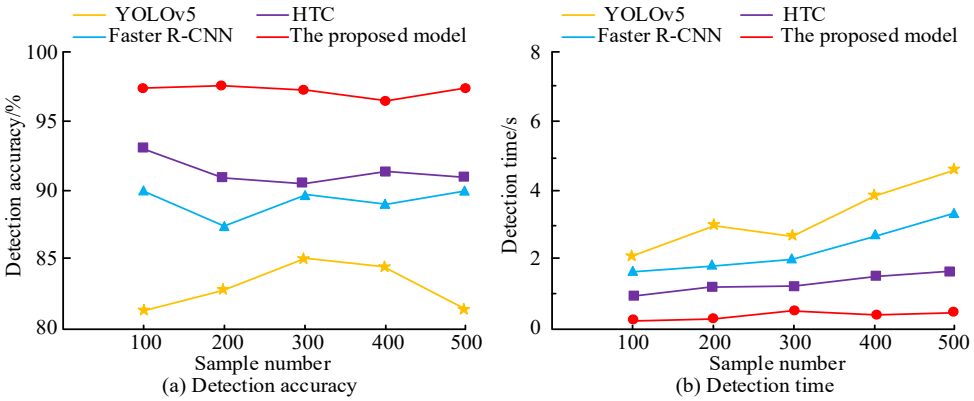


Figure 8. Detection accuracy and detection time for the four employed algorithms

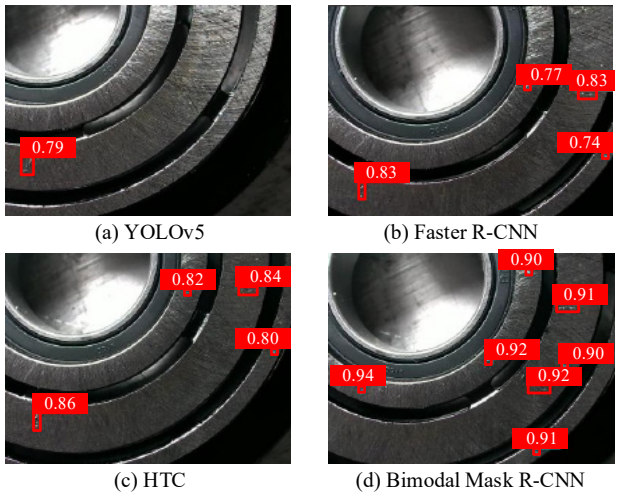


Figure 9. Defect detection images for the four models (original figure created by the authors)

As shown in Figure 9, the proposed model can detect more locations for defects related to the car's unidirectional sensor, while YOLOv5 can only detect more obvious defects. Overall, the proposed model performs the best with regard to detection time, accuracy, and actual detection tasks. This advantage primarily stems from the depth information augmentation provided by bimodal data fusion, which enables this model to more accurately identify the target area when confronted with intricate surface textures and nuanced structural alterations. In addition, CBAM adaptively adjusts important features at different feature layers, further improving the model's sensitivity to key defect locations.

5. Conclusion

To enhance the precision of target recognition and AUDDD in RAG operations, this study proposed a bimodal Mask R-CNN detection model that integrates RGB images and depth images. In the experiment carried out, the MAE and MSE of the proposed model were as low as 0.26 and 0.42, respectively and its ROC curve area was much larger than those of the YOLOv5, Faster R-CNN, and HTC algorithms. This indicated that this model could accurately identify target objects and had good classification ability. For the RAG task, the recognition accuracy of the proposed model reached as high as 98.84%, and the recognition time was as low as 0.01s. For the defect detection task, the detection accuracy of the research model was as high as 98.25%, and its detection time was only 0.24s.

Although the proposed dual-modal Mask R-CNN model performs well in robot arm grasping and automobile one-way device defect detection tasks, it still has certain limitations and room for improvement. For example, when image data is affected by noise or occlusion, it may affect the detection results. Future research could consider introducing more modal data, such as temperature

ICI Bucharest © Copyright 2012-2025. All rights reserved

images or infrared images, to further enhance the applicability of the model in harsh environments and expand its application range. Also, the current experimental verification is only focused on industrial manufacturing scenarios, and the adaptability of the model in other non-industrial fields (such as medical image recognition, agricultural plant disease detection, logistics object recognition, and security monitoring etc.) has not been empirically verified. Although the model integrated RGB and depth image information and introduced the CBAM to enhance the multi-scale feature extraction capability, and although it has a good target detection capability and defect localization potential, it may still face challenges such as image modality differences, the feature distribution transfer, and differences between the annotation methods in cross-domain applications. Future research could optimize the model structure and input modality in a targeted manner based on specific application requirements, and conduct migration experiments in one or more non-industrial fields to systematically evaluate its robustness and generalization capabilities in complex scenarios, thereby expanding its application scope to smart healthcare, smart agriculture, and multi-scenario automatic detection.

REFERENCES

Abdrakhmanov, R., Elemesova, M., Zhussipbek, B. et al. (2023) Mask R-CNN Approach to Real-Time Lane Detection for Autonomous Vehicles. *International Journal of Advanced Computer Science and Applications*. 14(5), 549-556. https://doi.org/10.14569/IJACSA.2023.0140558.

Abdullah, A. B., Jawahar, M., Manogaran, N. et al. (2024) Leather Image Quality Classification and Defect Detection System using Mask Region-based Convolution Neural Network Mode. *International Journal of Advanced Computer Science and Applications*. 15(4), 526-536. https://doi.org/10.14569/IJACSA.2024.0150455.

Aydin, I., Sevi, M., Akin, E. et al. (2023) A Deep Learning-Based Hybrid Approach to Detect Fastener Defects in Real-Time. *Tehnički vjesnik* [*Technical gazette*]. 30(5), 1461-1468. https://doi.org/10.17559/TV-20221020152721.

Benmiloud, I., Kouzi, K. & Ameur, A. (2023) Sensorless vector control of Dual Star induction Motor Associated with Fuzzy Mutual - MRAS estimator in electric vehicle. *Revista Română de Informatică și Automatică* [*Romanian Journal of Information Technology and Automatic Control*]. 33(2), 135–146. https://doi.org/10.33436/v33i2y202310.

Boyle, M. J., Goldman, Y. E. & Composto, R J. (2023) Enhancing Nanoparticle Detection in Interferometric Scattering (iSCAT) Microscopy Using A Mask R-CNN. *The Journal of Physical Chemistry B*. 127(16), 3737-3745. https://doi.org/10.1021/acs.jpcb.3c00097.

Chen, S., Liu, M., Hou, X. et al. (2023) Wafer Map Defect Pattern Detection Method Based on Improved Attention Mechanism. *Expert Systems with Applications*. 230(C), art. no. 120544. https://doi.org/10.1016/j.eswa.2023.120544.

Diaz, P. M. & Tittus, P. (2023) Fast detection of wind turbine blade damage using Cascade Mask R-DSCNN-aided drone inspection analysis. *Signal, Image and Video Processing*. 17(5), 2333-2341. https://doi.org/10.1007/s11760-022-02450-6.

Ding, M., Wu, B., Xu, J. et al. (2022) Visual Inspection of Aircraft Skin: Automated Pixel-level Defect Detection by Instance Segmentation. *Chinese Journal of Aeronautics.* 35(10), 254–264. https://doi.org/10.1016/j.cja.2022.05.002.

Djavadifar, A., Graham-Knight, J. B., Körber, M. et al. (2022) Automated visual detection of geometrical defects in composite manufacturing processes using deep convolutional neural networks. *Journal of Intelligent Manufacturing*. 33(8), 2257-2275. https://doi.org/10.1007/s10845-021-01776-1.

Fişne, A., Kalay, A. & Eken S. (2024) Fast and efficient computing for deep learning-based defect detection models in lightweight devices. *Journal of Intelligent Manufacturing*. 35(9), 1895-1910. https://doi.org/10.1007/s10845-024-02487-z.

Gheisari, M., Hamidpour, H., Liu, Y. et al. (2022) Data Mining Techniques for Web Mining: A Survey. *Artificial Intelligence and Applications*. 1(1), 3-10. https://doi.org/10.47852/bonviewAIA2202290.

Goyal, K., Kumar, P. & Verma, K. (2023) AI-based fruit identification and quality detection system. *Multimedia Tools and Applications*. 82(16), 24573-24604. https://doi.org/10.1007/s11042-022-14188-x.

Han, S., Choi, S. K. & Choi, H. J. (2024) A Mask R-CNN based process monitoring system for fabricating high density ceramic parts using photo-polymerization. *Journal of Mechanical Science and Technology*. 38(9), 4571-4577. https://doi.org/10.1007/s12206-024-2411-z.

He, K., Gkioxari, G., Dollár, P. et al. (2017) Mask R-CNN. In: *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, *22-29 October 2017, Venice, Italy*. New York, USA, IEEE. pp. 2961-2969.

Hou, M., Huo, D., Yang, Y. et al. (2024) Using Mask R-CNN to rapidly detect the gold foil shedding of stone cultural heritage in images. *Heritage Science.* 12, art. no. 46. https://doi.org/10.1186/s40494-024-01158-9.

Jotić, G., Štrbac, B., Toth, T. et al. (2023) The Analysis of Metrological Characteristics of Different Coordinate Measuring Systems. *Tehnički vjesnik* [*Technical gazette*]. 30(1), 32-38. https://doi.org/10.17559/TV-20220204091212.

Kasinathan, T & Uyyala S R. (2023) Detection of fall armyworm (*spodoptera frugiperda*) in field crops based on mask R-CNN. *Signal, Image and Video Processing*. 17(6), 2689-2695. https://doi.org/10.1007/s11760-023-02485-3.

Macías-Macías, M., Sánchez-Santamaria, H., Garcia Orellana, C. J. et al. (2023) A. Mask R-CNN for quality control of table olives. *Multimedia Tools and Applications*. 82(14), 21657-21671. https://doi.org/10.1007/s11042-023-14668-8.

Mani, V., Yarlagadda, S. R., Ravipati, S. et al. (2023) ANN optimized Hybrid Energy Management Control System for electric vehicles. *Studies in Informatics and Control*. 32(1), 101–110. https://doi.org/10.24846/v32i1y202310.

Moccia, S., Fiorentino, M. C. & Frontoni, E. (2021) Mask-R^2 CNN: a distance-field regression version of Mask-RCNN for fetal-head delineation in ultrasound images. *International Journal of Computer Assisted Radiology and Surgery*. 16(10), 1711-1718. https://doi.org/10.1007/s11548-021-02430-0.

O`Donovan, C., Popov, I., Todeschini, G. et al. (2023) Ladle pouring process parameter and quality estimation using Mask R-CNN and contrast-limited adaptive histogram equalisation. *The International Journal of Advanced Manufacturing Technology*. 126(3), 1397-1416. https://doi.org/10.1007/s00170-023-11151-4.

Sahin, M. E, Ulutas, H., Yuce, E. et al. (2023) Detection and classification of COVID-19 by using faster R-CNN and mask R-CNN on CT images. *Neural Computing and Applications*. 35(18), 13597-13611. https://doi.org/10.1007/s00521-023-08450-y.

Tilki, U & Ölgün, M. (2023) Terminal and Backstepping Sliding Mode Control with Genetic Algorithms for Robot Manipulators. *Studies in Informatics and Control*. 32(2), 117–126. https://doi.org/10.24846/v32i2y202311.

Wang, S., Zhou, Z. & Kan, Z. (2022) When Transformer Meets Robotic Grasping: Exploits Context for Efficient Grasp Detection. *IEEE Robotics and Automation Letters.* 7(3), 8170–8177. https://doi.org/10.1109/LRA.2022.3187261.

Woo, S., Park, J. Lee, J. Y. et al. (2018) CBAM: Convolutional Block Attention Module. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, *8–14 September 2018, Munich, Germany*. Berlin, Heidelberg, Germany, Springer-Verlag. pp. 3-19.

Yang, L., Xu, S., Fan, J. et al. (2022) A pixel-level deep segmentation network for automatic defect detection. *Expert Systems with Applications.* 215, art. no. 119388. https://doi.org/10.1016/j.eswa.2022.119388.

Yasir, S. M. & Ahn, H. (2023) Faster Metallic Surface Defect Detection Using Deep Learning with Channel Shuffling. *Computers, Materials and Continua*. 75(1), 1847-1861. https://doi.org/10.32604/cmc.2023.035698.

Zhou, Q. H., Qiu, Y. H., Liu, H. S. et al. (2023) Performance Study on Structural Parameters of Centre-Axle-Trailer Combinations. *International Journal of Simulation Modelling*. 22(1), 168-179. https://doi.org/10.2507/IJSIMM22-1-CO5.

ICI Bucharest © Copyright 2012-2025. All rights reserved



This is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License.