Optimization Design of Robot Grasping Based on Lightweight YOLOv6 and Multidimensional Attention

Junyan Niu*, Guanfang Liu
Henan College of Transportation, Zhengzhou 450000, China

***Abstract*—To address the computational redundancy and robustness limitations of industrial grasping models in complex environments, this study proposes a lightweight capture detection framework integrating Mobile Vision Transformer (MobileViT) and You Only Look Once version 6 (YOLOv6). Three innovations are developed: 1) A cascaded architecture fusing convolution and Transformer to compress parameters; 2) A multidimensional attention mechanism combining channel-pixel dual enhancement; 3) A Pixel Shuffle-Receptive Field Block (PixShuffle-RFB) decoder enabling sub-pixel localization. Experiments demonstrate that the model achieves 0.88 detection accuracy with 66 Frames Per Second (FPS) in simulations and 90.04% grasping success rate in physical tests. The lightweight design reduces computational costs by 37% versus conventional models while maintaining 93.54% segmentation efficiency (2.85 milliseconds inference). This multidimensional attention-driven approach effectively improves industrial robot adaptability, advancing capture detection applications in high-noise manufacturing scenarios.**

***Keywords*—Capture detection; YOLOv6; multidimensional attention; MobileViT; industrial robot; lightweight**

I. Introduction

With the rapid growth of demand for industrial production automation, robots have become a key force driving productivity leaps and factory automation. Robot grasping detection technology combines machine vision with robots, and can improve object recognition and grasping efficiency on the production line through intelligent algorithms. According to the differences in grasping algorithm logic, robot grasping detection can be broken into rule-based grasping design and learning-based grasping design [1]. Rule-based grasping detection utilizes geometric models and physical properties to determine the optimal grasping point by analyzing object shape and force closure conditions [2]. Learning-based grasping detection relies on a large amount of data training to automatically learn object features and grasping strategies, adapting to unknown objects and complex environments [3]. However, with the increasingly complex production environment and processing tasks, traditional robot grasping and detection methods are no longer able to meet practical needs. For example, support vector machines have low efficiency in processing large-scale data and poor detection accuracy for complex shaped objects [4]. The random forest decision tree model is too large, resulting in poor real-time performance and making it difficult to deploy applications on embedded devices [5]. Gaussian mixture models are sensitive to initial parameters, have long training times, and are difficult to quickly adapt to environmental changes [6]. These issues seriously affect the accuracy and real-time performance of robot grasping. Therefore, the current industrial grasp detection faces a dual challenge: 1) the traditional model has insufficient feature discriminative power under complex background interference, leading to the accumulation of localization bias; 2) there is a significant contradiction between real-time detection demand and model computational load, and it is difficult for the existing methods to balance accuracy and efficiency. This restricts the ability of automated production lines to efficiently process shaped workpieces, and there is an urgent need to establish a new paradigm of lightweight and highly robust gripping detection.

In response to the above challenges, starting from the effective acquisition of object position and optimization of grasping pose, the research focuses on the basic logic and problems of single-stage real-time object detection algorithm You Only Look Once Version 6 (YOLOv6) and Multi-Dimensional Attention Fusion Network (MDAFN) modules, and improves them by proposing a Lightweight YOLOv6 with MDAFN for Robotic Grasping Detection (L-YOLOv6-MA). The research aims to: 1) establish a lightweight feature extraction framework to solve the contradiction between real-time performance and accuracy of traditional models; 2) strengthen the feature discrimination ability for the complex texture interference problem; 3) realize sub-pixel level grasping bit-position estimation to provide an end-to-end solution with both high accuracy and low latency for dynamic industrial scenes. The innovations of the research are: 1) establishing Mobile Vision Transformer (MobileViT) and YOLOv6 hybrid architecture, realizing the complementary advantages of MobileViT and YOLOv6; 2) designing the channel-space dual-domain attention mechanism to enhance the physical-semantic correlation of feature representations; 3) developing a multi-scale receptive field decoder to overcome the problem of dynamic balance between the local features of the grasping point and the global context information, and providing a solution for industrial inspection and detection. The research is structured into four sections. The first section introduces the current research on the logic and algorithms of robot grasping detection worldwide. The second section starts from modules such as YOLOv6 and MDAFN to establish a precise and real-time robot grasping detection model. The third section provides numerical examples and practical application analysis of the proposed algorithm model to verify its reliability. The final

*Corresponding Author

section provides a comprehensive summary and analysis of the article.

II. Related Works

With the quick growth of information technology and the scaling up of various industries, the application of robots in fields such as workshop transportation and assembly line processing is showing a rapidly increasing trend. Robot grasping detection is the core of achieving factory automation and fine operations, and it is also an important application direction that smart industry needs to continuously expand and deepen. However, in practical work, the performance of robot grasping detection for complex tasks is not stable, so many researchers are improving this problem. Wang S et al. raised a Transformer-based robot vision grasping model for object feature capture and long-range dependency modeling. By combining local window attention mechanism to obtain local contextual information, the model could simultaneously handle local information and long-range visual concept relationships in complex scenes [7]. In response to the demand for grasping posture and quality evaluation in robot grasping tasks, Yu S et al. proposed a novel squeeze excitation residual U-shaped network, which combines residual blocks with channel attention mechanism to generate grasping postures and predict the quality score of each posture, improving grasping accuracy and time efficiency [8]. To address the issues of accurate and reliable estimation of grasping posture for complex shaped objects, Cheng H et al. designed a vision-based depth grasping detector, which uses a densely connected feature pyramid network and multiple two-stage detection units to achieve dense grasping posture, achieving accurate grasping posture detection and gripper opening measurement [9]. Jiang J et al. proposed a new framework for visually-guided tactile detection to solve the problem of robots grasping transparent objects. The segmentation network was utilized to predict the horizontal upper region on the transparent object as the detection area, which is detected by a high-resolution haptic sensor to obtain the precise contour, improving the detection accuracy and grasping success rate of the transparent object [10]. Aiming at the problem that industrial robotic arms lack high-precision visual recognition ability, Wu Y proposed a visual recognition optimization method based on neural network and Transformer model. By combining the feature extraction ability of the deep learning model and the attention mechanism, the object recognition and grasping localization accuracy of the robot could be improved, and the autonomous operation ability and adaptability of the robotic arm in industrial scenes could be enhanced [11].

In addition, for the problem of robot grasping pose estimation for complex objects in unstructured environments, Cheng H et al. proposed a novel depth model for anchorless fully convolutional grasping pose detection. The grasping pose was considered as a rotating bounding box on the image plane, and the six-channel image was directly output to represent the key points and geometric information of the grasping rectangle, which improved the accuracy and efficiency of the grasping detection [12]. Regarding the problem of robot grasping in chaotic scenes, Yu S et al. proposed a chaotic grasping network, which used a dual branch squeezing excitation residual network as the skeleton, utilized multi-scale features and refined the grasping area to improve the success rate of robot grasping in chaotic scene tasks [13]. Cao H et al. raised a novel grasping detection network to balance the accuracy and inference speed of deep learning models in general object grasping detection. The network used a grasping representation method based on Gaussian kernel to highlight the center point with the highest grasping confidence. By suppressing noise features and highlighting object features, the network improved the grasping success rate while ensuring the model running speed [14]. To solve the problem of significant object detection in robot visual perception under complex interference environment, Song K et al. raised a novel three mode image fusion strategy. By constructing an image acquisition system under variable lighting scenes, and using multi-level weighting to suppress interference, effective cross modal feature fusion was achieved, enabling the robot to quickly and accurately complete the target capture task [15]. Aiming at the problems of limited 2D grasping direction and poor real-time performance of 3D point cloud, Hui N M et al. proposed a grasping detection algorithm that fuses 2D image and 3D point cloud. An improved single-stage multi-frame detector network is used to optimize the a priori frame scaling strategy to improve the target localization accuracy, and the target spatial position is extracted by the view cone transformation and point cloud segmentation algorithms, which improves the success rate and real-time performance of the capture, and reduces the time-consumption of the capture at the same time [16]. Aiming at the problem of differentiating color, shape and size for object sorting in industrial automation, Abdullah-Al-Noman M et al. proposed a robotic arm gripping system based on computer vision. Using PixyCMU camera and OpenCV image processing technology, combined with Arduino Mega controller and servo motor drive, the system realized multi-color object recognition and geometric feature detection. The system improved the color recognition accuracy and shape classification accuracy [17].

In summary, numerous researchers worldwide have noticed the problems that exist in robot grasping detection during operation and have conducted multiple research efforts to address these issues. However, the existing models have limited perception of multi-scale targets, rely on a single attention mechanism, and have insufficient global-local feature dynamic balancing ability, which restricts accurate grasping in industrial scenarios. In addition, accurate and real-time completion of robot grasping detection is a prerequisite for expanding the scale of robot use in environments such as factory workshops, and its importance is self-evident. However, in the above studies, there have been few optimizations focused on the computational complexity of model object detection and the noise processing of grasping detection. YOLOv6 has fast inference speed, high detection accuracy, and is suitable for various embedded platforms, with flexible deployment [18]. MDAFN can suppress noise, highlight object features, enhance target perception in complex backgrounds, and improve detection accuracy [19]. Therefore, based on YOLOv6 and MDAFN, combined with lightweight network MobileViT, Pixel Shuffle (PixShuffle), etc., an L-YOLOv6-MA robot grasping detection model is established. The research fuses lightweight YOLOv6 and MobileViT to achieve parameter compression, enhances the physical-semantic association of features through channel-pixel dual-domain attention, and balances the local grasping points and global context information by combining the multi-scale sensing field decoder, to construct an end-to-end lightweight detection framework, which effectively improves the feature expression and localization accuracy under complex interference. The research aims to provide comprehensive and innovative solutions to the accuracy and efficiency issues of robot grasping and detection in actual factories or other environments.

III. Methods and Materials

This section is divided into two parts. The first part provides a detailed explanation of YOLOv6, Efficient Repetitive Backbone (ERB), and MobileViT, and proposes an object detection module. The second part takes MDAFN as the core, combines multi-scale receptive field Receptive Field Block (RFBs), Pixshuffle, etc., proposes a grasping detection module, and finally constructs the L-YOLOv6-MA robot grasping detection model to improve the robot grasping performance under model control.

A. Object Detection Module Based on YOLOv6

An efficient and accurate target detection strategy is the key to achieving real-time object recognition and tracking performance in complex scenes, and it is also a prerequisite for achieving robot grasping detection performance. However, the target detection strategy of traditional robot grasping detection models usually has high computational complexity, slow response speed, and is difficult to adapt to rapid changes in dynamic environments. YOLOv6 enables efficient deployment on embedded devices, providing real-time object detection while maintaining high accuracy and low latency. Therefore, the research builds an object detection module based on YOLOv6 framework, and the basic architecture of YOLOv6 is shown in Fig. 1.

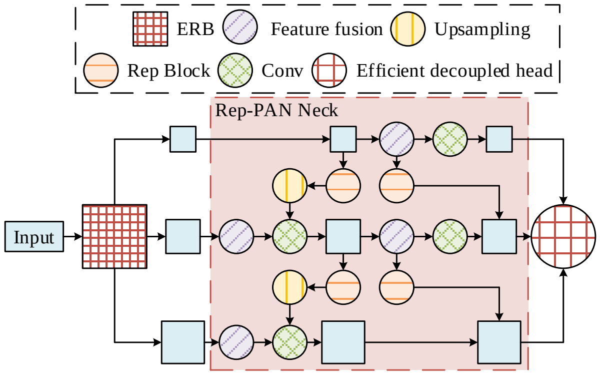


Fig. 1. The architecture of YOLOv6 networks.

In Fig. 1, the YOLOv6 backbone network adopts an ERB structure, which improves feature extraction capability and simplifies the model structure by using a simple repeated parameterized visual geometry group network structure. During training, ERB adopts a multi-branch structure to enhance performance, while during inference, it transforms into a single branch structure of Re-parameterized Block (Rep Blocks) to accelerate the prediction process [20]. The Neck section introduces a Re parameterized Path Aggregation Network (Rep-PAN) to enhance the ability of multi-scale feature fusion. The head adopts an efficient decoupling head design to separate classification and regression tasks, further improving detection accuracy and convergence speed [21]. However, when deploying YOLOv6 on small devices, there are problems such as large model size and high computational cost, which will lead to a decrease in its detection performance in low-resource environments. MobileViT combines the local feature extraction advantages of convolutional neural networks with the global information processing capabilities of visual transformers, enabling both lightweight design and efficient performance. Therefore, the study combines MobileViT for lightweight optimization of YOLOv6, and the network structure of MobileViT is shown in Fig. 2.

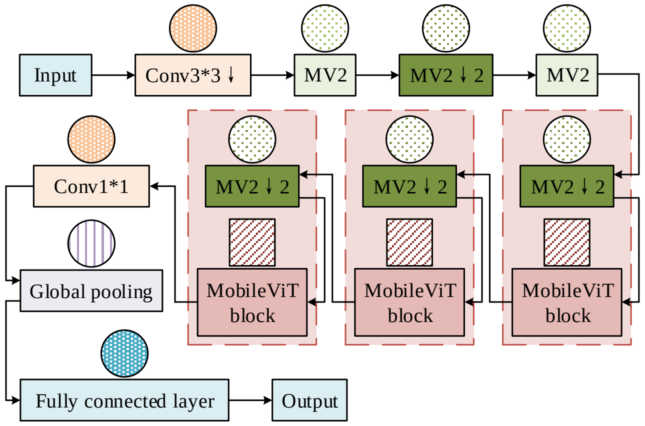


Fig. 2. The structure of MobileViT network.

As shown in Fig. 2, the MobileViT network consists of ordinary convolutional layers, MV2, and MobileViT components. The ordinary convolutional layer is responsible for preprocessing the input image and extracting low-level features. MV2 is the inverse residual structure in MobileNetV2, used for efficient downsampling operations in the network. It extracts features through 1 * 1 convolution for dimensionality enhancement and 3 * 3 deep convolution, and then compresses and expands features through 1 * 1 convolution for dimensionality reduction. The MobileViT component is the core of MobileViT, consisting of multiple Transformers, including three steps: local feature extraction, global feature modeling, and feature fusion [22]. Among them, the expansion factor of MV2 module is 6, which is responsible for controlling the proportion of channel dimensioning. Too small an expansion factor will limit the feature expression ability, while too large a factor will increase the model complexity. The number of stacked Transformer layers in MobileViT is 3, which needs to be considered as a balance between global modeling capability and computational efficiency. In addition, the global pooling layer reduces the dimensionality of the feature map to obtain global features. The fully connected layer maps these global features to the final prediction output. Therefore, the proposed object detection module architecture is shown in Fig. 3.

In Fig. 3, the input image is first subjected to feature extraction through the MobileViT network, which consists of multiple MV2 modules. Each module is followed by a MobileViT component to downsample the feature map. The MobileViT component utilizes its lightweight design to effectively extract image features while maintaining a low number of parameters. After being processed by the MobileViT network, the feature maps enter YOLOv6 and undergo further feature fusion and processing through Simplified Spatial Pyramid Pooling-Fast (SimSPPF) and other convolutional layers and residual connections. The SimSPPF module is located in the Neck structure and replaces traditional parallel structures with serial pooling operations, reducing redundant calculations and improving the network's detection capability for targets of different sizes. The SimSPPF module is located in the Neck structure, which reduces redundant computations by replacing the traditional parallel structure with serial pooling operations. Its pooling kernel size is set to [5,9,13], where too large a kernel size blurs the small target details, while too small a kernel size does not effectively cover the large target context. The Neck structure of YOLOv6 adopts a multi-scale feature fusion strategy, which enhances the network's detection ability for targets of different sizes by horizontally connecting feature maps of different scales. Finally, the feature maps processed by Neck enter the efficient coupling head for object detection tasks.

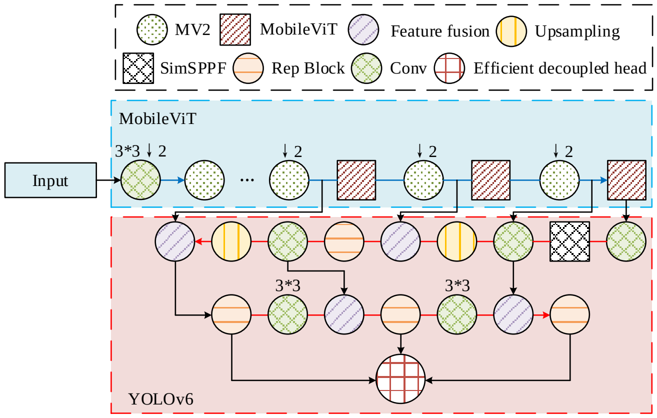


Fig. 3. The architecture of the object detection module.

B. Construction of Grab Detection Module and Robot Grab Detection Model

Object detection provides visual information for robots by identifying objects in images and providing bounding boxes and categories. The proposed object detection module can achieve efficient object detection under low computing resource conditions. However, it cannot directly perform the grasping detection function. MDAFN can suppress noise features, enhance effective features, and improve the accuracy and robustness of capture detection during the fusion process of shallow and deep features. Therefore, the research focuses on MDAFN as the core and constructs a grasping detection module. The basic structure of MDAFN is denoted in Fig. 4.

In Fig. 4, MDAFN is divided into two layers: pixel attention subnetwork and channel attention subnetwork. The pixel attention subnetwork utilizes a convolutional kernel size of 3 * 3 convolutional layers. The convolutional kernel size needs to be weighed against the spatial context-awareness capability and computational overhead. A larger kernel enhances the perceptual field but increases the number of parameters. Through the convolutional layers and Sigmoid activation function, the pixel attention subnetwork assigns weights to each pixel to highlight key visual information. The channel attention subnetwork enhances important channels in the feature map through global average pooling and fully connected layers. Finally, the subnetwork weighted feature map is combined with the original input feature map to integrate pixel and channel level attention information through element wise multiplication, suppressing noise [23]. However, MDAFN has limited performance when dealing with complex backgrounds or overlapping targets, while RFB can provide richer contextual information. Therefore, research is being conducted to optimize the input of MDAFN using RFB.

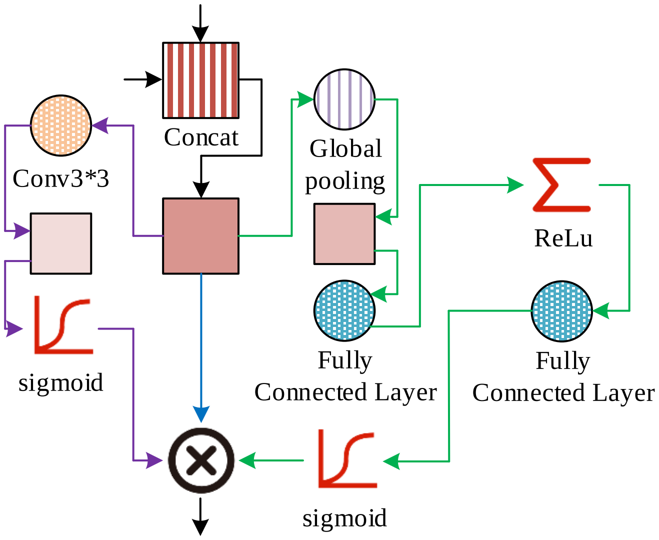


Fig. 4. The basic structure of MDAFN.

RFB aims to enhance the adaptability of the network to multi-scale characteristics by constructing convolutional layers of different scales. The operation process is as follows: RFB first adjusts the number of channels through a 1 * 1 convolution operation, and then extracts multi-scale features through convolution kernels and dilated convolutions of different scales. Its expansion rate is set to [1,3,5] and the number of multibranch channels is configured as [64,128,256], respectively. The feature maps of different scales are then merged to obtain feature representations with rich multi-scale information [24]. In addition, Pixshuffle can achieve efficient upsampling operations while preserving image details and texture information. The operation process is as follows: Pixshuffle first uses convolutional layers to increase the number of channels in the feature map to the square of the target resolution multiple. Afterwards, the channels are rearranged and each pixel's multi-channel is converted into a corresponding image block to achieve an increase in resolution [25]. Therefore, the proposed grasping detection module architecture is shown in Fig. 5.

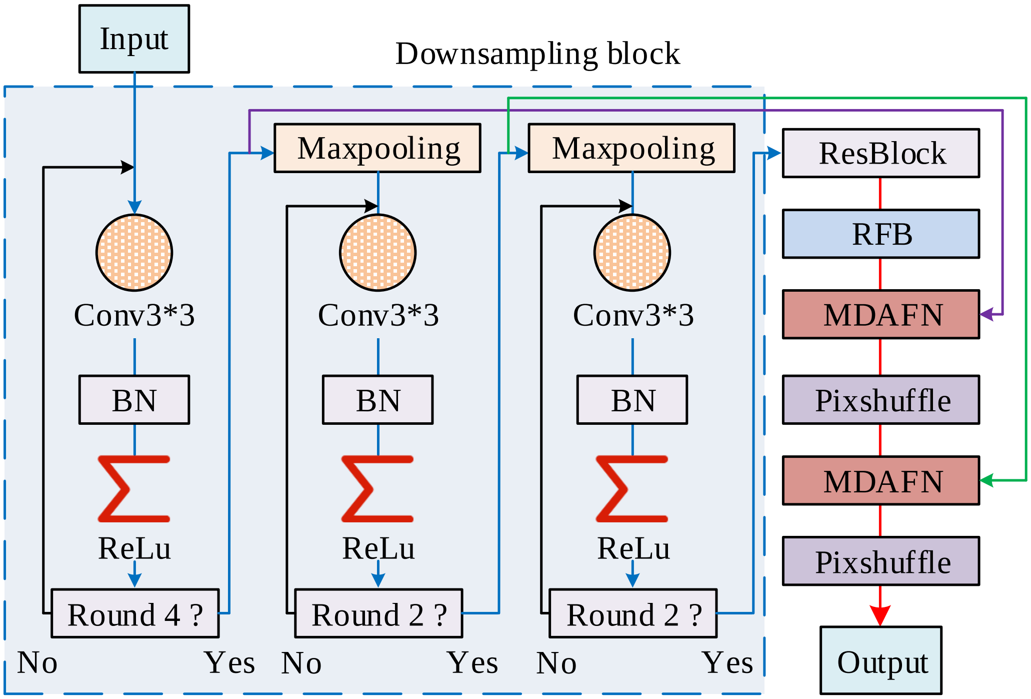


Fig. 5. The architecture of the grasp detection module.

In Fig. 5, the network mainly contains downsampling blocks and a backbone network. In the downsampling block, the input image first passes through a 3 * 3 convolutional layer, followed by a Batch Normalization (BN) layer and a ReLU activation function, and then enters a round of decision-making. If the conditions are met, it enters the max pooling layer and enters the above structure again. After three iterations, the feature outputs that meet the conditions will enter the backbone network. Residual Block (ResBlock) is the first layer of the backbone network, which works together with RFB to extract more discriminative and robust features. Afterwards, the features enter MDAFN and fuse shallow and deep semantic features. Pixshuffle serves as an upsampling layer for the capture detection module to increase feature resolution. In summary, the overall operational process of L-YOLOv6-MA, which combines the object detection module and the grasping detection module, is shown in Fig. 6.

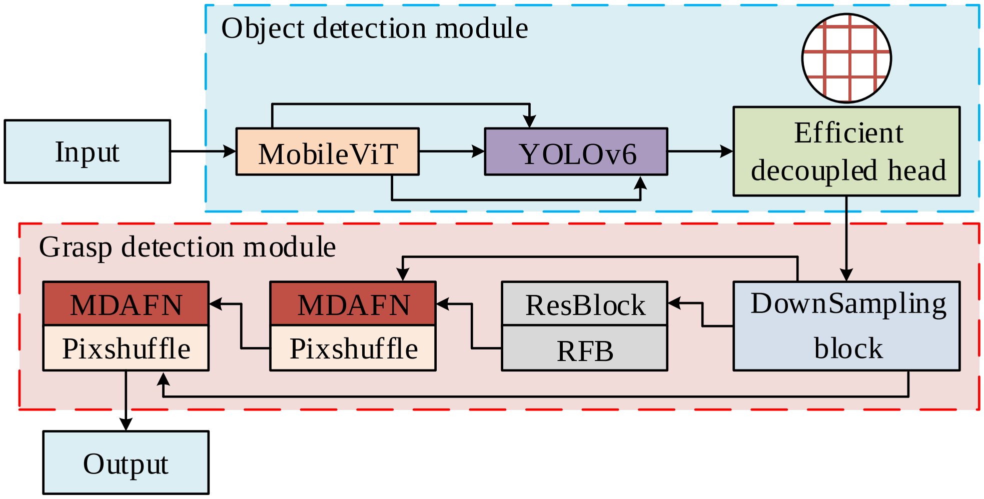


Fig. 6. The overall operation flow of L-YOLOv6-MA.

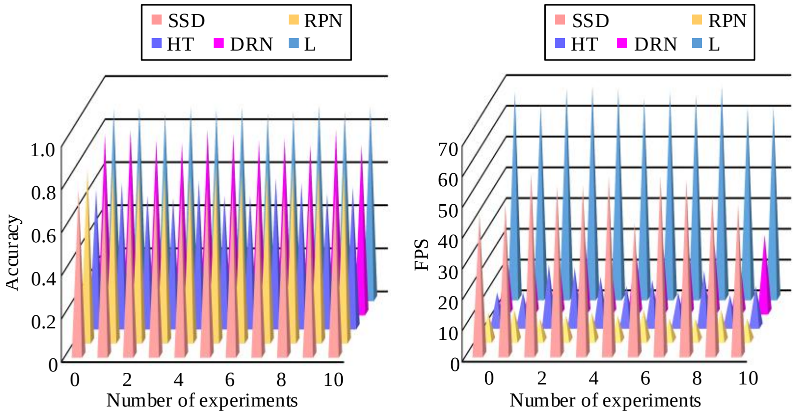
As shown in Fig. 6, the operation process of the L-YOLOv6-MA model includes two stages: object detection and grasping detection. In the object detection stage, YOLOv6 serves as the basic framework and achieves model lightweighting through MV2, MobileViT components, etc., reducing model complexity and computational costs. Fast and accurate recognition of objects in an image is achieved by convolutional operations such as SimSPPF. In the capture detection stage, MDAFN is used as the core to enhance key information in the feature map through pixel and channel attention subnetworks, thereby improving the accuracy of capture detection. By combining structures such as RFB, Pixshuffle, and downsampling blocks, the adaptability and resolution recovery performance of the network to targets of different scales and complex backgrounds are enhanced. The model ultimately outputs the predicted grasp quality, angle and width.

IV. Results

To prove the performance and superiority of the proposed L-YOLOv6-MA model, simulation experiments and actual model performance experiments were conducted based on the theoretical foundation and algorithm analysis mentioned above. The study analyzed the experimental results in detail and compared their performance such as detection accuracy and real-time performance.

A. *Simulation Operation Experiment*

In the simulation experiment, Windows 10 was chosen as the operating system, and the NVIDIA Isaac Sim simulation platform was used to simulate the robot grasping task environment. Moreover, the study constructed a simulated robot using the Gazebo simulator and robot operating system. The study introduced Single Shot MultiBox Detector (SSD), Region Proposal Network (RPN), Hough Transform (HT), and Deep Residual Network (DRN), and compared them with the proposed L-YOLOv6-MA model, which was named L. The study first used the Microsoft Universal Object Context dataset as the object of capture detection, and conducted object detection efficiency experiments by comparing the object detection accuracy and Frames Per Second (FPS) of different algorithms. The experimental results are denoted in Fig. 7.

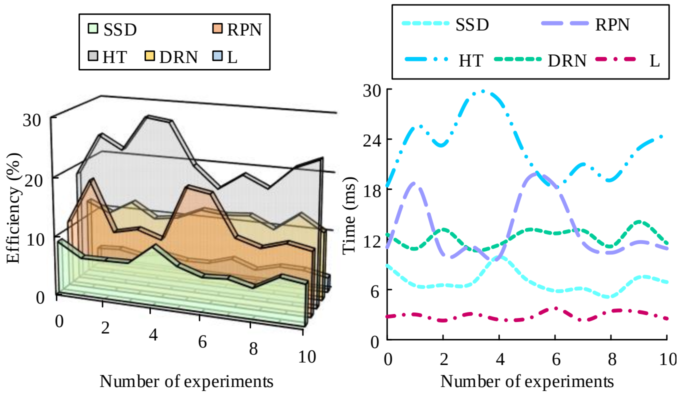


(a) Object detection accuracy

(b) FPS of Object detection

Fig. 7. Comparison of detection accuracy and FPS.

According to Fig. 7(a), the target detection accuracy of HT was the lowest, between 0.60 and 0.69. Next was SSD, with an accuracy between 0.75 and 0.78. The average accuracy of RPN and DRN was 0.80 and 0.82, respectively. The target detection accuracy of L was the highest, ranging from 0.86 to 0.90. As shown in Fig. 7(b), L also had the highest FPS, with an average FPS of 66.00. Next was SSD, with an average FPS of 52.82. The FPS ranges of HT and DRN were 10.00 to 19.00 and 20.00 to 30.00, respectively. The FPS of RPN was relatively low, with an average FPS of 8.36. The experimental findings indicated that the target detection efficiency of the proposed model was much higher than that of traditional methods. On this basis, the study explored the model's capture detection performance by comparing the image segmentation efficiency and running time of different algorithms. The experimental results are denoted in Fig. 8.

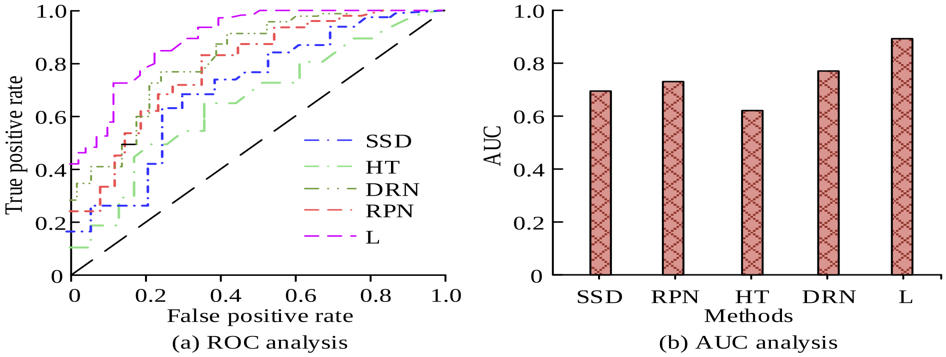


(a) Segmentation efficiency

(b) Running time

Fig. 8. Differences in segmentation efficiency and running time.

According to Fig. 8(a), the image segmentation efficiency of L was relatively high, ranging from 90.96% to 95.35%. Next was DRN, with an efficiency ranging from 82.32% to 88.41. The average efficiencies of SSD and RPN were 77.85% and 84.06%, respectively. The image segmentation efficiency of HT was the lowest, ranging from 65.09% to 73.94%. According to Fig. 8(b), HT had the longest running time, with an average time of 22.96ms. Next was RPN, with an average time of 13.01ms. The running times of SSD and DRN were between 5.14ms and 9.91ms, and 10.75m and 14.11ms, respectively. The running time of L was the shortest, with an average time of only 2.85ms. The experiment findings denoted that the image segmentation efficiency of the proposed model was far superior to other methods. Subsequently, the Receiver Operating Characteristic Curve (ROC) and Area Under the Curve (AUC) of the comparative model were studied to further investigate the performance of the model. The experiment findings are shown in Fig. 9.



(a) ROC analysis

(b) AUC analysis

Fig. 9. Differences in ROC curves and AUC values.

As shown in Fig. 9(a), when the false positive rate (FPR) was between 0 and 0.1, the growth rate of the model's true positive rate (TPR) was the highest. Afterwards, the growth of TPR gradually slowed down and reached its maximum value after the FPR was 0.6. When the FPR was the same, the TPR of L was the highest, and the TPR of HT was the lowest. For example, when the FPR was 0.5, the TPR of L was 0.99. At this time, the TPRs of SSD, RPN, HT, and DRN were 0.79, 0.85, 0.71, and 0.89, respectively. According to Fig. 9(b), the AUC value of L was as high as 0.91. The AUC value of DRN was slightly lower, at 0.79. The AUC values of SSD and RPN were 0.70 and 0.74, respectively. The AUC value of HT was the lowest, only 0.64. The experiment findings denoted that the comprehensive effectiveness of the raised model was much higher than traditional methods.

B. *Actual Model Performance Experiment*

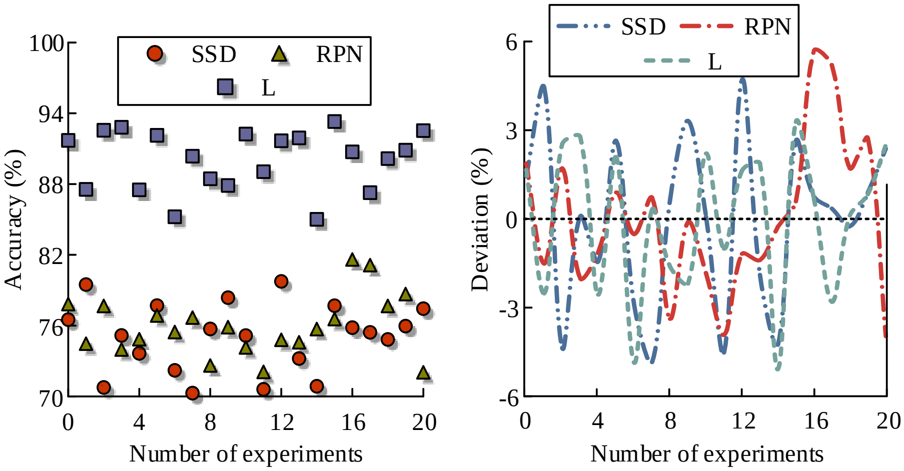
Simulation running experiments are an important reference for measuring robot grasping models. However, due to the complexity and randomness of the factory environment and grasping task behavior, there are often differences between the actual performance of the model and the simulation results. Therefore, the study selected SSD and RPN as comparative algorithms for actual model performance experiments. The study selected a certain parts processing workshop as the actual experimental environment, and verified its practical promotion potential by exploring the performance of the model in the actual environment. The study first explored the actual target detection and image segmentation performance of the robot under model control for screwdrivers. The experiment results are denoted in Table I.

TABLE I. ACTUAL DETECTION AND SEGMENTATION FOR SCREWDRIVER

Number of experiments	**Detection accuracy**			**Segmentation efficiency (%)**		
	SSD	**RPN**	**L**	**SSD**	**RPN**	**L**
1	0.69	0.72	0.76	75.53	79.71	85.01
2	0.74	0.74	0.78	74.75	74.54	79.32
3	0.74	0.74	0.76	72.12	69.09	84.15
4	0.71	0.75	0.78	74.49	73.78	81.62
5	0.70	0.73	0.75	70.09	78.15	78.44
6	0.71	0.73	0.82	72.71	73.08	81.43
7	0.69	0.75	0.83	70.83	72.22	80.48
8	0.70	0.76	0.79	73.61	74.61	86.45
9	0.73	0.75	0.82	68.21	74.69	80.12

10	0.73	0.76	0.82	75.33	75.25	83.95
11	0.74	0.73	0.77	74.89	77.13	82.03
12	0.75	0.75	0.83	68.47	73.75	80.93
13	0.68	0.77	0.77	74.46	71.33	83.35
14	0.69	0.75	0.79	72.60	74.77	82.09
15	0.70	0.72	0.79	76.50	75.00	83.80
16	0.69	0.72	0.83	74.55	76.88	80.24
17	0.70	0.74	0.83	78.81	75.57	79.17
18	0.69	0.74	0.81	72.83	74.07	83.92
19	0.73	0.71	0.77	73.69	75.61	84.82
20	0.75	0.71	0.75	72.73	76.74	85.83
Mean	0.71	0.74	0.79	73.36	74.80	82.36

According to Table I, the actual object detection accuracy of SSD was relatively low, ranging from 0.68 to 0.75, with an average accuracy of 0.71. The actual accuracy range of RPN was 0.71 to 0.77, with an average accuracy of 0.74. The actual accuracy of L was relatively high, ranging from 0.75 to 0.83, with an average accuracy of 0.79. In addition, SSD had the lowest actual image segmentation efficiency, ranging from 68.21% to 78.81%, with an average efficiency of 73.36%. The actual efficiency of RPN was 69.09% to 79.71%, with an average efficiency of 74.80%. The actual image segmentation efficiency of L ranged from 78.44% to 86.45%, with an average efficiency of 82.36%. The experiment findings denoted that the actual performance of the proposed model was much higher than traditional methods. On this basis, the grasping performance of robots controlled by comparative models on screwdrivers was studied, and the experimental results are shown in Fig. 10.



(a) Grasping accuracy

(b) Grasping accuracy deviation

Fig. 10. Grasping accuracy and deviation for the screwdriver.

As shown in Fig. 10(a), the success rates of SSD and RPN were relatively close when controlling the robot to grab the screwdriver. The success rates of the two were divided between 70.28% and 79.75%, and 72.04% and 81.61%. At this point, the success rate range of L was 85.01% to 93.30%. According to Fig. 10(b), the success rate deviation of RPN was the smallest, ranging from -3.93% to 5.64%. Next was L, with a deviation range of -5.03% to 3.26%. The deviation of SSD was relatively large, ranging from -4.80% to 4.67%. The experiment findings denoted that under the proposed model control, the robot had the highest grasping success rate and relatively stable performance. Finally, the study designed robots controlled by different models to grasp 50 screws and explored the successful grasping times of different models. The experimental results are shown in Fig. 11.

According to Fig. 11(a), when controlling the robot to grab screws, the SSD had the lowest success rate, between 25 and 33. Next was RPN, with a success rate of 27 to 35. The success rate of L was the highest, between 35 and 43 times. According to Fig. 11(b), the absolute deviation of RPN success times was the lowest, between 0.48 and 3.52. Next was L, with an absolute deviation range of 0.14 and 4.14. The absolute deviation of SSD was the largest, ranging from 0.05 to 4.05. The experiment findings denoted that under the proposed model control, the robot had the highest grasping efficiency and was relatively stable for smaller objects. From the above, the performance of the proposed model was much higher than traditional methods and had the potential for promotion and application.

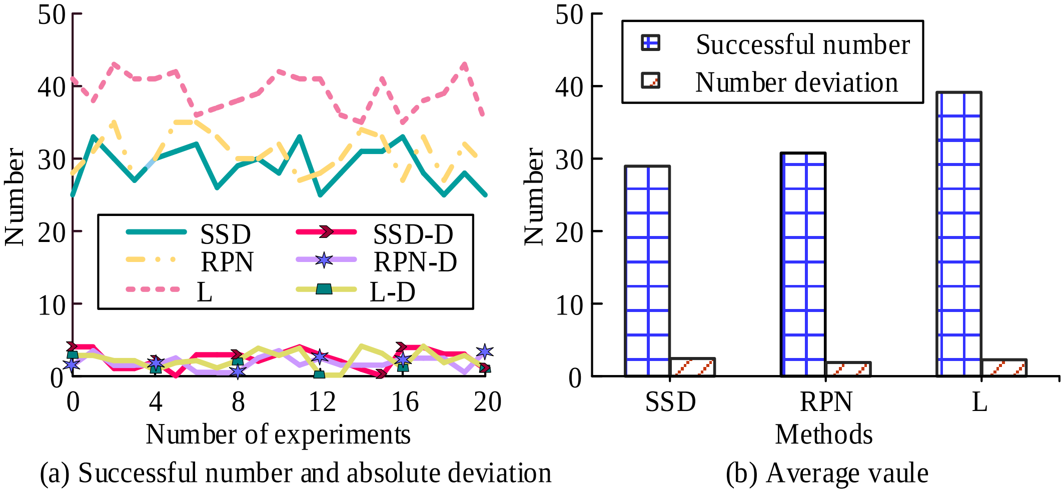


Fig. 11. Grasping number and deviation for screw.

V. Discussion

Aiming at the problem of low performance of traditional robot grasping detection models, this study focused on YOLOv6 and MDAFN as the core, constructed object detection modules and grasping detection modules, and proposed the L-YOLOv6-MA model by combining the two. The study introduced components such as MobileViT and Pixshuffle to achieve lightweight design of the model while reducing environmental noise and improving model accuracy. The experiment findings denoted that the simulated object detection accuracy and FPS of the proposed model were between 0.86 and 0.90, and 62.00 and 69.00, respectively. The average accuracy and FPS of other methods were 0.76 and 25.00, respectively. The simulation image segmentation efficiency and running time of the model were between 90.96% and 95.35%, and 2.28ms and 3.75ms, respectively. The average efficiency and running time of other methods were 79.01% and 13.80ms, respectively. The AUC value of the model was 0.91. The average AUC value of other algorithms was 0.72. In addition, the actual object detection accuracy and image segmentation efficiency range of the proposed model were 0.75 to 0.83 and 78.44% to 86.45%, respectively. The average accuracy and efficiency of other algorithms are 0.73% and 74.08%, respectively. The gripping rate and deviation range of the screwdriver under model control were between 85.01% and 93.30%, and -5.03% and 3.26%, respectively. The average grasping rate and absolute deviation of other methods were 75.52% and 2.15%, respectively. Moreover, the successful grasping times and absolute deviation range of screws under model control were 35 to 43 and 0.14 to 4.14, respectively. The average success rate and absolute elimination of other methods were 29.86 and 2.16, respectively. In summary, the core innovations of the L-YOLOv6-MA model are: 1) establishing a synergistic architecture between YOLOv6 and MobileViT to achieve efficient feature extraction in dynamic environments through lightweight reorganization; 2) constructing a channel-pixel dual-domain attentional mechanism to strengthen the ability of grasping feature discrimination under complex background interference; 3) designing a multiscale fusion decoder that combines sense-of-field extension and subpixel localization to improve the accuracy of grasping position estimation.

VI. Conclusion

The contribution of the L-YOLOv6-MA model is that it effectively solves the problem of grasping robustness in complex scenarios by establishing a synergistic mechanism of lightweight adaptive feature extraction and multidimensional attention. The detection framework breaks through the efficiency bottleneck of traditional staged processing, provides a high-precision and low-latency solution for shaped part grasping, and significantly improves the flexible adaptation capability and deployment efficiency of automated production lines.

While demonstrating notable advancements, this study has limitations: 1) Experimental validation primarily targets standard screw-type workpieces, requiring extended verification for reflective/flexible materials; 2) Synthetic data-based training lacks real-world physical parameter integration; 3) Hardware-specific deployment limits cross-platform adaptability, and there is insufficient coordination exists between visual detection and robotic motion control. Future work will focus on: 1) Developing multi-material grasping datasets enhanced by transfer learning to address generalization gaps; 2) Establishing a digital twin framework combining virtual simulation and physical parameters to refine predictive accuracy; 3) Creating hardware-agnostic deployment solutions for efficient edge computing adaptation across devices; 4) Implementing visual-force closed-loop coordination to enable real-time grip adjustment and slip compensation. These improvements aim to bridge the simulation-to-reality gap while optimizing dynamic control synchronization, ultimately supporting robust industrial deployment across diverse production scenarios.

REFERENCES

[1] Dong M, Zhang J. A review of robotic grasp detection technology. Robotica, 2023: 1-40.

[2] Zeng C, Li S, Chen Z, Yang C, Sun F, Zhang J. Multifingered robot hand compliant manipulation based on vision-based demonstration and adaptive force control. IEEE Transactions on Neural Networks and Learning Systems, 2022, 34(9): 5452-5463.

[3] Zhou Z, Zuo R, Ying B, Zhu J, Wang Y, Wang X, Liu X. A sensory soft robotic gripper capable of learning-based object recognition and force-controlled grasping. IEEE Transactions on Automation Science and Engineering, 2022, 21(1): 844-854.

[4] Zeng H, Luo J. Construction of multi-modal perception model of communicative robot in non-structural cyber physical system environment based on optimized BT-SVM model. Computer Communications, 2022, 181: 182-191.

[5] Ali W, Kolyubin S A. Emg-based grasping force estimation for robot skill transfer learning. Russian Journal of Nonlinear Dynamics, 2022, 18(5): 859-872.

[6] Lee H, Park S, Jang K, Kim S, Park J. Contact state estimation for peg-in-hole assembly using gaussian mixture model. IEEE Robotics and Automation Letters, 2022, 7(2): 3349-3356.

[7] Wang S, Zhou Z, Kan Z. When transformer meets robotic grasping: Exploits context for efficient grasp detection. IEEE Robotics And Automation Letters, 2022, 7(3): 8170-8177.

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

[9] Cheng H, Wang Y, Meng M Q H. A vision-based robot grasping system. IEEE Sensors Journal, 2022, 22(10): 9610-9620.

[10] Jiang J, Cao G, Butterworth A, Do T T, Luo S. Where shall i touch? vision-guided tactile poking for transparent object grasping. IEEE/ASME Transactions on Mechatronics, 2022, 28(1): 233-244.

[11] Wu Y. Research on grasping model based on visual recognition robot arm. Applied and Computational Engineering, 2024, 41: 11-21.

[12] Cheng H, Wang Y, Meng M Q H. A robot grasping system with single-stage anchor-free deep grasp detector. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-12.

[13] Yu S, Zhai D H, Xia Y. Cgnet: Robotic grasp detection in heavily cluttered scenes. IEEE/ASME Transactions on Mechatronics, 2022, 28(2): 884-894.

[14] Cao H, Chen G, Li Z, Feng Q, Lin J, Knoll A. Efficient grasp detection network with Gaussian-based grasp representation for robotic manipulation. IEEE/ASME Transactions on Mechatronics, 2022, 28(3): 1384-1394.

[15] Song K, Wang J, Bao Y, Huang L, Yan Y. A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception. IEEE/ASME Transactions on Mechatronics, 2022, 28(3): 1558-1569.

[16] Hui N M, Wu X H, Han X W, Wu B J. A robotic arm visual grasp detection algorithm combining 2D images and 3D point clouds. Applied Mechanics and Materials, 2024, 919: 209-223.

[17] Abdullah-Al-Noman M, Eva A N, Yeahyea T B, Khan R. Computer vision-based robotic arm for object color, shape, and size detection. Journal of Robotics and Control (JRC), 2022, 3(2): 180-186.

[18] Chen H, Wan W, Matsushita M, Kotaka T, Harada K. Automatically prepare training data for yolo using robotic in-hand observation and synthesis. IEEE Transactions on Automation Science and Engineering, 2023, 21(3): 4876-4892.

[19] Ren G, Geng W, Guan P, Cao Z, Yu J. Pixel-wise grasp detection via twin deconvolution and multi-dimensional attention. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 4002-4010.

[20] Shen X, Wang H, Li Y, Gao T, Fu X. Criss-cross global interaction-based selective attention in YOLO for underwater object detection. Multimedia Tools and Applications, 2024, 83(7): 20003-20032.

[21] Sharma P, Tyagi R, Dubey P. Optimizing real-time object detection-a comparison of YOLO models. International Journal of Innovative Research in Computer Science & Technology, 2024, 12(3): 57-74.

[22] Núñez Montoya B, Valarezo Añazco E, Guerrero S, Valarezo-Añazco M, Espin-Ramos D, Jiménez Farfán C. Myo transformer signal classification for an anthropomorphic robotic hand. Prosthesis, 2023, 5(4): 1287-1300.

[23] Yang L, Zhang C, Liu G, Zhong Z, Li Y. A model for robot grasping: integrating transformer and CNN with RGB-D fusion. IEEE Transactions on Consumer Electronics, 2024, 70(2): 4673-4684.

[24] Wu Y, Fu Y, Wang S. Real-time pixel-wise grasp affordance prediction based on multi-scale context information fusion. Industrial Robot: the international journal of robotics research and application, 2022, 49(2): 368-381.

[25] Hou X X, Liu R B, Zhang Y Z, Han X R, He J C, Ma H. NC2C-TransCycleGAN: Non-contrast to contrast-enhanced CT image synthesis using transformer CycleGAN. Healthcraft Front, 2023, 2(1): 34-45.