https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

# A visual grasping method for collaborative robots in unstructured scenes based on object detection and grasp pose estimation

Junxiao Liu[1], Jun Qian[1], Yunkai Tan[1], and Rong Zhou[1]

[1]School of Mechanical Engineering, Hefei University of Technology, Hefei 230009, China

**Correspondence:** Jun Qian (qianjun@hfut.edu.cn)

**Abstract.** In unstructured multi-object scenes, unclear target regions and interference from backgrounds and adjacent objects can affect grasp point selection. To address these problems, this paper proposes a visual grasping method for collaborative robots based on improved object detection and grasp pose estimation networks. In the object detection stage, the corresponding convolution and upsampling modules in YOLOv8n are replaced with RFAConv and DySample, respectively, to improve the detection of target boundaries and local features. In the grasp pose estimation stage, a residual attention structure and a region-guidance mechanism are incorporated into GR-ConvNet to enhance the stability of grasp prediction in complex backgrounds. To further reduce interference from non-target regions, the detected bounding box is used as a bounding box prompt for MobileSAM to generate a target mask. The target mask is then used to constrain the grasp quality map output by the improved GR-ConvNet and restrict grasp point search to the target region. The proposed method is validated on a collaborative robot visual grasping experimental platform using RGB-D images as input. Experimental results show a target detection success rate of 98 % and an overall grasp success rate of 95 %. The proposed method reduces interference from non-target regions during grasp point selection and improves the stability of target grasping in unstructured multi-object scenes.

## 1 Introduction

Autonomous robotic grasping is widely used in intelligent manufacturing, warehouse sorting, service robots, and other applications (Wang et al., 2025; Dong et al., 2024). Unlike structured industrial environments, unstructured multi-object scenes involve considerable variations in target categories, sizes, poses, and spatial distributions and are susceptible to interference from backgrounds and adjacent objects (Wang et al., 2025; Tian et al., 2023). Traditional methods based on taught trajectories or predefined rules often struggle to adapt to complex environments. In contrast, vision-based methods can localize targets and estimate grasp poses, thereby improving the autonomous grasping capability of robotic manipulators (Tian et al., 2023).

In recent years, deep-learning-based grasp pose estimation methods have developed rapidly. Networks such as GG-CNN and GR-ConvNet can predict graspable regions from depth information in RGB-D images and have achieved good performance on public datasets and in real grasping experiments (Morrison et al., 2018; Kumra et al., 2020). However, these methods typically search for grasp points across the entire image and are susceptible to high-response regions corresponding to the background and adjacent objects in multi-object scenes. As a result, the selected grasp point may deviate from the target object, reducing target grasping accuracy.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

To achieve target grasping, a positional constraint on the target object needs to be established before grasp pose estimation. YOLOv8n is lightweight and provides fast detection, which makes it suitable for online visual perception tasks in robotic applications (Sha et al., 2026). However, localization errors may still occur when target objects are small, target boundaries are indistinct, or multiple objects are closely spaced (Chen et al., 2025). A bounding box only provides a rectangular region and cannot accurately describe the actual contour of the target object. Background and adjacent-object regions within the bounding box may still interfere with subsequent grasp point search. Therefore, a more precise target region constraint is required.

SAM and its lightweight model MobileSAM can generate target masks from bounding box prompts, thereby providing a finer representation of the target region than bounding boxes (Kirillov et al., 2023; Zhang et al., 2023). Object detection results as input prompts for the segmentation model can reduce interference from background and adjacent-object regions within the bounding box during target region extraction. This allows the grasp point search region to better match the actual contour of the target object (Han et al., 2026).

To address unclear target regions and interference from backgrounds and adjacent objects during grasp point selection in unstructured multi-object scenes, this paper proposes a visual grasping method for collaborative robots based on improved object detection and grasp pose estimation networks. The method first uses the improved YOLOv8n to identify the target object and determine its position. The detected bounding box is then used as a bounding box prompt for MobileSAM to generate a target mask, which restricts grasp point search to the target region. The improved GR-ConvNet then predicts the grasp pose to enable target grasping in unstructured multi-object scenes. A visual grasping experimental platform consisting of a 6-DOF collaborative robot and an RGB-D camera is established to validate the effectiveness of the proposed method.

## 2 Grasp pose estimation using an improved GR-ConvNet

### 2.1 Two-dimensional grasp pose representation

For two-dimensional grasp pose detection, a five-dimensional grasp representation is commonly used (Lenz et al., 2015) and is defined as Eq. (1):

$$g = (x, y, \theta, h, w), \tag{1}$$

where $(x, y)$ denotes the position coordinates of the grasp rectangle center in the image coordinate system $XOY$; $\theta$ denotes the orientation angle of the grasp rectangle relative to the $X$-axis; $h$ and $w$ denote the height and width of the grasp rectangle, respectively, as shown in Fig. 1.

Based on the five-dimensional grasp representation, a grasp pose estimation network performs grasp prediction on RGB-D images, and the grasp prediction output is defined in Eq. (2):

$$G = (Q, \Theta_{\cos}, \Theta_{\sin}, W), \tag{2}$$

where $Q$ denotes the grasp quality map, which represents the grasp quality score at each pixel; $\Theta_{\cos}$ and $\Theta_{\sin}$ denote the cosine and sine maps of the grasp angle; and $W$ denotes the grasp width map.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

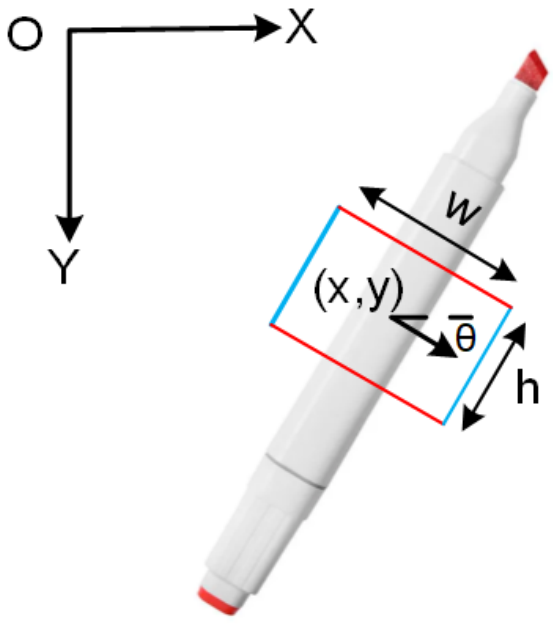



**Figure 1.** Five-dimensional grasp representation.

Based on the network output $G$, the position $(x_p, y_p)$ with the maximum response in the grasp quality map $Q$ is selected as the grasp center. The corresponding grasp angle and grasp width at this position are then used to obtain the two-dimensional grasp pose $g_p$ in the image coordinate system.

## 2.2 Improved GR-ConvNet grasp pose estimation network

GR-ConvNet takes RGB-D images as input and performs pixel-wise grasp pose prediction. The network has a simple structure and relatively fast inference speed, making it suitable for two-dimensional grasp pose detection tasks. However, in unstructured multi-object scenes, background textures, indistinct target boundaries, and adjacent objects can affect the response of the grasp quality map. This may lead to grasp point offsets or unstable grasp angle prediction. To improve the stability of grasp prediction in complex scenes, the feature enhancement module and prediction branch of GR-ConvNet are modified.

**(1) Improvement of the residual attention module**

The feature enhancement module of the original GR-ConvNet mainly employs conventional residual blocks. This structure facilitates feature propagation but does not sufficiently distinguish the importance of features in different channels and spatial locations. Unstructured multi-object scenes may include indistinct target boundaries, local occlusions, and complex backgrounds. Under these conditions, features from background regions and adjacent objects may also be retained, which may affect subsequent grasp prediction.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

This paper adopts the CBAM to enhance the network's ability to select effective grasp features. CBAM consists of a channel attention module and a spatial attention module, which adjust the response weights of feature channels and spatial locations, respectively, as shown in Fig. 2 (Woo et al., 2018). On this basis, CBAM is embedded in the residual block to construct the ResBlock_CBAM residual attention module, as shown in Fig. 3. This module retains the residual connection and incorporates attention weighting. It enhances features related to object contours, local geometric structures, and candidate grasp regions and reduces interference from backgrounds and adjacent objects during grasp prediction (Lu and Liu, 2022).

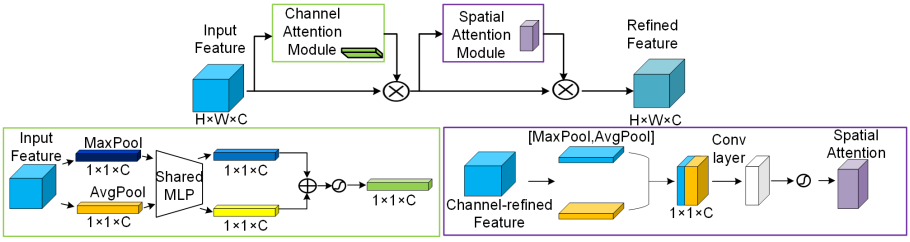



**Figure 2.** Structure of the CBAM.

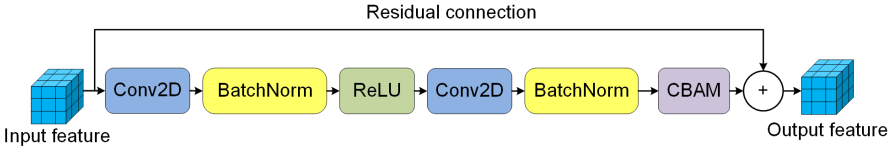



**Figure 3.** Structure of the ResBlock_CBAM module.

**(2) Improvement of the region-guided prediction module**

To further improve the network's ability to localize valid grasp regions, coarse-to-fine prediction and Gaussian regional guidance are added after the decoder (Jeng et al., 2021; Cao et al., 2023). In the coarse prediction stage, the decoder output feature $F_d$ is used to estimate the probability of a valid grasp pose at each pixel. The grasp probability heatmap $P$ is defined in Eq. (3):

$$P = f_c(F_d), \tag{3}$$

where $f_c(\cdot)$ denotes the input–output mapping of the coarse prediction branch.

In the fine prediction stage, an overlapping sliding-window traversal strategy is used to locate high-response regions in the probability heatmap $P$. For the $n$-th window $R_n$, the composite response value $s_n$ is defined as a weighted combination of the overall response and the maximum response within the window, with coefficients $\gamma_1$ and $\gamma_2$, as defined in Eq. (4):

$$s_n = \gamma_1 \sum_{(u,v)\in R_n} P(u,v) + \gamma_2 \max_{(u,v)\in R_n} P(u,v), \tag{4}$$

where $P(u,v)$ denotes the probability value at position $(u,v)$ in the probability heatmap $P$.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

According to $s_n$, the top $K$ high–response windows are selected and denoted by $R'_k$ $(k = 1, \ldots, K)$. The position with the maximum probability in each window is then determined as a candidate grasp position and denoted by $(u_k, v_k)$, as defined in Eq. (5):

$$(u_k, v_k) = \underset{(u,v)\in R'_k}{\arg\max} P(u, v), \tag{5}$$

The candidate grasp positions $(u_k, v_k)$ are used as centers. The corresponding Gaussian weight $D_k(u, v)$ at spatial position $(u, v)$ is defined in Eq. (6):

$$D_k(u, v) = \exp\left[-\frac{(u - u_k)^2 + (v - v_k)^2}{2\sigma^2}\right], \tag{6}$$

where $\sigma$ controls the spatial spread range of the Gaussian weights.

The Gaussian weights generated by all candidate grasp positions are summed at each spatial position and then subjected to maximum-value normalization to obtain the Gaussian region weight map $D$, as defined in Eq. (7):

$$D = \mathrm{Norm}_{\max}\left(\sum_{k=1}^{K} D_k(u, v)\right), \tag{7}$$

where $\mathrm{Norm}_{\max}(\cdot)$ denotes maximum-value normalization.

The Gaussian region weight map $D$ is broadcast along the channel dimension and then multiplied element-wise with the decoder output feature $F_d$ to obtain the fine prediction feature $F_f$, as defined in Eq. (8):

$$F_f = D \odot F_d, \tag{8}$$

where $\odot$ denotes element-wise multiplication.

The fine prediction feature $F_f$ is then fed into the fine prediction branch to obtain the grasp prediction result, as defined in Eq. (9):

$$G = f_p(F_f), \tag{9}$$

where $f_p(\cdot)$ denotes the input–output mapping of the fine prediction branch.

The overall structure of the improved GR-ConvNet is shown in Fig. 4. The CBAM attention mechanism is embedded in the original residual blocks to enhance the feature representation of object contours and local geometric features. Coarse-to-fine prediction and Gaussian regional guidance are added to the prediction stage. The Gaussian region weight map derived from the candidate grasp positions is used to weight the decoder features. This allows the fine prediction branch to focus on features near valid grasp regions and improves the stability of grasp pose prediction in complex backgrounds.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

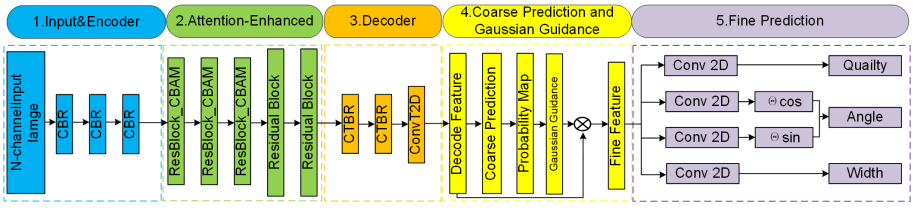



**Figure 4.** Structure of the improved GR-ConvNet grasp pose estimation network.

## 3 Improved YOLOv8n object detection network

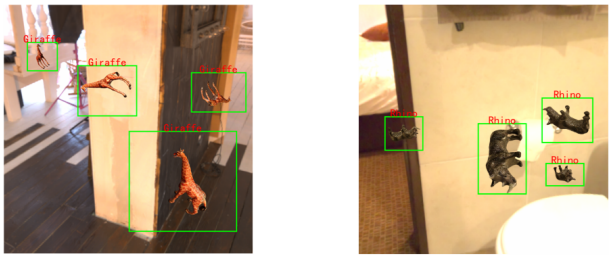
### 3.1 Construction of a synthetic object detection dataset

In robotic grasping scenes, object poses, scales, illumination, and background conditions vary considerably. Relying entirely on the collection and manual annotation of real images increases the cost of dataset construction and makes it difficult to cover complex scenes. Therefore, a synthetic object detection dataset is constructed to improve the adaptability of the detection model to unstructured grasping scenes.

The dataset contains 10 object categories: apple, banana, lemon, ball, marker pen, toothpaste box, cola can, elephant, giraffe, and rhino. During data generation, three-dimensional models from the Google Scanned Objects (GSO) dataset are used to construct synthetic scenes (Downs et al., 2022). Object positions, scales, rotation angles, camera viewpoints, and other parameters are randomly set (Tobin et al., 2017). High-resolution background images of indoor and outdoor scenes are used to enrich the synthetic scenes. The corresponding object class labels and bounding box annotations are generated simultaneously with the synthetic images, and the annotation files are organized in YOLO format. Examples of the synthetic object detection dataset are shown in Fig. 5.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS




(a)Synthetic image with an indoor background

(b)Synthetic image with an outdoor background

**Figure 5.** Examples of the synthetic object detection dataset.

### 3.2 Improved YOLOv8n network structure

YOLOv8n features a lightweight architecture and fast inference. It is suitable for online visual perception tasks in robotic applications. However, in unstructured multi-object scenes, targets vary considerably in scale, pose, and boundary characteristics. The original YOLOv8n remains limited in local structural feature extraction and multi-scale feature alignment. To address these limitations, the corresponding convolution modules in the backbone are replaced with RFAConv modules, while the two fixed upsampling operations in the neck are replaced with DySample modules (Zhang et al., 2026; Liu et al., 2023). The structure of the improved network is shown in Fig. 6.

In the backbone, standard convolution applies shared kernel parameters to receptive fields at different spatial locations, which limits its adaptability to variations in local content. When target boundaries are indistinct, local textures are weak, or adjacent objects are closely spaced, contour and local structural information may diminish during successive convolution and downsampling operations. For small or slender objects, boundary regions occupy only a small proportion of the feature maps and are therefore more susceptible to interference from backgrounds and adjacent objects.

To improve local feature extraction, the corresponding convolution modules in the backbone are replaced with RFAConv modules. Each RFAConv module constructs receptive-field spatial features from the input features and generates corresponding attention weights. These weights are used to reweight features at different positions within the receptive field. The reweighted features are then reshaped and passed to subsequent layers. This allows feature responses to adapt to variations in local content. It also enhances the representation of target boundaries and local structures while reducing interference from backgrounds and adjacent objects.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

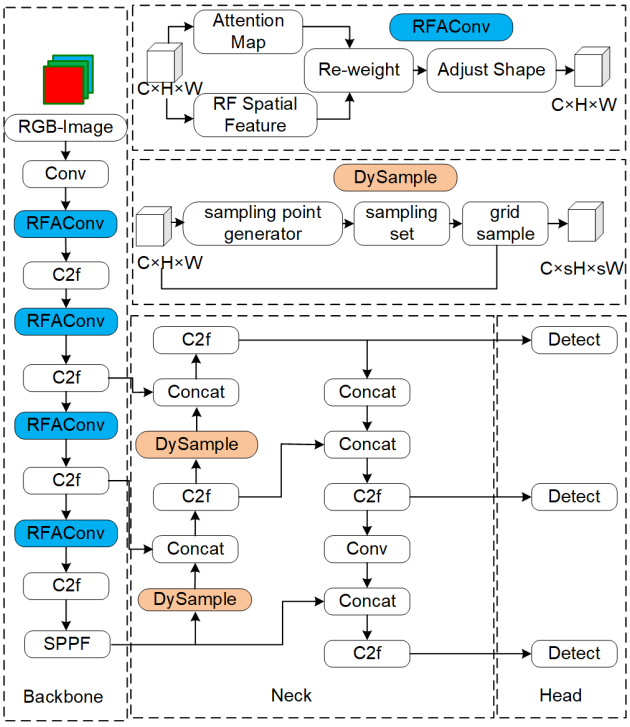



**Figure 6.** Structure of the improved YOLOv8n network.

In the neck, the original YOLOv8n uses fixed upsampling for high-level features and concatenates them with shallow features from the backbone. High-level features mainly contain semantic information about the targets, whereas shallow features retain more boundary and positional details. In unstructured multi-object scenes, considerable variations in target scale may cause spatial misalignment between features from different levels when fixed sampling locations are used. This can affect multi-scale feature fusion and bounding box prediction.

Accordingly, the two fixed upsampling operations in the neck are replaced with DySample modules. DySample predicts sampling offsets from the input features and adjusts the sampling coordinates based on a regular sampling grid. The upsampled features are obtained through grid sampling. These features are then concatenated with the corresponding shallow features from the backbone and further fused by the C2f modules. The sampling locations can therefore be adaptively adjusted according to the input features. This improves the spatial correspondence between high-level semantic features and shallow detail features.

The improved YOLOv8n uses RFAConv to enhance the extraction of target boundary and local structural features, while DySample improves spatial alignment during multi-scale feature fusion.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

## 4 Target grasping method based on target mask constraint

### 4.1 Grasping strategy based on target region constraint

In unstructured multi-object scenes, the object detection network identifies the target object and provides its bounding box. However, a rectangular bounding box cannot accurately represent the object boundary. When the target shape is irregular or adjacent objects are closely spaced, background and adjacent-object regions may be included within the bounding box and affect grasp point selection in the grasp quality map (Dong et al., 2021).

MobileSAM is a lightweight image segmentation model based on SAM and can generate target masks from prompts such as points and bounding boxes (Xiong et al., 2024). To obtain a more precise target region, the bounding box output by the object detection network is used as a bounding box prompt for MobileSAM to segment the target object and obtain the target mask $M_{sam}$. Compared with a rectangular bounding box, the target mask more closely matches the contour of the visible target region and can be used to constrain the grasp point search range.

To suppress grasp quality responses from non-target regions, the target mask $M_{sam}$ is multiplied element-wise with the grasp quality map $Q$ to obtain the mask-constrained grasp quality map $Q_m$, as defined in Eq. (10):

$$Q_m = Q \odot M_{sam} \tag{10}$$

During grasp point selection, the pixel with the maximum response in $Q_m$ is selected as the grasp center. The corresponding grasp angle and grasp width at this position are then used to obtain the two-dimensional grasp pose.

### 4.2 Visual grasping process for the collaborative robot

The visual grasping process is shown in Fig. 7. After the collaborative robot moves to the initial pre-grasp pose, the RGB-D camera captures an image of the current scene. The RGB image is fed into the improved YOLOv8n to detect the target object and obtain its bounding box. The RGB-D image is fed into the improved GR-ConvNet to obtain the grasp quality map $Q$, grasp angle map $\Theta$, and grasp width map $W$.

During object detection, the current grasping process is terminated if the target object is not detected. If the target object is detected, the detected bounding box is used as a bounding box prompt for MobileSAM to generate the target mask $M_{sam}$. The target mask $M_{sam}$ is then used to constrain the grasp quality map $Q$ to the target region and obtain the mask-constrained grasp quality map $Q_m$. The two-dimensional grasp pose $g_p$ corresponding to the selected grasp center is then obtained in the image coordinate system, and the depth value at the grasp center pixel $(x_p, y_p)$ is subsequently extracted from the depth image.

The grasp pose in the camera coordinate system is then obtained through the coordinate transformation from the image coordinate system to the camera coordinate system. Based on the hand-eye calibration result, the grasp pose is transformed into the robot base coordinate system to obtain the spatial grasp pose $G_r$ required for robot execution. The collaborative robot then performs the grasping operation to achieve target grasping in unstructured multi-object scenes. The spatial grasp pose $G_r$ is defined in Eq. (11):

$$G_r = T_{rc}\left(T_{ci}(g_p)\right), \tag{11}$$

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

where $T_{rc}$ denotes the coordinate transformation from the camera coordinate system to the robot base coordinate system, and $T_{ci}$ denotes the coordinate transformation from the image coordinate system to the camera coordinate system.

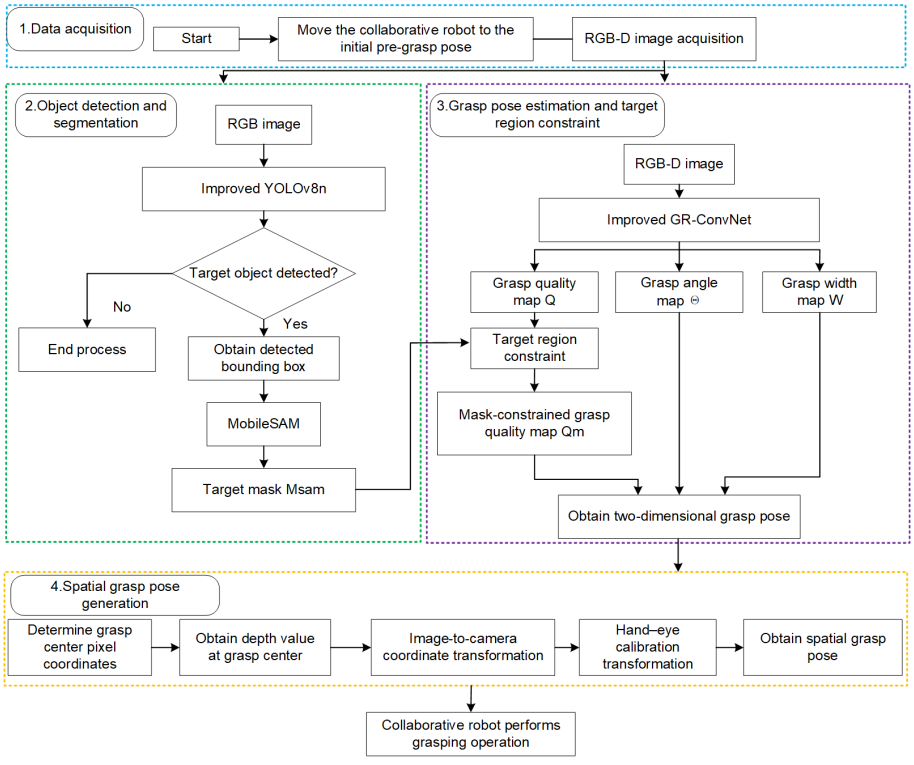



**Figure 7.** Visual grasping process for the collaborative robot.

## 5 Experimental results and analysis

### 5.1 Grasp pose estimation experiments

To evaluate the performance of the grasp pose estimation network, this paper uses the Cornell Grasping Dataset for training and testing. This dataset is a public dataset commonly used for two-dimensional grasp detection (Jiang et al., 2011). The model performance is evaluated using the rectangle metric commonly adopted for the dataset. A predicted grasp rectangle $G_{pre}$ is considered a correct grasp if it satisfies both the angle error and rectangle overlap criteria with respect to the ground-truth grasp rectangle $G_{gt}$. The angle error criterion is defined in Eq. (12):

$$|\theta_{pre} - \theta_{gt}| < \frac{\pi}{6}, \tag{12}$$

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

where $\theta_{pre}$ and $\theta_{gt}$ denote the predicted grasp angle and the ground-truth grasp angle, respectively. The rectangle overlap is measured using the Jaccard index, and the criterion is defined in Eq. (13):

$$\text{Jaccard}(G_{pre}, G_{gt}) = \frac{|G_{pre} \cap G_{gt}|}{|G_{pre} \cup G_{gt}|} > 0.25 \tag{13}$$

The grasp detection accuracy is calculated as the ratio of the number of correctly predicted samples to the total number of test samples. Experiments on the Cornell dataset commonly use two splitting strategies: image-wise(IW) splitting and object-wise(OW) splitting (Zou et al., 2023). Considering that unknown objects are more common in practical grasping scenes, the object-wise splitting is used to compare different network structures. The input image size is $224 \times 224$ , and the experimental results of different models are shown in Table 1.

**Table 1.** Comparative experiments of grasp pose estimation models.

| Model | ResBlock_CBAM | ResBlocks | CF-GRG | Inference time (ms) | Parameters (M) | OW (%) |
|---|---|---|---|---|---|---|
| GG-CNN | – | – | – | 12.28 | 0.07 | 84 |
| GR-ConvNet | 0 | 5 | – | 17.26 | 1.89 | 91 |
| GR-ConvNet-3CBAM-2Res | 3 | 2 | – | 18.96 | 1.91 | 96 |
| Improved GR-ConvNet | 3 | 2 | ✓ | 22.08 | 1.91 | 98 |

As shown in Table 1, GR-ConvNet achieves higher accuracy than GG-CNN. After the incorporation of ResBlock_CBAM, the accuracy is further improved with only a small change in the number of parameters. This indicates that the residual attention module can enhance the representation of effective grasp features. After the addition of coarse-to-fine prediction and Gaussian regional guidance, the improved model achieves the highest grasp detection accuracy. Although the inference time increases, it remains at the millisecond level. The prediction results of the improved model on the Cornell dataset are shown in Fig. 8. For objects with different sizes, shapes, and poses, the improved model produces effective grasp responses, and the predicted grasp rectangles adapt to variations in object position and orientation. A comparison of the prediction results before and after the improvement is shown in Fig. 9. Compared with the original GR-ConvNet, the improved model produces more concentrated responses in the target regions, with fewer redundant responses near image boundaries and in background regions. In multi-object scenes, the response regions corresponding to different objects are more clearly separated, which reduces interference from backgrounds and adjacent objects during grasp point selection.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

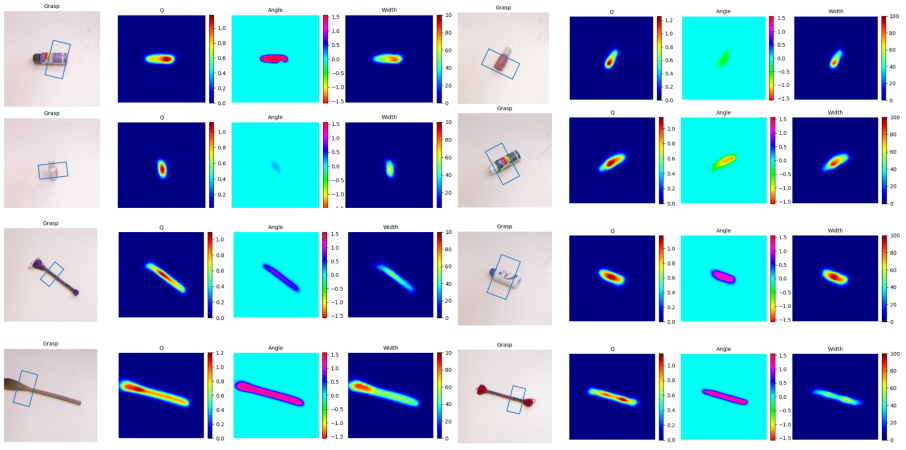



**Figure 8.** Prediction results of the improved GR-ConvNet on the Cornell dataset.

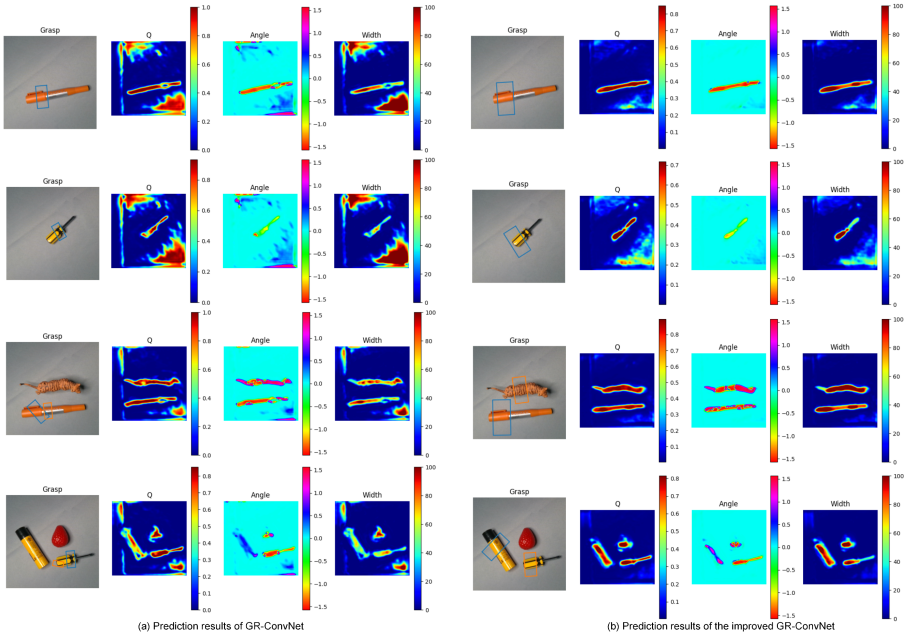



**Figure 9.** Comparison of grasp prediction results between GR-ConvNet and the improved GR-ConvNet.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

### 5.2 Object detection experiments and analysis

To validate the effectiveness of the improved YOLOv8n object detection network, this paper uses the synthetic object detection dataset for training and testing. The input image size is $640 \times 640$.

To analyze the effects of different modules on detection performance, four sets of ablation experiments are conducted under the same data split, input image size, and training parameters. The experimental results are presented in Table 2.

**Table 2.** Results of ablation experiments on the YOLOv8n improvement modules.

| Model | RFAConv | DySample | mAP50 (%) | mAP50:95 (%) | Params (M) | GFLOPs |
|---|---|---|---|---|---|---|
| YOLOv8n | – | – | 93.1 | 84.9 | 3.01 | 8.1 |
| YOLOv8n-RFAConv | ✓ | – | 94.4 | 86.7 | 3.03 | 8.3 |
| YOLOv8n-DySample | – | ✓ | 93.9 | 85.8 | 3.02 | 8.2 |
| Improved YOLOv8n | ✓ | ✓ | 94.9 | 87.5 | 3.05 | 8.3 |

As shown in Table 2, using either RFAConv or DySample individually improves the detection accuracy of YOLOv8n, with RFAConv providing a more pronounced improvement. When both modules are incorporated, the model achieves the highest mAP50 and mAP50:95. Compared with the original YOLOv8n, the improved model shows only slight increases in the number of parameters and computational complexity. This indicates that the two modules improve detection accuracy while maintaining low model complexity.

To evaluate the overall performance of the improved model, this paper selects several lightweight object detection models for comparison. The experimental results are presented in Table 3.

**Table 3.** Comparative results of object detection models.

| Model | mAP50 (%) | mAP50:95 (%) | Inference time (ms) | Params (M) | GFLOPs |
|---|---|---|---|---|---|
| YOLOv8n | 93.1 | 84.9 | 11.41 | 3.01 | 8.1 |
| YOLOv9t | 93.5 | 85.4 | 27.75 | 1.97 | 7.6 |
| YOLOv10n | 93.2 | 85.3 | 12.84 | 2.27 | 6.6 |
| YOLOv11n | 93.4 | 85.1 | 13.73 | 2.60 | 6.3 |
| Improved YOLOv8n | 94.9 | 87.5 | 13.42 | 3.05 | 8.3 |

As shown in Table 3, the improved YOLOv8n achieves the highest mAP50 and mAP50:95. This indicates that it has good object detection performance on the synthetic dataset used in this paper. Compared with the original YOLOv8n, the improved model shows only slight increases in the number of parameters and computational complexity. Although the inference time increases, it remains at the millisecond level and can meet the online processing requirements of the object detection stage.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

To improve the model's adaptability to real-scene images, this paper collects a small-scale real-world object detection dataset. The model parameters obtained from training on the synthetic dataset are used as pretrained weights, and the model is then fine-tuned on the real-world dataset. The detection results in real multi-object scenes are shown in Fig. 10. Compared with the original YOLOv8n, the improved model shows improved detection stability for some target objects and provides more accurate target regions for subsequent target mask extraction and grasp point search region constraint.

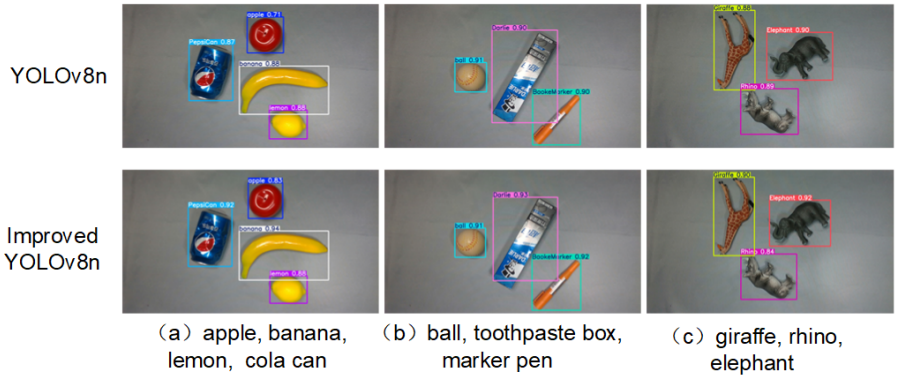



**Figure 10.** Comparison of object detection results in real multi-object scenes.

## 5.3 Analysis of target region constraint effectiveness

This paper uses a marker pen as the test object and selects three scenes: an isolated target, adjacent objects, and partial occlusion. Target mask extraction experiments are conducted in these scenes, and the results are shown in Fig. 11. In all three scenes, the target mask preserves the main visible region of the marker pen and reduces the inclusion of background and adjacent-object regions within the bounding box.

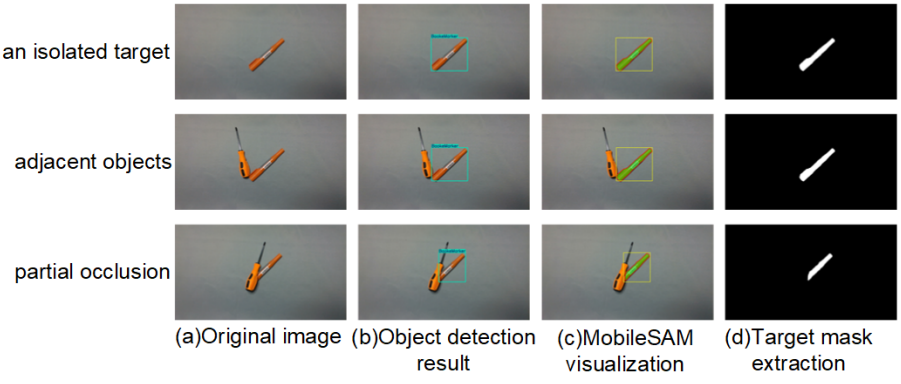



**Figure 11.** Target mask extraction results in different scenes.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

To quantify the effectiveness of the target region constraint, this paper uses images from real multi-object scenes for comparison. The rectangular region covered by the YOLO bounding box is denoted by $M_{box}$, the target mask region generated by MobileSAM is denoted by $M_{sam}$, and the manually annotated region is denoted by $M_{gt}$. Each of $M_{box}$ and $M_{sam}$ is used as the predicted region $M_{pred}$.

The intersection over union (IoU) is used to measure the overlap between the predicted region and the ground-truth target region. It is defined in Eq. (14):

$$\mathrm{IoU} = \frac{|M_{pred} \cap M_{gt}|}{|M_{pred} \cup M_{gt}|}, \tag{14}$$

When the region covered by the YOLO bounding box is used, $M_{pred} = M_{box}$; when the target mask region generated by MobileSAM is used, $M_{pred} = M_{sam}$.

To measure the proportion of background and adjacent-object regions within the predicted region, the background redundancy ratio $R_b$ is defined in Eq. (15):

$$R_b = \frac{|M_{pred}| - |M_{pred} \cap M_{gt}|}{|M_{pred}|}, \tag{15}$$

A smaller $R_b$ indicates a lower proportion of non-target regions within the predicted region.

The evaluation results of the target region constraint are shown in Table 4. Compared with the YOLO bounding box, the MobileSAM mask increases the mean IoU by approximately 55.5% and reduces the mean background redundancy ratio by approximately 86.5%. The results indicate that the target mask more accurately matches the visible target region and reduces the inclusion of background and adjacent-object regions.

**Table 4.** Evaluation results of target region constraint effectiveness.

| Method | Mean IoU | Mean background redundancy ratio |
|---|---|---|
| YOLO bounding box | 0.586 | 0.414 |
| YOLO + MobileSAM mask | 0.911 | 0.056 |

### 5.4 Collaborative robot grasping experiments

To validate the target grasping capability of the proposed method in multi-object scenes, this paper constructs an experimental platform using a UR5 collaborative robot and an Intel RealSense D435i RGB-D camera. The RGB-D camera is mounted on the robot end-effector. A marker pen, a lemon, and other objects are placed simultaneously in the experimental scene, and the marker pen and lemon are sequentially selected as target objects for grasping.

During the experiment, the collaborative robot moves to the pre-grasp pose, and the RGB-D camera captures an image of the current scene. The object detection network simultaneously identifies the marker pen and lemon in the scene. The marker pen is first selected as the target object, and its detected bounding box is used as a bounding box prompt for MobileSAM to generate the target mask. The target mask is then used to constrain the grasp quality map and obtain the mask-constrained

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

grasp quality map $Q_m$. The pixel with the maximum response in $Q_m$ is selected as the grasp center. The corresponding grasp angle and grasp width are then used to obtain the two-dimensional grasp pose. Based on the depth value corresponding to the grasp center and the hand–eye calibration result, the two-dimensional grasp pose is transformed into the robot base coordinate system to obtain the spatial grasp pose. The collaborative robot then completes the target grasping and placement operation. After the marker pen is grasped, the scene image is captured again. The lemon is then selected as the target object and grasped following the above process. The grasping process is shown in Fig. 12.

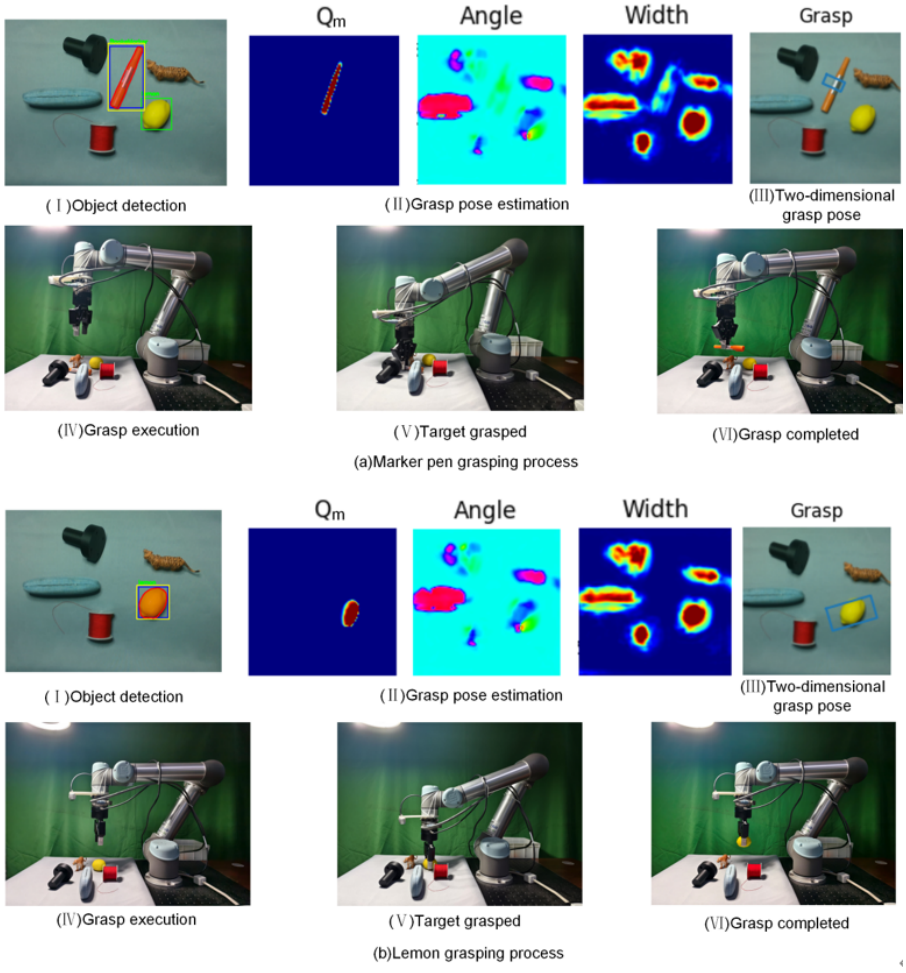



**Figure 12.** Target grasping process in an unstructured multi-object scene.

When multiple objects are present in the scene, the system can determine the target object according to the input object class. In unstructured multi-object scenes, this paper conducts 100 target grasping experiments using the UR5 collaborative robot. The experimental results are shown in Table 5. The target detection success rate is 98%. This shows that the improved YOLOv8n can reliably identify target objects. The few detection failures are mainly caused by complex backgrounds and

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

interference from adjacent objects. The overall grasp success rate is 95%. This shows that the proposed method can accomplish target grasping under interference from multiple objects. Among the successfully detected targets, the remaining grasp failures are mainly caused by errors in the predicted grasp center position or grasp angle, which prevent the gripper from accurately aligning with a suitable grasp position on the target object.

**Table 5.** Results of target grasping experiments.

| Object classes | Detections | Successful detections | Successful grasps | Grasp success rate (%) |
|---|---|---|---|---|
| marker pen, lemon | 20 | 20 | 19 | 95 |
| apple, banana | 20 | 20 | 20 | 100 |
| toothpaste box, cola can | 20 | 19 | 18 | 90 |
| rhino, elephant | 20 | 19 | 18 | 90 |
| ball, giraffe | 20 | 20 | 20 | 100 |

## 6 Conclusions

To address unclear target regions and interference from backgrounds and adjacent objects during grasp point selection in unstructured multi-object scenes, this paper proposes a visual grasping method for collaborative robots based on improved object detection and grasp pose estimation networks. A visual grasping experimental platform is also established to validate the proposed method. The main work is summarized as follows:

(1) In the grasp pose estimation stage, a residual attention structure and a region-guidance mechanism are incorporated into GR-ConvNet. This concentrates the grasp quality responses and reduces interference from complex backgrounds during grasp point prediction.

(2) In the object detection stage, the corresponding convolution and upsampling modules in YOLOv8n are replaced with RFAConv and DySample, respectively. This enhances the detection of target boundaries, local structures, and scale variations and provides bounding boxes for subsequent target mask generation.

(3) In the visual grasping stage, the detected bounding box is used as a bounding box prompt for MobileSAM to generate the target mask. The target mask is then used to constrain the grasp quality map and restrict the grasp point search to the target region. This reduces interference from background regions and adjacent objects during grasp point selection.

The visual grasping experimental results show that the proposed method can achieve target grasping based on the input object class. It provides a feasible solution for visual grasping by collaborative robots in unstructured multi-object scenes.

*Code and data availability.* The code and data used in this study are available from the corresponding author upon reasonable request.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

*Author contributions.* JL conducted the main research, performed the experiments, developed the code, and prepared the original manuscript. JQ supervised the research and revised the manuscript. YT assisted with the experiments. RZ provided research guidance.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.* This project was supported by Natural Science Foundation of Anhui Province (Grant No. 2208085ME127)

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

## References

Cao, H., Chen, G., Li, Z., Feng, Q., Lin, J., and Knoll, A.: Efficient grasp detection network with Gaussian-based grasp representation for robotic manipulation, IEEE/ASME Trans. Mechatron., 28, 1384–1394, https://doi.org/10.1109/TMECH.2022.3224314, 2023.

Chen, C., Li, J., Shuai, Z., Wang, Y., and Wang, Y.: A lightweight optimization framework for real-time pedestrian detection in dense and occluded scenes, Mech. Sci., 16, 877–886, https://doi.org/10.5194/ms-16-877-2025, 2025.

Dong, L., Ma, J., Cao, J., and Wang, D.: Serial–parallel cooperative assembly approach for precision micro-assembly of axial holes, Mech. Sci., 15, 653–665, https://doi.org/10.5194/ms-15-653-2024, 2024.

Dong, M., Wei, S., Yu, X., and Yin, J.: MASK-GD segmentation based robotic grasp detection, Comput. Commun., 178, 124–130, https://doi.org/10.1016/j.comcom.2021.07.012, 2021.

Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T. B., and Vanhoucke, V.: Google Scanned Objects: a high-quality dataset of 3D scanned household items, in: 2022 IEEE International Conference on Robotics and Automation (ICRA), pp. 2553–2560, https://doi.org/10.1109/ICRA46639.2022.9811809, 2022.

Han, D., Li, H., Li, Y., and Chen, S.: Real-time target-oriented grasping framework for resource-constrained robots, Sensors, 26, 645, https://doi.org/10.3390/s26020645, 2026.

Jeng, K.-Y., Liu, Y.-C., Liu, Z. Y., Wang, J.-W., Chang, Y.-L., Su, H.-T., and Hsu, W.: GDN: a coarse-to-fine (C2F) representation for end-to-end 6-DoF grasp detection, in: Proceedings of the 2020 Conference on Robot Learning, vol. 155 of *PMLR*, pp. 220–231, 2021.

Jiang, Y., Moseson, S., and Saxena, A.: Efficient grasping from RGBD images: learning using a new rectangle representation, in: 2011 IEEE International Conference on Robotics and Automation (ICRA), pp. 3304–3311, https://doi.org/10.1109/ICRA.2011.5980145, 2011.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R.: Segment Anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3992–4003, https://doi.org/10.1109/ICCV51070.2023.00371, 2023.

Kumra, S., Joshi, S., and Sahin, F.: Antipodal robotic grasping using generative residual convolutional neural network, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 9626–9633, https://doi.org/10.1109/IROS45743.2020.9340777, 2020.

Kumra, S., Joshi, S., and Sahin, F.: GR-ConvNet v2: a real-time multi-grasp detection network for robotic grasping, Sensors, 22, 6208, https://doi.org/10.3390/s22166208, 2022.

Lenz, I., Lee, H., and Saxena, A.: Deep learning for detecting robotic grasps, Int. J. Robot. Res., 34, 705–724, https://doi.org/10.1177/0278364914549607, 2015.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S.: Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936–944, https://doi.org/10.1109/CVPR.2017.106, 2017.

Liu, W., Lu, H., Fu, H., and Cao, Z.: Learning to upsample by learning to sample, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6004–6014, https://doi.org/10.1109/ICCV51070.2023.00554, 2023.

Lu, J. and Liu, Y.: A real-time and accurate detection approach for bucket teeth falling off based on improved YOLOX, Mech. Sci., 13, 979–990, https://doi.org/10.5194/ms-13-979-2022, 2022.

Morrison, D., Leitner, J., and Corke, P.: Closing the loop for robotic grasping: a real-time, generative grasp synthesis approach, in: Proceedings of Robotics: Science and Systems XIV, https://doi.org/10.15607/RSS.2018.XIV.021, 2018.

https://doi.org/10.5194/ms-2026-165
Preprint. Discussion started: 7 September 2026


mechanical sciences
DISCUSSIONS

Sha, A., Xue, F., Zhang, Y., Zang, X., and Zhao, J.: RIL-YOLO: a lightweight real-time object detection model on mobile devices for kart racing, Mech. Sci., 17, 371–379, https://doi.org/10.5194/ms-17-371-2026, 2026.

Tian, H., Song, K., Li, S., Ma, S., Xu, J., and Yan, Y.: Data-driven robotic visual grasping detection for unknown objects: a problem-oriented review, Expert Syst. Appl., 211, 118 624, https://doi.org/10.1016/j.eswa.2022.118624, 2023.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 23–30, https://doi.org/10.1109/IROS.2017.8202133, 2017.

Wang, J., Sun, X., and Wang, W.: Trajectory planning for manipulator grasping with obstacle avoidance and visual occlusion in a complex environment, Mech. Sci., 16, 445–455, https://doi.org/10.5194/ms-16-445-2025, 2025.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S.: CBAM: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19, https://doi.org/10.1007/978-3-030-01234-2_1, 2018.

Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F. N., Krishnamoorthi, R., and Chandra, V.: EfficientSAM: leveraged masked image pretraining for efficient Segment Anything, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16 111–16 121, https://doi.org/10.1109/CVPR52733.2024.01525, 2024.

Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S.-H., Lee, S., and Hong, C. S.: Faster Segment Anything: towards lightweight SAM for mobile applications, arXiv [preprint], https://doi.org/10.48550/arXiv.2306.14289, 2023.

Zhang, X., Liu, C., Song, T., Yang, D., Ye, Y., Li, K., and Song, Y.: RFAConv: receptive-field attention convolution for improving convolutional neural networks, Pattern Recognit., 176, 113 208, https://doi.org/10.1016/j.patcog.2026.113208, 2026.

Zou, M., Li, X., Yuan, Q., Xiong, T., Zhang, Y., Han, J., and Xiao, Z.: Robotic grasp detection network based on improved deformable convolution and spatial feature center mechanism, Biomimetics, 8, 403, https://doi.org/10.3390/biomimetics8050403, 2023.