THIS IS AN EARLY ACCESS ARTICLE.
This article has been accepted for publication in a future issue of this journal,
but has not been fully edited. Content may change prior to final publication.

BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES
DOI: 10.24425/bpasts.2027.7968

Automatic Assembly and Control Method of Automotive Transmission Gear Shaft Based on Improved SAM Model and SSD Algorithm

Fan Li

Academic Affairs Office, Shaanxi College of Communications Technology, Xi'an, 710018, China
E-mail: lifan.1982@outlook.com

Abstract: Automotive transmission gear shaft assembly demands high precision, stability, and efficiency. Traditional methods struggle with occlusions, lighting, and interference. This research proposes a novel joint assembly and control method using enhanced "segmentation anything" and "single-stage multi-box detector" models. The detection stage employs a dual backbone and feature pyramid for multi-scale fusion, boosting accuracy via an attention mechanism. The segmentation stage improves feature representation to compute assembly pose from centroids and spindle direction. Impedance control enables force-displacement coordination. Results show the improved detector achieves 91.5% average precision, with 90.1% IoU and 93.5% Dice score, surpassing mainstream models. In end-to-end tests, the method attained a 98.6% success rate, 0.43 mm alignment error, 49.7 N peak insertion force, 5.4 s cycle time, and a 1.8% repair rate. Optimal tolerance tests at 50 μm clearance yielded a 98.8% success rate, 1.4% mismatch alarm, and 103 ms adjustment time. The method demonstrates high accuracy, stability, and robustness, significantly advancing automation in complex industrial assembly.

Keywords: Image segmentation; Object detection; Automatic assembly; Transmission gear shaft; Attention mechanism

1. Overview

The automobile transmission is the core component of the power transmission system, and its assembly quality directly determines the transmission efficiency, running stability, and service life of the entire vehicle [1,2]. As a key component, gear shafts require extremely high assembly accuracy, and in industrial production, they need to satisfy the assembly requirements of high precision, low impact, and high efficiency simultaneously. However, traditional manual assembly and rule-based visual methods often exhibit problems such as low efficiency, poor accuracy, and susceptibility to interference in complex working conditions, making it difficult to meet the high reliability and high cycle requirements of modern intelligent manufacturing [3,4]. Currently, single-stage detection methods are broadly utilized for real-time detection due to their fast speed, but they are prone to accuracy degradation in complex backgrounds and multi-scale object recognition. The two-stage detection method has advantages in accuracy, but it has high computational complexity and is difficult to meet the industrial production pace [5,6]. The segmentation method can finely characterize assembly parts, but there are still shortcomings in boundary recognition, detail modeling, and adaptability to complex interference. Therefore, in the automatic assembly scenario of gear shafts, a single detection or segmentation method is hard to simultaneously meet the requirements of accuracy, robustness, and real-time performance [7,8]. In the actual assembly process, there are often issues of occlusion, lighting changes, and complex background interference between assembly parts, leading to unstable Object Detection (OD) results. At the same time, the recognition of part boundaries and interfaces requires extremely high segmentation accuracy, and traditional methods often suffer from missed detections and misclassifications in complex boundaries. In addition, during the assembly process, it is necessary to control the insertion force and trajectory deviation while maintaining high precision. Traditional position control and force control methods still have limitations in stability and safety [9,10].

At present, research on OD models and control methods has made certain progress in the field of automatic assembly on production lines. Dai et al. designed a novel detection method based on machine vision for identifying and removing large foreign objects in coal flows on conveyor belts. This method achieved effective screening of multi-category large volume foreign objects by optimizing the visual detection process, but its sensitivity to small or concealed foreign objects was still insufficient in complex working conditions [11]. Sampaio et al. used Computer-Aided Design (CAD) models to automatically generate composite images to strengthen the robustness of industrial OD. This method has

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

significant advantages in data augmentation and model generalization, but there is still a gap between the authenticity and intra domain consistency of the generated images [12]. Citlak et al. constructed a real-time object classification system for enamel coating production line based on Mask Region-based Convolutional Neural Network (Mask R-CNN), achieving online recognition and sorting of coating products. However, this method relies on high-quality annotated data, and its adaptability to occlusion and lighting changes needs to be enhanced [13]. Mei et al. designed an improved binocular vision 3D collaborative manufacturing platform for non-holonomic constrained industrial mobile robots, which demonstrated outstanding performance in robot collaborative operation and 3D perception accuracy. However, the real-time performance and stability of the system in dynamic scenes are still limited [14]. Miki et al. developed a robot interface based on coordinate position source data, realizing the automation application of articulated robotic arms in production line tasks. This method has potential in task automation and data-driven control, but its generalizability in multi-task parallel and flexible production environments still needs further validation [15].

To strengthen the accuracy and efficiency of automatic assembly of automotive transmission gear shafts, a fusion of Improved Single Shot MultiBox Detector (ISSD) and Improved Segmentation Anything Model (ISAM) is proposed for automatic assembly and control. This method introduces a dual backbone network and a Feature Pyramid Network (FPN) structure in the detection stage to enhance the ability of multi-scale feature extraction and small OD. Multi-scale feature enhancement and attention mechanism in the segmentation stage are integrated to achieve fine segmentation of gear shaft boundaries and assembly interfaces. Impedance control law in the control phase is introduced to achieve dynamic coupling adjustment of force and displacement. The innovation of the research lies in proposing a detection framework combining dual backbone networks with FPN, and introducing attention mechanism and multi-scale enhanced segmentation model to achieve accurate recognition of assembly interfaces and detailed features. In addition, the control method combining pose correction and mechanical constraints achieves collaborative optimization in assembly accuracy, insertion force control, and assembly rhythm. Research can help reduce the scrap rate of parts and the cost of manual intervention, thereby improving the overall economic benefits of the production line.

2. Methods and Materials

2.1 Construction of object detection model based on ISSD algorithm

In modern manufacturing, with the continuous improvement of product precision and output requirements, the efficiency and stability of manual assembly have become hard to meet the needs of large-scale production. In the area of automatic assembly, precise OD is the key to achieving efficient automated production [16,17]. The research adopts ISSD algorithm as the OD model to improve the accurate recognition and positioning ability during the assembly process of transmission gear shaft. Traditional SSD algorithms often face challenges such as occlusion between targets, scale differences, and complex backgrounds when dealing with multi-OD [18-20]. The ISSD algorithm improves the accuracy of OD in complex scenes by improving the network structure and loss function. Specifically, the proposed OD model introduces a dual backbone network to achieve parallel extraction of features at different levels and enhance the model's multi-scale adaptability. At the same time, combined with the feature pyramid model, target prediction is performed at multiple scales to enhance the detection ability of small targets. The structure diagram of the SSD OD model is denoted in Figure 1.

In Figure 1, the SSD OD model uses a Visual Geometry Group-16 (VGG-16) convolutional neural network as the backbone network for feature extraction. The image data is first processed through the first few layers of VGG-16 (Conv-3 layer) to extract preliminary features. Subsequently, the model further extracts multi-scale features through additional convolutional layers to adapt to OD of different sizes. Finally, a classifier is used to classify and predict each feature map, generating the category of each target and its corresponding bounding box. To enhance detection accuracy, the model uses non-maximum suppression techniques to eliminate redundant boxes [21,22]. To further enhance the detection accuracy and robustness of the model in multi-scale OD, an improved scheme based on parallel backbone network is proposed, as shown in Figure 2.

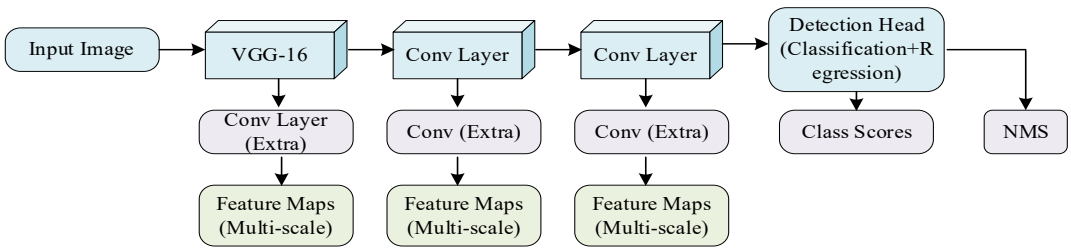


Figure 1 Structure diagram of SSD object detection model

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

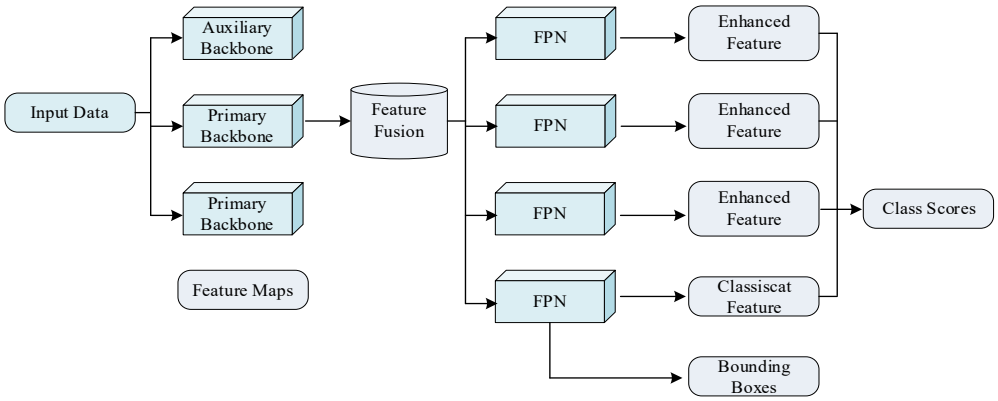


Figure 2 Parallel backbone network architecture diagram

Figure 2 shows the structure of a parallel backbone network, which consists of an auxiliary backbone network and a main backbone network. The input data is first subjected to feature extraction through an auxiliary backbone network, and a preliminary feature map is generated. The feature map of the auxiliary backbone network is fused with the feature map of the main backbone network. The main backbone network further extracts and processes higher-level features, ultimately generating outputs for OD. Through this mechanism, the network can simultaneously process information from different levels, thereby improving the accuracy and robustness of multi-scale OD. In the feature fusion stage, F_v refers to the feature map extracted by VGG16, and F_r represents the feature map extracted by ResNet50. After aligning the channels, the two are fused element by element weighted fusion, as shown in equation (1).

$$F = \alpha \cdot F_v + (1-\alpha) \cdot F_r \quad (1)$$

In equation (1), F represents the final fused feature map. $\alpha \in [0,1]$ denotes a learnable weight parameter used to dynamically balance the importance of two types of features. After optimizing the parallel backbone network, the FPN structure is further introduced to enhance the performance of the model in multi-scale OD. FPN improves the detection ability of small and large targets by combining feature maps of different scales [23-25]. The specific structure of FPN is denoted in Figure 3.

Figure 3 shows the structure of FPN, aimed at improving OD performance through Multi-Scale Feature Fusion (MSFF). The input feature maps in the figure are 19x19, 38x38, and 75x75 in size, and each feature map is processed by Conv2D convolution layers to extract higher-level features. The feature map adjusts its resolution through upsampling and downsampling operations to adapt to feature map fusion at different scales. Upsampling operation elevates low resolution feature maps to higher resolution, making it easier to combine with high-resolution feature maps. Downsampling reduces high-resolution feature maps to lower resolutions, enabling effective concatenation of feature maps at different scales. The expression for FPN MSFF is shown in equation (2).

$$P_l = Conv(F_l) + Up(P_{l+1}) \quad (2)$$

In equation (2), F_l is the feature map of the l th layer, and P_l is the fused feature map. Based on the above improvement plan, an ISSD OD model is proposed, and its structure is denoted in Figure 4.

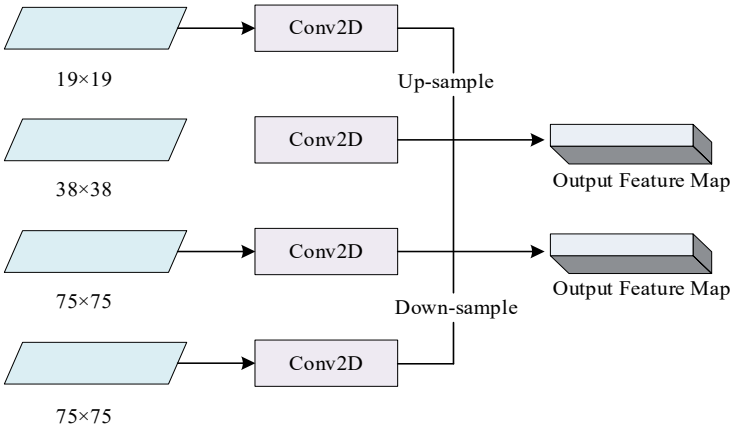


Figure 3 FPN structure diagram

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

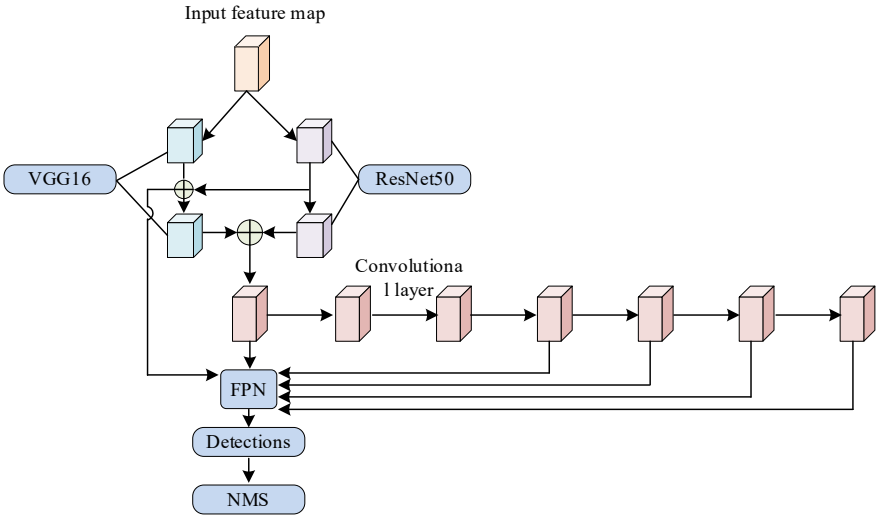


Figure 4 Structure diagram of ISSD object detection model

Figure 4 showcases the structure of the ISSD OD model. The input image first undergoes feature extraction through two backbone networks, VGG16 and ResNet50, to extract feature maps at different levels. VGG16 extracts low-level features through multiple convolutional layers, while ResNet50 further captures higher-level semantic information through its residual structure. These feature maps are fused at different network levels and integrate information from multiple scales through addition operations. The feature map is passed to the FPN module for MSFF, further enhancing the model's recognition ability for targets of different sizes. The fused feature map is used for OD tasks, including object category prediction and bounding box regression. The output of the model generates the final detection result through the decoding layer, ensuring accurate localization and classification of the target. The confidence calculation of ISSD candidate boxes is shown in equation (3).

$$s_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \quad (3)$$

In equation (3), s_i is the predicted probability of category i, z_i is the logit output by the classifier, and C is the total number of categories. The loss function is shown in equation (4).

$$L = L_{\text{cls}} + \lambda L_{\text{loc}} \quad (4)$$

In equation (4), L refers to the total loss function of ISSD. L_{cls} stands for classification loss, which measures the difference between predicted and true categories. L_{loc} is the regression loss, which measures the difference between the predicted bounding box and the true bounding box. λ is the balance coefficient that controls the relative importance of the two. The classification loss is shown in equation (5).

$$L_{\text{cls}} = -\sum_{c} p_c \log(\hat{p}_c) \quad (5)$$

In equation (5), p_c is the true category label of the target, $\hat{p}_c$ is the predicted probability of the target category, and c is the category index. The regression loss is shown in equation (6).

$$L_{\text{loc}} = \sum_{i} S_{L1}(t_i - \hat{t}_i) \quad (6)$$

In equation (6), t_i is the true position of the target box, and $\hat{t}_i$ means the predicted position of the target box. S_{L1} is the smoothed L1 loss, defined as equation (7).

$$S_{L1}(x) = \begin{cases} 0.5x^2, |x| < 1 \\ |x| - 0.5, |x| \geq 1 \end{cases} \quad (7)$$

In equation (7), x represents the difference between the predicted bounding box position and the true bounding box position. Non-maximum suppression measures the degree of overlap between candidate boxes by calculating their intersection to union ratio, as shown in equation (8).

$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (8)$$

In equation (8), A and B are two candidate bounding boxes, representing the pixel sets of the rectangular detection box.

2.2 Automatic assembly and control method of transmission gear shaft integrating ISAM model

After optimizing the OD module, further research will be conducted to introduce the ISAM model into the automatic assembly and control of the transmission gear shaft. Compared with SAM, ISAM can achieve finer boundary and interface segmentation in complex industrial scenarios through multi-scale feature enhancement and attention mechanisms, thereby avoiding boundary blurring problems [26-28]. ISAM can achieve fine segmentation based on the candidate regions detected, thus accurately distinguishing gear shafts from surrounding components in complex industrial scenes. The encoding end of ISAM adopts multi-scale feature enhancement to extract boundary and semantic information, and the decoding end combines attention

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

mechanism to enhance the response of key parts, ensuring that the output mask can clearly depict the contour of the gear shaft and the position of the assembly interface. The segmentation loss function of ISAM is defined as equation (9).

$$L_{seg} = -\frac{1}{|\Omega|}\sum_{j=1}^{|\Omega|}\left(y_j \log(p_j) + (1-y_j)\log(1-p_j)\right) \quad (9)$$

In equation (9), $|\Omega|$ represents the total amount of pixels in the mask area, y_j denotes the true category label of pixel j, and p_j means the predicted probability of the corresponding pixel. This loss is used to constrain the consistency between the segmentation results and the real mask. To further improve the accuracy of boundary segmentation, the introduction of Dice loss is studied, as shown in equation (10).

$$L_{dice} = 1 - \frac{2\cdot|M \cap \hat{M}|}{|M| + |\hat{M}|} \quad (10)$$

In equation (10), M stands for the predicted mask area, $\hat{M}$ indicates the real mask area, and $|M \cap \hat{M}|$ represents the number of overlapping pixels between the two. Dice loss can effectively measure mask similarity and avoid boundary blurring problems caused by relying solely on cross entropy. In the assembly process, the mask obtained by ISAM segmentation can be used to calculate the geometric parameters of the gear shaft. The calculation of its center of mass position is shown in equation (11).

$$(x_c, y_c) = \frac{1}{|\Omega|}\sum_{(x_j,y_j)\in\Omega}(x_j, y_j) \quad (11)$$

In equation (11), (x_j, y_j) represents the coordinates of the pixels in the mask, and (x_c, y_c) represents the centroid coordinates, which are used to estimate the position of the gear shaft. Furthermore, the main axis direction of the gear shaft is obtained through principal component analysis, as shown in equation (12).

$$\vec{d} = \arg\max_{\vec{v}} \frac{1}{|\Omega|}\sum_{(x_j,y_j)\in\Omega}((x_j, y_j) - \vec{\mu})^T \vec{v} \quad (12)$$

In equation (12), $\vec{\mu}$ is the centroid vector, $\vec{v}$ is the candidate direction vector, and $\vec{d}$ is the final determined principal axis direction. Based on the centroid and directional information, a homogeneous transformation matrix for the pose of the gear shaft can be constructed, as shown in equation (13).

$$T = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \quad (13)$$

In equation (13), R means the rotation matrix derived from the main axis direction, t denotes the translation vector determined by the centroid position, and T is used to describe the pose of the gear shaft in the assembly coordinate system. In the path planning stage, the robot assembly trajectory is optimized by minimizing the path deviation, as shown in equation (14).

$$L' = \sum_{k=1}^{K} \| p_k - \hat{p}_k \|^2 \quad (14)$$

In equation (14), L' is the minimum path deviation. p_k is the actual path point, $\hat{p}_k$ is the expected path point, and K is the number of path points. This loss ensures that the robot can approach the expected trajectory during insertion and installation. In the control phase, the system adopts impedance control law to achieve force displacement coupling, as shown in equation (15).

$$F' = K'(x_d - x') + D(\dot{x}_d - \dot{x}) \quad (15)$$

In equation (15), F' is the control force. x_d, and x' are the expected position and actual position, respectively. $\dot{x}_d$ and $\dot{x}$ respectively represent the expected speed and actual speed, K' represents the stiffness matrix, and D represents the damping matrix. Impedance control enables the system to make real-time adjustments during assembly based on contact force and pose deviation. The overall structure of ISAM is denoted in Figure 5.

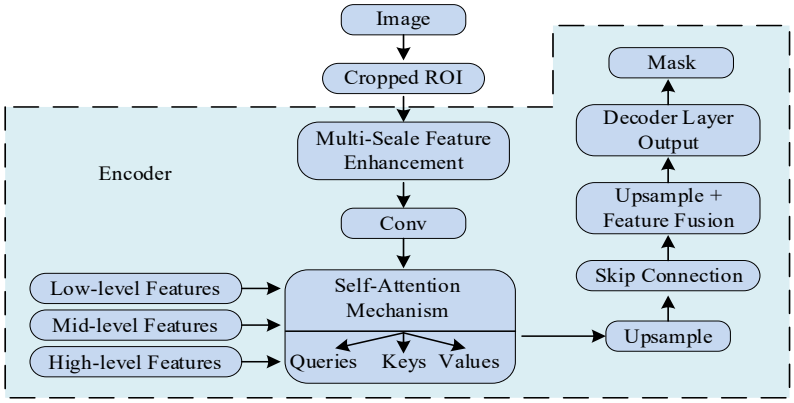


Figure 5 ISAM model structure diagram

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

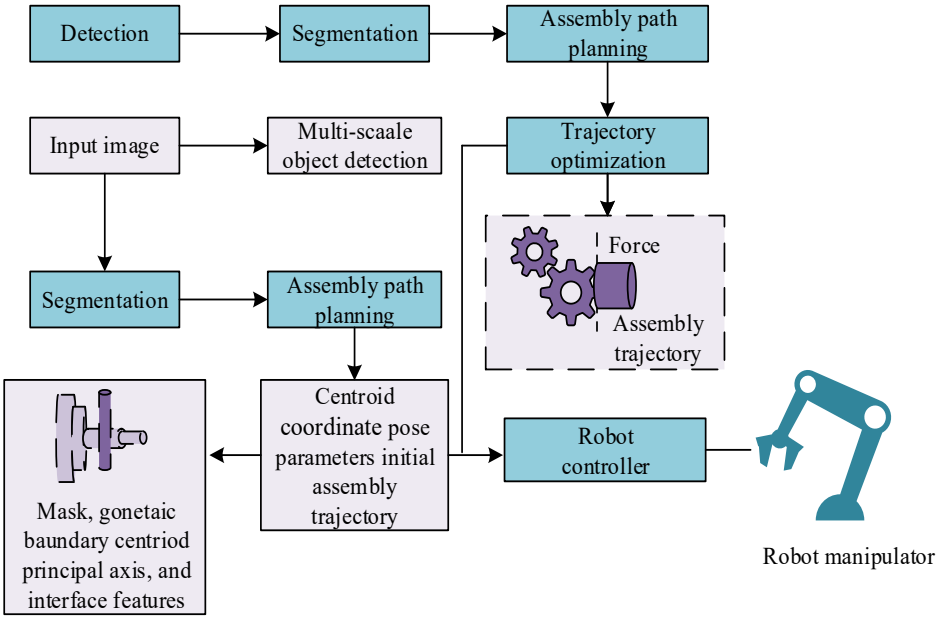


Figure 6 Automatic assembly process diagram of gear shaft based on ISAM-ISSD

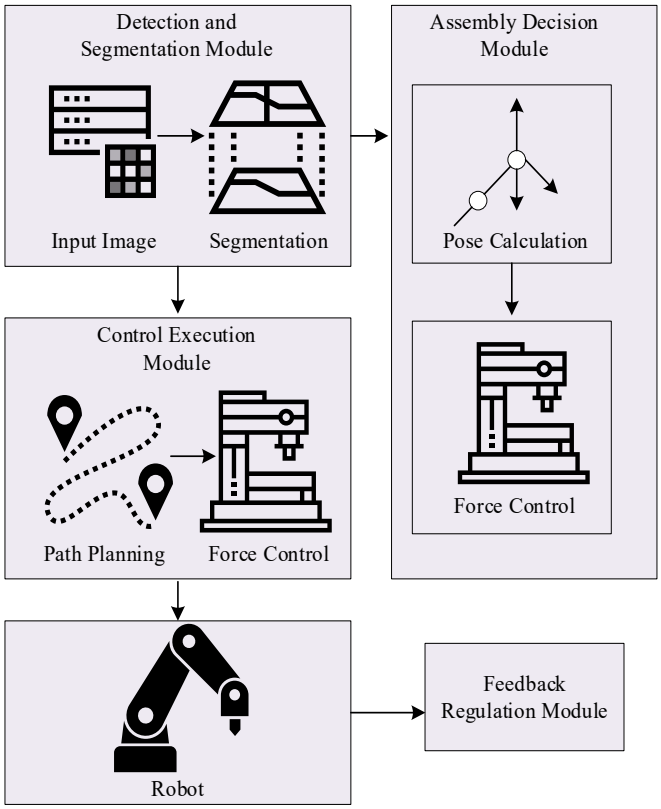


Figure 7 Framework diagram of automatic assembly and control system for gear shaft (Source from: https://www.svgrepo.com/svg/450985/image-server; https://www.svgrepo.com/svg/450984/image-segmentation; https://www.svgrepo.com/svg/322934/path-distance; https://www.svgrepo.com/svg/185065/industrial-robot-factory; https://www.svgrepo.com/svg/482984/robot-arm-1)

Figure 5 showcases the overall structure of the ISAM model in segmentation tasks. The input image is cropped into regions of interest through the detection module. At the encoding end, the model introduces a multi-scale feature enhancement module to integrate convolutional features from different scales to capture the edge details and global contour information of the gear axis. Low level features provide fine-grained descriptions of geometric boundaries, while high-level features depict semantic distinctions in complex backgrounds. The features are fed into the self attention mechanism module, which constructs a mapping relationship between queries, keys, and values, enabling the model to dynamically enhance the response of key components such as assembly interfaces within a spatial range. The decoding end adopts a step-by-step upsampling and skip connection method to restore the segmentation results layer by layer to the same

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

size as the input, and combines attention weighting mechanism to suppress the interference of irrelevant components. The final output mask not only contains the complete contour of the gear shaft, but also marks the key interface positions. The overall assembly process is denoted in Figure 6.

Figure 6 shows the complete process of automatic assembly of gear shafts based on the ISAM-ISSD model. Firstly, the detection module performs multi-scale OD on the input image to quickly locate candidate regions for gear shafts and related components. The segmentation module performs fine mask segmentation on the candidate regions, extracting the geometric boundaries, centroid, main axis direction, and interface features of the gear shaft. The assembly path planning module receives detection and segmentation results, combines position information with interface features, and generates an initial assembly trajectory by calculating centroid coordinates and pose parameters. The system optimizes the trajectory to ensure high alignment accuracy even in the presence of dimensional tolerances or slight offsets in the components. The optimized assembly parameters are transmitted to the robot controller, and the actuator drives the robotic arm to perform assembly actions. The framework of the control system is denoted in Figure 7.

In Figure 7, the system consists of a detection and segmentation module, an assembly decision-making module, a control execution module, and a feedback adjustment module. The detection module performs multi-scale object recognition on the input image to quickly locate the gear shaft and related components. The segmentation module performs fine mask segmentation on candidate regions to obtain accurate geometric contours and assembly interface features. These pieces of information are fed into the assembly decision-making module for calculating assembly paths, determining part postures, and aligning methods. In the control execution stage, the system combines assembly parameters with the robot kinematic model, and uses pose adjustment and force control constraints to achieve real-time control of the assembly process. The robot performs insertion operations based on the planned trajectory. When assembly resistance or position deviation is detected, the feedback adjustment module promptly corrects the motion trajectory and contact force to ensure the smoothness and safety of the assembly action.

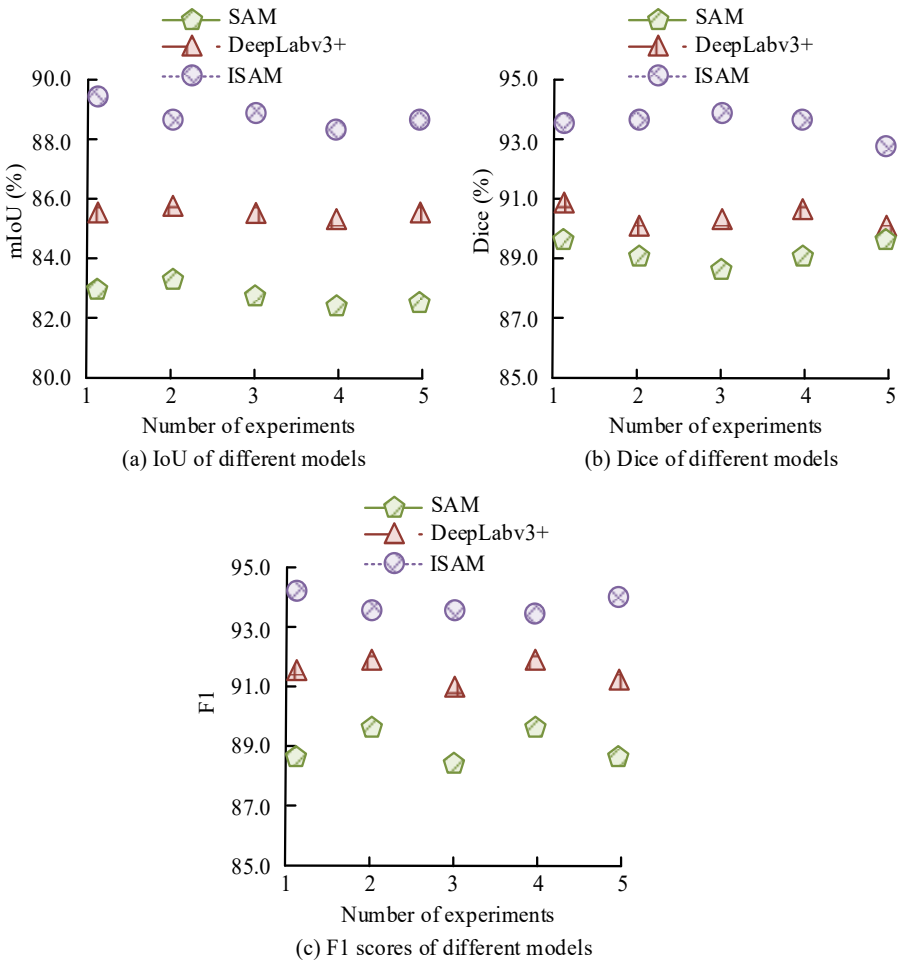
3. Results and Analyses

3.1 Performance analysis of ISAM-ISSD model

To verify the performance of the raised ISAM and ISSD algorithms, a systematic comparative experiment was designed and the model training and testing were completed in a unified experimental environment. In terms of hardware platform, the experiment used two NVIDIA RTX A6000 (48GB video memory) GPUs for parallel training, with Intel Xeon Gold 6338 (32 cores and 64 threads) CPU and 256GB system memory. The software environment configuration was Ubuntu 20.04 operating system, and the deep learning framework used PyTorch 1.13.1, CUDA version 11.7, combined with cuDNN 8.5 acceleration library to improve computing efficiency. During the training, mixed precision training and gradient pruning mechanisms were enabled to accelerate model convergence and minimize the risk of gradient explosion. The data preprocessing and enhancement part was developed and implemented based on OpenCV 4.6.0 and Python 3.9. The experiment used the gear-shaft subset of the publicly available Fraunhofer IPA Bin-Picking Dataset (https://www.bin-picking.ai/en/dataset.html). This subset contains 8,060 synthetic scenes and 230 real-world scenes with a resolution of 512 × 512 pixels and provides RGB images, depth images, point clouds, instance segmentation masks, visibility scores, and 6D pose annotations. The synthetic scenes were used for model training, while the real-world scenes were reserved for independent testing. The provided instance masks were used as the supervision for ISAM and were converted into the minimum enclosing rectangles required to train ISSD. The 6D pose annotations were used to verify the centroid and principal-axis-based pose calculation. Rotation, translation, scaling, illumination adjustment, partial occlusion, and random noise were applied to the training images to improve robustness under complex assembly conditions. The comparison algorithms selected were the Region-based Convolutional Neural Network with ResNet50 (Faster R-CNN) and You Only Look Once version 8 (YOLOv8). To comprehensively evaluate the detection and segmentation performance, the experiment used Mean Average Precision (mAP), Intersection over Union (IoU), Dice Similarity Coefficient (DSC), F1 score, Giga Floating Point Operations (GFLOPs) per second, and Frame Per Second (FPS) as indicators. The experimental results of ISAM segmentation performance comparison are shown in Figure 8.

As shown in Figure 8 (a), ISAM consistently scored higher than SAM and DeepLabv3+ in the mIoU index across five experiments, with an average of about 88.5%, significantly better than DeepLabv3+'s 85.4% and SAM's 82.1%. This indicates that ISAM has stronger advantages in overall segmentation accuracy and region overlap consistency, especially in the extraction of complex gear shaft boundary regions where it performs more stably. From Figure 8 (b), ISAM also maintained the best Dice coefficient, with an average score of about 93.2%, which was about 3.5% and 4.3% higher than DeepLabv3+ (89.7%) and SAM (88.9%), respectively. This indicates that ISAM has higher robustness in region overlap and boundary reconstruction, and can effectively alleviate common segmentation edge error problems. From Figure 8 (c), in terms of F1 score, ISAM was generally stable at over 93%, with the highest being close to 94.5%, which was better than DeepLabv3+ (about 91%) and SAM (about 89%). This indicates that ISAM is more excellent in detecting, recognizing, and segmenting small targets and detailed structures, and can better balance recall and accuracy. The experimental results showed that ISAM performed the best in the three indicators of mIoU, Dice, and F1 score, and the fluctuations were small in multiple experiments, verifying its dual advantages in segmentation accuracy and robustness. The comparative experimental results of ISSD detection performance are shown in Figure 9.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.



(a) IoU of different models

(b) Dice of different models

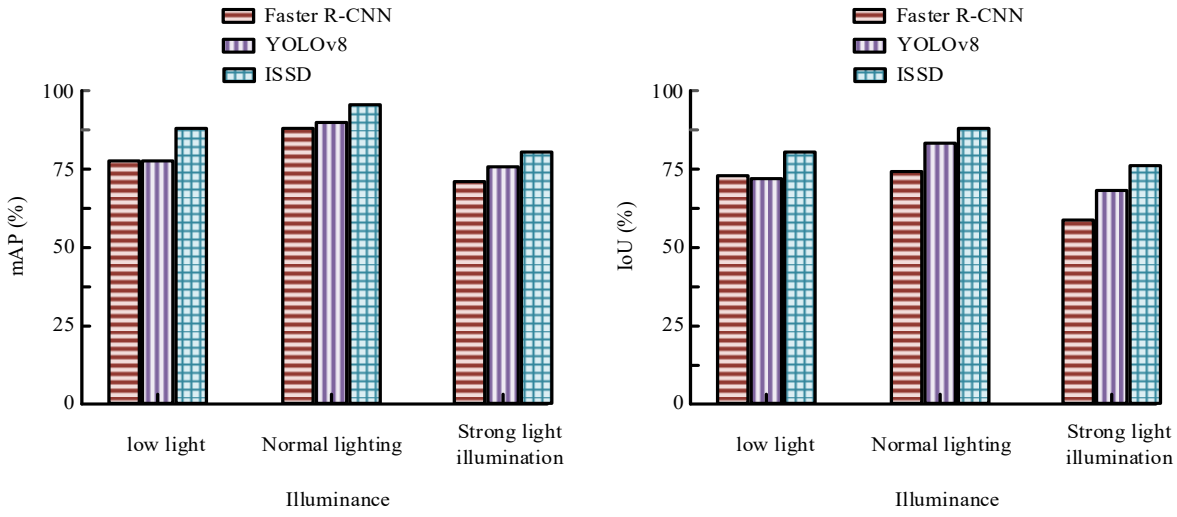
(c) F1 scores of different models

Figure 8 Comparison of ISAM segmentation performance experimental results

From Figure 9 (a), under different lighting conditions, the mAP performance of ISSD was always better than the comparison model, reaching the highest value of about 83% under normal lighting conditions, which was about 4% and 8% higher than YOLOv8 (about 79%) and Faster R-CNN (about 75%), respectively, indicating that ISSD has stronger robustness in OD accuracy. From Figure 9 (b), ISSD also maintained a leading position in the IoU index, with values above 80% under all three lighting conditions, especially in strong light environments, indicating its better adaptability to light interference. According to Figure 9 (c), YOLOv8 performed the best in FPS metrics, consistently maintaining above 70, significantly higher than ISSD (about 60) and Faster R-CNN (less than 20). The experimental results showed that ISSD had more advantages in accuracy and robustness, while YOLOv8 performed outstandingly in detection speed, and Faster R-CNN had significant performance disadvantages in complex lighting environments. The SAM-ISSD model ablation experiment is denoted in Table 1.

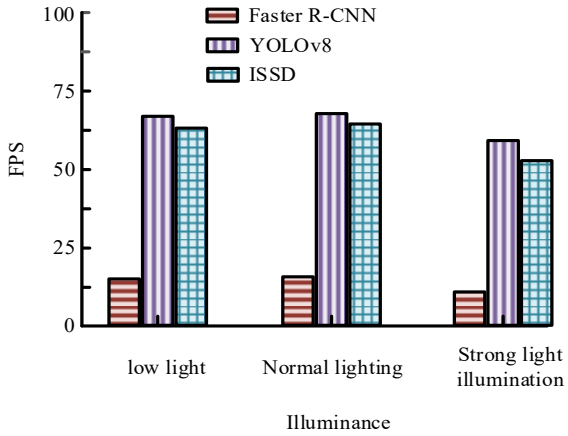
As shown in Table 1, after gradually introducing different modules, the overall performance of ISAM-ISSD continued to improve. In the combination of basic SSD and SAM, the mAP, IoU, Dice, and F1 scores were 87.9%, 85.6%, 88.1%, and 89.0, respectively. After introducing a dual backbone network, the four indicators improved to 89.7%, 87.3%, 89.4%, and 90.6% respectively, indicating that MSFF could enhance the representation ability of candidate regions and boundaries. When further incorporating FPN structure, mAP and IoU increased to 90.6% and 88.1%, respectively, and Dice and F1 score also increased to 90.2% and 91.4, indicating that multi-scale feature transfer effectively improved the fine-grained segmentation performance of complex targets. After adding attention mechanism, the four indicators of the model reached their highest values, with mAP increasing to 91.5%, IoU and Dice reaching 90.1% and 93.5% respectively, and F1 score reaching 94.0. The overall results indicate that each module has made significant contributions to performance improvement. The performance evaluation under different interference conditions is denoted in Figure 10.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.



(a) MAP under different lighting conditions

(b) IoU under different lighting conditions



(c) FPS under different lighting conditions

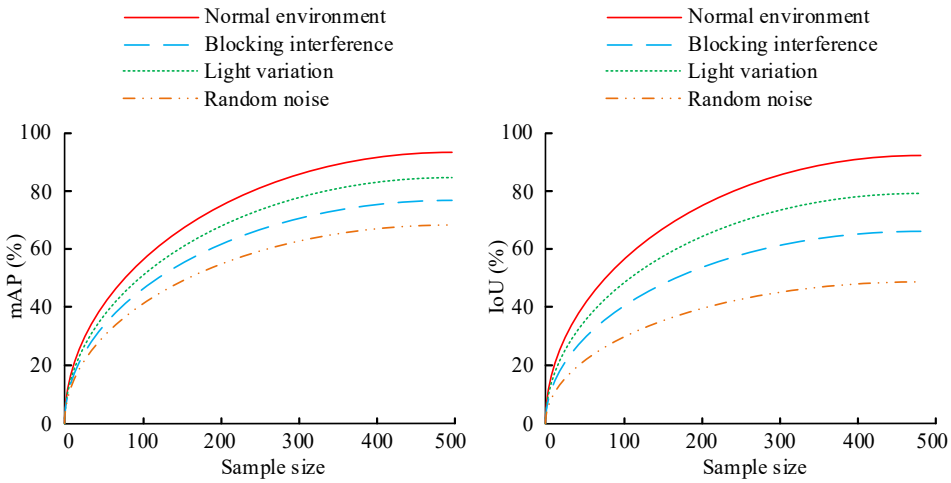
Figure 9 Comparison of ISSD detection performance experimental results

TABLE 1 ISAM-ISSD MODEL ABLATION EXPERIMENT

Experimental setup	mAP(%)	IoU(%)	Dice(%)	F1 score
Basic SSD+SAM	87.9	85.6	88.1	89.0
+Dual backbone network	89.7	87.3	89.4	90.6
+FPN structure	90.6	88.1	90.2	91.4
+Attention mechanism	91.5	90.1	93.5	94.0

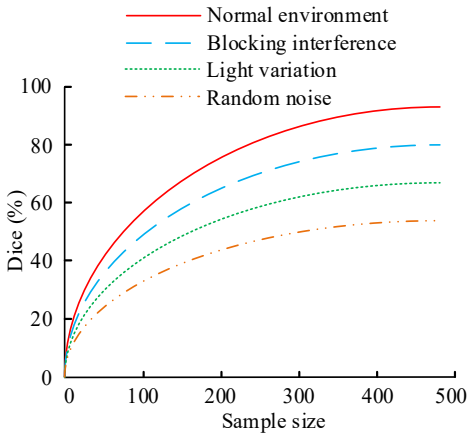
From Figure 10 (a), in the mAP index evaluation, the model performed best in normal environments. As the sample size gradually increased from small to large, the mAP value rapidly increased and tended to stabilize, eventually approaching 95%. In contrast, under the conditions of occlusion interference and lighting changes, mAP remained lower than normal environment, decreasing by about 8% and 10% respectively. However, it performed the worst under random noise conditions, with a maximum difference of over 15%, indicating that random noise had the most significant destructive effect on detection accuracy. As shown in Figure 10 (b), in terms of the IoU index, the curve remained the highest under normal conditions, and as the sample size increased, the IoU stabilized at around 90%. Both occlusion interference and lighting changes had a certain impact on IoU, with a greater decrease in lighting changes, indicating that the model had deviations in locating target boundaries under complex lighting conditions. Under random noise conditions, IoU was the lowest and could only be maintained at around 70%. According to Figure 10 (c), on the Dice metric, the model still performed the best in normal environments, with an average value close to 92%. The occlusion interference and lighting changes decreased by about 6% and 8% respectively, while Dice was the lowest under random noise conditions, only slightly exceeding 70% when the sample size increased to 500. This indicates that the noisy environment not only weakens the overall segmentation accuracy, but also affects the ability to distinguish the target area from the background.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.



(a) MAP of models under different interference conditions

(b) IoU of models under different interference conditions



(c) Dice of models under different interference conditions

Figure 10 Performance evaluation under different interference conditions

3.2 Analysis of gear shaft automatic assembly system based on ISAM-ISSD model

The study further analyzed the practical effectiveness of the ISAM-ISSD model in the automated assembly and control of gear shafts. The experiments introduced Key Performance Indicators (KPIs) in the assembly domain, including assembly accuracy, misalignment error, assembly success rate, and average assembly time, to comprehensively evaluate the feasibility and robustness of this method in industrial applications. All experiments were conducted on an industrial robot platform, with the assembly task involving the insertion of a gear shaft into a mating part. The Fraunhofer IPA dataset was used only to evaluate the visual detection, segmentation, and pose-calculation modules. The assembly success rate, alignment error, insertion force, cycle time, repair rate, and control stability reported in Figures 11–12 and Tables 2–3 were obtained from 500 independent physical assembly cycles conducted using automotive transmission gear shafts on the robotic platform. Testing was performed under various disturbance conditions and operational scenarios. The end-to-end assembly KPI comparison results are denoted in Table 2.

TABLE 2 END-TO-END ASSEMBLY KPI COMPARISON

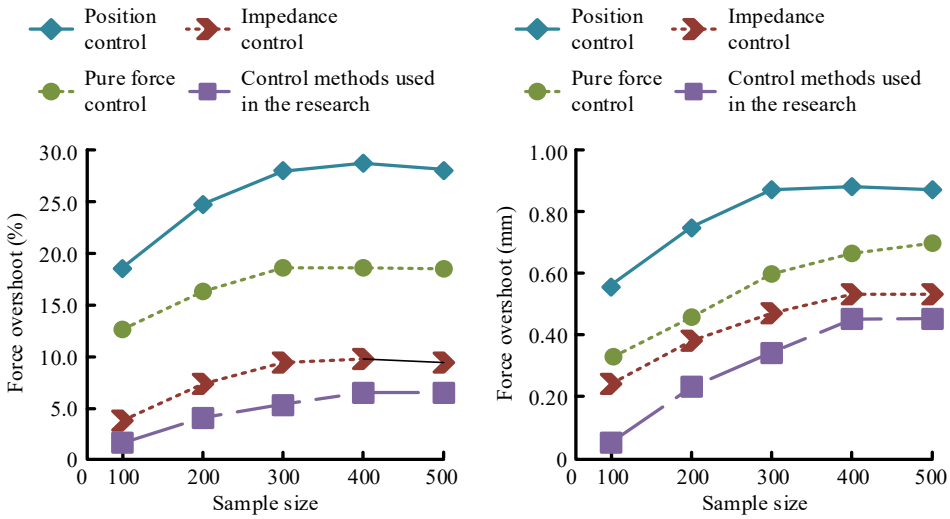
Plan	Assembly success rate (%)	Alignment error (mm)	Peak insertion force (N)	Single piece rhythm (s)	Repair rate (%)
Traditional rule vision+position control	92.4	0.92	66.8	6.8	5.7
YOLOv8+SAM+position time-sharing control	95.3	0.71	58.4	6.1	3.9
Method used in the research	98.6	0.43	49.7	5.4	1.8

According to Table 2, in terms of assembly success rate, the method used in the study reached 98.6%, which was 6.2%

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

higher than traditional rule-based visual+position control (92.4%) and 3.3% higher than YOLOv8+SAM+position time-sharing control (95.3%), indicating that this method has stronger success rate guarantee capability in end-to-end assembly tasks. In terms of accuracy indicators, the alignment error of the research method was only 0.43 mm, significantly lower than the traditional method's 0.92 mm and the comparative method's 0.71 mm, indicating that the designed method has higher alignment accuracy in feature extraction and pose correction. In terms of mechanical performance, the "peak insertion force" of the research method was controlled at 49.7 N, which was lower than the traditional method (66.8 N) and the comparative method (58.4 N), reducing the risk of mechanical impact and part damage during assembly. In terms of efficiency and quality, the "single piece cycle time" of the research method was shortened to 5.4 seconds, which was significantly optimized compared to the traditional rule-based method (6.8 seconds) and YOLOv8+SAM method (6.1 seconds). At the same time, the repair rate was only 1.8%, lower than the traditional scheme (5.7%) and the comparative scheme (3.9%). Experimental data showed that the end-to-end assembly method proposed by the research was significantly superior to existing methods in key KPIs such as assembly success rate, accuracy, mechanical control, and cycle and repair rate, reflecting its comprehensive advantages of high reliability and efficiency in practical production environments. The stability assessment results are shown in Figure 11.

As shown in Figure 11(a), under the force control overshoot index, the overshoot of the position control method was the most severe. As the sample size gradually increased from 100 to 500, the overshoot amplitude rapidly increased from about 18% to nearly 29%, and tended to saturate when the sample size is large. In contrast, the overshoot of pure force control and impedance control was significantly lower, ultimately stabilizing at around 19% and 11%, respectively. The control method adopted by the research performed the best, consistently maintaining below 5%, indicating that it can effectively suppress the instability caused by contact stiffness and force feedback during the assembly process. From Figure 11(b), in terms of trajectory deviation indicators, position control still performed the worst, with the highest deviation approaching 0.9mm, indicating that its trajectory tracking accuracy is relatively low when dealing with external interference. The deviation between pure force control and impedance control was relatively small under the same conditions, and ultimately stabilized at around 0.7mm and 0.55mm, which could balance pose and force to a certain extent. The control method proposed by the research had the lowest trajectory deviation, only about 0.45mm when the sample size reaches 500, significantly better than other methods, indicating that this method has a stronger suppression effect on trajectory errors under complex assembly conditions. The experimental results showed that there were significant shortcomings in both accuracy and stability of position control, while traditional pure force control and impedance control, although improved, still could not meet the requirements of high-precision gear shaft assembly. The research method maintained the lowest level in both force control overshoot and trajectory deviation indicators, verifying its advantages in improving the stability and safety of automatic assembly. The evaluation results of production line availability and cycle time are shown in Figure 12.

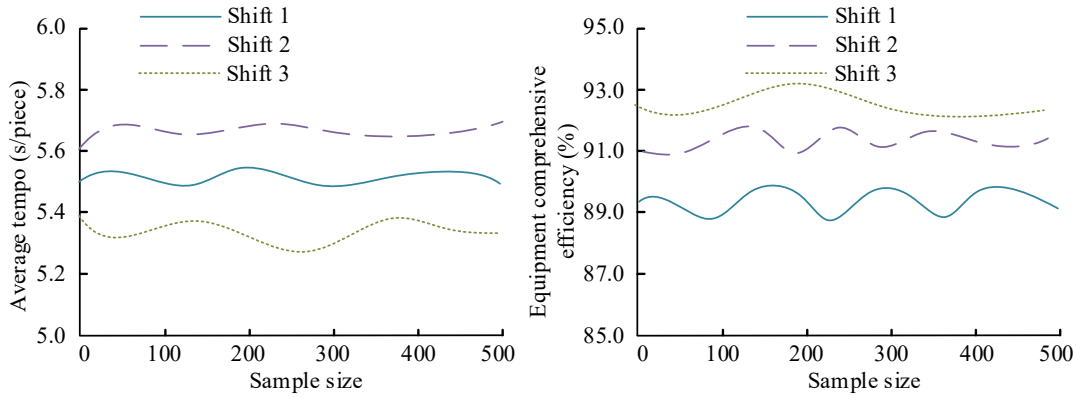


(a) Force overshoot of different control methods

(b) Trajectory deviation of different control methods

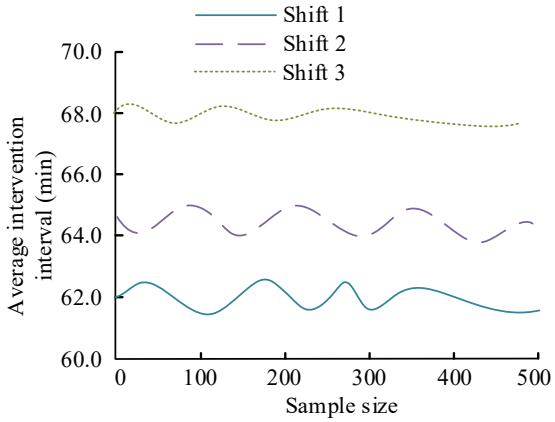
Figure 11 Results of control stability assessment

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.



(a) The average pace of different shifts

(b) Comprehensive efficiency of equipment in different shifts



(c) The average intervention interval for different shifts

Figure 12 Production line availability and cycle evaluation

From Figure 12 (a), shift 1 overall remained between 5.4-5.6 with small fluctuations, indicating a relatively stable production process. The average beat of shift 2 was slightly higher than shift 1, about 5.6-5.8, but the fluctuation amplitude was smoother, indicating good consistency in efficiency. Shift 3 was generally lower, ranging from 5.3 to 5.5, indicating that its output efficiency per unit time was better, but there were also local fluctuations. From Figure 12 (b), shift 3 performed the best in terms of overall equipment efficiency, maintaining an average of over 92%, and the curve trend was relatively stable, indicating high equipment utilization and less downtime due to malfunctions. The efficiency of shift 2 remained relatively stable at around 91%. The efficiency of shift 1 was the lowest, fluctuating only between 89% and 90%, and the fluctuation amplitude was relatively large, indicating that there may be some uncertainty in equipment use and maintenance. According to Figure 12 (c), the average intervention interval for shift 3 remained stable at 68-69 minutes, indicating that the equipment operated more reliably and had the lowest intervention frequency during this shift. The experimental results showed that different shifts had high performance in production line utilization and stability, and the proposed method had good application potential. The impact of assembly tolerance changes on the system is shown in Table 3.

TABLE 3 EFFECTS OF ASSEMBLY TOLERANCE CHANGES ON THE SYSTEM

Coordination gap (μm)	Success rate (%)	Average insertion force (N)	Mismatch alarm rate (%)	End adjustment duration (ms)
30	97.9	53.2	2.1	118
50	98.8	49.7	1.4	103
80	98.4	47.3	1.7	96
120	97.1	45.6	2.8	95

According to Table 3, in terms of assembly success rate, when the coordination gap was within the range of 30-80 μ m, the success rate remained above 98%, with the highest value reaching 98.8% at 50 μm. When the gap further increasef to 120 μm, the success rate dropped to 97.1%, indicating that excessive assembly tolerances could lead to reduced assembly stability. In terms of average insertion force, as the assembly gap increased, the insertion force showed a decreasing trend,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

gradually decreasing from 53.2 N at 30 μm to 45.6 N at 120 μm. This indicates that larger gaps help reduce the mechanical force required during the assembly process, but too low insertion force may lead to loose bonding of the assembly interface, thereby affecting the assembly quality. In terms of false positive rate, the setting of 50 μm performed the best, only 1.4%, which has decreased compared to 2.1% for 30 μm and 2.8% for 120 μm, indicating that a moderate tolerance range helps reduce assembly deviation and error accumulation. In terms of end adjustment time, as the assembly gap increased, the time gradually shortened from 118 ms at 30 μm to 95 ms at 120 μm, indicating that under larger tolerance conditions, the system reduced the computational load for end alignment and compensation, thereby shortening the adjustment time. The experiment findings denoted that the overall performance of the system was optimal when the assembly gap was 50 μ m. When the gap was too small or too large, although there were certain advantages in insertion force and adjustment time, the assembly success rate and false positive rate were both affected.

4. Discussion

ISAM-ISSD has shown significant advantages in experimental results for automatic assembly tasks of automotive transmission gear shafts in complex environments. In terms of detection and segmentation performance, the mAP of the improved detector reached 91.5%, and the IoU and Dice coefficients were 90.1% and 93.5%, respectively, which were better than the comparison model. In the end-to-end assembly experiment, the assembly success rate was as high as 98.6%, the alignment error was only 0.43 mm, the peak insertion force was reduced to 49.7 N, the single piece cycle time was shortened to 5.4 s, and the repair rate was only 1.8%. In the assembly tolerance experiment, when the coordination gap was 50 μm, the system achieved optimal performance, with an assembly success rate of 98.8% and a false positive rate of only 1.4%. Meanwhile, stability control experiments have shown that this method maintains the lowest level in both force control overshoot and trajectory deviation indicators, verifying its advantages in stability and safety. In related research, Dang et al. proposed a human-machine collaborative assembly method based on image processing, which achieved collaborative operation through visual recognition of assembly parts. However, this method relies too much on traditional image processing and geometric feature extraction, and lacks adaptability to multi-target occlusion and dynamic interference in complex environments, resulting in limited assembly accuracy [29]. Krasko et al. proposed a parameter optimization method for flexible automated assembly lines to ensure the efficiency and flexibility of the production line in mechanical assembly. This study performed well in terms of flexibility and parameter adaptation at the system level, but had weak dependence on OD and real-time feedback in the assembly process, and lacked direct modeling of visual complexity [30]. Compared with the above methods, the proposed framework introduced deep learning-driven MSFF and dynamic weighting mechanism, which not only effectively improved the detection and recognition accuracy in complex backgrounds, but also combined real-time correction mechanism for assembly deviation, achieving a balance between accuracy and efficiency in the assembly process. In summary, the ISAM-ISSD method proposed by the research not only achieves comprehensive improvements in detection accuracy, assembly stability, and efficiency in experiments, but also proposes a more comprehensive and robust solution, providing a new technological path for intelligent assembly of automotive transmission gear shafts.

5. Conclusion

The study proposes an ISAM-ISSD method that integrates ISAM and ISSD to address the issues of insufficient target recognition accuracy, high sensitivity to assembly tolerances, and unstable mechanical control in the automatic assembly process of automotive transmission gear shafts. This method introduces a dual backbone network and FPN in the detection stage, which enhances the multi-scale recognition capability in complex scenes. In the segmentation stage, multi-scale feature enhancement and attention mechanism are used to optimize the boundary expression, and the assembly pose is calculated by combining the centroid and main axis direction; Impedance control law in the control phase is introduced to achieve force displacement coupling regulation. The findings denote that this method significantly outperforms the comparison model in both detection and segmentation accuracy. The mAP of the improved detector reaches 91.5%, with IoU and Dice coefficients of 90.1% and 93.5%, respectively. In the assembly process, the success rate of assembly is 98.6%, the alignment error is only 0.43 mm, the peak insertion force is reduced to 49.7 N, the single piece cycle time is shortened to 5.4 s, and the repair rate is only 1.8%. The assembly tolerance experiment shows that when the coordination gap is 50 μm, the overall performance is the best, with an assembly success rate of 98.8%, a false positive rate of only 1.4%, and a shortened end adjustment time of 103 ms. The control stability experiment further verifies that the method maintains the lowest level in terms of force control overshoot and trajectory deviation. The findings denote that the ISAM-ISSD method proposed by the research achieves collaborative optimization of accuracy, efficiency, and stability in complex assembly environments, and has good potential for engineering applications. However, this method still exhibits certain performance degradation under extreme noise and strong interference conditions. Future research will focus on enhancing low signal-to-noise ratio environments, designing lightweight models, and developing cross scene transfer capabilities to further improve its adaptability and practicality in large-scale industrial production.

References

[1] G. Fantoni, S. Q. Al-Zubaidi, E. Coli, and D. Mazzei. "Automating the process of method-time-measurement," Int. J. Prod. Perform. Manag., vol. 70, no. 4, pp. 958-982, April, 2021, DOI: 10.1108/IJPPM-08-2019-0404.

[2] M. Gardiner, A. Truskovsky, G. Neville-Neil, and A. Mashatan. "Quantum-safe trust for vehicles: The race is already on," Commun. ACM, vol. 64, no. 9, pp. 54-61, September, 2021, DOI: 10.1145/3466174.

[3] J. Terfurth, M. Waldhof, M. Schmid, J. Weber, and M. Keller. "Design and evaluation of a spatially distributed and gear-integrated transverse

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

flux machine," IEEE Trans. Ind. Electron., vol. 70, no. 5, pp. 5065-5073, May, 2022, DOI: 10.1109/TIE.2022.3189082.
[4] Z. Liu, K. Wang, D. Liu, Q. Wang, and J. Tan. "A motion planning method for visual servoing using deep reinforcement learning in autonomous robotic assembly," IEEE/ASME Trans. Mechatronics, vol. 28, no. 6, pp. 3513-3524, June, 2023, DOI: 10.1109/TMECH.2023.3275854.
[5] Z. Li, H. Chen, B. Biggio, Y. He, H. Cai, and F. Roli. "Toward effective traffic sign detection via two-stage fusion neural networks," IEEE Trans. Intell. Transp. Syst., vol. 25, no. 8, pp. 8283-8294, August, 2024, DOI: 10.1109/TITS.2024.3373793.
[6] S. N. Katkade, V. C. Bagal, R. R. Manza, and P. L. Yannawar. "Advances in real-time object detection and information retrieval: A review," Artif. Intell. Appl., vol. 1, no. 3, pp. 139-144, March, 2023, DOI: 10.47852/bonviewAIA3202456.
[7] Y. W. Ke and A. C. Lin. "A rapid image comparison approach to automatic recognition and assembly of jigsaw puzzles," Int. J. Precis. Eng. Manuf., vol. 25, no. 7, pp. 1403-1414, July, 2024, DOI: 10.1007/s12541-024-00993-x.
[8] A. V. Byalonovych. "Digital material surface image analysis for evaluating geometric fatigue crack parameters," Strength Mater., vol. 55, no. 5, pp. 960-966, May, 2023, DOI: 10.1007/s11223-023-00587-4.
[9] T. Hao and D. Xu. "Automated control method of multiple peg-in-hole assembly for relay and its socket," IEEE Trans. Ind. Informat., vol. 21, no. 1, pp. 990-998, January, 2024, DOI: 10.1109/TII.2024.3476535.
[10] Y. M. Zamuragin, B. L. Salamandra, and K. B. Salamandra. "Adaptive system of position synchronization of tape materials in automatic assembly," J. Mach. Manuf. Reliab., vol. 52, no. 7, pp. 791-799, July, 2023, DOI: 10.1134/S1052618823070270.
[11] L. Dai, X. Zhang, P. Gardoni, H. Lu, X. Liu, and G. Krolczyk. "A new machine vision detection method for identifying and screening out various large foreign objects on coal belt conveyor lines," Complex Intell. Syst., vol. 9, no. 5, pp. 5221-5234, May, 2023, DOI: 10.1007/s40747-023-01011-9.
[12] I. G. B. Sampaio, J. Viterbo, and J. Guerin. "Improving robustness of industrial object detection by automatic generation of synthetic images from CAD models," Comput. Intell., vol. 39, no. 3, pp. 415-432, March, 2023, DOI: 10.1111/coin.12572.
[13] M. E. Sahin, H. Ulutas, E. Yuce, and M. F. Erkoc. "Detection and classification of COVID-19 by using faster R-CNN and mask R-CNN on CT images," Neural Comput. Appl., vol. 35, no. 18, pp. 13597-13611, March, 2023, DOI: 10.1007/s00521-023-08450-y.
[14] W. Mei, Y. Zheng, and G. Gu. "Design of intelligent 3D collaborative manufacturing platform for non-holonomic mobile industrial robots based on improved binocular vision," Int. J. Intell. Robot. Appl., vol. 7, no. 4, pp. 740-751, December, 2023, DOI: 10.1007/s41315-023-00293-z.
[15] K. Miki, F. Nagata, K. Furuta, K. Arima, T. Shimizu, and T. Ikeda. "Development of a hyper CLS data-based robotic interface for automation of production-line tasks using an articulated robot arm," Artif. Life Robot., vol. 27, no. 3, pp. 547-553, September, 2022, DOI: 10.1007/s10015-022-00778-3.
[16] B. Zhou, J. Zhang, and Q. Fei. "Bi-objective green in-house transportation scheduling and fleet size determination in mixed-model assembly lines with mobile robots," Eng. Comput., vol. 39, no. 7, pp. 2630-2654, July, 2022, DOI: 10.1108/EC-08-2021-0483.
[17] Y. Feng, C. He, Y. Gao, H. Zheng, and J. Tan. "A method to assess design complexity of modular automatic assembly system in design phase," Assem. Autom., vol. 42, no. 1, pp. 28-39, January, 2022, DOI: 10.1108/AA-04-2021-0038.
[18] Q. Gao, Y. Chen, and Z. Ju. "Oropharynx visual detection by using a multi-attention single-shot multibox detector for human-robot collaborative oropharynx sampling," IEEE Trans. Hum.-Mach. Syst., vol. 53, no. 6, pp. 1073-1082, December, 2023, DOI: 10.1109/THMS.2023.3324664.
[19] Y. Wang, X. Liu, and R. Guo. "An object detection algorithm based on the feature pyramid network and single shot multibox detector," Cluster Comput., vol. 25, no. 5, pp. 3313-3324, May, 2022, DOI: 10.1007/s10586-022-03560-z.
[20] G. Liu, Z. Cao, S. Liu, B. Song, and Z. Liu. "An improved SSD method for infrared target detection based on convolutional neural network," J. Comput. Methods Sci. Eng., vol. 22, no. 4, pp. 1393-1408, August, 2022, DOI: 10.3233/JCM-226112.
[21] W. Song, Y. Tang, W. Tan, and S. Ren. "ISHD: Intelligent standing human detection of video surveillance for the smart examination environment," CMES-Comput. Model. Eng. Sci., vol. 137, no. 1, pp. 1-16, January, 2023, DOI: 10.32604/cmes.2023.026933.
[22] H. Han. "A novel single shot-multibox detector based on multiple Gaussian mixture model for urban fire smoke detection," Comput. Sci. Inf. Syst., vol. 20, no. 4, pp. 1819-1843, October, 2023, DOI: 10.2298/CSIS221218032H.
[23] Y. Zhu, X. Sun, M. Wang, and H. Huang. "Multi-modal feature pyramid transformer for RGB-infrared object detection," IEEE Trans. Intell. Transp. Syst., vol. 24, no. 9, pp. 9984-9995, September, 2023, DOI: 10.1109/TITS.2023.3266487.
[24] J. Wang, J. Yu, and Z. He. "ARFP: A novel adaptive recursive feature pyramid for object detection in aerial images," Appl. Intell., vol. 52, no. 11, pp. 12844-12859, November, 2022, DOI: 10.1007/s10489-021-03147-y.
[25] Y. Wu, J. Jiang, Z. Huang, and Y. Tian. "FPANet: Feature pyramid aggregation network for real-time semantic segmentation," Appl. Intell., vol. 52, no. 3, pp. 3319-3336, March, 2022, DOI: 10.1007/s10489-021-02603-z.
[26] Y. Guo, Y. Xu, H. Cui, M. Dang, and S. Li. "Segment anything model-based crack segmentation using low-rank adaption fine-tuning," Struct. Health Monit., vol. 24, no. 4, pp. 2579-2591, April, 2025, DOI: 10.1177/14759217241261089.
[27] K. Song, W. Cui, H. Yu, X. Li, and Y. Yan. "SAMEra: Can it segment any industrial surface defects?," Comput. Mater. Continua, vol. 78, no. 3, pp. 1-16, March, 2024, DOI: 10.32604/cmc.2024.048451.
[28] W. Ji, J. Li, Q. Bi, T. Liu, and W. Li. "Segment anything is not always perfect: An investigation of SAM on different real-world applications," Front. Inform. Technol. Electron. Eng., vol. 21, no. 4, pp. 617-630, April, 2024, DOI: 10.1007/s11633-023-1385-0.
[29] A. T. Dang, Q. C. Hsu, and Y. S. Jhou. "Development of human-robot cooperation for assembly using image processing techniques," Int. J. Adv. Manuf. Technol., vol. 120, no. 5, pp. 3135-3154, May, 2022, DOI: 10.1007/s00170-022-08968-w.
[30] A. S. Krasko, M. A. Filin, and I. I. Bazarov. "Development of a methodology for ensuring optimal parameters of flexible automated lines of mechanical assembly production," J. Mach. Manuf. Reliab., vol. 53, no. 8, pp. 971-979, August, 2024, DOI: 10.1134/S1052618824700717.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.