INTERNATIONAL JOURNAL OF SOFTWARE ENGINEERING AND COMPUTER SYSTEMS
E-ISSN: 2180-0650
VOLUME 10, ISSUE 1, 2024, pp. 20 – 31
DOI: https://doi.org/10.15282/ijsecs.10.1.2024.2.0120

RESEARCH ARTICLE

Partial occlusion object detection based on improved Mask-RCNN

Liu QingChuan[1,3]*, Muhammad Azmi Ayub[1], Fazlina Ahmat Ruslan[1], Mohd Nor Azmi Ab Patar[1], Shuzlina Abdul-Rahman[2]

[1]College of Engineering, UiTM Shah Alam, 40450 Shah Alam Selangor, Malaysia
[2]School of Computing Sciences, College of Computing, Informatics and Media, UiTM Shah Alam, 40450 Shah Alam Selangor, Malaysia
[3]HeBei Institute of Mechanical and Electrical Technology, XingTai, China

ABSTRACT - In the grasping task of industrial robots, multi-target objects are often placed in disorder or even partially occlusion or stacked, which brings certain difficulties to visual detection such as accuracy and real-time. The traditional Mask-RCNN algorithm can achieve high detection accuracy in scene which the target objects are neatly placed, but in the complex scenarios such as disorderly placed or partially occlusion is still have space for improvement in accuracy and speed. Mask-RCNN introduces the mask head structure to achieve pixel level segmentation mask prediction, it achieves high detection accuracy but increases the amount of calculation, this cause the detection speed is limited. To deal the above problems, this paper proposes an indirect frame subtraction for loss function to improved Mask-RCNN, which uses adjacent frames as comparison templates to find differences for image, that is, after one recognition, the previous recognition result is used as the background, and the next change is used as the target. thereby improving the recognition accuracy, reducing the repetitive estimation of regions, and improving the detection accuracy and running speed. Through experiments on self-made datasets, it is proved that new method can improve the image recognition accuracy about 2.3%, another, image recognition time has been reduced by 9% and the FPS value is improved by 6 frames which indicate the speed was improved. The research has important reference significance for the realization of robot flexible grasping task in intelligent manufacturing environment.

ARTICLE HISTORY
Received : 19 March 2024
Revised : 11 August 2024
Accepted : 25 August 2024
Published : 2 September 2024

KEYWORDS
Vision detection
Mask-RCNN
Recognition
Accuracy
Speed

1.0 INTRODUCTION

With the continuous progress of science and technology, visual inspection has ushered in a new wave of development worldwide, occupying an increasingly large market size and playing an important role in many industries [1]. Especially in repetitive motion scenes such as grasping classification and assembly in the industrial field, visual inspection has been widely applied. However, when performing visual detection tasks, it is common to encounter disorderly placement of multiple target objects, even partial occlusion or stacked of the target object bring certain difficulties to the accuracy and real-time performance of visual detection.

The combination of visual detection and deep learning has become a hot research topic both domestically and internationally. Compared with traditional fixed-point, manual guidance, or simple visual recognition localization detection methods [2], the combination of visual detection and deep learning has higher accuracy, flexibility, and application value. Mask-RCNN is the latest research achievement in this field, which mainly completes three tasks: object detection, object classification, and pixel level object segmentation. This deep learning network is inherited from CNN, RCNN, Fast-RCNN, Faster-RCNN [3].

CNN was proposed by Hubel in the 1960s, which shortens the training time of the network by compressing high-dimensional features, making it more practical [4]. Convolution neural networks start from the most basic edge features and extract the most prominent features in each network layer to obtain translation, rotation, or other forms of features, thereby obtaining high in-variance of deformations such as translation, rotation, and scaling [5]. Girshick [6] proposed the Region based Convolution Neural Network (RCNN) detection algorithm by applying CNN models in the area of target detection. It can be said that it is the first successful model algorithm to apply deep learning to the area of object detection. In traditional object detection methods, the idea is mostly based on image recognition, using exhaustive methods to select all region boxes containing objects on the image, and extracting and classifying features on these region boxes. Finally, the results are output through Non-maximum suppression (NMS) method. However, due to the fact that this method takes up a large amount of disk space when extracting features corresponding to all candidate regions and storing them, as well as the image deformation caused by cropping and scaling operations during the normalization process, it has a series of problems such as low efficiency and longtime consumption, and RCNN has not been widely applied. Fast-RCNN obtains candidate regions through selective search algorithms and then uses ROI projection operations to proportionally reduce the candidate regions in the original image and map them to the regions corresponding to convolution features [7][8]. Then, it extracts features from these regions on the convolution layer through the ROI

***CORRESPONDING AUTHOR | L. QingChuan** | ✉ 411533094@qq.com

pooling layer (ROI pooling 1ayer). Compared with RCNN, Fast-RCNN could improve training and inference speed, and no longer occupies a large amount of disk space, reducing the required space for training. However, this algorithm uses selective search algorithms to extract candidate boxes, resulting in poor real-time performance. Therefore, Faster-RCNN algorithm has emerged by introducing candidate region generation. The advantage of using Region Proposal Network (RPN) to extract candidate region boxes is that it can eliminate the separation problem caused by selective search [9]. Faster-RCNN simplifies the object detection process and improves computational efficiency, However, the algorithm still uses the ROIPooling network to convert feature maps of different sizes into the same size. This operation rounds the feature maps twice, causing pixels in the feature map to deviate from those in the real image, reducing the accuracy of object recognition.

Mask-RCNN has been proposed by replacing ROI Pooling with ROI Align, avoiding the rounding operation of ROI Pooling in feature scaling, reducing the loss of spatial symmetry, and thus maintaining the refinement of feature pixels [10]. In addition, Mask-RCNN introduces the Mask Head structure to enhance the prediction performance of the model, achieving pixel level segmentation mask prediction. The mask branch predicts the $_{m\times m}$ binary mask output of *K* types(where $_{m\times m}$ represents the size of the mask and *K* represents the number of categories), applies a sigmoid function to each pixel point, and the overall loss is defined as the average binary cross loss entropy, which improves recognition accuracy to a certain extent. It is precisely because the network achieves high detection accuracy through compare the detection results with the labeled image pixel by pixel to achieve pertinence learning of image features, but it increases the computational complexity of the image detection process, which to some extent affects the detection speed. On the other hand, this algorithm also has room for improvement in image recognition performance especially in the presence of partial occlusion or stacking of multiple target objects.

The existing algorithms may result in missing features extracted from partially occluded or stacked multiple target objects, reducing the accuracy of target object localization. At the same time, they are more sensitive to the threshold of NMS, which can easily lead to misjudgment of target categories. At present, the accuracy of the most advanced object detectors and deep neural network-based classifiers under partial occlusion will decrease. Many studies focus on improving network structure and improving the accuracy of existing methods for identifying and locating occluded objects. Qi et al. proposed a Multi-Layer Coding (MLC) network, which uses an occlusion classification branch to improve the modal perception ability to infer occlusion parts. DeVries and Taylor proposed a regularization technique called Cutout, which enhances training data with partially occluded images. For each input image, randomly select a pixel as the center point of a fixed-size zero mask to remove adjacent parts of the image. However, increasing training data by improving the network structure can lead to a slow detection speed.

To solve this problem, this article proposes an improved Mask-RCNN algorithm, and then uses 4 cuboids as detection targets and verifies the accuracy and speed of the improved algorithm through experiments when the 4 target objects are placed in disorder or even partially occluded.

This article mainly introduces the following content: Section 1 mainly introduces the development of deep learning algorithms. Section 2 introduces the basic architecture and principle of Mask-RCNN network which is widely used nowadays. In section 3 an optimization algorithm with an improved loss function is proposed based on Mask-RCNN algorithm. In section 4 do the experiment and analysis the experimental data. The final section provides a summary of the paper.

2.0 RELATED WORKS

Mask-RCNN is an instance segmentation algorithm in target detection [11], Figure 1 is the Mask-RCNN overall architecture. The target object image is pre-processed, and then classified, regressed, and segmented. The process is as follows: First, the pre-trained feature extraction network (RESNET+FPN) extracts the features to obtain the corresponding feature map; Second, Input the feature map to the area suggestion network (RPN) to generate multiple target candidate areas (ROIs) [12]. Then, input the feature map and generate multiple ROIs to the RoIAlign layer together, so that each target candidate region's ROIs are normalized to the same size. This procedure guarantees that the pixels in the feature map are aligned with the pixels in the primitive image. Finally, the relevant target features are collected from the feature map, and then output to the FClayers and FCN for target classification and instance segmentation. Through decoupling the relationships between multiple subtasks, the accuracy of target object detection in complex backgrounds (multi-objects placed disorderly or partial occlusion) can be improved [13]. This makes Mask-RCNN significantly superior to early template matching algorithms and other deep learning object detection algorithms in target recognition accuracy [14][15].

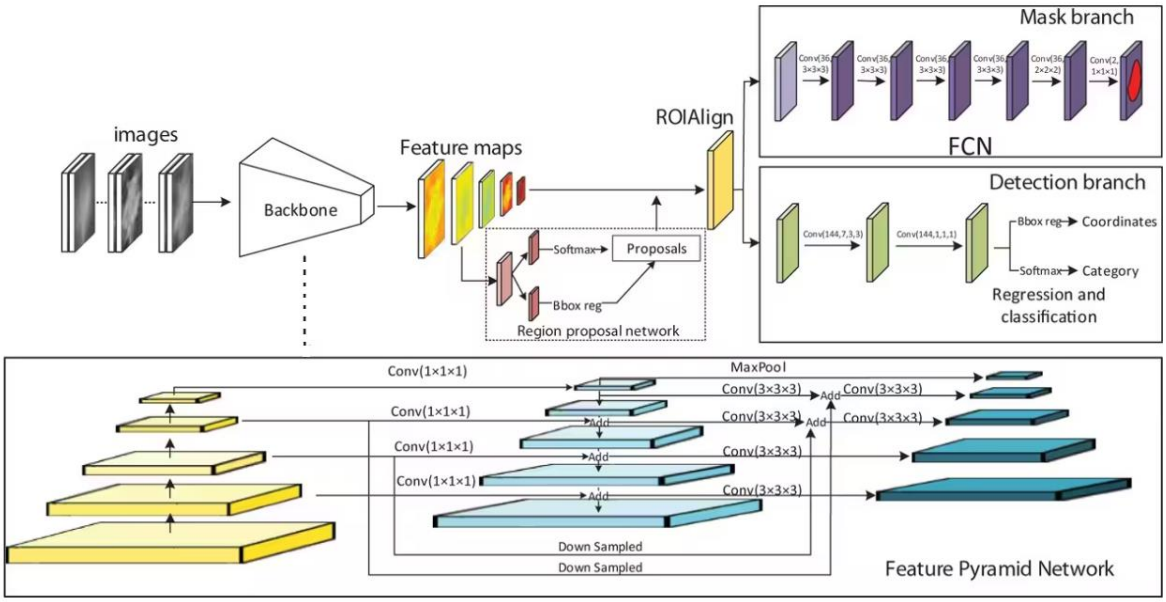


Figure 1. Mask- RCNN pipeline

The Mask-RCNN algorithm mainly improves the Faster-RCNN, replacing the region of interest pooling operation (ROIPooling) with the region of interest alignment operation (ROIAlign), in the ROIAlign operation, bi-linear interpolation was used to avoided the rounding operation of ROIPooling operation in feature scaling, reducing the loss of spatial symmetry, and thus maintaining the refinement of feature pixels.

We get the feature map for the ROIAlign of the Mask-RCNN by formula (1):

$$\mathrm{k} = \mathrm{k}_0 + \log_2(\sqrt{\mathrm{wh}}/224) \tag{1}$$

In the equation, k_0, w, and h are respectively the area, width and height of the feature map. Here 224 represents the size of the image used for pre-training. Different from Faster-RCNN , ROIAlign of Mask-RCNN quantize the region of interest use bi-linear interpolation [16]. As shown in Figure 2, if we know the points of A_{11}, A_{12}, A_{21}, and A_{22}, then Mask-RCNN obtains points of B_1 and B_2 by linear interpolation, and then interpolates the obtained points of B_1 and B_2 , and finally we can get the interpolation point P, that is:

$$f(P) \approx \frac{y_2 - y}{y_2 - y_1} \cdot f(B_1) + \frac{y - y_1}{y_2 - y_1} \cdot f(B_2) \tag{2}$$

$$f(B_1) \approx \frac{x_2 - x}{x_2 - x_1} \cdot f(A_{11}) + \frac{x - x_1}{x_2 - x_1} \cdot f(A_{21}) \tag{3}$$

$$f(B_2) \approx \frac{x_2 - x}{x_2 - x_1} \cdot f(A_{12}) + \frac{x - x_1}{x_2 - x_1} \cdot f(A_{22}) \tag{4}$$

$$\begin{aligned} f(x,y) = {} & \frac{f(A_{11})}{(x_2 - x_1)(y_2 - y_1)} \cdot (x_2 - x)(y_2 - y) + \frac{f(A_{21})}{(x_2 - x_1)(y_2 - y_1)} \cdot (x - x_1)(y_2 - y) \\ & + \frac{f(A_{12})}{(x_2 - x_1)(y_2 - y_1)} \cdot (x_2 - x)(y - y_1) + \frac{f(A_{22})}{(x_2 - x_1)(y_2 - y_1)} \cdot (x - x_1)(y - y_1) \end{aligned} \tag{5}$$

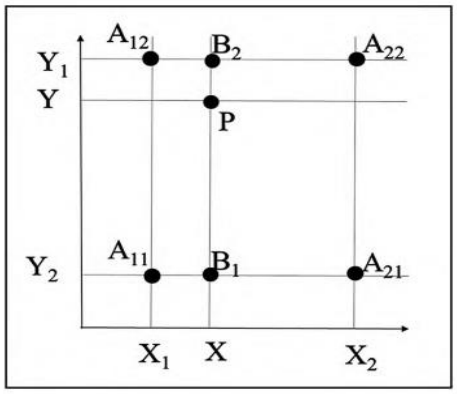


Figure 2. ROIAlign bi-linear interpolation

The entire ROIAlign process did not use quantization operations, so did not introduce errors, meaning that the pixels are completely aligned without deviation between the original image and feature map.

Mask-RCNN is a two-stage instance segmentation method. The first stage is feature extraction by backbone network, and the second stage is classification, box regression, and Mask prediction for each ROI by head network as shown in Figure 3 below.

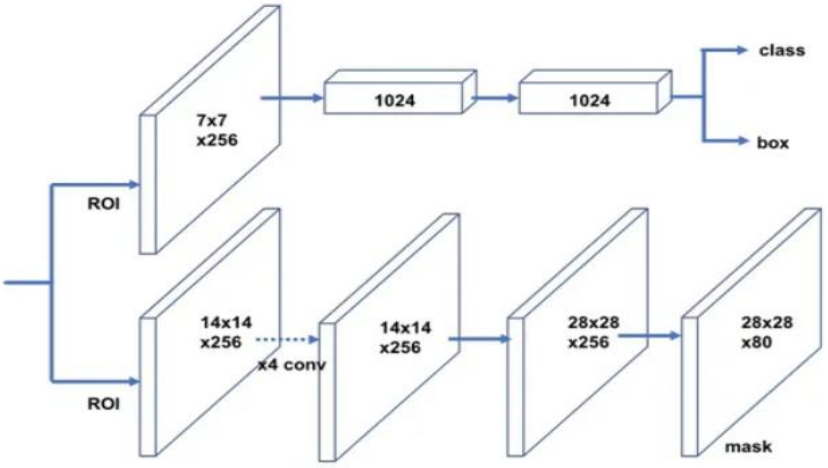


Figure 3. Head network for Mask-RCNN

Backbone uses pre-trained ResNet to extract features from input images to get ROI. Then, there are two branches. The up-branch function is classification and regression, while the bottom branch function generates corresponding masks. First, the ROI is changed to the feature of $14\times14\times256$, then the same operation is carried out for 5 times [17], and then the deconvolution operation is carried out to output the Mask of $28\times28\times80$, that is, a larger and finer Mask is output.

During the training process, in order to match the ROIAlign characteristic distribution of the target in the image, Define a multi-task loss function for each ROI during training. $L_{mask-rcnn}$ can be written as equation (6):

$$L_{mask-rcnn} = L_{cls} + L_{box} + L_{mask} \tag{6}$$

$$L_{mask} = -\frac{1}{m^2}\sum_{k=1}^{K}\sum_{i=1}^{m^2}\log\left[p_{ki}^{*}P_{ki} + (1-p_{ki}^{*})(1-P_{ki})\right] \tag{7}$$

Where, L_{cls}, L_{box} and L_{mask} are classification loss, regression box loss and segmentation loss respectively [18]. m is the image length and width of ROI processed by the dimension matching layer, k is the type number of the detection target, K is the total number of targets for model detection, p_{ki}^{*} is the value of the i^{th} pixel in the true Mask image of the k^{th} detection target. If this pixel belongs to the k-type detection target, its value is 1; otherwise, it is 0. P_{ki} is the value of the i^{th} pixel in the predictive Mask image of the k^{th} detection target [19][20]. If the model considers that this pixel belongs to the k-type detection target, its value is 1; otherwise, it is 0. According to equations (6) and (7), it can be seen that this function realizes targeted learning of image features pixel-by-pixel by comparing the detection results with the labeled

image, which is the root cause for detection accuracy of Mask-RCNN is superior to the general target detection algorithm [21].

3.0 METHODS AND MATERIAL

Although Mask-RCNN achieves high detection accuracy, but in scenarios where multiple objects are partially occluded or stacked may result in missing features extracted from target objects, reducing the accuracy of target object localization. At the same time, they are more sensitive to the threshold of NMS, which can easily lead to misjudgment of target categories, and pixel level segmentation mask prediction increases data computational, so the detection speed is affected to some extent. So, this paper gives a new idea, that is, propose an indirect frame subtraction (IFS) to improve it, after one recognition, the previous recognition result is used as the background, and the next change is used as the target. This can reduce computational complexity and improve detection speed. On the other hand, the original target detection is considered to be equivalent to increasing to four times, and it is progressive layer by layer, which can further improve image recognition accuracy. The method we take is divide a detection cycle into four times t_0, t_1, t_2 and t_3. The time t_0 is the target prediction Mask of Mask-RCNN, while the time t_1, t_2 and t_3 take the real Mask P'_{ki} ,P''_{ki} ,P'''_{ki} of the previous acquisition time as the background.

Now, the equation (7) is modified as follows:

$$\begin{cases} L_{mask} = -\dfrac{1}{m^2}\sum_{k=1}^{K}\sum_{i=1}^{m^2} log\,[p^*_{ki}P_{ki} + (1-p^*_{ki})(1-P_{ki})] & t_0 \\ L'_{mask} = -\dfrac{1}{m^2}\sum_{k=1}^{K}\sum_{i=1}^{m^2} log\,[p'_{ki}P''_{ki} + (1-p'_{ki})(1-P''_{ki})] & t_1 \\ L''_{mask} = -\dfrac{1}{m^2}\sum_{k=1}^{K}\sum_{i=1}^{m^2} log\,[p''_{ki}P'''_{ki} + (1-p''_{ki})(1-P'''_{ki})] & t_2 \\ L'''_{mask} = -\dfrac{1}{m^2}\sum_{k=1}^{K}\sum_{i=1}^{m^2} log\,[p'''_{ki}P''''_{ki} + (1-p'''_{ki})(1-P''''_{ki})] & t_3 \end{cases} \tag{8}$$

Where:t_0, t_1, t_2 and $t_3$are the four moments in a cycle. P'_{ki} ,P''_{ki} ,P'''_{ki} and P''''_{ki} are the real mask of the acquisition time t_0, t_1, t_2 and t_3. So, in one complete cycle, we can get the modified loss function L^*_{mask} as equation (9).

$$L^*_{mask} = L_{mask}+L'_{mask}+L''_{mask}++L'''_{mask} \tag{9}$$

After improving Mask- RCNN, the loss function is equation (10).

$$L_{mask-rcnn} = L_{cls} + L_{box} + L^*_{mask} \tag{10}$$

According to the idea of the improved algorithm and the above mathematical formula, we can get the flow chart of the IFS method as shown in Figure 4. When the masks input to the network, reshape for simplicity to merge first two dimensions into one, permute predicted masks to the form [N, num_classes, height, width], only positive ROIs and only the class specific mask of each ROI contribute to the loss. Gather the masks (predicted and true) that contribute to loss, calculate the loss for the first time, which is the normal mask loss function, loop through each calculation, and y_true is the target masks of the previous group, this function divides target masks into four groups based on the batch. Finally, add the losses to get the total mask loss.

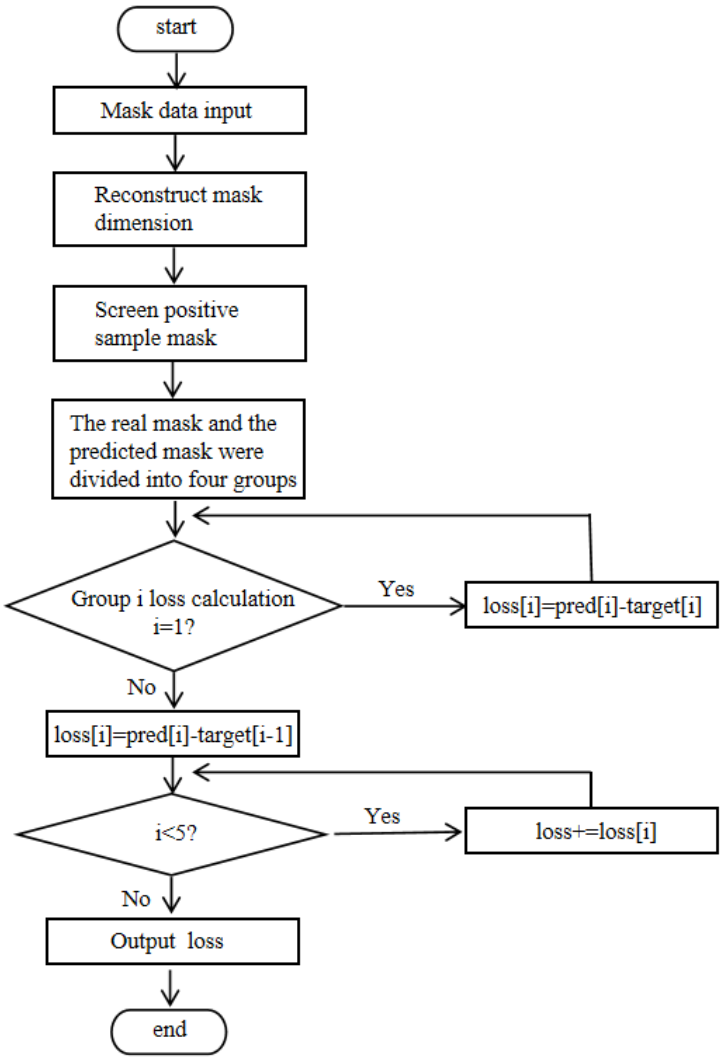


Figure 4. Flowchart of improved loss function

Due to the Figure 4, flowchart of improving the loss function, the corresponding program code is written. Figure 5 is part of the core code.

```
    # Calculate the loss for the first time, which is the normal Mask loss function.
    loss_0 = K.switch(tf.size(y_true) > 0,K.binary_crossentropy(target=y_true_splited[0],
output=y_pred_splited[0]),tf.constant(0.0))
    loss_0 = K.mean(loss_0)
    loss_sum += loss_0
    # Loop, each time you calculate, y_true is the target_masks of the previous group
    for i in range(1, b):
    loss_i = K.switch(tf.size(y_true) > 0,K.binary_crossentropy(target=y_true_splited[i-1],
output=y_pred_splited[i]),tf.constant(0.0))
    loss_i = K.mean(loss_i)
    loss_sum += loss_i
    loss_avg = loss_sum / b
    return loss_avg
def split_into_4 parts(data, b):
    # Split data into 4 parts
    if is instance(data, tf.Tensor):
    length_eachpart = b // 4
    data_splited = tf.split(data, b, axis=0)
    elif is instance(data, np.ndarray):
    data_splited = np.split(data, b, axis=0)
    else:
    raise Not Implemented Error
    assert len(data_splited) == b
    return data_splited
```

Figure 5. Part code of improved loss function

4.0 RESULTS AND DISCUSSION

This experiment takes 4 small cuboids as detection targets, the size of which is $20 \times 15 \times 10$ mm. The 4 target objects are placed out of order and partially occluded or stacked. The target objects are divided into two categories, labeled by "*TOP*" and "*BOTTOM*", where "*TOP*" refers to objects above occluded or stacked, and "*BOTTOM*" refers to objects that are occluded or regularly placed. A complete system is used to recognize real images to verify the feasibility of the segmentation technique in complex cases.

4.1 Experimental environment and network parameters

The network is trained on Intel Core i7-9750H processor, 64-bit Windows 10, and NVIDIA GeForce RTX 2060 14GB GPU operating system. Because of the difference in the data set, when training need to fine-tune network parameters. The network parameters are set in Table 1.

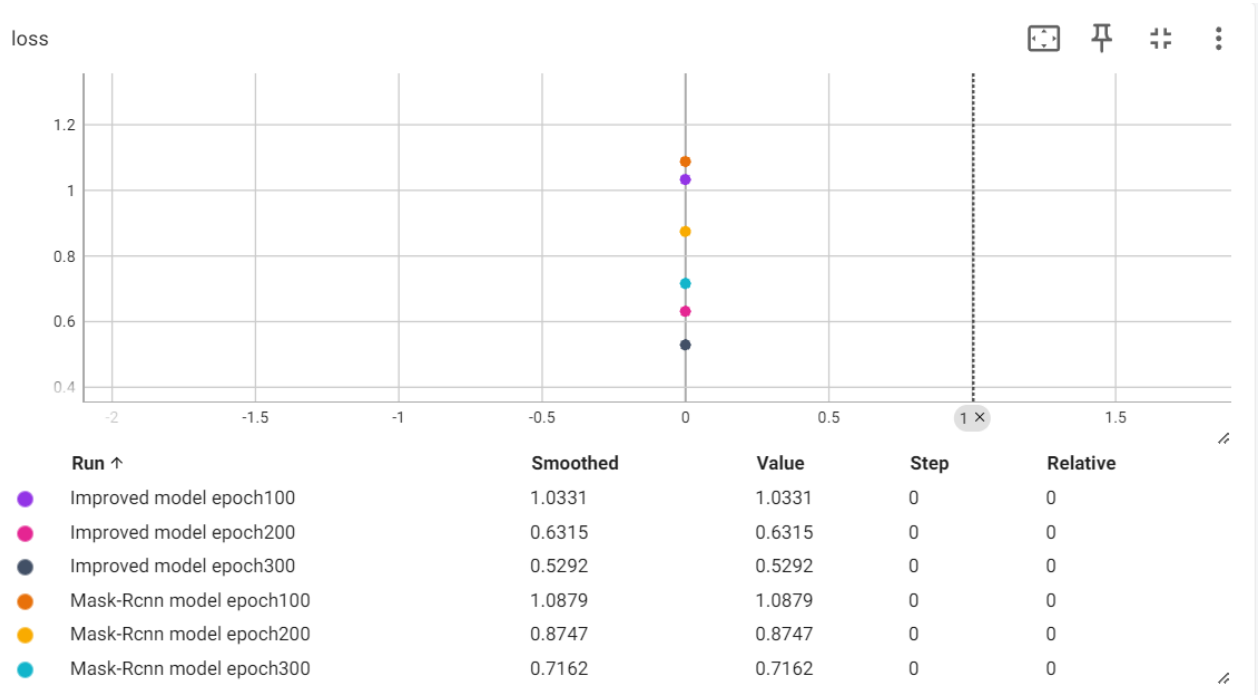
Table 1. Network parameters

Parameters	Value
NUM_CLASSES	3(1+2)
IMAGE_MAX_DIM	1024
IMAGE_MIN_DIM	320
RPN_ANCHOR_SCALES	(32,64,128,256,512)
RPN_ANCHOR_RATIOS	[0.5,1,2]
MAX_GT_INSTANCES	5
DETECTION_MIN_CONFIDENCE	0.8
WEIGHT_DECAY	0.0001
LEARNING_RATE	0.001
STEPS_PER_EPOCH	300

NUM_CLASSES is the total number of categories, in this case 3, including 2 categories and 1 background category; Set IMAGE_MAX_DIM as the maximum side length of the image, and IMAGE_MIN_DIM as the minimum side of the image. Through the experiment, it is best to set this parameter to a multiple of 64, this can guarantee the feature mapping is smoothly scaled up and down at the 6 levels of the FPN pyramid, mismatched settings may be affected by the minimum image scale and the loss of result information. Since the number of objects contained in each required recognition image is 4, in order to maintain the training speed, the maximum number of instances MAX_GT_INSTANCES is set to 5. The DETECTION_MIN_CONFIDENCE is set to 0.8. This parameter is the confidence threshold which is considered a valid key point only when the confidence level of the detected feature is higher than DETECTION_MIN_CONFIDENCE. If the set value is too high, it may cause some effective key points to be filtered out, thereby affecting the accuracy of the model. If the set value is too low, some inaccurate key points may also be added to the model, thereby reducing the accuracy of the model. During model training, the value can be adjusted to select the most suitable parameter. WEIGHT_DECAY is a regularization technique that suppresses overfitting in the model, thereby improving its generalization, it can adjust the numerical value to alter the model's performance, here we set to 0.0001, WEIGHT_DECAY is the learning rate and here we set to 0.001, STEPS_PER_EPOCH is the number of training steps per epoch we set to 300, but we can adjust the epoch to gain different model weight.

4.2 Data analysis

After setting the network parameters, the model was trained. The loss function values in Tensor board are shown in Figure 6. It can be seen when increase the network training epoch, the loss value of the network continuously decreases and tends to remain unchanged. Moreover, under the same number of epochs, the loss value of the improved algorithm is smaller than that of the original model, indicating that the improved algorithm model has achieved better operational performance.



Run ↑	Smoothed	Value	Step	Relative
Improved model epoch100	1.0331	1.0331	0	0
Improved model epoch200	0.6315	0.6315	0	0
Improved model epoch300	0.5292	0.5292	0	0
Mask-Rcnn model epoch100	1.0879	1.0879	0	0
Mask-Rcnn model epoch200	0.8747	0.8747	0	0
Mask-Rcnn model epoch300	0.7162	0.7162	0	0

Figure 6. The loss value of the two algorithms under different epoch

To verify the performance of the algorithm, 20 images are randomly selected for the experiment under the two detection models, Equation (11) is used to evaluate the precision of the model, and the detailed information is shown in Table 2. We can see the detection precision of Mask-RCNN is 85%, and our algorithm is 100%, the detection precision is higher than the original algorithm.

$$P = TP/(TP + FP) \tag{11}$$

Where: P is the precision of the algorithm, TP is the algorithm predicts the correct positive samples, FP is the algorithm predicts the false positive samples. Figure 7 shows part of the target object was not identified, Figure 8 shows that the categories of some target objects is misjudged, in both cases we call it FP, and Figure 9 shows the target object detection is successful, we call it TP.

Figure 7. Part of the target object was not identified

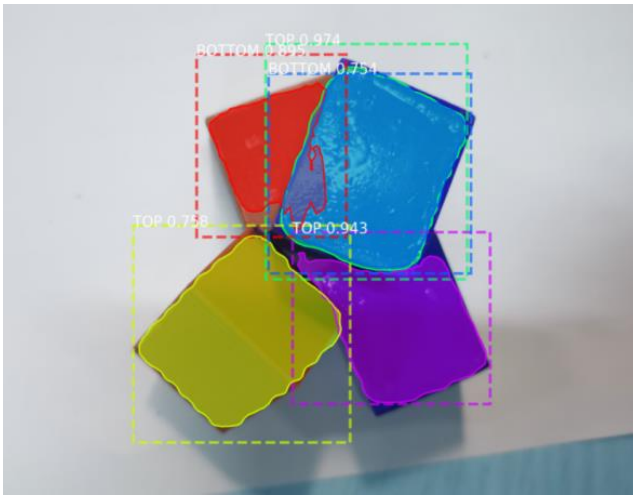


Figure 8. Categories Misjudgment

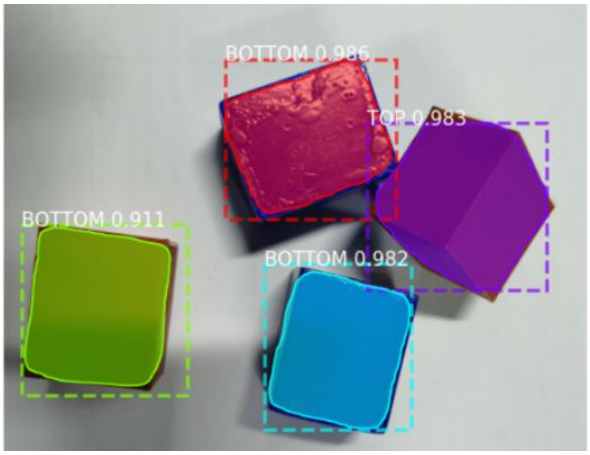


Figure 9. Identify success

Table 2. Image recognition precision

MODEL	Quantity of test	Not identified number(*FP*)	Categories Misjudgments(*FP*)	Success (*TP*)	precision
Mask-RCNN	20	1	2	17	85%
Improved Mask-RCNN	20	0	0	20	100%

In addition, to compare the pixel recognition accuracy of the two algorithms, 10 images are randomly selected, and the image recognition scores obtained under the two algorithm models are calculated by Equation (12). The comparison of target object recognition accuracy is shown in Table 3. We can see that the average pixel recognized score by the Mask-RCNN algorithm for the same 10 photos is 0.913, while our improved algorithm has an average recognition score of 0.936, which improves the recognition accuracy about 2.3% compared to the original algorithm. Figure 10 shows the two original pictures, Figure 11 and Figure 12 show the recognition results under the Mask-RCNN and our improved algorithm model respectively.

$$M = \sum_{k=1}^{10} \sum_{i=1}^{4} C_i / n \tag{12}$$

Where:
M: The average score of object recognition in the image
C: Recognition score for each object in the image
i: Number of target objects in each image
k: Number of images in the experiment
n: Number of target object in experimental image,40 in here.

Figure 10 (a). Scenario 1

Figure 10 (b). Scenario 2

Figure 10. Original pictures of scenario 1and scenario 2

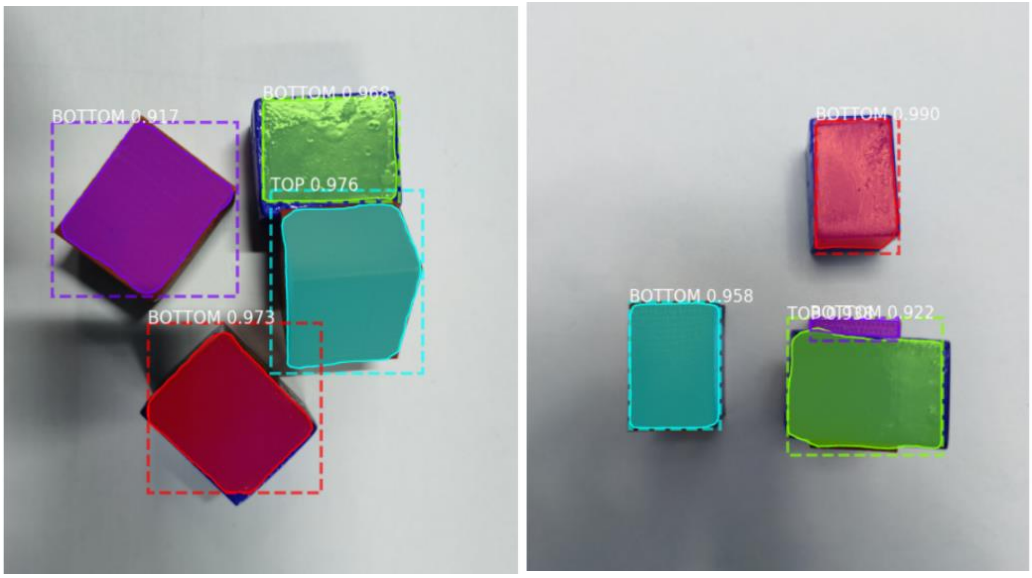


Figure 11(a). Scenario 1 Figure 11(b). Scenario 2

Figure 11. Detection results of Mask-RCNN algorithm for scenario 1and scenario 2.

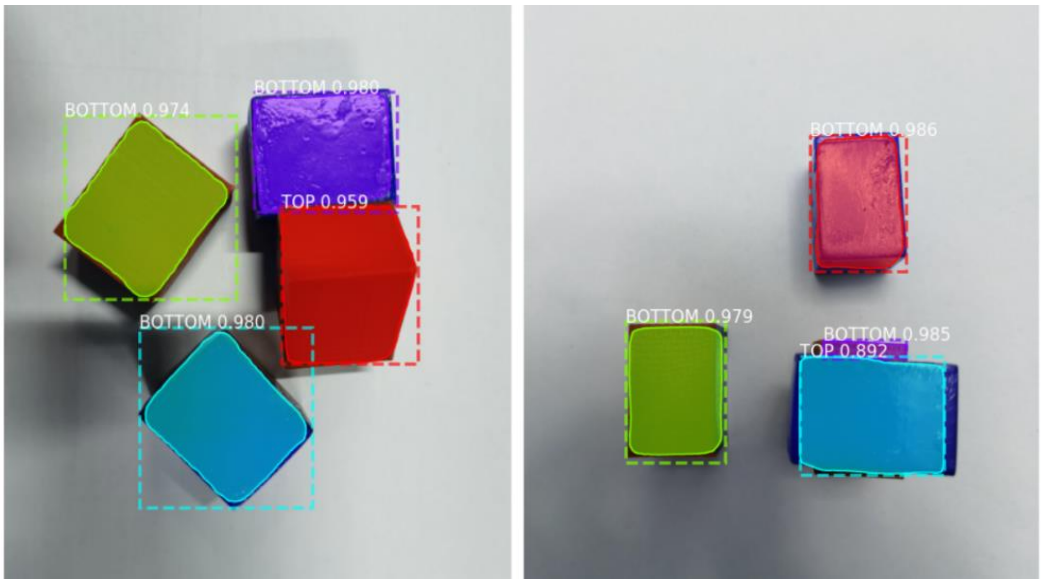


Figure 12(a). Scenario 1 Figure 12(b). Scenario 2

Figure 12. Detection results of Improved Mask-RCNN algorithm for scenario 1 and scenario 2.

Table 3. Target object recognition accuracy comparison

MODEL	M(average score)
Mask-RCNN	0.913
Improved Mask-RCNN	0.936

Figure 13 shows the time taken by the Mask‑RCNN algorithm to recognize a picture of 1024 *320, and Figure 14 shows the time taken by our algorithm proposed in the paper to recognize the same picture. Through the object detection experiment on 20 images under the current hardware experimental conditions, the average elapsed time of Mask-RCNN algorithm is 3.26s, the improved algorithm's average elapsed time is 3.05s, our algorithm saves 9% in average recognition time compared with the original algorithm, detailed data comparison is shown in Table 4. The improved speed can also be seen in another indicator, Mask-RCNN's frame per second (FPS) is 26, while the FPS of our algorithm is 32 FPS, and the FPS value is improved to 6 FPS.

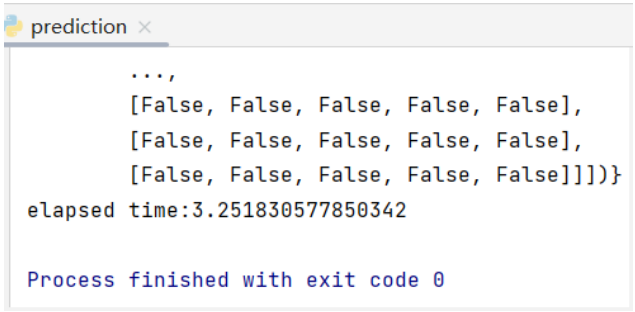


Figure 13. Elapsed time of Mask-RCNN

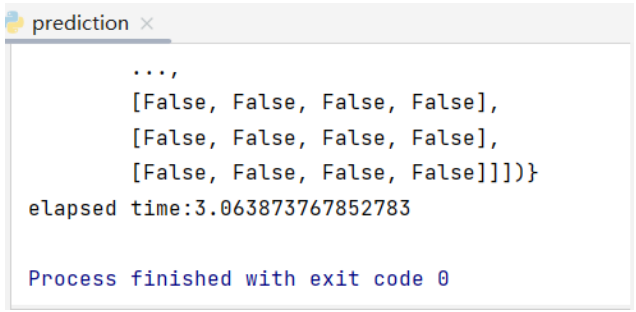


Figure 14. The elapsed time of Improved Mask-RCNN

Table 4. The average time required for recognition of 20 images under two models respectively

	Mask-RCNN	Improved Mask-RCNN
Average elapsed time(s)	3.26	3.05

As can be seen from the above, the target object detection realized based on the improved Mask-RCNN algorithm can achieve higher object detection precision and accuracy. Another, the improved algorithm has significantly reduced the image detection time and increased the value of FPS, which proves that our algorithm can improve the speed of image recognition.

5.0 CONCLUSIONS

This article proposes a method of using indirect frame subtraction (IFS) to improve the loss function of the Mask-RCNN, which uses adjacent frames as comparison templates to get differences for images, reducing the repetitive estimation of regions, the experimental results show that this method can improve the accuracy and precision of object detection, especially in scenes where multiple target objects are partially occluded or stacked. The accuracy and precision of recognition have been improved compared to the original algorithm. More importantly, compared with the original network, the object detection speed has also been significantly improved. However, the target objects in this article are all of the same shape and size, and they are all symmetrical objects. There is no validation for more types of objects or on the published dataset, which is also what this algorithm needs to further verify in the future. Nevertheless, this article proposes a new approach to further improve the detection accuracy and speed of the Mask-RCNN algorithm, exploring the realization of accurate and rapid identification of products in production sites. The results of this research have addressed the problem statement and achieved the research objective by using the improved Mask-RCNN algorithm to detect 4 cuboids targets objects. The results show and verify that the accuracy and speed of the improved algorithm through experiments when the 4 target objects are placed in disorder or even partially occluded.

ACKNOWLEDGEMENTS

This work was supported by the Research Management Center, Shah Alam, Selangor, Malaysia for the financial support from 600-IRMI/FRGS 5/3 (461/2019) through the Ministry of Higher Education grant (FRGS/1/2019/ICT02/UITM/02/10), Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia for the research support. The authors would also like to thank the Hebei Institute of Mechanical and Electrical Technology, Xingtai, China, for providing the experimental equipment.

AUTHORS CONTRIBUTION

Liu QingChuan(Conceptualisation; Methodology; Validation;Writing - original draft)

Muhammad Azmi Ayub (Methodology, Formal analysis; Data curation;Supervision)

Fazlina Ahmat Ruslan (Formal analysis; Investigation; Resources; Software; Visualisation)

Mohd Nor Azmi Ab Patar (Conceptualization; Formal analysis; Writing - review & editing)

Shuzlina Abdul-Rahman (Funding acquisition; Project administration; Supervision)

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

[1] Fu Donghao, “Based on visual target recognition with seven degrees of freedom robot grasping study,” Shenyang university of technology, 2023. doi: 10.27322 /, dc nki. Gsgyu. 2023.000388.

[2] WEI.H,PAN.S,MA.G,et al, “Vision-guided hand-eye coordination for robotic grasping and its application in tangram puzzles,”AIVolume 2, Issue 2. PP 209-228, 2021,doi:10.3390/AI2020013.

[3] Yun.Ke Chen et al,“K-TIG welding seam tracking based on Mask-RCNN model Systematic research,”Master's thesis of South China University of Technology,pp.9-12,2021,doi:10561TG409.

[4] Zhu.Jiahui,“Research on cross-weather road scene re-recognition algorithm,” Beijing University of Posts and Telecommunications, pp.35-36,2018,doi:TP391.41.

[5] Ma Hao,“Research on crowd counting algorithm based on Convolutional Neural Network,” University of Science and Technology of China, pp.15,2018,doi:10.27791/d.cnki.ghegy.2023.000360.

[6] Girshick,Ross B.et al,“Rich Feature Hierarchies for Accurate Object Detection and Semantic Segm entati on,” IEEE Conference on Computer Vision and Pattern Recognition （ 2014 ） ,pp.580-587,2014,doi:10.1109/CVPR.2008.4587671.

[7] Salti.S, Tombari.F, Di.Stefano L. SHOT, “Unique sign atures of histograms for surface and texture description,” Computer Vision and Image Understan ding, 125251-264, 2014,doi:10.1007/978-3-642-15558-1_26.

[8] Xiang.Y, Schmidt.T, Narayan.an V, et al,“PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes,” 14th Conference on Robotics -Science and Sy stems, pp.123-136,2018,doi:10.15607/rss.2018.xiv.019.

[9] Tejani A, Tang D H, Kouskouridas R, et al, “Latent-Class Hough Forests for 3D Object Detection and Pose Estimation,” 13th European Conferen ce on Computer Vision(ECCV),pp.462-477,2014 .doi:10.1109/TPAMI.2017.2665623.

[10] LENZ.I,LEE.H,SAXENA.A,“Deep learning for detecting robotic grasps,”The International Journal of Robotics Research,34(4),pp.705‑724.2015,doi:10.1177/0278364914549607.

[11] Li Longyan,“Based on visual intelligent fetching of mechanical arm technology research,” North China university of water conservancy and hydropower, 2023. doi: 10.27144 /, dc nki. GHBSC. 2023.000205.

[12] Kang.Wei H, Yan.Jun.L,“Mask-RCNN Segmentation Based on 3D Points Cloud Matching for Fish Dimension Measurement,”39th Chinese Control Conference . doi: 10.26914 / Arthur c. nkihy. 2020.038310.

[13] Yan.Na L, DanYang D,“Design and research of bridge crack detection method based on Mask-RCNN,”Applied Optics 43.01(2022):100-105+118. doi: CNKI: SUN: YYGX. 0.2022-01-016.

[14] Da.Song L, “Automatic crack detection of aircraft structure based on improved Mask‑RCNN method,”Journal of Vibration Measurement & Diagnosis,2021, (03),pp.487-494,2021,doi:10.16450/j.cnki.issn.1004-6801.2021.03.010.

[15] Zeng.A, Song.S, Yu.K,“Robotic Pick-and-Place of Novel Objects in Clutter with Multi-Affordance Grasping and Cross-Domain Image Matching,” 2018 EEE International Conference on Robotics and Autom action (ICRA),pp.3750-3757,2018,doi:10.1109/ICRA.2018.8461044.

[16] Jiang.Chuan.Y,“Pavement crack detection based improved Mask-RCNN,”Television Technology 46.06 (2022):7-9+19. doi: 10.16280 / j.v ideoe 2022.06.003.

[17] Yang.Hua.T, Hong.Jian.W,“Feature Extraction for Side Scan Sonar Image Based Deep Learning,” Harbin engineering university, 2022. doi: 10.27060 /, dc nki. Ghbcu. 2022.002320.

[18] Bo.Sen F, Yun Bo Z, “Road information detection algorithm based on improved Mask-RCNN,”Journal of Information Science and Technology University of Beijing (Natural Science Edition). 37(03),pp.19-23,2022,doi:10.11772/j.issn.1001-9081.2020030357.

[19] Qi.C, Rui.Cheng L,“Power line identification method based on improved Mask-RCNN,”Journal of Shantou University (Natural Science Edition). 37(02),pp.65-68,2022,doi:1001 - 4217（2022）02 - 0043 - 07.

[20] Bin.W,“Research on robot capture detection algorithms based depth image and deep learning,”Zhejiang University(2019).

[21] Na.L,Tao.H,“Building extraction from Mask-RCNN high-resolution remote sensing images,” Remote Sensing Information (03),1-6. doi:CNKI:SUN:YGXX.0.2022-03-001.