

Parts Grasping by Intelligent Industrial Robots Based on Image Matching Algorithms and CNNs

Tao Li

Lecturer, School of Equipment Engineering, Shanxi Vocational University of Engineering Science and Technology, Jinzhong, 030619, China, Email: taoli71@outlook.com

Production Management

Received January 7, 2026; revised March 19, 2026; August 2, 2026; accepted September 12, 2026

Available online September 20, 2026

Abstract: This study proposes an intelligent industrial robot part-grasping optimization model that integrates image-matching algorithms with convolutional neural networks. The model employs SuperPoint-based feature matching and introduces multi-scale convolution and a channel attention mechanism to enhance feature extraction. A generative convolutional neural network module further predicts the optimal grasping posture end-to-end. Experimental results show that the proposed image matching model achieves a matching accuracy of 97.8%, a repeatability of 92.5%, and a feature detection speed of 12.1ms per frame. The model maintains stable performance under complex lighting conditions and multi-view industrial environments, demonstrating strong robustness and real-time capability. These improvements enable more reliable perception and decision-making in robotic grasping tasks, thereby supporting flexible automation and the stable deployment of industrial robots in real manufacturing scenarios.

Keywords: Intelligent industrial robot; image matching algorithm; convolutional neural network; part grasping; feature fusion.

Copyright © Journal of Engineering, Project, and Production Management (EPPM-Journal).
DOI 10.32738/JEPPM-2026-20

1. Introduction

With the implementation of intelligent manufacturing and the advancement of Industry 4.0, industrial robots are increasingly used in automated production lines, precision assembly stations and online inspection processes and have gradually become an important technical foundation for improving efficiency and achieving flexible transformation in the manufacturing industry. Part gripping is a core part of the robot's operation process. Whether the target recognition is accurate and the gripping action is stable often directly determines the production cycle and product qualification rate. In the past, visual recognition methods that relied on manually designed features could meet the needs in ideal or simple scenarios, but once they entered complex environments with significant changes in lighting conditions, large differences in part postures, or partial occlusion, the feature extraction and matching process was prone to distortion, and the recognition results fluctuated, leading to gripping deviation or even gripping failure (Gtifa and Sakly, 2023). Meanwhile, there are many types of parts in actual industrial sites, and factors such as strong reflection of metal surfaces and irregular structural shapes further increase the difficulty of robot visual perception and stable gripping. In recent years, convolutional neural networks (CNNs) have performed exceptionally well in image recognition and spatial perception, providing a new technical path for tackling the above challenges. Compared with traditional algorithms based on handcrafted features, CNNs can automatically extract spatial morphology and texture information of targets through hierarchical feature learning while maintaining high detection accuracy and good generalization performance in complex backgrounds (Sunil and Afzal, 2023). Zhang (2025) combined AI-based image recognition with blockchain technology to address identity fraud by detecting forged documents. This method provided a novel solution for cross-border financial transactions. Wang et al. (2024) proposed a meta-learning model containing feature calibration modules. The model was superior on three commonly used datasets. To address the poor performance of ViT on small and medium-sized datasets, Lu et al. (2025) proposed a causal visual transformer that integrated causal relationships and inductive bias. The research results showed that the causal visual transformer achieved a Top-1 accuracy of 80.6% on ImageNet-1k, which was 2.7% higher than the ViT benchmark, with 92% fewer parameters, and consistent performance on smaller datasets, with a close causal relationship. To address misalignment when the current facial model is fitted to the target image, Li and Da (2025) proposed a sensor-geometry supervision method driven by image matching, which treated face alignment as a matching problem and established pixel-level correspondences to improve

reconstruction. The outcomes demonstrated that the approach achieved an advanced level among self-supervised methods on the new test set and obtained results comparable to the benchmark test. However, regarding Multi-Scale Feature (MSF) fusion and gradient transfer stability, some networks still have bottlenecks, making it difficult to simultaneously account for real-time performance and accuracy. Therefore, this study proposes an intelligent method for grasping industrial robot parts based on image-matching algorithms and CNNs. The research aims to provide a high-precision, high-efficiency, and low-latency approach to visual perception and decision-making for industrial robotic grasping systems in intelligent manufacturing environments. The novelty of the study lies in the adaptive modeling of complex industrial part geometries through multi-scale enhancement and attention fusion, and the accurate prediction of grasping posture via end-to-end regression.

To further clarify the scientific objectives of this study, the following research questions are addressed:

RQ1: How can image matching algorithms based on deep feature extraction improve the robustness and accuracy of part detection in complex industrial environments?

RQ2: How can the integration of MSF enhancement and attention mechanisms improve feature representation and matching efficiency for robotic grasping tasks?

RQ3: To what extent can the proposed GGCNN-CSPNet architecture improve grasping accuracy, execution speed, and operational stability compared with existing grasping models?

The contribution of this study mainly lies in the engineering integration and application-oriented optimization of existing vision and grasping techniques. By combining image matching algorithms with CNN-based grasp prediction and introducing multi-scale enhancement, attention fusion, and CSP-based feature optimization, the proposed framework improves the robustness and efficiency of robotic grasping in practical industrial environments. Rather than proposing a new theoretical learning paradigm, the study focuses on enhancing the reliability and practical applicability of robotic perception and grasping systems in real manufacturing scenarios.

2. Methods

2.1. Image Matching Model Based on SuperPoint

In intelligent industrial robot parts grasping, accurate identification and positioning of target parts are crucial, and the image matching algorithm is the core for achieving efficient grasping. Traditional feature point methods are not sufficiently stable in complex backgrounds, lighting changes and multi-view conditions (Vite-Chávez et al., 2023; Chen et al., 2023). To improve matching accuracy, the SuperPoint deep-feature point detection and description model is introduced. Through unsupervised pre-training and end-to-end feature learning, key point detection and description are unified within a single network framework. The model is trained on self-supervised corner data and enhances the geometric consistency of real images through Homographic Adaptation, thereby improving feature stability and matching accuracy (Hong et al., 2024). By combining multi-view images of parts captured by industrial cameras, this study further optimizes the feature pyramid structure and the SuperPoint-based matching filtering strategy, enabling them to adapt to changes in lighting, occlusion, and pose in complex industrial scenarios. Its structure is presented in Fig. 1.

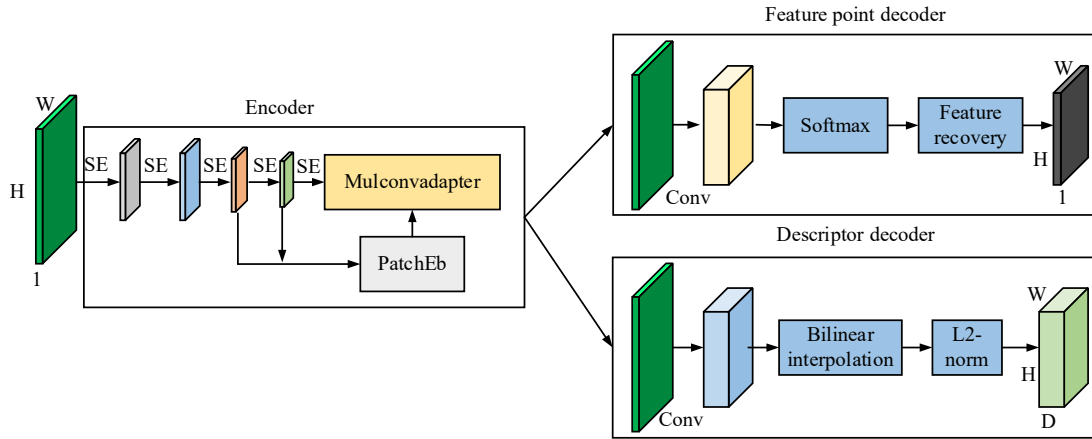


Fig. 1. SuperPoint structural analysis

In Fig. 1, the overall network framework comprises three parts: the encoder, the feature point decoder and the descriptor decoder. The input image first passes through an improved encoder containing a Squeeze-and-Excitation (SE) module to extract multi-layer semantic features. The SE module dynamically fine-tunes the channel weights via a channel attention mechanism, as expressed in Eq. (1).

$$\mathbf{s} = \sigma(W_2 \delta(W_1 \cdot \text{GAP}(\mathbf{F}))) \quad (1)$$

In Eq. (1), $\mathbf{F}$ is the input feature map, $GAP(\square)$ represents global average pooling (GAP), W_1 and W_2 are the weight matrices of the two fully connected layers, $\delta(\square)$ is the ReLU function, $\sigma(\square)$ is the Sigmoid function, and the output s is employed to weight each channel. The improved encoder also introduces the Multi-scale Convolution Adapter (MulConvAdapter) and PatchEb modules to enhance the multi-scale convolutional feature fusion and local structure texture perception capabilities. Subsequently, the feature point decoder makes predictions on the downsampled feature map and obtains the pixel-level feature point probability distribution via Softmax normalization, as expressed in Eq. (2).

$$P_{ij} = \frac{e^{z_{ij}}}{\sum_k e^{z_{ik}}} \quad (2)$$

In Eq. (2), z_{ij} is the response value of type j at position i and P_{ij} is the confidence score of the feature point at the corresponding position. The descriptor decoder upsamples the encoded features, restores the spatial resolution through bilinear interpolation, and generates a feature descriptor matrix. To improve matching stability, the output features are L2-regularized, and their expression is given in Eq. (3) (Narang and Tallur, 2022).

$$\hat{D}_{ij} = \frac{D_{ij}}{\|D_{ij}\|_2} \quad (3)$$

In Eq. (3), D is the feature descriptor matrix, and $\|\cdot\|_2$ represents the Euclidean norm. To further boost the model's capability, a multi-scale enhancement module is established, whose structure is presented in Fig. 2.

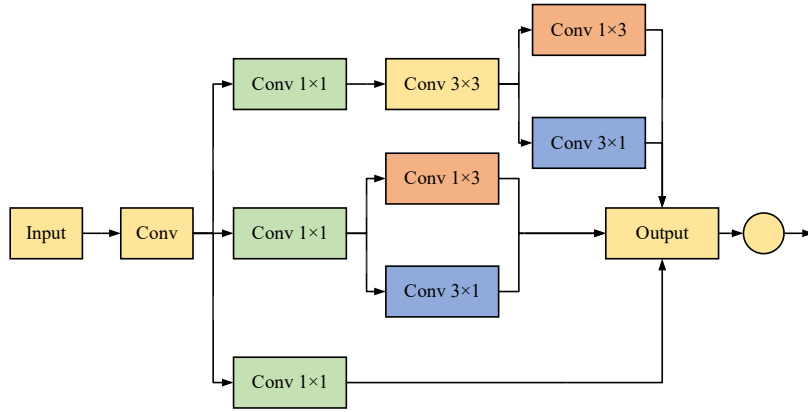


Fig. 2. Multi-scale enhancement module

In Fig. 2, the right branch uses a single-layer 1×1 convolution to retain the channel information of the original features while reducing redundant feature dimensions (Lora-Rivera et al., 2023). The outputs of each branch are then fused through a weighted aggregation mechanism, and the fused feature is combined with the input feature through a residual connection. This design strengthens the interaction between shallow texture information and deep semantic features, thereby improving the stability of feature extraction under illumination variations and viewpoint changes. As demonstrated in Table 1, the enhanced feature representation contributes to higher matching accuracy (97.8%) and faster feature detection (12.1ms per frame), indicating that the multi-scale enhancement module improves both robustness and computational efficiency in practical industrial environments. The final model structure is presented in Fig. 3.

As shown in Fig. 3, an SE module is embedded after convolution to achieve adaptive channel weighting. The response intensity of each channel is computed using global average pooling, and the Sigmoid function produces channel attention weights that emphasize informative channels while suppressing redundant responses. This attention-guided feature selection improves the discriminability of key points in complex industrial scenes, particularly when parts exhibit reflections, occlusions, or pose variations. The improved representational capability enables the SuperPoint-based model to maintain stable matching performance, as reflected in a lower localization error of 0.93px and a higher repeatability of 92.5% reported in Table 1. BLOCK4 continues to use max pooling for feature compression and maintains the 128-channel convolutional configuration to extract high-level semantic information. Finally, the ReductionB module performs feature dimensionality reduction and information fusion. The design of the entire network structure enables step-by-step feature extraction and attention control from shallow texture to deep semantics, making the SuperPoint model more accurate at detecting feature points in complex industrial environments and more robust to lighting changes and viewpoint shifts.

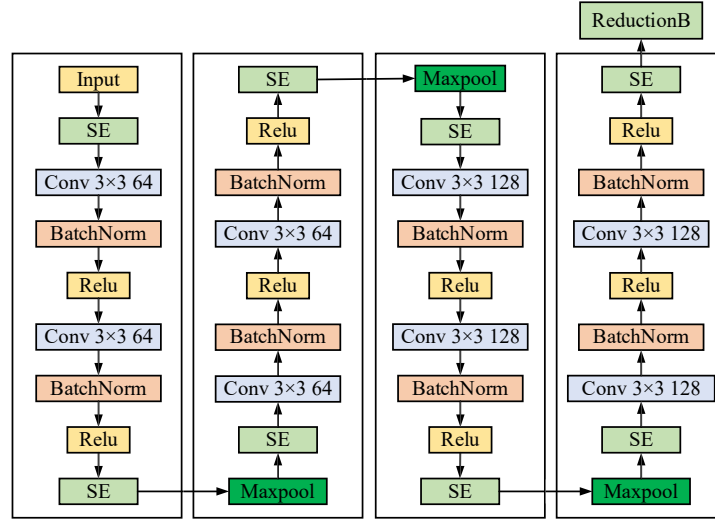


Fig. 3. Image matching model structure based on improved SuperPoint

2.2. Parts Grasping Model Based on Improved CNN

The SuperPoint-based image-matching model achieves robust feature point detection and description, laying a visual foundation for accurate identification and localization of industrial parts. However, relying solely on two-dimensional image features remains insufficient to handle complex situations such as changes in part pose, uneven illumination, and occlusions (Kalaierasu and Natarajan, 2022; Yang et al., 2022). To address this, the study further develops a part-grasping model based on the Generative Grasping CNN (GGCNN). By introducing an MSF enhancement module and an attention fusion mechanism, end-to-end learning from feature extraction to grasping decision is achieved (Wang and Huang, 2025). The GGCNN structure is presented in Fig. 4.

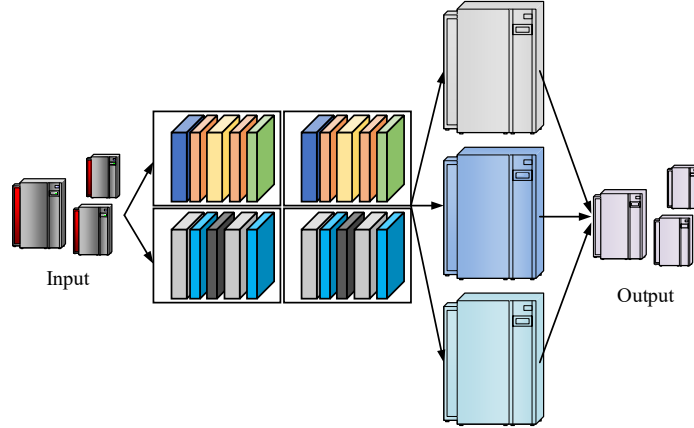


Fig. 4. GGCNN structure

As shown in Fig. 4, the network contains five main convolutional layers with kernel sizes of 9×9 , 5×5 , 3×3 , 5×5 , and 9×9 , corresponding to 32, 16, 8, and 32 channels, respectively, thus achieving a multi-layer representation from low-level geometric edges to high-level structural features (Pilato et al., 2022; Gao and Zhou, 2022). The output of GGCNN consists of three key maps: the grasping quality map Q_θ , the grasping orientation map W_θ , and the grasping opening angle map Φ_θ . The grasping quality map is used to evaluate the feasibility of each pixel as a grasping point; its calculation is given by Eq. (4).

$$Q_\theta(x, y) = \sigma(f_Q(x, y)) \quad (4)$$

In Eq. (4), $Q_\theta(x, y)$ represents the capture score predicted by the convolutional network, and $\sigma(\cdot)$ is the Sigmoid function used to normalize the output to the $[0, 1]$ interval (Li and Quan, 2023). The capture direction pattern represents the local normal-angle distribution, as given in Eq. (5).

$$W_\theta(x, y) = \arctan\left(\frac{G_y(x, y)}{G_x(x, y)}\right) \quad (5)$$

In Eq. (5), G_x and G_y represent the network's gradient responses in the horizontal and vertical directions, reflecting changes in the grasping direction. The expression for the grasping opening angle is presented in Eq. (6) (Kumar and Paul, 2023; Li et al., 2023).

$$\Phi_\theta(x, y) = \pi \cdot \tanh(f_\phi(x, y)) \quad (6)$$

In Eq. (6), $f_\phi(x, y)$ is the predicted grasping-angle mapping value used to control the robot arm's closing range. The final grasping posture vector is presented in Eq. (7).

$$\mathbf{G}_\theta = \{Q_\theta, W_\theta, \Phi_\theta\} \quad (7)$$

In Eq. (7), $\mathbf{G}_\theta$ represents the final grasping pose vector. To further boost the model's capabilities, the CSPNet module is introduced. The CSPNet structure is presented in Fig. 5.

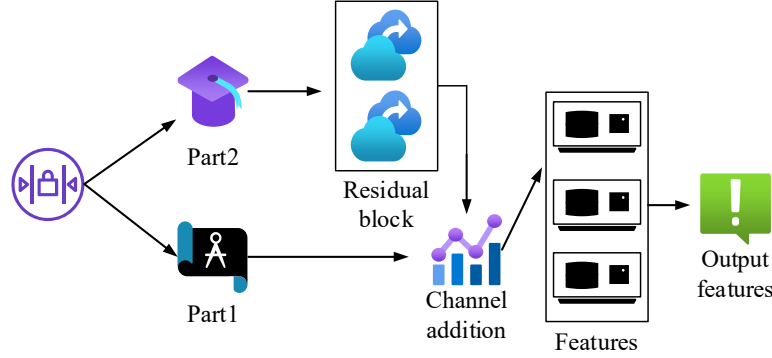


Fig. 5. CSPNet structure

As shown in Fig. 5, this module improves gradient flow efficiency and feature representation capability through feature splitting and cross-stage fusion. By preserving shallow features while simultaneously transforming deep features, the CSPNet structure reduces redundant computation and enhances feature diversity. This design stabilizes the network training process and improves the accuracy of grasp prediction. As reflected in Table 2, the model integrating CSPNet achieves a grasping success rate of 94.8% and a positioning accuracy of 0.93 mm, demonstrating that the cross-stage feature fusion effectively enhances the robustness of grasp prediction (Mann et al., 2022; Preethi and Mamatha, 2023). The input features are first categorized into two parts: Part 1 and Part 2. Part 1 directly performs channel transfer to retain shallow information, while Part 2 enters the residual block for deep feature extraction and nonlinear transformation. The residual block employs residual learning, and its computation is given by Eq. (8).

$$Y = F(X, W) + X \quad (8)$$

In Eq. (8), X represents the input feature map, $F(X, W)$ represents the feature output after convolution and activation function transformation, W is the convolution kernel weight, and Y is the output feature of the residual block. This structure can introduce new deep semantic information while preserving the original features, preventing gradient vanishing, and improving the model's convergence stability. Subsequently, the shallow features from Part 1 and the deep features from Part 2, after the residual block, are fused along the channel dimension to form a comprehensive representation, as shown in Eq. (9).

$$Z = \text{Concat}(Y_1, Y_2) \quad (9)$$

In Eq. (9), Y_1 represents the features from Part 1, Y_2 represents the features after passing through the residual block, $\text{Concat}(\square)$ represents the channel concatenation operation, and Z represents the output feature map after fusion. Through this partial residual connection method, CSPNet effectively reduces computational redundancy, improves feature diversity and information flow, and enables the network to achieve stronger feature expression and generalization capabilities while maintaining a lightweight design, thereby providing efficient and stable feature support for subsequent CNN feature extraction and part-grasping posture prediction. The final structure is presented in Fig. 6.

As shown in Fig. 6, the input image is processed through an initial 9×9 convolutional layer to extract low-level geometric and texture features. A subsequent 5×5 convolutional layer further enhances spatial awareness, extracting edge and shape contour information of the part. A third 3×3 convolutional layer captures local details before entering the cross-stage partial connections (CSP) module. This module achieves cross-fusion of high and low-level semantic information through feature splitting and residual connections, thus reducing computation while maintaining gradient stability. The network then employs a three-layer deconvolutional structure with 3×3 and 9×9 convolutional kernels to restore the feature-space resolution and upsample the feature map to the grasping pose mapping. Each TConv layer has 8 or 16 convolutional kernels to progressively restore positional information and the depth feature distribution in the target grasping region. The network's final output layer generates a grasping prediction tensor through a 9×9 convolution, containing information such as grasping

position, angle, and confidence. Regression optimization is then used to obtain the final optimal grasping pose. By combining MSF extraction with CSP-based feature fusion, the network can accurately identify graspable regions while maintaining efficient inference speed. This architectural design enables the model to achieve an average grasping time of 0.84 s and improved grasp stability compared with the baseline models, as reported in Table 2. This demonstrates its suitability for real-time robotic manipulation in industrial automation scenarios.

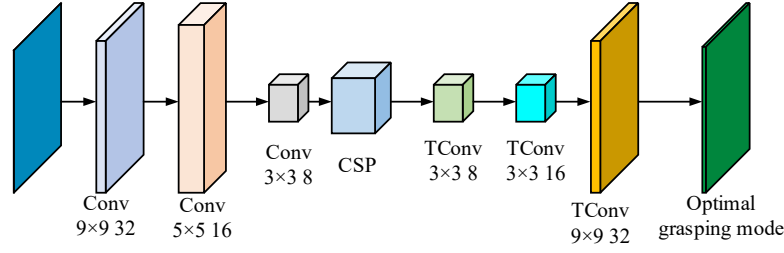


Fig. 6. Structure of part grasping model based on GGCNN

3. Results

3.1. Performance Analysis of Image Matching Model Based on SuperPoint

The experimental hardware configuration consisted of an Intel Core i5-8750H CPU, an NVIDIA GeForce GTX 2080 Ti GPU with 8 GB of VRAM, and 16 GB of RAM. The experiments were implemented using the PyTorch deep learning framework. The dataset used in this study was the publicly available Dex-Net dataset, which contains approximately 6.7 million synthetic depth images and grasping candidate samples generated from simulation environments. The dataset was divided into training, validation, and testing subsets in a 70%:15%:15% ratio to ensure a fair evaluation of model performance. During training, the network parameters were optimized using the Adam optimizer with an initial learning rate of 0.001, a batch size of 32, and 50 training epochs. Early stopping based on validation loss was employed to avoid overfitting and improve training stability. The dataset consists of three parts: ① an object model library containing approximately 1500 standardized 3D object meshes. ② point cloud and depth map data rendered under different viewpoints, noise, and occlusion conditions. ③ grasping candidates and quality labels, providing feasibility and success-probability scores for each candidate grasp pose. Dex-Net provides depth images rendered from different viewpoints, with varying noise levels and object poses, making it suitable for evaluating grasp prediction algorithms. In this study, depth images were used as the primary input modality for grasp prediction, while RGB images were primarily used in the feature-matching stage of the SuperPoint-based visual perception module. Although the Dex-Net dataset was generated in simulation and did not directly simulate lighting variations, the robustness of the proposed perception module under complex lighting conditions was evaluated through image-matching experiments using RGB images captured in industrial environments. In addition to the dataset-based evaluation, a robotic grasping platform was used to verify the practical feasibility of the proposed method. The experimental platform consisted of a six-axis industrial robot equipped with a parallel gripper and an industrial RGB camera mounted above the working table. The camera captured images of the target parts, and the proposed visual perception and grasp prediction framework was used to determine the optimal grasp pose. The predicted grasp parameters were then transmitted to the robot controller, which executed the grasping operation in real time. Several types of industrial parts, including gears, nuts, and connecting rods, were used in the experiments to evaluate the robotic system's grasping performance under different orientations and positions. The experimental results confirmed that the proposed method could be effectively deployed on a robotic grasping platform and achieve stable grasping performance. The study selected Scale-Invariant Feature Transform (SIFT) and Oriented Fast and Rotated BRIEF (ORB) as comparative models, and the outcomes are presented in Fig. 7. To ensure reliable evaluation, the model performance was assessed using both accuracy-related metrics and error-based indicators, including matching accuracy, repeatability, Mean Squared Error (MSE), Mean Absolute Error (MAE), and localization error. The validation set was used for hyperparameter tuning and model selection, while the final performance reported in Tables 1 and 2 was obtained from the independent testing set.

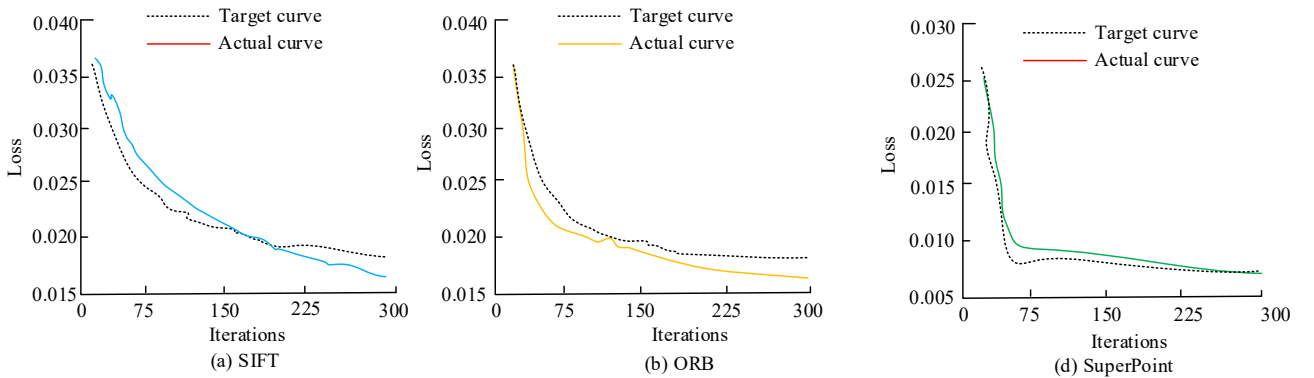


Fig. 7. Analysis of model loss variation

Fig. 7(a) shows the loss change process of the SIFT model in 300 iterations, Fig. 7(b) shows the training loss curve of the ORB model, and Fig. 7(c) shows the loss change trend of the SuperPoint model. As shown in Fig. 7(a), the initial loss value was approximately 0.038. The model's final minimum loss was 0.003 higher than the target curve. As shown in Fig. 7(b), the initial loss was 0.037, indicating faster convergence than SIFT. The convergence accuracy still deviated from the target curve by 0.002 in the later stages. As shown in Fig. 7(c), the initial loss was 0.028, almost coinciding with the target curve. SuperPoint effectively improved the discriminability of feature points. The error analysis of each model is presented in Fig. 8.

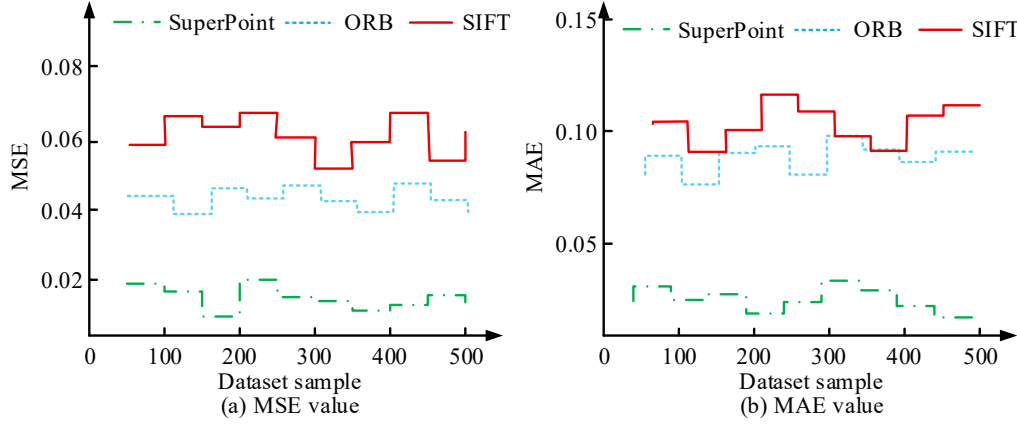


Fig. 8. Model error analysis

Fig. 8(a) shows the MSE of the three models under various data samples, and Fig. 8(b) shows the trend of the MAE of the three models under the same data samples. As shown in Fig. 8(a), the SIFT model exhibited MSE values ranging from 0.055 to 0.065. The ORB model had a relatively low MSE value. The SuperPoint model's curve was smooth and converged quickly. As shown in Fig. 8(b), the SIFT model had MAE values concentrated between 0.10 and 0.12. The ORB model had an MAE of 0.07-0.09. The SuperPoint model had an MAE stable between 0.02 and 0.04. The proposed model had the best performance. The performance analysis results of each model are presented in Table 1.

In Table 1, SuperPoint achieved a matching accuracy of 97.8%. Regarding repetition rate, SuperPoint achieved 92.5%, significantly higher than SIFT and ORB. Regarding localization accuracy, SuperPoint's average localization error was only 0.93 pixels. In terms of computational efficiency, SuperPoint's feature detection took only 12.1 milliseconds per frame, and its matching time was 14.8 milliseconds. SuperPoint demonstrated the best performance in regard to accuracy, speed, stability, and resource utilization.

3.2. Performance Analysis of Parts Grasping Model Grounded in Improved CNN

To further confirm the model's capability, CNN-CSPNet and GGCNN were selected as comparison models, and the results are presented in Fig. 9.

Table 1. Alignment model performance analysis

Model	SIFT	ORB	SuperPoint
Matching accuracy (%)	87.6	91.3	97.8
Repeatability (%)	79.4	83.7	92.5
Mean localization error (px)	1.82	1.46	0.93
MSE	0.058	0.041	0.019
MAE	0.112	0.085	0.034
Feature detection speed (ms/frame)	35.7	18.4	12.1
Feature matching time (ms/frame)	41.3	22.6	14.8
Number of keypoints	1650	1420	1260
Memory usage (MB)	182	109	95

Fig. 9(a) shows the accuracy trends of the three models under different iteration numbers, and Fig. 9(b) shows the root mean square error (RMSE) of the three models as a function of iteration numbers. In Fig. 9(a), the GGCNN-CSPNet model consistently achieved the highest accuracy, reaching 0.75 after 10 iterations and increasing to 0.96 after 50 iterations. After introducing the CSPNet module, the network's MSF fusion and gradient flow optimization mechanism significantly enhanced the model's feature representation capabilities. As shown in Fig. 9(b), the GGCNN-CSPNet model exhibited the most significant error reduction, rapidly decreasing from 0.32 to 0.06. CNN-CSPNet decreased from 0.36 to 0.09, while

GGCNN decreased slowly from 0.42 to 0.15. The CSPNet module made the model more likely to converge during training and reduced prediction bias. Fig. 10 shows the grasping rates for eight different part types.

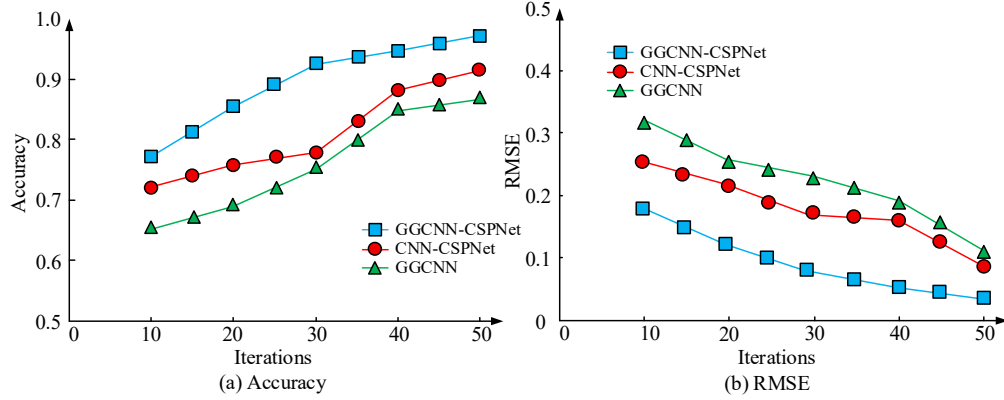


Fig. 9. Model accuracy and root MSE analysis

Figs. 10(a), 10(b), and 10(c) show a comparison of the three models' success rates in grasping eight types of parts. As shown in Figs. 10, the GGCNN-CSPNet model's grasping rates for regular parts such as gears, nuts, and connecting rods were 97.6%, 96.2%, and 95.8%, respectively, demonstrating extremely high stability. In contrast, the CNN-CSPNet model achieved an average grasping rate of 89.4%, whereas the traditional GGCNN model achieved only 82.7%. GGCNN-CSPNet had the shortest overall execution time, averaging 0.84 seconds, while CNN-CSPNet and GGCNN took 1.52 seconds and 4.06 seconds, respectively.

Fig. 11 compares the grasp prediction results of three models in the same scene, where black objects represent industrial parts and red line segments indicate the predicted grasp positions and orientations. In Fig. 11(a), the proposed GGCNN-CSPNet model produced grasp lines that aligned well with the objects' geometric structure, and the predicted grasp centers were located in stable contact regions. This indicated that integrating CSPNet and MSF fusion enhanced feature representation and improved the accuracy of grasp pose estimation. In Fig. 11(b), the CNN-CSPNet model also generated reasonable grasp predictions for most parts; however, slight deviations in grasp angle and position could be observed for certain objects. This suggested that although CSPNet improved feature extraction and training stability, the absence of a dedicated generative grasp prediction structure limited its ability to handle complex object poses. In Fig. 11(c), the traditional GGCNN model exhibited greater deviations in grasp orientation and position across several parts, particularly for objects with irregular shapes or rotated poses. This reflected the limitations of the baseline network in extracting robust features under complex scene conditions. Overall, the comparison demonstrated that the GGCNN-CSPNet model yielded the most accurate and stable grasp predictions among the three methods, indicating that the proposed architectural improvements effectively enhanced robustness and reliability in grasping for robotic manipulation tasks. The overall performance of each model was analyzed, and the outcomes are given in Table 2.

Table 2 shows that GGCNN-CSPNet achieved a 94.8% success rate in grasping. In terms of real-time performance, GGCNN-CSPNet's average grasping time was 0.84 seconds. In conclusion, this model achieved optimal performance in grasping accuracy, speed, stability, and computational efficiency.

4. Conclusion

This study developed an engineering-oriented robotic grasping framework that integrates SuperPoint-based image matching with an improved CNN grasp prediction model. The model demonstrated rapid convergence and stable performance improvement. It showed the most significant error reduction, rapidly decreasing from 0.32 to 0.06. Its grasping rates for regular parts such as gears, nuts, and connecting rods were 97.6%, 96.2%, and 95.8%, respectively. The proposed framework demonstrated clear advantages in grasping accuracy, execution speed, and operational stability, highlighting its practical value for industrial robotic manipulation and flexible automation systems. Despite the encouraging experimental results, several limitations should be acknowledged. First, the primary evaluation was conducted using the Dex-Net dataset, which mainly contained simulated depth images. Although this dataset is widely used for robotic grasping research, simulated environments may not fully capture the complexity of real industrial production scenarios. Second, the experiments primarily focused on single-object grasping tasks under controlled conditions, whereas real production lines often involve cluttered scenes, dynamic object arrangements, and continuous-operation constraints. Third, the current framework mainly relies on visual perception, and the integration of multimodal sensing, such as force feedback or tactile sensing, has not been explored. Future research will focus on extending the proposed framework to more complex industrial environments, including cluttered object scenes and continuous production workflows. In addition, integrating multimodal sensing information and conducting large-scale real-world deployment experiments could further improve the reliability and adaptability of robotic grasping systems. Beyond algorithmic performance improvements, the proposed framework also demonstrated important practical implications for industrial robotic production systems. By improving feature detection stability and grasp pose prediction accuracy, the system could reduce grasping failures and repeated manipulation attempts in automated production lines. This contributed to higher operational efficiency and more stable production cycles. Furthermore, integrating image matching with CNN-based grasp prediction enabled the robotic system to adapt to parts with varying shapes, orientations, and surface characteristics, enhancing the flexibility of robotic manipulation in flexible manufacturing environments. This

capability is particularly valuable in modern production systems where small-batch and multi-type manufacturing tasks are increasingly common. In addition, the improved perception robustness and faster execution time of the proposed framework can enhance the reliability of robotic grasping operations in practical deployment scenarios. By maintaining stable grasping performance under different viewpoints and environmental variations, the system can reduce unexpected interruptions in automated workflows and improve the overall stability of robotic production systems.

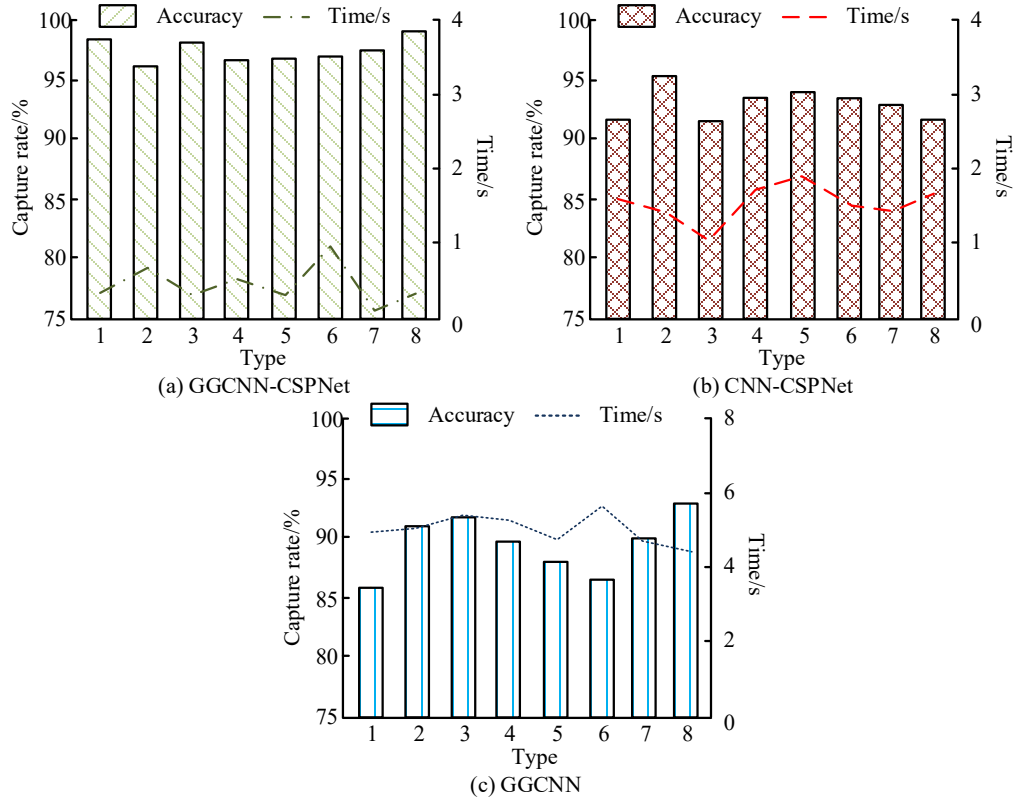


Fig. 10. Analysis of model capture rate and capture time

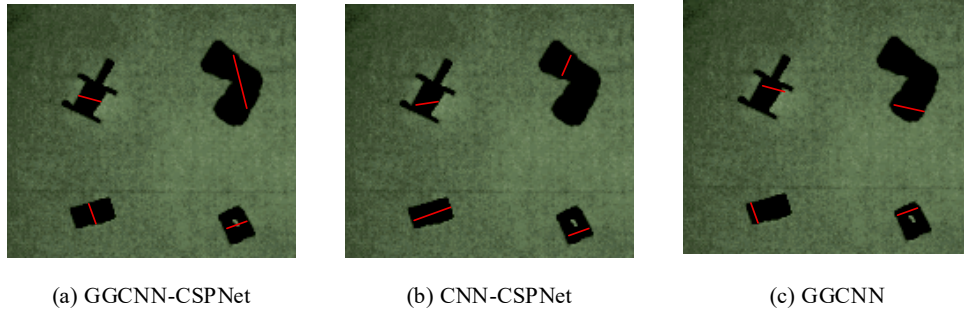


Fig. 11. Simulation experiment diagram

Table 2. Model performance analysis

Model	GGCNN	CNN-CSPNet	GGCNN -CSPNet
Grasping success rate (%)	82.7 00	89.4 00	94.8 00
Grasping positioning accuracy (mm)	1.84 0	1.27 0	0.93 0
RMSE	0.142	0.098	0.063
MAE	0.118	0.081	0.049
Average grasping time (s)	1.26 0	1.020	0.840
Pose prediction accuracy (%)	86.300	91.700	96.500
Repeat grasping stability (%)	81.500	88.600	93.200
Model parameters (M)	8.300	6.900	6.400
Computational resource usage (%)	92.000	78.000	69.000

Funding

The research is supported by: University level research fund, research and application of BP neural network in machine vision image processing (No. KJ202331)

Institutional Review Board Statement

Not applicable.

Declaration of AI and AI-Assisted Tools

The authors confirm that no AI tools were used in the preparation of this manuscript.

References

- Chen, L., Gao, J., and Lopes, A. M. (2023). Adaptive fractional-order genetic-particle swarm optimization Otsu algorithm for image segmentation. *Applied Intelligence*, 53(22), 26949-26966. DOI:10.1007/s10489-023-04969-8
- Gao, Y. and Zhou, Z. (2022). Automatic Recognition and Repair System of Mural Image Cracks Based on Cloud Edge Computing and Digitization. *Mobile Information Systems*, 22(31), 1534596.1-1534596.12. DOI:10.1155/2022/1534596
- Gtifa, W. and Sakly, A. (2023). Integrating Xilinx FPGA and intelligent techniques for improved precision in 3D brain tumor segmentation in medical imaging. *Journal of Real-Time Image Processing*, 20(6), 115.1-115.15. DOI:10.1007/s11554-023-01372-x
- Hong, H., Choi, D., Kim, N., Lee, H., and Kang, B. (2024). Survey of convolutional neural network accelerators on field-programmable gate array platforms: architectures and optimization techniques. *Journal of Real-Time Image Processing*, 21(3), 156-167. DOI:10.1007/s11554-024-01442-8
- Kalaiaarasu, S. and Natarajan, S. (2022). Conducted Electromagnetic Interference Mitigation on Two-Stage Cascaded Boost (TSCB) DC-DC Converter Using FPGA Based DCPWM Technique for EV Applications. *Journal of Electrical Engineering & Technology*, 18(3), 2003-2013. DOI:10.1007/s42835-022-01264-3
- Kumar, K. P. K. and Paul, V. (2023). Complementary spatial transformer network for real-time 3D object recognition A tiny deep learning model in target space. *Journal of Real-Time Image Processing*, 20(5), 88.1-88.12. DOI:10.1007/s11554-023-01340-5
- Li, C., and Da, F. (2025). Refined dense face alignment through image matching. *The Visual Computer*, 41(1), 157-171. DOI:10.1007/s00371-024-03316-3
- Li, Y., Jiang, L., Li, X., and Wang, F. (2023). Non-contact clothing anthropometry based on two-dimensional image contour detection and feature point recognition. *Industria Textila*, 74(1), 67-73. DOI:10.35530/IT.074.01.202279
- Li, X., Quan, H. (2024). MVPCL: Multi-view prototype consistency learning for semi-supervised medical image segmentation. *The Visual Computer*, 41(3), 1841-1854. DOI:10.1007/s00371-024-03497-x.
- Lora-Rivera, R., Oballe-Peinado, O., and Vidal-Verdu, F. (2023). Texture Detection With Feature Extraction on Embedded FPGA. *IEEE Sensors Journal*, 23(11), 12093-12104. DOI:10.1063/5.0065642
- Lu, F., Jia, K., and Zhang, X. (2025). CRViT: Vision transformer advanced by causality and inductive bias for image recognition. *Applied Intelligence*, 55(1), 1-16. DOI:10.1007/s10489-024-05910-3
- Mann, F. A., Galonska, P., Herrmann, N., and Kruss, S. (2022). Quantum defects as versatile anchors for carbon nanotube functionalization. *Nature Protocols*, 17(3), 727-747. DOI:10.1038/s41596-021-00663-6
- Narang, S. and Tallur, S. (2022). Field-programmable gate array (FPGA) based programmable digital emulator of vibratory microelectromechanical systems (MEMS) gyroscopes. *The Review of Scientific Instruments*, 93(3), 035003. DOI:10.1063/5.0065642.
- Pilato, C., Fang, Z., Hara-Azumi, Y., and Hwang, J. (2022). Introduction to the Special Section on High-level Synthesis for FPGA: Next-generation Technologies and Applications. *ACM Transactions on Design Automation of Electronic Systems*, 27(4), 29-3. DOI:10.1145/3519279
- Preethi, P. and Mamatha, H. R. (2023). Region-Based Convolutional Neural Network for Segmenting Text in Epigraphical Images. *Artificial Intelligence Applications*, 1(2), 119-127. DOI:10.47852/bonviewAIA2202293
- Sunil, K. and Afzal, S. (2023). A modified probabilistic roadmap algorithm for efficient mobile robot path planning. *Engineering Optimization*, 55(9), 1616-1634. DOI:10.1080/0305215X.2022.2104840
- Vite-Chávez, O., Flores-Troncoso, J., and Olivera-Reyna, R. (2023). Improvement Procedure for Image Segmentation of Fruits and Vegetables Based on the Otsu Method. *Image Analysis and Stereology*, 42(3), 185-196. DOI:10.5566/ias.2939
- Wang, Y., Ji, Y., and Wang, W. (2024). Bi-channel attention meta learning for few-shot fine-grained image recognition. *Expert Systems with Applications*, 242(15), 4-7. DOI:10.1016/j.eswa.2023.122741
- Wang, Y., Huang, G., Lu, Z. (2025). HEDN: Multi-oriented hierarchical extraction and dual-frequency decoupling network for 3D medical image segmentation. *Medical & Biological Engineering & Computing*, 63(1), 267-291. DOI:10.1007/s11517-024-03192-y.
- Yang, Y., Shen, Z., Zhu, X., and Wang, Z. (2022). FPGA-based electronic system for the control and readout of superconducting quantum processors. *The Review of Scientific Instruments*, 93(7), 74701.1-74701.14. DOI:10.1063/5.0085467
- Zhang, H. (2025). Application of Artificial Intelligence Image Recognition and Blockchain Technology in Enhancing the Security of Cross-Border Financial Transactions. *Traitement du Signal*, 42(2), 414-425. DOI:10.18280/ts.420207



Tao Li is a lecturer at the School of Equipment Engineering, Shanxi Vocational University of Engineering Science and Technology, Jinzhong, Shanxi, China. She graduated from the North University of China, majoring in Pattern Recognition and Intelligent Systems. She has been teaching robotics for many years. Her research interests include industrial robot programming technology, machine vision, and robot system integration and application.