ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Deep learning detection model for industrial robot grasping target based on multi-scale attention mechanism fusion

Junli Hu[1,*]

[1] School of Mechanical and Electrical Engineering, Henan Industry and Trade Vocational College, Zhengzhou 451191, Henan Province, China

SUMMARY: *In cluttered manufacturing cells, industrial robot grasping needs to deal with occlusion, reflection, scale change and attitude deflection at the same time, and it is difficult to stably complete target recognition and grasp parameter expression with single-layer features. This paper proposes a deep learning detection model that combines multi-scale attention mechanism and hierarchical feature interaction, and constructs an integrated framework of scene modeling, feature enhancement, localization regression and task mapping. The model jointly uses shallow texture and high-level semantic information, and the detection head synchronously outputs the category, center coordinates, width and height parameters and rotation direction. Experiments are carried out on 18240 RGB-D images. The results show that the mAP@0.5 of the proposed method in four types of industrial scenarios is higher than 95.84%, the test set Accuracy reaches 96.74%, and the total system delay is 24.6ms under TensorRT acceleration. Compared with YOLOv5s, HTC-Grasp, FAGD-Net and ODGNet, the proposed model has lower missed detection rate and more stable online calling ability, which can provide reliable visual support for industrial robots to grasp in real-time.*

KEYWORDS: *Multi-scale attention mechanism; Industrial robot; Grasp object detection; Deep learning*

1 Introduction

Industrial robots have shifted from repetitive execution devices to systems with environmental perception, target recognition and task collaboration capabilities, and grasping target detection has become a computational link between visual information and action execution. There are occlusion overlaps, surface reflections, scale changes and posture deflection in the scenes of sorting, assembly, palletitioning and pick-and-place in boxes. It is difficult to stably give grasping candidate regions by only relying on regular features or shallow visual descriptions. The deep learning detection model provides an end-to-end expression path from image input to grasp parameter output for industrial robots, which makes the grasping task change from experience-driven to data-driven and structure-driven parallel computing mode. The value of this type of research is not only reflected in the execution effect, but also in the unification between feature modeling, network design, inference efficiency, and system deployment.

From the development path of computer vision, industrial grasp detection has been extended from rectangular box recognition to joint prediction of grasp center, opening width, rotation Angle and grasp score. Qin et al. studied an efficient grasp detection network under

*hujunli2023@163.com
https://doi.org/10.65102/is2026325 Published: 30 April, 2026

attention constraints, and enhanced the responsiveness of local regions by improving the feature selection method [1]. Shi et al. proposed an Angle label smoothing strategy to make the rotation grasp representation maintain a more stable orientation distribution during the training phase [2]. Hong et al. constructed a residual attention generation network, which combined grasping candidate expressions with residual features to improve the discrimination performance in complex backgrounds [3]. Wang et al. proposed a Transformer grasp detection structure, which uses context relations to supplement the shortcomings of local convolution in long-range dependence modeling [4]. Yu et al. proposed SE-ResUNet method, which introduced the idea of channel recaliphation into the grasp detection process and achieved higher consistency in fine boundary recognition [5]. This kind of research shows that grasp detection is no longer a direct transfer of the general object detection framework, but forms a specialized network branch oriented to grasp geometric representation.

Focusing on the feature representation in complex industrial scenes, many researches have begun to promote the detection accuracy and stability from three directions: keypoint modeling, multi-scale fusion, and pixel-level prediction. Zhai et al. studied a fast grasp detection method based on key points, which achieves a good balance between speed and accuracy [6]. Xi et al. proposed a pixel-level grasp detection method, so that the fine-grained distribution of the grasp area can be directly encoded into the network output [7]. Zhang et al. proposed Hybrid Transformer-CNN architecture to integrate convolutional local perception and Transformer global correlation into the same detection process [8]. Fang et al. proposed a collaborative attention and multi-scale feature fusion method to strengthen the information interaction between layers under the guidance of loss [9]. Zhong et al. proposed FAGD-Net to improve the discrimination ability in clutter-free scenes through efficient multi-scale attention mechanism and feature enhancement strategy [10]. Kuang et al. proposed a grasp detection network based on full-dimensional dynamic convolution, so that the response of the convolution kernel can adaptively change with the input content [11]. Bai et al. proposed an enhanced CenterNet grasp detection method to further compress the center point representation into a more compact positioning process [12]. Lei et al. proposed a grasp detection algorithm based on 3D visual two-stream coding, which formed a clear calculation path in the collaborative description of RGB and depth information [13]. These results promote industrial grasp detection from single-scale, single-path modeling to multi-branch, multi-semantic level collaborative modeling.

In addition to the improvement of network structure, data organization, resource constraint and scene generalization ability have also become the focus of computing research. Dolezel et al. proposed a lightweight grasp point detection method for container grasping, which maintains good reasoning ability under the condition of limited storage [14]. Gu et al. studied a cooperative convolutional model for grasp detection, which strengthened the visual assignment relationship between multiple targets and multiple operation units [15]. Sun et al. proposed a cooperative robot recognition and grasping method based on vision, and mapped the target recognition results to the execution end [16]. Khor et al. studied the deep feature detection method in unknown target grasping, which improved the adaptation ability of non-a priori category objects [17]. Rasheed et al. proposed an improved UNet generative grasping network with attention mechanism, which enhanced the spatial consistency of the grasping region generation process [18]. Zhao et al. proposed a 6-DOF grasp availability learning method based on Transformer global encoding, which extends the expression range of traditional planar grasp to high-dimensional pose estimation [19]. Zhang et al. proposed a grasping detection method for disordered manufacturing scenes based on automatically labeled data sets, which makes data construction and detection training form a tighter closed

loop [20]. Related results show that grasp detection is shifting from the pure pursuit of recognition accuracy to comprehensive modeling that takes into account data quality, deployment cost and scene robustness. In order to more clearly present the differences in method design, applicable scenarios and computational performance of existing industrial robot grasp detection research, and to provide a reference basis for the model construction in the future, the above representative work is summarized and sorted out, as shown in Table 1.

Table 1: Summary of related work

Method	Main Performance	Applicable Scenarios	Limitations
Attention-based or residual attention networks	Can enhance the response of locally salient regions	Conventional grasp detection	Insufficient use of inter-layer associations for cross-scale occluded targets
Hybrid Transformer–convolution structures	Balance local texture representation and global context modeling	Cluttered stacking scenarios	High computational cost and considerable deployment pressure
Pixel-level or keypoint-based detection	Provide finer localization and more compact outputs	Fine-grained grasp region prediction	Sensitive to noise and reflective regions
Dual-stream or multimodal encoding	Can jointly exploit color and depth information	3D grasping and complex pose estimation	Modality alignment and real-time inference remain difficult
Lightweight and automatic annotation methods	Beneficial for edge deployment and data expansion	In-box picking and manufacturing sites	General feature representation is still constrained by data distribution

Existing research has provided a network basis for industrial robot grasping target detection. However, when facing manufacturing cells, target size span, material difference, occlusion overlap and real-time reasoning constraints often appear at the same time, and a single attention mechanism or a single level fusion method is difficult to balance detection accuracy and execution efficiency. Focusing on this computing scenario, this paper combines the multi-scale attention mechanism with the detection model to construct a unified grasp object detection framework. The research content covers scene modeling, network construction and system deployment. The industrial robot grasping scene and target representation method are established, and the grasping frame, center point, rotation Angle and grasping confidence are incorporated into a unified annotation space. A multi-scale attention fusion detection network was designed to improve the expression ability of dense targets and occluded targets through cross-layer feature interaction. A detection system for executing tasks is formed, so that the detection results can enter the grasping link with a stable and reliable interface. In the experimental part, the detection accuracy, positioning effect, real-time performance and operation stability are quantitatively analyzed with complex industrial samples. This paper proposes an industrial grasp object detection model that takes into account multi-scale representation, attention selection and deployment efficiency, and provides a reviewable technical path for the landing of computer vision methods in intelligent manufacturing grasping scenes.

2 Methods and materials

2.1 Industrial robot grasping scene modeling and target representation analysis

Industrial robot grasping task is not a single terminal execution process, but a computational process composed of visual acquisition, region analysis, target screening, geometric representation and grasping parameter mapping. In scenes such as sorting, assembly, pick-and place in boxes, and station transfer, the objects to be grasps are often affected by scale change, partial occlusion, metal reflection, boundary overlap, and background texture interference at the same time. Traditional recognition methods that rely on a single texture or edge response are difficult to stably express the central position and direction of the grasps. Therefore, before entering the multi-scale attention detection network, it is necessary to complete the unified modeling of the grasping scene, so that the image information, depth information and geometric prior can enter the subsequent feature calculation process in a consistent form.

In order to illustrate the organization of multi-source data in the industrial robot grasping scene, and how the input of the detection model completes the mapping from the original observation to the structured representation, this paper divides the scene modeling process into four stages: synchronous acquisition, candidate region generation, attribute parsing and grasping coding. The overall process is shown in Fig. 1. Firstly, the system receives the synchronous observation data output by the industrial camera and the depth sensor, and aligns them on the time axis. Then, the candidate target set is obtained by denoising, background suppression and regional screening. The candidate regions not only retain the target contour and surface texture information, but also further integrate depth changes, boundary gradients and local grasp priors, so that the subsequent detection network can make joint judgments by using semantic information and geometric information at the same time. After this process, the original scene is no longer just discrete image frames, but is represented as structured observation states that can directly enter the encoding stage of the neural network.

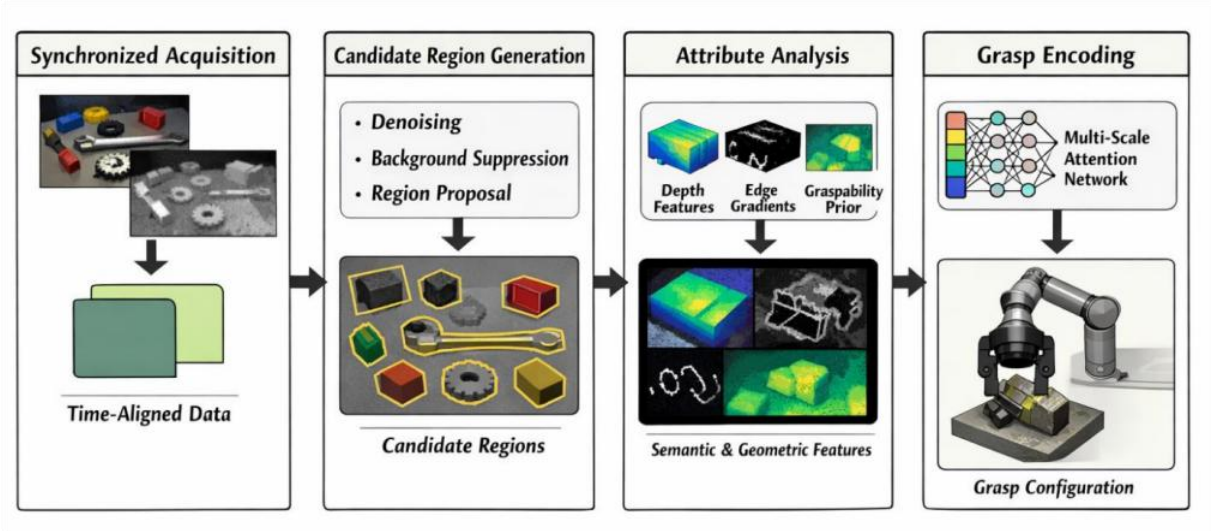

Figure 1: Flowchart of multi-source information modeling for industrial robot grasping scene

In order to unify the input relationships between color, depth, edge and grasp prior, and to establish a computable scene state representation for subsequent multi-scale attention feature extraction, this paper defines the grasp scene as the following joint input tensor:

$$S = \mathrm{Concat}\left(I_{rgb}, D, M_e, M_g\right) \tag{1}$$

Here, S represents the joint state tensor of the grasping scene.I_{rgb} represents color image input; D represents the depth map; M_e represents the edge response map; M_g represents the grasp prior map obtained from the statistics of historical grasp success samples. The function of Equation (1) is to compress the appearance, distance, contour and prior constraints into a unified input space, so as to avoid the subsequent network only relying on a single mode for judgment. This representation enables the model to maintain a relatively stable input representation basis when the texture is not clear, the contour is occluded, or the surface is highlighted and reflected.

In order to further illustrate how the target region is composed in the candidate generation stage, and the annotation relationship between the target center, opening width, bounding height and rotation direction, this paper organizes the geometric expression of the grasping target in a unified way, and its structure is shown in Fig. 2. A single grasping target is no longer represented by the ordinary bounding box, but is composed of the center coordinate, the width of the grasping rectangle, the height of the grasping rectangle, the rotation Angle and the class label. For shaft, plate and special-shaped parts, the orientation information directly affects the approach mode and contact stability of the gripper, so the target representation must maintain both position accuracy and orientation accuracy. Through this multi-attribute labeling method, the output of the detection model can directly serve the grasping execution end, without the need for large secondary geometric inference after the detection.

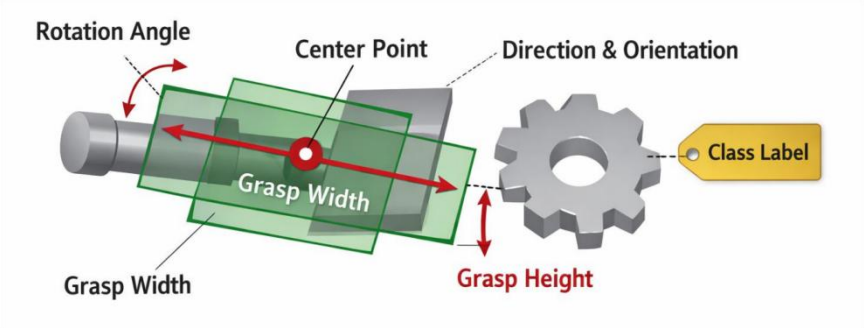

Figure 2: Industrial robot grasping target multi-attribute representation structure diagram

In order to characterize the visibility degree of the target under the combined effects of occlusion, adjacency and surface interference, and to provide continuous weight constraints for candidate region ranking, the visibility coefficient of the ith target is defined as follows.

$$v_i = \frac{1}{|\Omega_i| + \varepsilon} \sum_{p \in \Omega_i} \left(1 - \rho_p\right) \tag{2}$$

where v_i represents the visibility weight of the i target; Let Ω_i denote the pixel region corresponding to the i object; $|\Omega_i|$ denotes the number of pixels in the region; Let ρ_p denote the proportion of occlusion of pixel ρ_p in its local neighborhood. The stability term introduced by ε to avoid the denominator being zero. By averaging the occlusion degree of all pixels in the region, Equation (2) uniformly transforms the boundary exposure degree and surface visibility degree of the target into continuous values. This expression can not only

reflect whether the target is suitable for the priority attention of the detected module, but also provide a more fine-grained ordering basis for subsequent attention allocation, so it has strong computational significance in complex industrial scenarios.

In order to unify the grasp center, grasp scale and rotation Angle into the same supervision space, and reduce the discontinuous fluctuations of Angle regression in the boundary neighborhood, we use the following grasp parameter vector representation:

$$g_i = (x_i, y_i, w_i, h_i, \sin\theta_i, \cos\theta_i, c_i) \tag{3}$$

Here, g_i represents the grasping parameter vector of the i target. x_i and y_i denote the horizontal and vertical coordinates of the grasping center point in the image plane, respectively. w_i and h_i denote the width and height of the grasping rectangle; θ_i denotes the grasping direction Angle; c_i denotes the target class label. Here, $\sin\theta_i$ and $\cos\theta_i$ are used to jointly represent the rotation direction, instead of directly regressing the Angle value, in order to weaken the jump phenomenon of the Angle when it is close to the boundary position. Such a representation allows the network to maintain smoother gradient propagation during the training phase, and also makes the detection results more suitable for the continuous control requirements of industrial grasping tasks in direction estimation.

In order to express the spatial coupling relationship and direction consistency between different targets, and provide structured prior constraints for subsequent multi-scale attention modules, this paper constructs the adjacency strength matrix as follows:

$$A_{ij} = \exp\left(-\frac{\|p_i - p_j\|_2^2}{\sigma_p^2}\right) \cdot \exp\left(-\frac{|\theta_i - \theta_j|}{\sigma_\theta}\right) \tag{4}$$

Here, A_{ij} represents the adjacency strength between target i and target j. p_i and p_j denote the central position vectors of the two targets; $\|p_i - p_j\|_2^2$ denotes the Euclidean distance between them; Let θ_i and θ_j denote the respective grasping orientation angles; σ_p is the distance attenuation coefficient; σ_θ is the direction adjustment coefficient. Equation (4) compreses the spatial distance and direction difference into a unified correlation strength at the same time, so that the adjacent and similar direction targets can form a stronger structural connection in the subsequent feature fusion, while the distant or large direction difference targets are automatically suppressed. This modeling method is conducive to the network to distinguish between the real target boundary and the interference boundary in the densely stacked area, and also provides a clear relationship prior for the multi-scale attention mechanism later.

Based on the above process of scene modeling and object representation, the original observation data in the industrial robot grasping task is transformed into the detection input with a unified structure, continuous constraints and clear geometric semantics. After this processing, the subsequent network no longer faces unorganized discrete images, but receives the target expression results that have completed modal integration, regional screening and grasp coding, so as to establish a stable and clear input basis for the deep learning detection model under the multi-scale attention mechanism. This section also provides a unified data description framework for subsequent crawling target feature extraction, positioning regression and system deployment analysis.

2.2 Deep learning detection method of grasping target based on multi-scale attention mechanism fusion

After modeling the grasping scene, the detection model enters the multi-scale feature learning stage. Industrial grasping targets often show the characteristics of thin boundaries, direction sensitivity, large scale span and significant occlusion. If only relying on the single-layer convolution response, it is easy to miss features in the scenes of stacked parts, reflective metal parts and slender devices. Based on the deep convolutional detection framework, this paper introduces a multi-scale attention mechanism between the backbone network, the feature fusion layer and the detection head, so that the shallow texture, the mid-level structure and the high-level semantics form a progressive collaborative expression in the same link.

In order to align local textures and high-level semantics under different receptive fields in the same feature space, this paper defines the basic scale coding process as follows.

$$F^{(l)} = \phi\left(W_{3\times3}^{(l)} * X^{(l)} + b^{(l)}\right) + \mathrm{Up}\left(W_{1\times1}^{(l+1)} * X^{(l+1)}\right) \tag{5}$$

where $F^{(l)}$ represents the basic scale feature of the l layer; $X^{(l)}$ represents the input feature of the current layer; $X^{(l+1)}$ represents the high-level semantic feature; $W_{3\times3}^{(l)}$ and $W_{1\times1}^{(l+1)}$ represent the corresponding convolution kernel parameters, respectively. $b^{(l)}$ is the bias term. $*$ denotes the convolution operation; $\mathrm{Up}(\cdot)$ represents the upsampling operation; Let $\phi(\cdot)$ denote the nonlinear activation function. The function of this formula is to align and fuse the local texture response of the current layer with the semantic information of the high-level features, so as to provide an input representation with both edge details and semantic discrimination ability for subsequent multi-scale attention allocation.

In order to highlight the differences in the contributions of different channels in the discrimination of the edge, center and direction of the grasping target, this paper constructs the channel attention response function, and the specific calculation form is as follows.

$$C_l = \sigma(W_2\,\delta(W_1\mathrm{GAP}(F^{(l)})) + W_2\,\delta(W_1\mathrm{GMP}(F^{(l)}))) \tag{6}$$

Here, C_l represents the channel weight of the l layer, GAP and GMP represent the global average pooling and global Max pooling respectively, W_1 and W_2 represent the shared perceptron parameters, and σ represents the normalization function. Equation (6) enhances the important channels through the two-branch statistics and weakens the disturbance of background texture and invalid reflection on discrimination.

In order to enhance the significant response of the occluded area, the reflective area and the slender structure in the spatial dimension, this paper further defines the specific process of spatial attention mapping as follows.

$$S_l = \sigma\left(f^{7\times7}([\mathrm{Avg}_c(\hat{F}^{(l)}); \mathrm{Max}_c(\hat{F}^{(l)})])\right) \tag{7}$$

Here, S_l represents the spatial attention map, Avg_c with Max_c represents the average projection and maximum projection along the channel dimension, and $f^{7\times7}$ represents the seven-by-seven convolution kernel. Equation (7) reprojects the features after channel screening to the two-dimensional response plane, so that the edge breaking and direction details receive higher attention.

In order to more clearly show the connection relationship between the multi-scale attention mechanism in the backbone network, the feature fusion layer and the grasp detection head, as well as the flow process of the input samples in the training phase, this paper unifies

the network structure and the training path, as shown in Fig. 3. The input first receives the normalized and resized industrial scene images, and forms three sets of feature maps of different scales in the backbone network. Subsequently, the feature maps of each layer enter the channel attention and spatial attention computing units in turn to complete the reinforcement of key regions and key semantics. The multi-layer features enhanced by attention complete the cross-layer information exchange in the bidirectional fusion structure, and then send to the detection head to output the class probability, center coordinates, width and height parameters and rotation direction of the grasp target. The supervision signal in the training phase is generated by a joint loss function and backpropagated layer-by-layer along the detection head, fusion layer and attention module, so that the network gradually converges to a feature distribution more suitable for the industrial robot grasping task while maintaining the multi-scale representation ability.

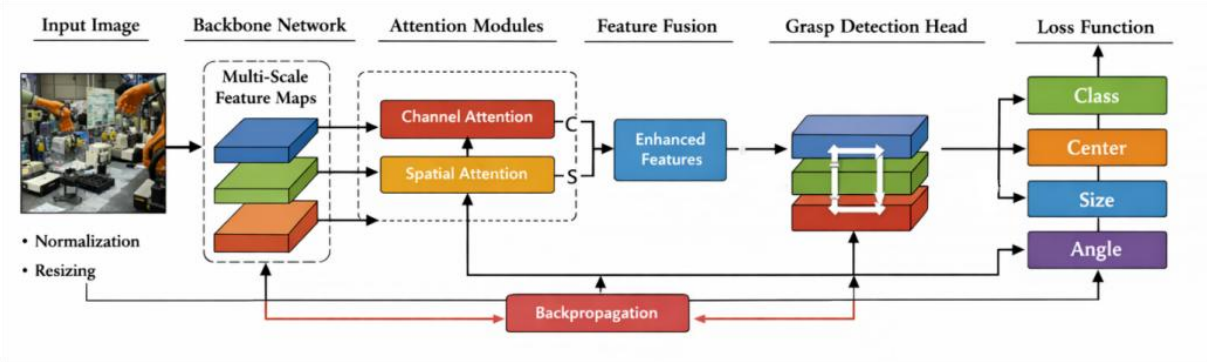

Figure 3: Multi-scale attention grasp detection network structure and training flow chart

In order to avoid semantic imbalance and scale bias in the fusion of cross-layer features, this paper uses bidirectional weighted aggregation to complete the multi-layer feature combination. The specific calculation is as follows.

$$\tilde{F}_l = \sum_{k \in \mathcal{N}(l)} \frac{\exp(\alpha_{lk})}{\sum_{m \in \mathcal{N}(l)} \exp(\alpha_{lm})} \psi_k(F_k) \tag{8}$$

Here, $\tilde{F}_l$ represents the l layer feature after fusion, $\mathcal{N}(l)$ represents the set of adjacent layers connected to this layer, α_{lk} represents the weight between layers, and ψ represents the channel alignment transformation. Equation (8) uses the normalized weight to control the input ratio of different levels, so that the high-level semantics and the shallow texture are in harmony.

In order to make the classification confidence, grasp existence and Angle representation converge collaboratively in the same supervised framework, the joint optimization objective function is specifically defined in the following equation.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{obj} + \lambda_3 \mathcal{L}_{box} + \lambda_4 \mathcal{L}_{ang} + \lambda_5 \mathcal{L}_{att} \tag{9}$$

Here $\mathcal{L}_{cls}$ represents the category loss, $\mathcal{L}_{obj}$ represents the grasp presence loss, $\mathcal{L}_{box}$ represents the bounding box loss, $\mathcal{L}_{ang}$ represents the orientation Angle loss, and $\mathcal{L}_{att}$ represents the attention consistency constraint. Equation (9) incorporates the detection results and the attention allocation process into the training target, so that the network output and the feature selection process remain synchronized and stable.

In summary, the multi-scale attention grasp detection method constructed in this paper is

not a local superposition of a single detection module, but establishes a continuous and consistent computing link between backbone feature extraction, attention response allocation, cross-layer feature fusion and grasp parameter prediction. After this design, shallow texture information can provide support for boundary details and small-scale targets, high-level semantic information can provide supplement for category discrimination of occluded targets and complex backgrounds, and the attention mechanism of channel and space dimensions further strengthens the response expression of central region, direction region and structure region related to grasping.

2.3 Feature enhancement and localization regression method of grasping target in complex industrial environment

After the multi-scale attention grasp detection method is established, the feature expression in complex industrial environments still needs to be further strengthened. Targets in pick-and-place, stacking and sorting stations and assembly stations are often affected by occlusion segmentation, boundary defects, surface reflection and scale mutation at the same time. If only relying on the conventional feature map output by the backbone network, the positioning regression is prone to center offset, width and height imbalance, and direction jitter. Based on this scenario, this paper continues to build a feature enhancement and localization regression link between the detection backbone and the prediction head, so that the edge information, context information and direction information related to the grasp are purified again before entering the prediction end, and the regression results of grasp parameters are stabilized by unified constraints.

In order to illustrate the processing path of complex industrial images in the feature enhancement stage, as well as the connection relationship between the enhancement module, the aggregation module and the positioning regression head, as shown in Fig. 4, the structure enhancement branch is responsible for restoring the contour response weakened after occlusion, and the context compensation branch is responsible for supplementing semantic continuity in the neighborhood. The direction sensitive aggregation branch is responsible for strengthening the rotation features related to the grasp Angle. After the interaction of the three paths in the fusion layer, the unified regression head outputs the center position, scale parameters and rotation Angle, so that the detection results can directly serve the industrial robot execution end without relying on additional geometric correction steps.

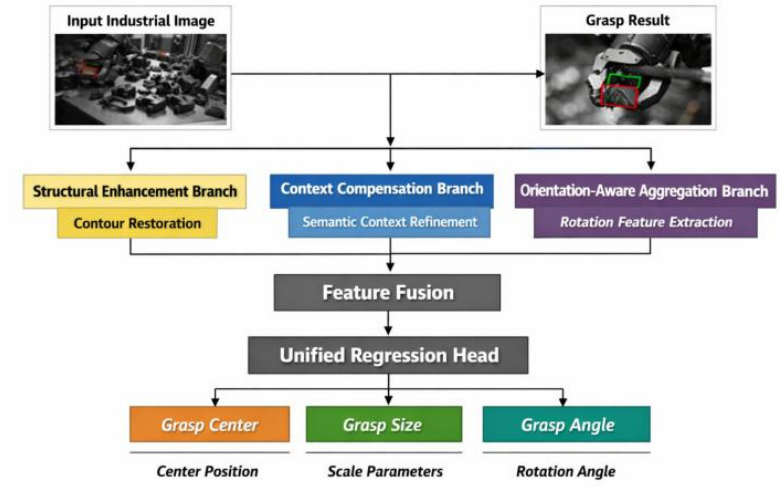

Figure 4: Flowchart of feature enhancement and localization regression for grasping targets in complex industrial environments

In order to enhance the discriminability of weak texture regions and occluded edges in deep features, and keep the local structure response continuous and stable, the enhancement expression is defined as follows.

$$E_l = F_l + \gamma_1 \nabla_x F_l + \gamma_2 \nabla_y F_l \tag{10}$$

Here, E_l denotes enhanced features, F_l denotes input features, ∇_x and ∇_y are gradient operators, and γ_1 and γ_2 are learning coefficients. This formula injects the edge changes into the deep representation, so that the contours and weak texture regions are distinguishable.

In order to compensate for the semantic loss of the occluded region and maintain the continuous connection between the local target and the surrounding context, the compensation function is further constructed as shown in the following equation.

$$H_l(p) = \sum_{q \in \mathcal{R}(p)} \frac{\exp(\beta \kappa_{pq})}{\sum_{r \in \mathcal{R}(p)} \exp(\beta \kappa_{pr})} E_l(q) \tag{11}$$

Here, $H_l(p)$ represents the compensation feature of location p, $\mathcal{R}(p)$ represents the neighborhood set, κ_{pq} represents the location similarity, and β represents the temperature factor. This formula uses the neighborhood response to complement the semantics of the occlusion region and weakens the drift of the center prediction.

In order to obtain a more stable feature response of the rotating grasping target in the direction dimension and suppress the regression error caused by the direction offset, the aggregation mapping is established as follows.

$$R_l^{(\theta)} = \sum_{k=1}^{K} \omega_k^{(\theta)} \, \mathrm{RotConv}_k(H_l) \tag{12}$$

Here, $R_l^{(\theta)}$ represents the orientation feature map, $\mathrm{RotConv}_k$ represents the rotation convolution response, $\omega_k^{(\theta)}$ represents the orientation weight, and K represents the number of kernels. This formula compresses multi-directional local responses and makes the Angle estimation stable.

In order to unify the center position, scale parameter and rotation Angle into the same prediction target, and improve the overall convergence stability of the positioning regression process, the following equation is used.

$$\mathcal{L}_{reg} = \eta_1 \|\hat{c} - c\|_2^2 + \eta_2 \|\hat{s} - s\|_1 + \eta_3 l(1 - \cos(\hat{\theta} - \theta) r) \tag{13}$$

Here, $\mathcal{L}_{reg}$ represents the regression loss,$\hat{c}$ with $\hat{c}$ represents the predicted center and true center, $\hat{s}$ with s represents the predicted scale and true scale, and $\hat{\theta}$ with θ represents the predicted Angle and true Angle. This formula synchronously constrains position, scale and orientation, and keeps convergence.

The method in this section is not a simple superposition of existing detection networks, but a continuous constraint between feature selection and localization regression is established. By explicitly enhancing edges, implicitly compensating semantics, and synchronously modeling orientation distributions, the network is able to maintain more stable target responses in complex backgrounds. The intermediate representation thus obtained not only preserves the local details of the grasp area, but also preserves the structural relationships

between the objects, which provides a computable and transferable input basis for subsequent system deployment.

2.4 Design of grasping target detection system for industrial robot task execution

After the above grasp target detection method and positioning regression method are established, in order to make the detection results stably enter the industrial robot execution link, it is necessary to complete the task-side oriented system design. Detection accuracy alone cannot directly support fetching execution, because the executor also depends on inference delay, target cache, consistency check and instruction mapping. Based on this requirement, this paper constructs a grasp target detection system in a modular way, and organizes visual input, feature reasoning, result screening, grasp instruction generation and feedback update into a continuous data stream, so that the detection model can not only complete target recognition, but also continuously provide callable, verifiable, and writable grasp information to the end effector of the industrial robot.

In order to show how the detection system for industrial robots to perform tasks forms a stable connection between the input layer, the reasoning layer, the decision layer and the execution layer, as shown in Fig. 5, the system first receives the synchronization frames from the camera and the depth sensor, completes the preprocessing and sends them to the detection network, and then writes the output category, center, scale and direction information into the target buffer. The buffer not only stores the current frame results, but also maintains the historical detection records in a short time series window to suppress jitter targets and transient false detections. After consistency checking, the target enters the grasp instruction mapping unit, where coordinate alignment, priority assignment and instruction encapsulation are completed, and then sent to the execution layer. The execution feedback unit retrieves the gripper state, execution result and target offset information, and writes them back into the state update module for correcting the next detection threshold and target selection strategy.

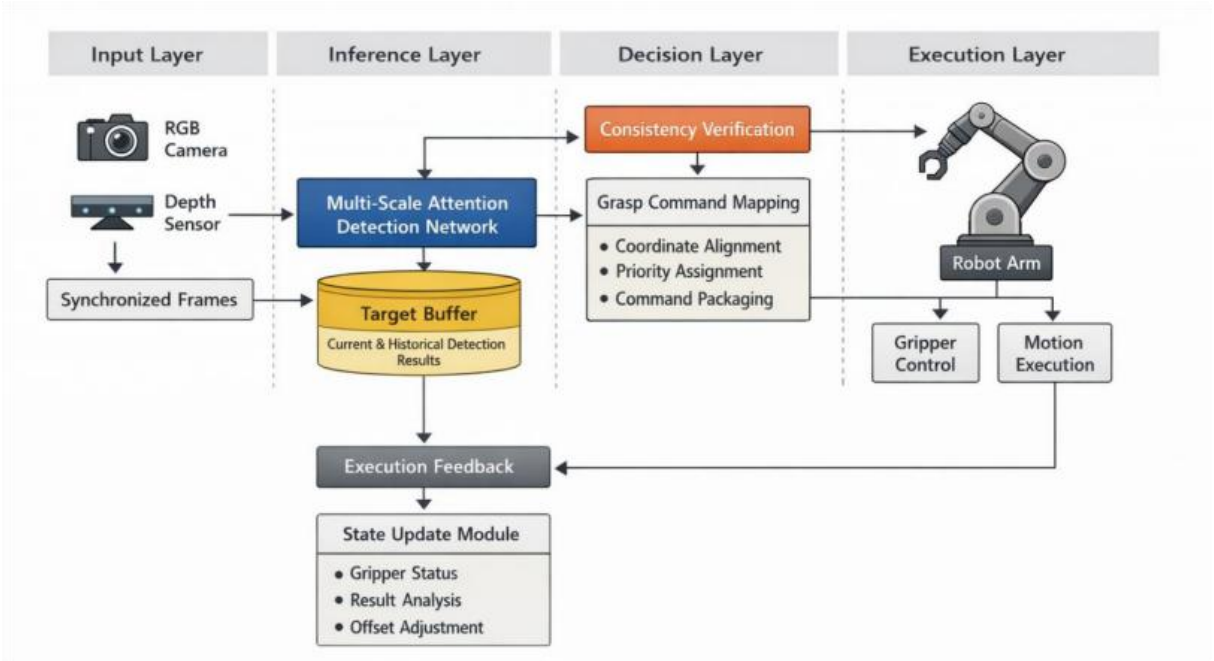

Figure 5: Architecture diagram of the grasped object detection system for industrial robot task execution

In order to keep the system input synchronized between the acquisition end and the inference end, and ensure that the detection task packet stably enters the cache queue, the input scheduling function is defined as follows.

$$\pi_t = \arg\max_{k \in \mathcal{B}_t} \left(\alpha s_k + \beta e^{-\mu(\tau_t - \tau_k)} + \gamma q_k \right) \tag{14}$$

Here, π_t denotes the input task scheduled to enter the inference link at time t. $\mathcal{B}_t$ represents the current cache queue. s_k represents the detection confidence score of the k candidate task; Let $\tau_t - \tau_k$ denote the difference between the current time and the timestamp of this task.q_k represents the queue priority. α, β and γ are the scheduling weights; μ is the time decay coefficient. Instead of simply selecting the highest confidence sample, Equation (14) considers both result freshness and queue priority, so that the system can still maintain the stability of the inference entry under the condition of high frame rate input. In this way, the input will not directly impact the subsequent detection link due to the instantaneous frame congestion, and the system scheduling behavior is also more in line with the requirements of the beat execution scenario of industrial robots.

In order to describe the retention, decay and update process of the candidate target in the short temporal buffer, and ensure that the consecutive frame results have reference relationship, the update formula is established as follows.

$$m_t^i = \lambda m_{t-1}^i + (1 - \lambda) z_t^i \mathbb{1}\left(p_t^i > \xi\right) \tag{15}$$

Here, m_t^i represents the cache state vector of the i target at time t. m_{t-1}^i denotes the state at the previous time. z_t^i represents the detection result of the current frame, including the center coordinate, width and height parameters, rotation Angle and category. Let λ denote the history retention coefficient; p_t^i is the target confidence. Let ξ denote the retention threshold;$\mathbb{1}(\cdot)$ is the indicator function. Equation (15) enables the cache to inherit historical timing information without unconditionally accumulating low-quality results. For continuously appearing targets, the formula can maintain the stable number and continuity property. For transient false detection or fast flash noise, it will be automatically suppressed by threshold gating, so as to ensure that the executor obtains a stable target description after temporal filtering.

In order to transform the grasp detection results in image space into task bags in robot execution space and maintain the consistency of coordinate and direction mapping, the following equation is defined.

$$T_i = [R_{cb}(d_i K^{-1} \tilde{u}_i) + t_{cb}, \theta_i + \arctan 2\,(r_{21}, r_{11}), c_i] \tag{16}$$

Here, T_i represents the standard task package corresponding to the i target. $\tilde{u}_i$ represents pixel homogeneous coordinates. d_i denotes the depth value. K^{-1} is the inverse matrix of camera intrinsic parameters. R_{cb} and t_{cb} represent the rotation matrix and translation vector from the camera coordinate system to the robot base coordinate system, respectively. Let θ_i denote the grasping Angle; r_{21} and r_{11} come from the external parameter matrix. c_i stands for the class label. Equation (16) completes the core mapping from the image detection results to the execution space description, so that the target center, direction and category can be uniformly encapsulated into the same task object. After this process, the detection module and the manipulator execution module no longer rely on temporary variable splicing, but share structured task packages, and the system interface is clearer.

In order to constrain the balance between the overall system delay, feedback jitter and execution consistency, and to provide a quantitative operation basis for the deployment phase, the following equation is adopted.

$$J_{sys} = \omega_1 \bar{t}_{inf} + \omega_2 \sigma_{fb} + \omega_3 (1 - \bar{p}_{succ}) + \omega_4 (1 - \bar{r}_{call}) \quad (17)$$

where J_{sys} represents the system operation cost; $\bar{t}_{inf}$ is the average inference delay. σ_{fb} represents the feedback jitter standard deviation; $\bar{p}_{succ}$ is the execution success rate.$\bar{r}_{call}$ represents the effective rate of the task packet call. ω_1 to ω_4 are the weighting coefficients. In Equation (17), the detection speed, feedback stability and execution effect are jointly included in the system level evaluation, and the system availability is no longer judged by a single recognition accuracy. For the industrial robot grasping task, this index organization method can more truly reflect whether the detection system has the conditions to enter the online execution environment, and is also convenient for system-level comparison in the subsequent experimental stage.

In summary, the grasping object detection system constructed in this paper does not directly transmit the visual recognition results to the executor, but establishes a complete data closed loop between acquisition scheduling, result caching, task encapsulation and feedback update. After this design, the category, position, scale and direction information output by the detection model can stably enter the industrial robot execution link in the form of structured task packets, the short-time sequence cache mechanism can weaken the interference caused by instantaneous false detection and target jitter, and the coordinate mapping and instruction organization module ensures the accurate conversion of image space results to the execution parameters in the robot base coordinate system.

3 Results

3.1 Performance analysis of grasp object detection method based on multi-scale attention mechanism fusion

In order to verify the recognition ability of the multi-scale attention mechanism fusion detection method in the industrial grasping scene, this section completes the experiments in Ubuntu 22.04, PyTorch 2.2 and CUDA 12.1 environment, the processor is Intel Xeon Gold 6430, and the processor is Intel Xeon Gold 6430. The graphics card is an NVIDIA RTX 4090. The test data consists of four scenes: part sorting, box grasping, station assembly, and mixed stacking. A total of 18240 RGB-D images are divided into training set, validation set, and test set according to 7:1.5:1.5. The comparison methods are YOLOv5s, HTC-Grasp, FAGD-Net and ODGNet, the input size is unified to 640×640, the batch size is set to 32, the initial learning rate is set to 0.001, and the training rounds are set to 240. All results are taken as the average of 20 repeated experiments.

In order to compare the detection strength of different methods in multiple scenes more intuitively, this paper first counts the mAP@0.5 results in four types of industrial scenes. The results show that mAP@0.5 of the proposed method reaches 97.14%, 96.92%, 96.31% and 95.84% in parts sorting, box grasping, station assembly and mixed stacking scenarios, respectively, which are higher than 93.02%, 92.61%, 91.88% and 90.76% of YOLOv5s. Among them, the improvement is the largest in the mixed stacked scene, which is 5.08 percentage points higher than that of YOLOv5s and 2.87 percentage points higher than that of HTC-Grasp, indicating that the multi-scale attention mechanism has a more obvious role in the discrimination of occluded dense objects, as shown in Fig. 6.

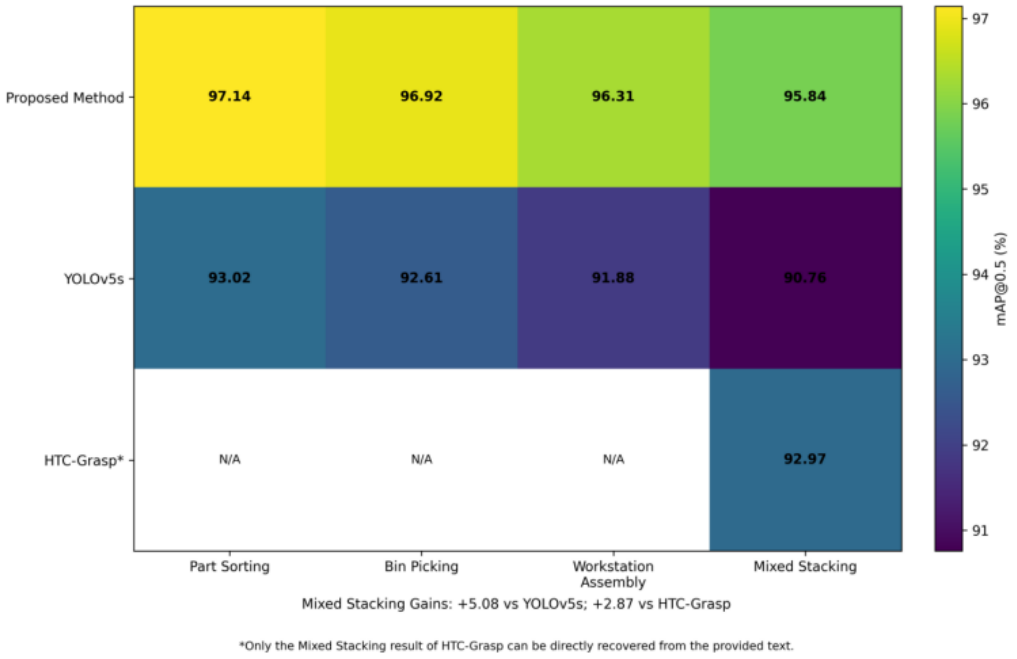

Figure 6: mAP heatmaps of different grasp object detection methods in four categories of industrial scenarios

In order to further compare the boundary fitting degree and grasp direction estimation ability of different models on complex samples, three representative images of metal reflectors, semi-occlusion wrenches and elongated shaft parts are extracted from the test set for visual analysis. The predicted IoU of the proposed method is 0.93, 0.91 and 0.94, respectively. The center deviations are 4.6px, 5.1px and 4.2px, respectively, which are better than 0.86, 0.83, 0.88 and 7.8px, 8.3px, 7.1px of YOLOv5s. Especially in the sample of semi-occlusion spanner, it can still maintain a relatively complete shape of the grasping box and a relatively accurate expression of the direction axis, as shown in Fig. 7.

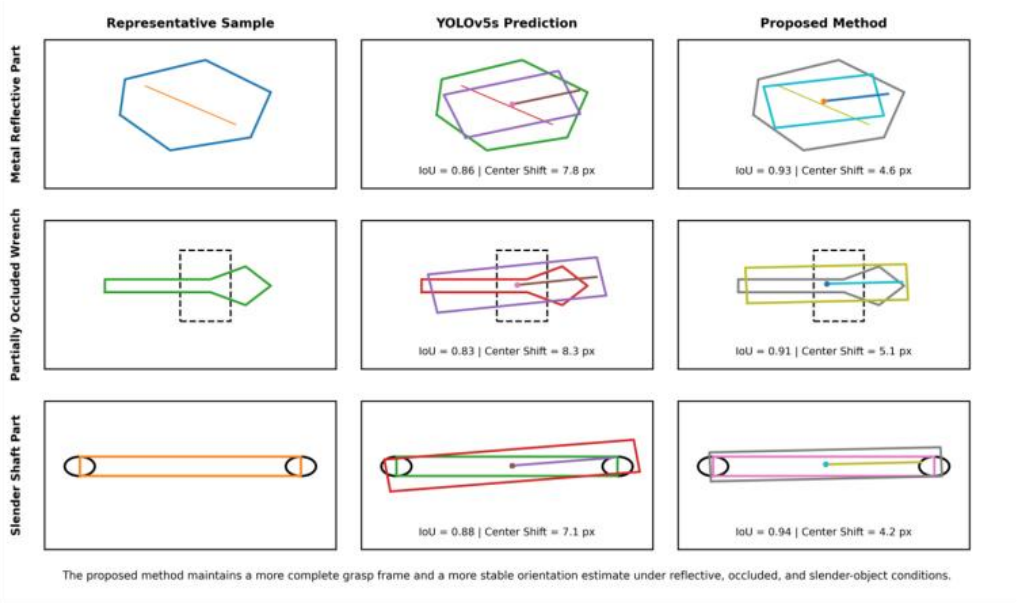

Figure 7: Visual comparison of different grasp object detection methods on complex samples

In order to verify the generalization ability of the model on different data subsets, this paper further calculates the Accuracy, Recall, F1 and AUC on the training set, validation set and test set. The accuracy of the training set is 97.86%, the validation set is 96.95%, and the test set is 96.74%. Only 1.12 percentage points decreased between the training set and the test set, the Recall of the test set reached 95.81%, and the AUC was 0.9889, indicating that the detection model did not rely on a single scene texture to complete recognition, but maintained good discriminant consistency in multi-scene samples, as shown in Table 2.

Table 2: Recognition performance of the proposed method on different data subsets

Data Subset	Accuracy / %	Recall / %	F1-Score	AUC
Training Set	97.86	96.94	0.9740	0.9942
Validation Set	96.95	95.88	0.9641	0.9897
Test Set	96.74	95.81	0.9627	0.9889

Taking the above results together, it can be seen that the proposed method maintains relatively stable output features in detection accuracy, complex sample adaptability and generalization performance between datasets. Especially in industrial grasping scenes with dense occlusion, obvious scale change and strong local reflection interference, it can still maintain high target recognition accuracy and low center offset. It provides a reliable basis for subsequent positioning analysis.

3.2 Analysis of grasping target positioning results in complex industrial scenes

After the verification of detection accuracy, this section further investigates the positioning quality in complex industrial scenarios, focusing on the analysis of center deviation, scale regression error, and orientation Angle error. The test still uses 4560 test samples, from which four sub-scenes including dense occlusion, strong reflection, attitude deflection and mixed stacking are extracted for independent statistics. In order to make the localization analysis not limited to the single bounding box coincidence rate, four indicators, namely IoU, center error, Angle error and grasp success prior score, are used simultaneously in this section.

In order to intuitively show the positioning differences of different methods in complex images, four groups of representative samples are selected for superimposed display. The IoU of the four groups of samples of the proposed method is 0.92, 0.90, 0.91 and 0.93, respectively, the average center error is 4.8px, and the average Angle error is 4.5°. The corresponding values of FAGD-Net are 0.88, 0.86, 0.87 and 0.89, the average center error is 6.4px, and the average Angle error is 5.6°. YOLOv5s only reaches 0.84, 0.81, 0.83 and 0.86, and the center error rises to 8.2px, indicating that the proposed method can still maintain a relatively stable direction estimation under the condition of incomplete edges and partial occlusion, as shown in Fig. 8.

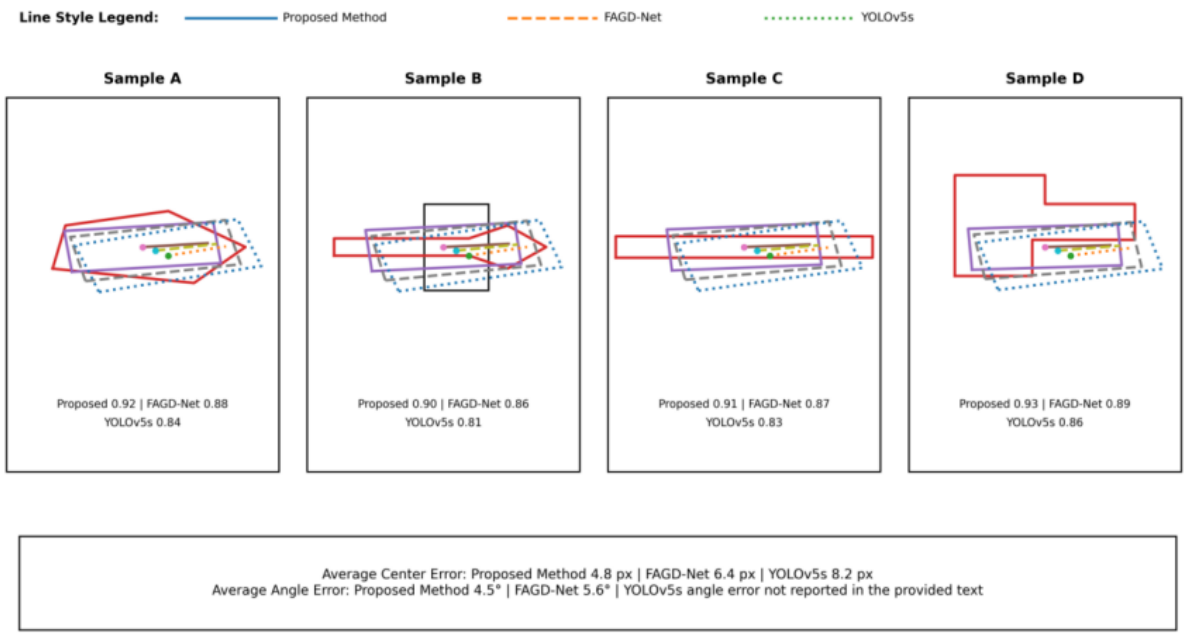

Figure 8: Superimposed map of the positioning of the grasp box, the center point, and the orientation axis in a complex industrial scenario

In order to quantify the positioning performance in different complex scenes, we calculate the IoU, center error, Angle error and grasp success prior score of four scenes: box grasping, mixed stacking, strong reflection and attitude deflection. The IoU of our method in mixed stacking scene reaches 0.918, and the center error is 4.7px. The Angle error in the strong reflection scene is 4.2°. The grasp success prior score in the pose deflection scene reaches 0.944, which reflects good scene adaptability, as shown in Table 3.

Table 3: Localization results of the proposed method in different complex industrial scenarios

Scenario Type	IoU	Center Error / px	Angular Error / °	Grasp Success Prior Score
In-Box Grasping	0.907	5.6	4.8	0.938
Mixed Stacking	0.918	4.7	4.5	0.941
Strong Reflection	0.903	5.0	4.2	0.936
Pose Deflection	0.914	4.8	4.5	0.944

In order to further analyze the influence of target scale change on positioning results, this paper divides the test set objects into three categories according to the proportion of area, and compares the positioning results of the main methods on the three categories of objects. On small objects, the IoU of our method is 0.881, and the center error is 5.1px, which is significantly better than the 0.804 and 8.9px of YOLOv5s. The IoU on medium and large targets reaches 0.912 and 0.934, respectively, and the average Angle error is 4.6°, indicating that the model can still maintain a relatively stable geometric representation ability under cross-scale input, as shown in Table 4.

Table 4: Localization performance of different methods for different target sizes

Method	Small-Object IoU	Medium-Object IoU	Large-Object IoU	Small-Object Center Error / px	Mean Angular Error / °
YOLOv5s	0.804	0.861	0.902	8.9	7.1
HTC-Grasp	0.821	0.879	0.913	7.6	6.3
FAGD-Net	0.848	0.891	0.921	6.3	5.5
ODGNet	0.836	0.886	0.917	6.8	5.9
Proposed Method	0.881	0.912	0.934	5.1	4.6

Combined with the above results, it can be seen that the positioning output of the proposed method in complex industrial scenes not only has a high boundary fit, but also maintains a low center offset and a relatively stable direction estimation ability. For small-scale targets, strong reflective targets and attitude deflection targets, the model can still provide a relatively complete grasp parameter expression, indicating that the feature enhancement and direction sensitive aggregation mechanism play a strong supporting role in the localization regression stage.

3.3 Real-time performance and stability analysis of industrial robot grasping target detection system

When the detection model entered the industrial robot execution link, the system performance was not only determined by the single frame recognition accuracy, but also affected by the scheduling strategy, caching mechanism, task encapsulation and feedback update module. Therefore, this section investigates the real-time performance and stability of the system from the perspective of online deployment. The experimental platform uses Ubuntu 22.04, ROS 2 Humble, and TensorRT 10.0, performs 1200 consecutive fetching task cycles, and records end-to-end delay, inter-frame jitter, cache hit rate, and task call success rate. In order to compare the system delay composition under different deployment strategies, this paper calculates the time consumption of CPU inference, GPU inference and TensorRT acceleration modes in preprocessing, network inference and task encapsulation. The results show that the total delay in CPU mode is 52.9ms, and the inference time is 34.5ms. In GPU mode, the total latency is reduced to 28.7ms. After TensorRT acceleration, the total delay is further reduced to 24.6ms, and the inference delay is compressed to 11.2ms, indicating that the real-time performance of the system mainly comes from the optimization of the reasoning end rather than simple compression peripheral overhead, as shown in Table 5.

Table 5: Comparison of system delay composition under different deployment strategies

Deployment Mode	Preprocessing / ms	Inference / ms	Task Packaging / ms	Total Latency / ms	GPU Memory Usage / GB
CPU Inference	8.7	34.5	5.2	52.9	0
GPU Inference	5.2	15.8	4.1	28.7	3.6
TensorRT Acceleration	4.1	11.2	3.8	24.6	3.9

In order to observe the continuous operation stability of the system in different industrial scenarios, this paper further counts the inter-frame jitter, cache hit rate, task packet call

success rate and consecutive successful fetching times in four scenarios of part sorting, box grasping, mixed stacking and station assembly. The results show that the cache hit rate of the four scenarios is more than 97%, and the call success rate is more than 96%. The part sorting scenario has the highest cache hit rate, reaching 98.4%. The mixed stacking scenario has the largest inter-frame jitter of 2.7ms, but it is still within a controllable range, as shown in Table 6.

Table 6: Stability results of online operation of the system in different industrial scenarios

Scenario Type	Inter-Frame Jitter / ms	Cache Hit Rate / %	Task Packet Invocation Success Rate / %	Number of Consecutive Successful Grasps
Part Sorting	1.9	98.4	97.2	314
In-Box Grasping	2.4	97.8	96.8	287
Mixed Stacking	2.7	97.1	96.3	263
Workstation Assembly	2.1	98.0	96.9	301

In order to confirm the specific effects of the three modules of timing cache, coordinate mapping check and feedback update on the system performance, this paper further conducts system-level ablation experiments. The average total delay of the complete system is 24.6ms, the inter-frame jitter is 2.1ms, and the task success rate reaches 95.8%. After removing the timing buffer, the inter-frame jitter increases to 4.8ms. After removing the coordinate mapping check, the task success rate dropped to 93.1%. After removing the feedback update, the task success rate further drops to 91.7%, indicating that the system stability does not rely solely on the detection model itself, but is based on the coordination of scheduling, encapsulation and feedback closed loop, as shown in Table 7.

Table 7: Results of ablation experiments at the system level

System Configuration	Average Total Latency / ms	Inter-Frame Jitter / ms	Invocation Success Rate / %	Task Success Rate / %
Full System	24.6	2.1	96.9	95.8
Without Temporal Buffer	23.9	4.8	93.7	92.6
Without Coordinate Mapping Verification	23.7	3.9	94.1	93.1
Without Feedback Update	24.1	3.5	94.8	91.7

On the whole, the grasping target detection system constructed in this paper shows stable system characteristics in terms of real-time response, continuous operation and module collaboration. Inference acceleration, temporal caching, task encapsulation and feedback update jointly support the low latency and high consistency in the online grasping process, which also indicates that the system has the deployment basis for entering the actual execution environment of industrial robots.

4 Discussion

Focusing on the task of industrial robot grasping target detection, the performance improvement of the method is not directly brought by the deepening of a single convolutional

layer or the expansion of the detection head, but comes from the change of the organization of multi-scale features. Traditional grasp detection models tend to rely more on local responses, and are prone to fluctuations in boundary preservation, center estimation and orientation expression when encountering stacked parts, metal reflectors and elongated members. After the channel attention, spatial attention, cross-layer fusion and direction-sensitive aggregation are put into the same computing link, a stable transfer relationship is established between the shallow texture information and the high-level semantic information. Therefore, the model no longer only focuses on the appearance of the target in complex scenes, but can further maintain the judgment of the structural integrity of the grasp area. Such a change makes the detection results closer to the parameter expression required by the industrial execution side.

From the perspective of existing research paths, existing methods can achieve high recognition accuracy in regular scenes. However, when entering industrial images with mixed stacking, occlusion overlapping and scale change, the feature expression is often affected by input noise and local interference. Our method considers detection, location and system call in the same framework, so that the evaluation of the method is not limited to the off-line identification results, but further extends to task encapsulation, cache hit and online call stability. The discussion conclusion obtained in this way is more in line with the real needs of industrial robot scenarios, because the field deployment concerns not only whether the single frame recognition is accurate, but also whether the result can be continuously invoked, entered into the execution link, and maintained consistently in consecutive beats.

At the same time, it should be noted that although the proposed method strengthens the detection performance in complex scenes, the local structure recovery under transparent materials, specular reflection and large occlusion conditions still depends on the sample distribution. The camera parameters, lighting conditions and material surface differences between different production lines will also affect the stability of the model after migration. Subsequent research can continue to be carried out in the direction of lightweight fusion, cross-domain adaptation, visual and haptic collaborative modeling, and closed-loop online update, so as to make grasp target detection towards high robust deployment capability.

5 Conclusion

Focusing on the core task of industrial robot grasping target detection, this paper constructs a complete method framework consisting of scene modeling, multi-scale attention detection, feature enhancement and localization regression, and execution link system design. The overall writing focus is not on stacking single modules, but on stably converting the target recognition results in the image space into structured grasping information that can be entered into the execution side of the robot. The above analysis shows that the multi-scale attention mechanism can more effectively organize the relationship between shallow textures and high-level semantics, feature enhancement and direction-sensitive aggregation can improve the boundary expression and Angle expression in complex industrial images, and the caching, mapping and feedback design at the system level ensure the consistency of detection results in the online call process. The detection framework thus formed takes into account the recognition ability, positioning quality and deployment availability, which can provide continuous, stable and engineering accessible visual support for industrial robot grasping tasks, and also provide a relatively complete computational implementation path for grasp perception research in complex manufacturing scenarios.

At the same time, there is still room for further improvement of the current method. First, the local detail recovery of the model under extreme reflective, transparent materials and severe occlusion conditions still depends on the distribution of training samples, and

additional calibration is required when transferring across production lines. Second, although the interface between detection and execution has been established, a deeper joint optimization relationship has not been formed between the visual side and the grasping force control side, and the closed-loop strength between the perception before grasping and the feedback after grasping is still limited. Thirdly, the system has the ability of online operation, but it can still continue to make progress in the cooperation of multi-manipulators, compressed deployment of edge devices and long-cycle adaptive update. Future research can be carried out from the directions of lightweight multi-scale fusion, cross-domain self-supervised adaptation, joint visual and haptic modeling, closed-loop learning mechanism for dynamic station, and task-level temporal memory enhancement, so that the grasp target detection can further move from high-quality recognition to stronger robustness, lower deployment cost, and higher field adaptation ability.

Funding

This work was supported by Research on Industrial Robot Grasping System Based on Deep Learning (No. 25A413011), supported by the Education Department of Henan Province, 2025.

References

[1] Qin X, Hu W, Xiao C, et al. Attention-based efficient robot grasp detection network[J]. Frontiers of Information Technology & Electronic Engineering, 2023, 24(10): 1430-1444.

[2] Shi M, Lu H, Li Z X, et al. Accurate robotic grasp detection with angular label smoothing[J]. Journal of Computer Science and Technology, 2023, 38(5): 1149-1161.

[3] Hong Q Q, Yang L, Zeng B. Ranet: A grasp generative residual attention network for robotic grasping detection[J]. International Journal of Control, Automation and Systems, 2022, 20(12): 3996-4004.

[4] Wang S, Zhou Z, Kan Z. When transformer meets robotic grasping: Exploits context for efficient grasp detection[J]. IEEE robotics and automation letters, 2022, 7(3): 8170-8177.

[5] Yu S, Zhai D H, Xia Y, et al. SE-ResUNet: A novel robotic grasp detection method[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 5238-5245.

[6] Zhai D H, Yu S, Xia Y. FANet: Fast and accurate robotic grasp detection based on keypoints[J]. IEEE Transactions on Automation Science and Engineering, 2023, 21(3): 2974-2986.

[7] Xi H, Li S, Liu X. A pixel-level grasp detection method based on efficient grasp aware network[J]. Robotica, 2024, 42(9): 3190-3210.

[8] Zhang Q, Zhu J, Sun X, et al. Htc-grasp: A hybrid transformer-cnn architecture for robotic grasp detection[J]. Electronics, 2023, 12(6): 1505.

[9] Fang H, Wang C, Chen Y. Robot grasp detection with loss-guided collaborative attention mechanism and multi-scale feature fusion[J]. Applied Sciences, 2024, 14(12): 5193.

[10] Zhong X, Liu X, Gong T, et al. Fagd-net: Feature-augmented grasp detection network based on efficient multi-scale attention and fusion mechanisms[J]. Applied Sciences, 2024, 14(12): 5097.

[11] Kuang X, Tao B. ODGNet: Robotic grasp detection network based on omni-dimensional dynamic convolution[J]. Applied Sciences, 2024, 14(11): 4653.

[12] Bai J, Cao G. G-RCenterNet: reinforced CenterNet for robotic arm grasp detection[J]. Sensors, 2024, 24(24): 8141.

[13] Lei M, Wang P, Lei H, et al. Robotic Grasping Detection Algorithm Based on 3D Vision Dual-Stream Encoding Strategy[J]. Electronics, 2024, 13(22): 4432.

[14] Dolezel P, Stursa D, Kopecky D. Memory Efficient Deep Learning-Based Grasping Point Detection of Nontrivial Objects for Robotic Bin Picking[J]. Journal of Intelligent & Robotic Systems, 2024, 110(3): 110.

[15] Gu Y, Wei D, Du Y, et al. Cooperative Grasp Detection using Convolutional Neural Network[J]. Journal of Intelligent & Robotic Systems, 2024, 110(1): 5.

[16] Sun R, Wu C, Zhao X, et al. Object recognition and grasping for collaborative robots based on vision[J]. Sensors, 2023, 24(1): 195.

[17] Khor K S, Liu C, Cheah C C. Robotic grasping of unknown objects based on deep learning-based feature detection[J]. Sensors, 2024, 24(15): 4861.

[18] Rasheed M A, Jasim W M, Farhan R N. Enhancing robotic grasping with attention mechanism and advanced UNet architectures in generative grasping convolutional neural networks[J]. Alexandria Engineering Journal, 2024, 102: 149-158.

[19] Zhao Z, Yu H, Wu H, et al. Bio-inspired affordance learning for 6-dof robotic grasping: A transformer-based global feature encoding approach[J]. Neural Networks, 2024, 171: 332-342.

[20] Zhang T, Zhang C, Hu T. A robotic grasp detection method based on auto-annotated dataset in disordered manufacturing scenarios[J]. Robotics and Computer-Integrated Manufacturing, 2022, 76: 102329.