Multi-Target Vision Detection and Grasping of Electronic Devices Using DL-SORT

Hong Wang
Zhengzhou Vocational College of Industrial Safety, Zhengzhou 451192, China
E-mail: 18638583266@163.com

Keywords: Object vision detection; Simple online and realtime tracking; Automatic recognition; Deep learning; Recursive labeling

Received: February 24, 2025

Many production industries are increasingly dependent on intelligent electronic devices. However, traditional methods for recognizing and detecting electronic devices are inefficient and consume a large amount of production resources. This study introduces a hybrid DL-SORT model for automatic recognition and grasping of electronic devices, integrating deep learning with Simple Online and Realtime Tracking (SORT) to enhance object detection performance. In the model, Recursive labeling and Binary Robust Independent Elementary Features are also employed for key point detection and domain selection. Experimental results show that the hybrid algorithm outperforms Single Shot MultiBox Detector and Discriminative Correlation Filter with Channel and Spatial Reliability in terms of object detection performance, with loss values of 0.15, 0.24, and 0.23, respectively. Additionally, empirical analysis of the constructed hybrid model reveals that the proposed automatic recognition and grasping model for electronic devices achieves an accuracy of 0.95 in tape recognition, demonstrating good recognition accuracy. Testing in obstructed environments shows that the success rate of part detection remains above 80%, with minimal performance degradation. These results suggest that the hybrid model can detect multiple targets and improve production efficiency. This study contributes to the future development of drones and industrial robots in the automation field, enabling the acquisition of precise target location information.

Povzetek: Raziskava uvaja hibridni model DL-SORT za večciljno vizualno zaznavo in prijem elektronskih naprav, s čimer bistveno izboljša natančnost prepoznavanja, sledenja in učinkovitost industrijskih robotov.

1 Introduction

In recent years, scientific and technological innovations have significantly contributed to urban construction, with applications in smart transportation, smart logistics, and smart healthcare [1, 2]. However, with the rapid development of technology, traditional electronic technologies can no longer meet people's needs. For example, traditional automated robotic sorting is costly, and unmanned aerial vehicle vision detection is inefficient [3]. As a result, many experts have improved the characteristics of robots for different environments. Among these improvements are waste grabbing robots that search and plan global roadways, and robots used in power control to identify risk vulnerabilities, enhancing global infrastructure in the transportation sector [4]. Despite their contributions, these robots still face practical challenges such as unclear recognition and poor object grasping capabilities. As a result, many experts have also conducted research on the object vision detection algorithms used in these systems. Object vision detection algorithms can recognize and locate images or moving objects, with strong computational power [5]. The most widely used algorithms currently are those based on traditional machine learning, which can accurately judge image edges and lines and are often applied in facial recognition access control systems and intelligent transportation systems [6]. However, traditional object detection algorithms cannot detect multiple targets and fail to make visual judgments when there are obstructions. Therefore, this study proposes a hybrid model for automatic recognition and grasping of electronic devices, combining Deep Learning (DL) with the Simple Online and Realtime Tracking (SORT) to enhance object tracking and vision detection capabilities. DL is applied in electronic device recognition and mesh damage image detection to extract features such as device contours and textures using convolutional networks, so as to accelerate the convergence speed and accuracy of feature extraction. In the construction of this model, Recursive labeling and Binary Robust Independent Elementary Features (BRIEF) are also used for key point detection. The research innovation lies in the integration of DL-SORT and BRIEF to form continuous motion planning for target detection, tracking and grasping, and improve the grasping accuracy and efficiency of electronic devices. The research aims to improve the degree of completion of automatic identification and grasping tasks of electronic devices, meet the needs for grasping small targets and ensuring realtime responsiveness, and promote the evaluation

standardization of automatic detection in the electronic manufacturing industry.

2 Related works

To enhance the safety and performance of electronic devices during automatic identification and detection processes, research on automatic identification and grasping of electronic components is essential. Zhang et al. raised an enhanced pose estimation algorithm to address issues such as slow pose estimation speed and poor robustness during robot sorting and feeding processes. Experimental results showed that the method achieved an average distance error of 2.04 mm, an average angle error of 2.72 degrees, and an average robot grasping success rate of 97.08% [7]. Lin et al. proposed a robot grasping method based on object shape approximation and LightGBM to solve the challenge of heavy dependence on datasets for object grasping planning. Experimental results indicated that the method achieved a classification accuracy of 94.5%, with a detection time of 0.0003 s, and an average success rate of 91.81% in grasping new objects [8]. Wang et al. put forward a multi-modal dynamic collaborative fusion network to improve the robot's capability of detecting flat-surface object grasping. Experimental results demonstrated that the network achieved a grasping success rate of 98.8% in single-object scenarios [9]. Liu et al. introduced an industrial robot-based transplant workstation to overcome limitations in handling the root systems of mature old and young seedlings in transplanting machines. Experimental results showed that the method achieved a recognition accuracy exceeding 97.68%, and the success rate of transplanting and replanting reached 95% [10]. Yan et al. proposed a new lightweight grasping detection model to address issues such as low detection accuracy and large model parameters. Experimental results showed that the convolutional block attention module in the model could recognize multiple attributes of objects, achieving a detection accuracy of 98.44% in Image-wise segmentation [11].

The continuous improvement of electronic devices also indicates an increasing demand for advanced technological intelligence. It is not only required that electronic devices automatically recognize and grasp objects, but image acquisition must also be more precise. Therefore, object vision detection algorithms have been the focus of related studies. Li et al., to address the issues of multiple fabric defect types and small defect sizes, proposed an improved fabric defect detection algorithm. The experimental results showed that the mean Average Precision (mAP) of the improved algorithm was 65.1%, which was an increase of 8.3% and 3.2% compared to the original model [12]. Wang et al., to solve the problems of false positives and missed detections of small targets in the detection of aircraft skin defects under complex backgrounds, proposed an aircraft skin defect detection model, which achieved a detection accuracy of 97.9%, which was 7.3% higher than the baseline model, and the detection speed reached 139 FPS [13]. Ji et al., to enhance the stability of power systems, proposed an engineering machinery risk management intelligent detection algorithm based on visual perception for intrusion into transmission lines. The results showed that the model's average accuracy improved by 6.3%, its precision increased by 3.7%, and its recall rate increased by 3.1% [14]. Xiong et al., in order to solve the problem of low target recognition ability, researched and proposed a method for small dynamic target detection by combining YOLO and background subtraction. The accuracy of this method was improved by 2.3% and the recall rate increased by 3.5% [15].

In summary, both domestic and international scholars have conducted detailed studies on electronic device target recognition, and have made significant progress in object recognition and grasping for electronic devices. However, there are few studies on the combination of electronic devices and object vision detection algorithms. Therefore, a DL-SORT electronic device automatic recognition and grasping hybrid model is constructed, which can detect targets in complex scenes. The model also adopts a recursive labeling method for semantic analysis, effectively labeling images and further improving the model's recognition performance. The relevant worksheets are shown in Table 1.

Table 1: Related literature worksheet

Domain	Technology	Advantage	Performance index	Shortcoming	Author
Automatic recognition	Enhanced attitude estimation algorithm	Grasp objects with good accuracy	Mean distance error 2.04 mm, Average Angle error 2.72, Average grasping success 97.08%	Sensitive to occlusion	Zhang et al. [7]
Robot grasping	Robot grasping method based on object shape approximation and LightGBM	Fast detection time and high grasping accuracy	Classification accuracy 94.5%, Time 0.0003 s, ASR 91.81%	Parameter tuning is complex, Limited adaptability	Lin et al. [8]

Robot grasping	Multi-mode dynamic collaborative fusion network	The capture success rate of single scene is high	Grasp success rate 98.8%	Mode synchronization is difficult	Wang et al. [9]
Robot recognition	Transplanting workstation based on UR5 industrial robot lifts restrictions	High recognition accuracy	Recognition accuracy 97.68%	Positioning accuracy decline	Liu et al. [10]
Automatic recognition	Lightweight grab detection model	Multifaceted recognition	Detection accuracy 98.44%	Feature extraction ability is limited	Yan et al. [11]
Target visual detection	FD-YOLOv5 algorithm for fabric defect detection	High detection accuracy	mAP 65.1%	Limited adaptability to complex defect scenarios	Li et al. [12]
Target visual detection	Aircraft skin defect detection model based on YOLOv8n	High detection accuracy	Detection accuracy 97.9%	Small target detection is not high	Wang et al. [13]
Target visual detection	Intelligent detection algorithm for risk management of construction machinery intrusion transmission line based on visual perception	High detection accuracy	The average accuracy is improved by 6.3%, The recall rate increased by 3.7%	High requirements on hardware	Ji et al. [14]
Target visual detection	A method for detecting small dynamic targets using YOLO and background subtraction	High detection accuracy	Accuracy improved by 2.3%, Recall rate increased by 3.5%	Lack of detail	Xiong et al. [15]

3 DL-SORT optimized electronic device recognition and grasping strategy

3.1 Design of DL-SORT multi-target vision detection algorithm

With the continuous development of information technology, devices such as city cameras and drone detection are increasingly applied in public life, contributing to the development of smart cities [16]. However, these electronic devices face issues such as large amounts of missed detections and errors when processing vast amounts of image and video data, reducing the realtime performance of practical applications [17]. Therefore, the study proposes a DL-SORT multi-target vision detection algorithm to address problems such as low execution efficiency and inaccurate target detection in electronic devices, improving the accuracy of the detection system. Among them, DL algorithm captures target details at different scales and effectively detects tiny electronic components and mechanical parts [18]. Therefore, the study uses DL for target detection and classification, with the prediction error minimized as shown in Equation (1).

$$H = W \times x + b \tag{1}$$

In Equation (1), $y_{i'}$ and y_i represent the predicted results and actual targets, W is the weight matrix, and x and b represent the input values and bias values, respectively. In order to enable electronic mobile devices to promptly identify intruders during target search, the study also improves the grid image damage detection in DL, with the improved framework shown in Figure 1.

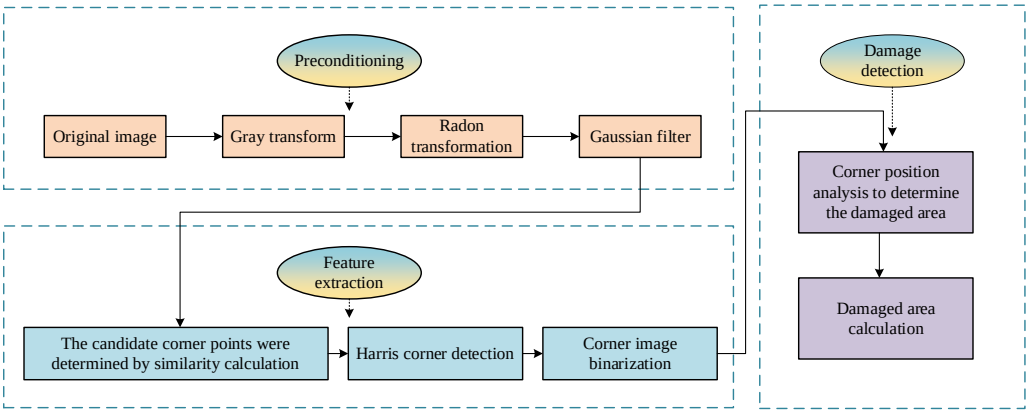


Figure 1: Grid image damage detection framework based on DL

As shown in Figure 1, the grid image damage detection framework based on DL is divided into three parts: image data preprocessing, feature data extraction, and damage detection. In the preprocessing part, the original image is first transformed to grayscale, followed by spatial projection using Radon transform for easier detection of light and dark spots. Gaussian filtering is then applied for noise reduction. After preprocessing, the data undergoes feature extraction, during which Harris corner detection is performed. Finally, the damage area is calculated. The Gaussian function for filtering and denoising is expressed in Equation (2) [19].

$$g(x) = e^{-x^2/2z^2} \quad (2)$$

In Equation (2), z represents the Gaussian distribution parameter of filter denoising, while $g(x)$ is the one-dimensional mean Gaussian function. Furthermore, to further improve the accuracy of the DL detection algorithm, depthwise separable convolution is used to reduce computational load and model parameters. The computational dimensions of depthwise separable convolution are shown in Equation (3) [20].

$$D_K * D_K * D_F * D_F * M * N, D_K * D_K * M * N \quad (3)$$

In Equation (3), $D_K * D_K$ represents the standard convolution, M, N denotes the thickness input of the output feature map, D_F refers to the length and width of the feature map to be extracted, and $D_F * D_F * M$ and $D_F * D_F * N$ represent the dimensions of the feature map input and output, respectively. The performance of the depthwise separable convolution is shown in Figure 2.

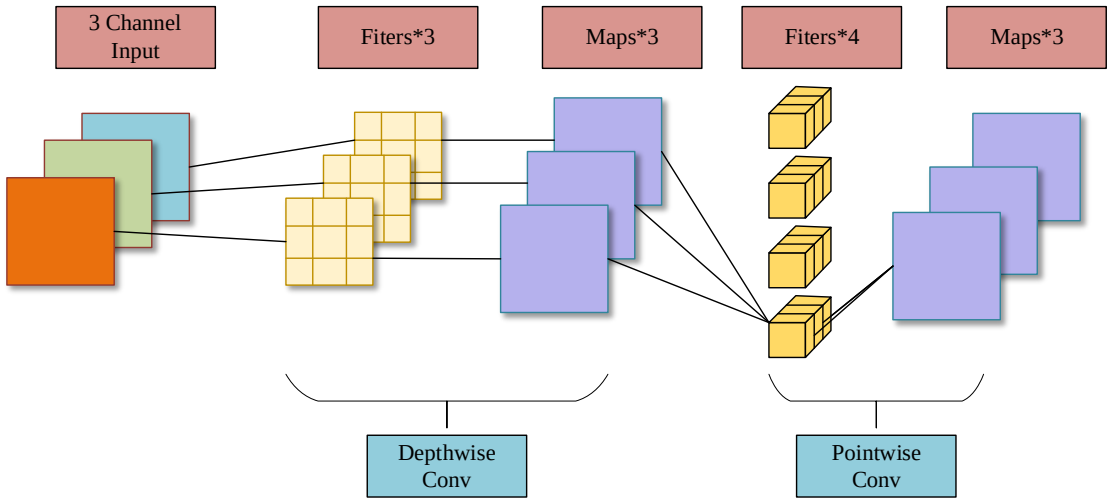


Figure 2 Schematic diagram of depth-wise separable convolution

As shown in Figure 2, the depthwise separable convolution combines 1×1 point convolution and depth convolution with the same kernel size, then delivers each convolution kernel to the channel. The number of output convolution kernels corresponds to the number of convolution channels. When the number of 2D convolution kernels in each group is equal to the number of channels, it indicates that the input channels and the number of convolution kernels match. Two-dimensional convolution is mainly used for feature extraction, and its kernel operation is shown in Equation (4).

$$T_{DWConv} + PWConv = D_k^2 \times channel \times D_f^2 + N \times M \times D_f^2 \quad (4)$$

In Equation (4), T_{DWConv} stands for depth-separable convolution computation, $PWConv$ represents the

computation of a point-by-point convolution, M indicates the parameter between the number of channels, D_k and D_f represent the width and height of the convolution kernel and input, respectively. N represents the number of convolution kernels, and FF denotes the number of channels. Although the improved DL can perform certain detections on specific targets, it still has some computational limitations when the number of targets increases or the state of moving targets becomes unstable. Therefore, based on the improved DL, SORT is integrated to complete multi-target prediction and tracking. SORT is simple and efficient, capable of updating the target state in realtime [21]. The target state model updated by SORT is expressed in Equation (5).

$$x(t)=[u,v,s,r,\dot{u},\dot{v},\dot{s}] \quad (5)$$

In Equation (5), u represents the position of the target in the horizontal direction in the image, v represents the vertical position, and s and r are the scale range and aspect ratio of the target's bounding box, respectively. To enhance the brightness of the captured target image, the study also integrates brightness detection into DL-SORT. The processed grayscale image is shown in Equation (6).

$$Gray(i,j)=0.299\times R(i,j)+0.578\times G(i,j)+0.114\times B(i,j) \quad (6)$$

In Equation (6), $Gray(i,j)$ represents the grayscale image, and $R(i,j)$ is the red channel coordinates of the color image, with $G(i,j)$ and $B(i,j)$ representing the green and blue channel coordinates, respectively. The specific process of the DL-SORT multi-target visual detection algorithm is shown in Figure 3.

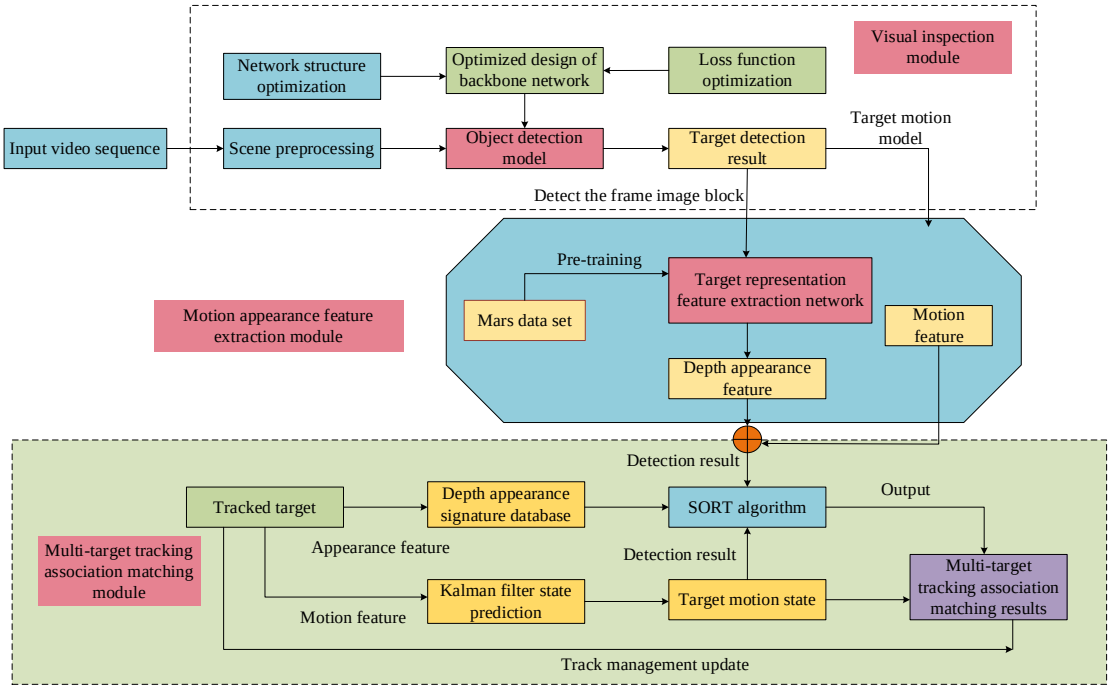


Figure 3: DL-SORT-based multi-target visual detection algorithm flow chat

As shown in Figure 3, the DL-SORT-based multi-target visual detection algorithm process is divided into three sections. The visual detection module requires the input of video sequences, followed by scene preprocessing. Then, the target detection model optimized by the network structure is used to obtain the detection results. These results are then passed to the motion appearance feature extraction module, where the target features are extracted by detecting different image blocks and pre-trained. Finally, the obtained deep appearance features are tracked and associated using SORT. The matched targets are compared with the deep appearance feature library, and during the matching process, the Kalman filter state is used for prediction. The final trajectory update is made by combining the predicted and detected results. In addition, this algorithm is suitable for processing single images as well. Compared with video sequences, the algorithm uses DL to automatically learn features such as image edges and textures during feature extraction, and low-dimensional feature vectors are generated from the target for subsequent frame judgment.

3.2 DL-SORT electronic device recognition and grasping model optimization

Currently, traditional intelligent robots not only have low production efficiency but also lack sufficient recognition ability. Therefore, researchers urgently need to optimize the automatic recognition and grasping of robots and other electronic devices [22]. Although the DL-SORT multi-target visual detection algorithm can detect multiple target parameters, there are still some shortcomings in the automatic recognition and grasping of electronic devices. To address the inaccuracy in automatic recognition and grasping of electronic devices, a hybrid method based on DL-SORT and recursive labeling is proposed. The recursive labeling method performs semantic analysis and

labels images effectively during the search process. The image centroid feature in the connected domain pixel detection based on recursive labeling is shown in Equation (7) [23].

$$\gamma(\bar{x},\bar{y})\times\left[\frac{\sum_{x=1}^{M}\sum_{y=1}^{N}xf(x,y)}{\sum_{x=1}^{M}\sum_{y=1}^{N}f(x,y)},\frac{\sum_{x=1}^{M}\sum_{y=1}^{N}yf(x,y)}{\sum_{x=1}^{M}\sum_{y=1}^{N}f(x,y)}\right]=\left[\frac{\sum_{x=1}^{M}\sum_{y=1}^{N}\bar{x}\gamma(\bar{x},\bar{y})}{\sum_{x=1}^{M}\sum_{y=1}^{N}\gamma(\bar{x},\bar{y})},\frac{\sum_{x=1}^{M}\sum_{y=1}^{N}\bar{y}\gamma(\bar{x},\bar{y})}{\sum_{x=1}^{M}\sum_{y=1}^{N}\gamma(\bar{x},\bar{y})}\right] \quad (7)$$

In Equation (7), γ represents the labeled image, and $(\bar{x},\bar{y})$ denotes the infrared image coordinates marked using the recursive labeling method, $f(x,y)$ represents the pixel value of the grayscale image. To allow electronic devices to perform fine-grained calculations of the surrounding areas during the automatic recognition process, the study also applies BRIEF on top of the recursive labeling method for local feature representation calculation. BRIEF can perform keypoint detection and domain selection. The improved feature descriptor formula is shown in Equation (8).

$$\tau(p;x,y)=\begin{cases}1, P(x)<P(y)\\ 0, other\end{cases} \quad (8)$$

In Equation (8), P_x and P_y represent the grayscale values of randomly selected pixels, while x and y denote the feature coordinates. The transformation matrix form of the rotated descriptor is shown in Equation (9).

$$S_r=R_\theta S=\begin{bmatrix}\cos\theta,-\sin\theta\\ \sin\theta,\cos\theta\end{bmatrix}\begin{bmatrix}x_1x_2...x_n\\ y_1y_2...y_n\end{bmatrix} \quad (9)$$

In Equation (9), S_r represents the rotated matrix, and n refers to the corresponding points around the extracted feature points. The specific process of the visual feature recognition method based on the recursive labeling and BRIEF improvement is shown in Figure 4.

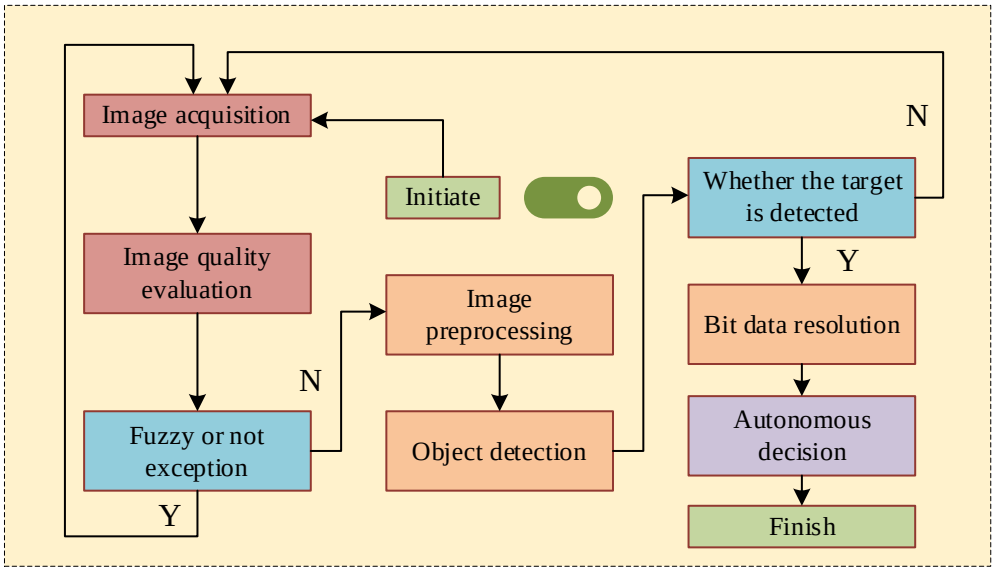


Figure 4: Process of visual feature recognition based on recursive labeling and BRIEF

As shown in Figure 4, the visual system process for electronic devices first captures the image, then detects and evaluates images for potential quality degradation. After the evaluation, the system determines whether the image requires further recognition and checks for any blur anomalies. If no anomalies are detected, the image undergoes preprocessing. Once preprocessing is complete, the image is subjected to target detection, and the system checks whether the targets are correctly identified. If no target is detected, the process returns to the image capture step. If the target is successfully detected, pose estimation is performed, and finally, autonomous decision-making is implemented. The grayscale morphological correction of the electronic device parameter image is shown in Equation (10).

$$A\cdot B=(A\Theta B)\oplus B \quad (10)$$

In Equation (10), A represents the image, and B refers to the structural element. By performing an erosion operation on image A, removing small bright noise and shrinking object boundaries, and then performing expansion operation to fill dark holes and expand boundaries, the final result is an image modified by $A\cdot B$ gray scale morphology processing. By defining the structural element, grayscale morphological correction can be performed, enhancing the edge features of the electronic device and eliminating influences generated during subsequent recognition. The feature extraction model for electronic devices is shown in Figure 5.

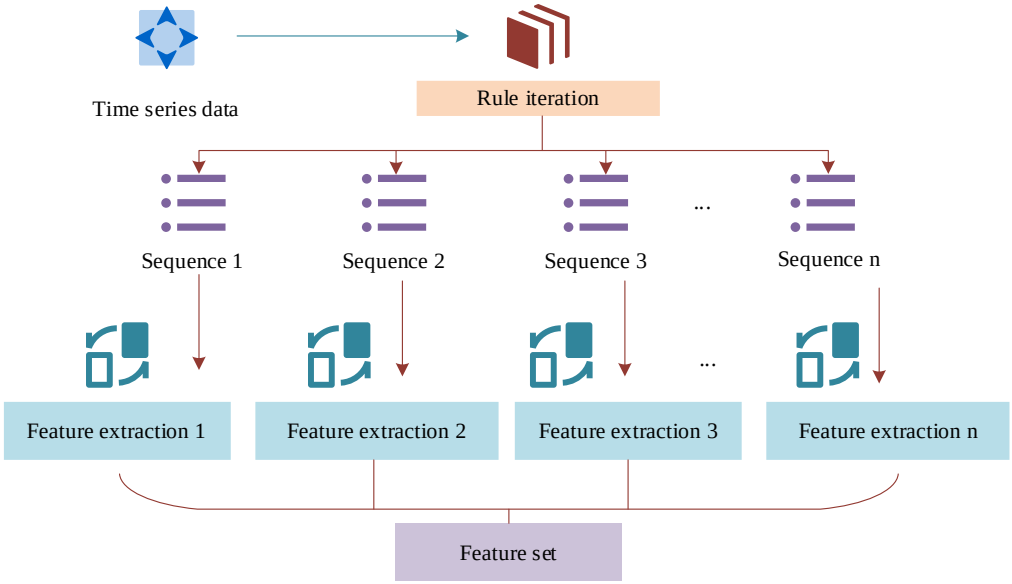


Figure 5: Workflow diagram of the electronic device feature extraction model

As shown in Figure 5, the feature extraction model for electronic devices uses time series data for rule iteration. By applying the rule iteration method, spatial data transformation is performed, extracting corresponding features from sequences 1 to n, forming a new feature sequence. Finally, the target change rule is recorded in detail through the time series, the grasping trajectory of different object shapes is generated, and the grasping path of the robot arm is predicted. This completes the entire process from time sequence data encoding to robot action decision mapping and realtime control. The distance calculation formula for object or image recognition and grasping in electronic devices is shown in Equation (11).

$$\begin{cases} \bar{d} = \dfrac{1}{n}\sum_{i=1}^{n}\bar{d}_i \\ \sigma = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}\left(\bar{d}_i - \bar{d}\right)^2} \\ Q = \left\{ p_i \in P \middle| \bar{d}_i \le \bar{d} + \lambda \cdot \sigma \right\} \end{cases} \tag{11}$$

In Equation (11), $\bar{d}$ represents the average distance mean, σ is the standard deviation, Q denotes the point cloud after removing certain boundary points, and λ is the scaling factor. The specific calculation formula for target search in electronic devices is shown in Equation (12).

$$f^*(n) = g^*(n) + h^*(n) \tag{12}$$

In Equation (12), n represents the electronic device's status, $f^*(n)$ is the minimal cost estimate from the initial state to the target state during the search process, and $h^*(n)$ and $g^*(n)$ are the minimal estimated cost for the path and the state, respectively. In conclusion, the electronic device automatic recognition and grasping method based on recursive labeling and BRIEF improvements is capable of detecting and locating targets. Therefore, the study combines this method and proposes a hybrid model for electronic device automatic recognition and grasping based on the DL-SORT multi-target visual detection algorithm to improve realtime multi-target detection and evaluation. The specific recognition and grasping process flow diagram for robots using this hybrid model is shown in Figure 6.

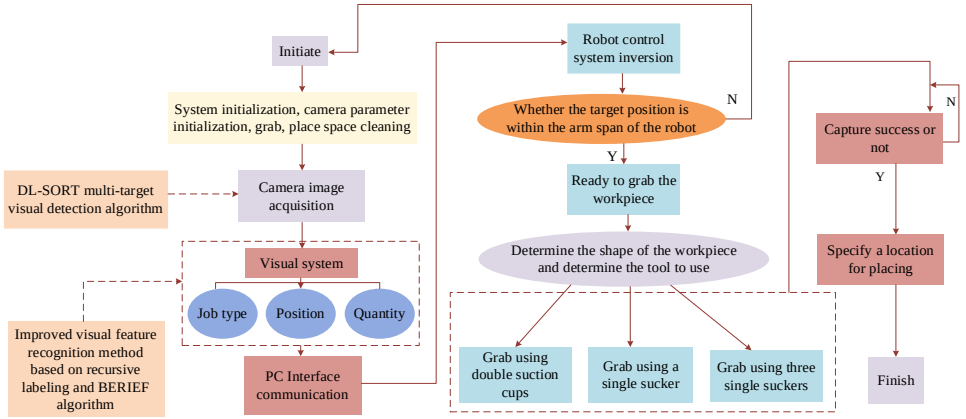


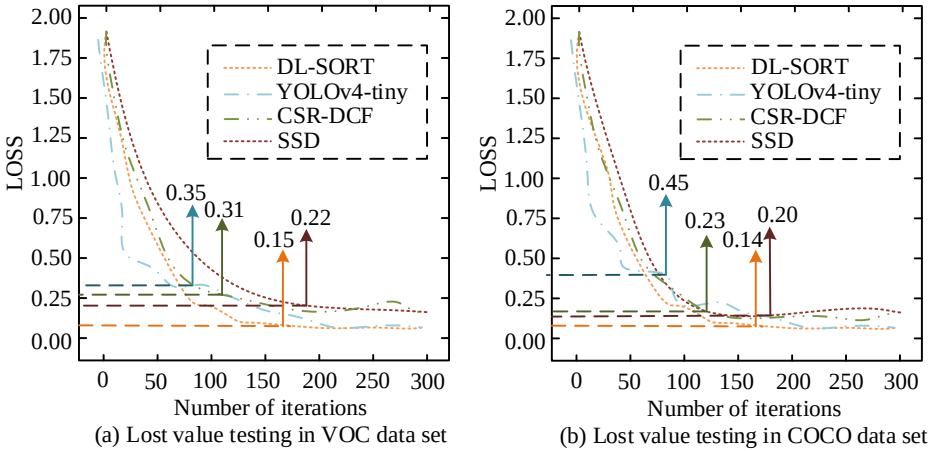
Figure 6: Robot recognition and grasping flow chart

As shown in Figure 6, the robot uses the DL-SORT automatic recognition and grasping model for workpiece handling. The system needs to be initialized, and the camera parameters must be adjusted accordingly. Then, DL-SORT is used for image acquisition, and the recursive labeling and BRIEF-improved visual feature recognition method strengthens the vision system, processing information such as the type and position of the workpiece. This information is then transmitted to the control system. When a static image or realtime video stream is input, the key component SORT mainly predicts the target motion trajectory based on Kalman filter and associates the detection frame of adjacent frames. Recursive marking is mainly used for hierarchical marking of the segmented work area to distinguish overlapping targets. BRIEF generates binary feature descriptors for objects to quickly match similar artifacts. The control system determines the shape of the workpiece and selects the appropriate tool based on the type of workpiece. For slender workpieces, the distance needs to be adjusted and a dual-suction cup tool is used. For small workpieces, the tool posture is adjusted to use a single suction cup. For flat workpieces, both dual-suction and single-suction cups are adjusted together. If the workpiece is successfully grasped, it is placed in the designated position, if unsuccessful, the system will return and attempt to grasp again.

4 Performance validation of DL-SORT

4.1 DL-SORT hybrid algorithm performance comparison

To demonstrate the superior performance of the DL-SORT hybrid algorithm, the study compared it with Single Shot MultiBox Detector (SSD),Discriminative Correlation Filter with Channel and Spatial Reliability (CSR-DCF), and the lightweight version of YOLOv4 (You Only Look Once v4, YOLOv4-tiny). The experimental setup used Ubuntu 18.04 LTS as the operating system, 16 GB of memory, STM32F405RGT6 as the controller module, and an AMD Ryzen R7 5700U CPU@1.80GHz. The experiments were conducted using the Keras deep learning framework. Among them, the DL learning rate is set to 0.01, Dropout is set to 0.3, batch size is set to [16, 22], and the number of iterations is set to within 500 times. The optimal parameters are obtained through Bayesian optimization. The SORT momentum is set to 0.3 and the IOU threshold to 0.3. In order to ensure the reliability and effectiveness of the experiment, the experimental data set is trained and tested using VOC and COCO data sets, both of which are suitable for target detection, classification and segmentation to improve the detection clarity of the algorithm. The total number of images in VOC and COCO data sets is 11500 and 330000 respectively, the number of training and test images is 5717 and 20288 respectively, and the image size is normalized to 512×512. In this study, four target detection algorithms were tested for the loss value of different targets, and the loss value was the quantified error index between the predicted results of the target detection algorithm and the real labeling during the training process. The loss value of the test algorithm was the sum of classification loss, positioning loss and confidence loss. The data sets used for this test were VOC and COCO data sets, and the test results were shown in Figure 7.



(a) Lost value testing in VOC data set (b) Lost value testing in COCO data set

Figure 7: Loss function test result diagram

As shown in Figure 7(a), when the DL-SORT hybrid algorithm was tested on the VOC dataset, the loss value reached 0.15 after 170 iterations. The loss value gradually stabilized between 100 and 150 iterations. In contrast, YOLOv4-tiny achieved a loss value of 0.35 after 70 iterations, while the DL-SORT hybrid algorithm's loss value was 0.27. As shown in Figure 7(b), the loss curve of YOLOv4-tiny fluctuated significantly before 50 iterations, and the loss value was 0.45 at 70 iterations. SSD reached a loss value of 0.20 after 170 iterations, and its loss value increased after 250 iterations. The loss value of each algorithm on the VOC

and COCO data sets vary significantly, because the test image scale of these two data sets is different. To highlight the accuracy performance of the DL-SORT hybrid algorithm, the study compared it with YOLOv4-tiny in terms of accuracy error and precision variation. Since YOLOv4-tiny involved significant optimizations in its network structure and training strategy, this algorithm was considered a representative model in terms of detection accuracy. The comparison results are shown in Figure 8.

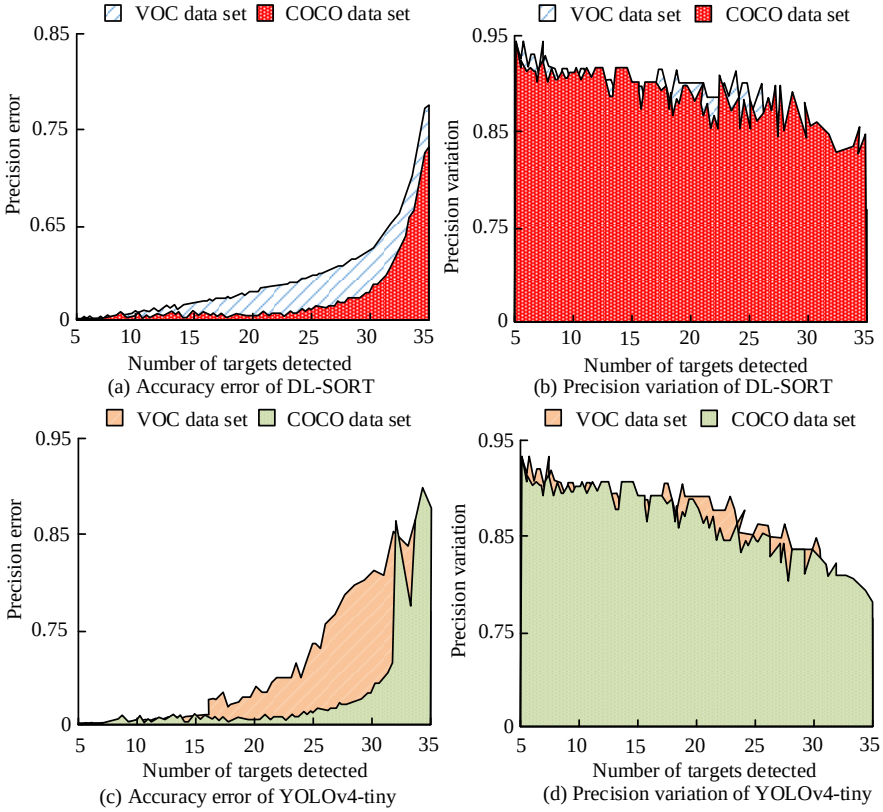


Figure 8: Accuracy error and precision variation test results

As shown in Figure 8(a), in the error variation test, the DL-SORT hybrid algorithm's error increased as the number of target detections increased. In Figure 8(b), DL-SORT achieved a detection accuracy of 0.91 on the COCO dataset when the number of target detections was 15. In Figure 8(c), when the number of target detections was 35, YOLOv4-tiny showed an error value of 0.80 on the VOC dataset, which was relatively higher compared to DL-SORT. In Figure 8(d), YOLOv4-tiny's detection accuracy fluctuated significantly when the number of target detections ranged from 20 to 25. These results indicate that DL-SORT demonstrated smaller changes in accuracy during target detection. To further enhance the comparison, the study also included Fully Convolutional You Only Look Once (FC-YOLO) and Feature Pyramid Network (FPN). The selected train test images are respectively from 1200 train images in VOC data set and 1300 train images in COCO data set. The comparison results for detection and recognition accuracy, and average precision for different targets are shown in Table 2.

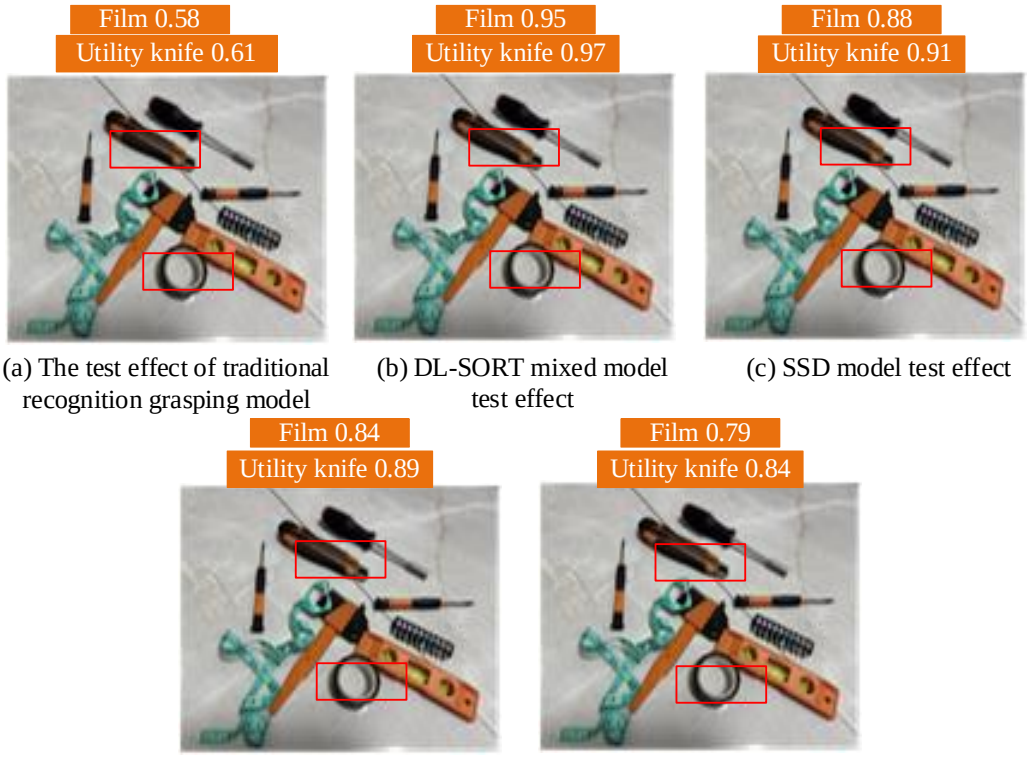
Table 2: Comparison results of detection and recognition accuracy and average precision (%)

Dataset	Algorithm	Person	Bicycle	Electric vehicle	Train	Truck	Mean Average Precision	Mean response time/ms
VOC	DL-SORT	92.6	91.8	93.4	95.8	94.2	93.5	15
	SSD	90.0	89.2	88.3	87.6	84.3	87.8	20
	CSR-DCF	78.2	78.6	79.1	74.6	71.3	76.3	35
	YOLOv4-tiny	84.5	88.7	86.1	84.2	86.3	85.9	32
	FC-YOLO	78.3	80.5	81.4	85.6	84.1	81.9	19
	FPN	79.2	78.4	79.6	80.2	81.3	79.7	37
COCO	DL-SORT mix	92.8	95.6	96.5	90.1	92.3	93.4	17
	SSD	88.8	89.4	90.1	84.3	83.6	87.2	26
	CSR-DCF	78.7	78.6	74.1	76.3	75.5	76.6	39
	YOLOv4-tiny	85.2	84.3	84.6	84.7	87.9	85.3	34
	FC-YOLO	79.3	80.1	83.6	82.5	87.4	82.5	24
	FPN	75.6	79.8	77.4	81.2	78.6	78.5	29

As shown in Table 2, when tested on the VOC dataset, the DL-SORT hybrid algorithm achieved a detection accuracy of 92.6% for people, 95.8% for trains, and an average detection accuracy of 93.5% for five different categories of targets. SSD achieved a detection accuracy of 88.3% for electric vehicles, with an overall average detection accuracy of 87.8%. The average image processing response time of DL-SORT hybrid algorithm is 15ms, while the average image processing response time of YOLOv4-tiny and SSD target detection algorithms is 32 ms and 20 ms respectively. The proposed algorithm is faster in image processing. When tested on the COCO dataset, FC-YOLO achieved a detection accuracy of 87.4% for trucks, 82.5% for trains, with an overall average detection accuracy of 82.5%. In summary, DL-SORT hybrid algorithm has strong multi-object processing capability and meets the realtime requirements of industrial detection.

4.2 Performance analysis of the hybrid model for electronic device recognition and grasping

After verifying the DL-SORT hybrid algorithm, in order to demonstrate the advantages of automatic recognition accuracy and robustness of the hybrid model for automatic recognition of electronic devices based on DL-SORT hybrid algorithm, it is also compared with the electronic device automatic recognition grasping model composed of SSD, CSR-DCF and YOLOv4-tiny target detection algorithms. The experimental environment used Visual Studio to create a visual platform for target recognition, equipped with a Cortex M7 chip and the STM32F103C8T microcontroller as the main controller. The image of the accuracy of the research test is mainly from the custom data set, which contains 1000 images and supports multi-tasks such as detection, segmentation and key points. The accuracy index of the test synthesizes the classification accuracy and positioning accuracy of the model electronic devices, and mainly calculates the intersection ratio between the sample prediction frame and the real frame. The study performed parts recognition and grasping tests with the four electronic device automatic recognition and grasping models, and the test results are shown in Figure 9.



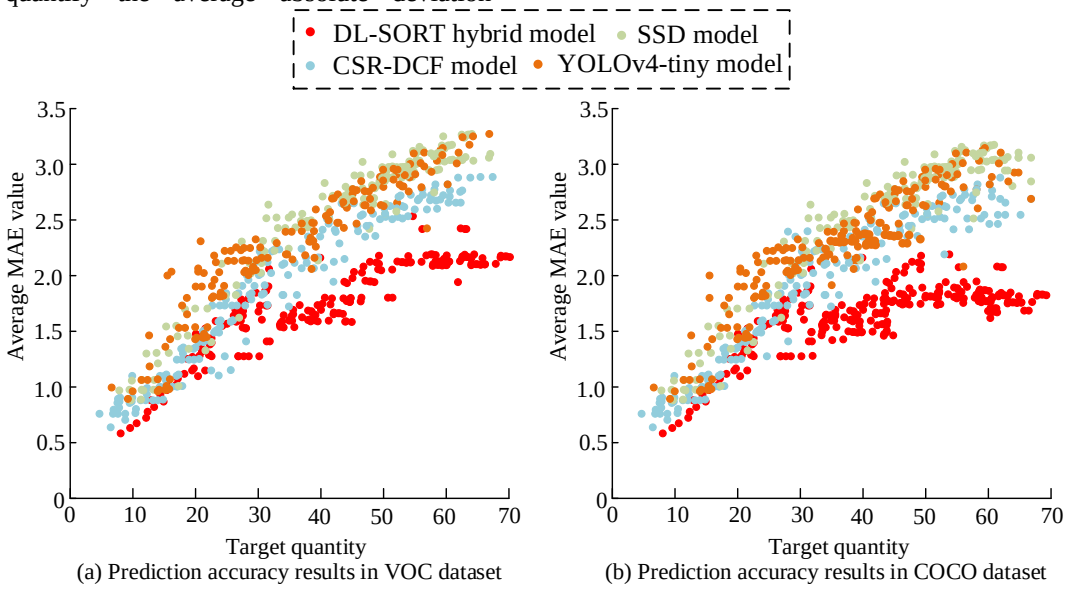
(a) The test effect of traditional recognition grasping model (b) DL-SORT mixed model test effect (c) SSD model test effect

(d) CSR-DCF model test effect (e) YOLOv4-tiny model test effect

Figure 9: Parts recognition and grasping test results

From Figure 9(a), it can be seen that the traditional recognition and grasping model achieved an accuracy of 0.61 for tool identification and 0.38 for tape recognition and grasping. In Figure 9(b), the DL-SORT model demonstrated recognition accuracy above 0.90 for both tool and tape, demonstrating superior recognition and grasping performance overall. As shown in Figure 9(c), SSD achieved an accuracy of 0.88 for tape recognition, with a 0.07 difference in accuracy compared to DL-SORT. In Figure 9(d), CSR-DCF achieved recognition accuracy above 0.83 for both tape and tool. Figure 9(e) shows that YOLOv4-tiny achieved an accuracy below 0.80 for tape

recognition. Overall, the DL-SORT model demonstrated superior recognition and grasping performance, with better part recognition capability. Additionally, the study compared the prediction performance of the four models—DL-SORT, SSD, CSR-DCF, and YOLOv4-tiny. The average absolute error was selected as the evaluation index because it can directly quantify the average absolute deviation between the grasping position predicted by the model and the real position. The average absolute error can directly reflect the small deviation of the electronic device and capture the center of the device, avoiding the limitation of IoU only focusing on the overlap area of the frame. The results of Mean Absolute Error (MAE) value of prediction accuracy are shown in Figure 10.



(a) Prediction accuracy results in VOC dataset (b) Prediction accuracy results in COCO dataset

Figure 10: Prediction accuracy MAE value result diagram

As shown in Figure 10(a), as the number of targets increased, the error of the DL-SORT model also increased proportionally. When the target count reached 70, the MAE value was 2.12, this meets the grasping accuracy requirements for electronic devices. The average absolute error of prediction of the model constructed by SSD, CSR-DCF and YOLOv4-tiny is 3.12, 2.82 and 3.01, which are inferior to the fusion model. The overall MAE value of SSD was higher, with its accuracy lower than that of DL-SORT. In Figure 10(b), when the target number ranged from 40 to 50, the MAE value of DL-SORT ranged from 1.72 to 1.84. YOLOv4-tiny showed an MAE value of 2.23 to 2.41 when the target count was between 20 and 30. The overall MAE value of CSR-DCF was higher than that of DL-SORT, with its MAE value ranging from 2.71 to 2.98 when the target count was between 50 and 60. To further showcase the grasping performance of the DL-SORT model under different environments, the study conducted target grasping experiments with targets numbered 0 to 7. The results of these experiments are presented in Table 3.

Table 3: Target grabbing experiment test results

Evaluation index	0	1	2	3	4	5	6	7
Number of successful attempts without occlusion	49	48	49	45	49	47	48	49
Average grab period without occlusion/s	11.5	11.2	11.8	11.2	11.4	12.1	11.1	11.2
Success rate without occlusion	98%	96%	98%	90%	98%	94%	96%	98%
Number of successful attempts under occlusion	42	43	41	42	41	40	43	42
Average grab period under occlusion/s	12.8	12.4	12.1	12.6	12.4	11.9	12.6	12.4
Success rate under occlusion	84%	86%	82%	84%	82%	80%	86%	84%

As shown in Table 2, when the DL-SORT model performed grasping detection on part 0, the success rate under the non-obstructed condition was 98%, with 49 successful attempts. For part 5, the average grasping cycle time under the non-obstructed condition was 11.5 s, which was the longest among all test parts. Overall, the success rate for grasping in the non-obstructed environment remained above 90%. When tested on part 1 in an obstructed environment, the DL-SORT model achieved 43 successful attempts, with a success rate of 84%, a 14% decrease compared to the non-obstructed environment. For part 5, the success rate dropped to 80%,

with an average grasping cycle time of 11.9 s. In conclusion, the DL-SORT-based electronic device automatic recognition and grasping hybrid model demonstrated a high success rate and better time control when grasping different parts. In order to better demonstrate the electronic device target recognition accuracy of DL-SORT electronic device automatic recognition and grasp hybrid model, the ablation experiment was conducted with six electronic device automatic recognition and grasp models, namely YOLOv4, SSD, CSR-DCF, DL(no SORT), SORT(no DL) and BRIEF. The experimental results are shown in Table 4.

Table 4: Comparison results of device recognition accuracy of different models (mAP±Standard deviation, %)

Model	Class A device	Class B device	Class C device	mAP	t	*P*
DL-SORT	92.6±0.5	91.8±0.4	93.4±0.3	93.5±0.3	/	/
DL(no SORT)	89.1±0.6	88.3±0.5	90.2±0.4	89.9±0.4	4.72	<0.001**
SORT(no DL)	75.3±1.2	74.8±1.1	76.5±1.0	75.4±1.1	8.92	<0.001**
BRIEF	74.1±0.8	71.6±1.1	72.3±1.3	74.2±1.3	3.89	0.002**
YOLOv4	84.5±0.8	88.7±0.7	86.1±0.6	85.9±0.8	8.91	<0.001**
SSD	90.0±0.7	89.2±0.6	88.3±0.5	87.8±0.7	5.34	<0.001**
CSR-DCF	78.2±1.2	78.6±0.9	79.1±0.8	76.3±1.2	7.36	<0.001**

*Note: ** represents a significant difference from the research model.*

As can be seen from Table 4, mAP of the DL(without SORT) electronic device automatic recognition capture model decreased by 3.6% compared with the research construction model, indicating that SORT contributed significantly to the tracking stability ($p<0.001$**). The mAP of the SORT (without DL) model for electronic device automatic recognition and grasping is 75.4%±1.1%. The mAP of the BRIEF electronic device automatic recognition capturing model is 74.2%±1.3%, and the accuracy of the recognition capturing device is significantly lower than that of the constructed model. Compared with the mAP model of YOLOv4 electronic device automatic recognition and capture, the constructed model has improved by 7.6%, especially for Class A devices. In summary, DL-SORT demonstrates the best overall performance among the hybrid model for automatic recognition and grasp of electronic devices, and SORT module and DL contribute the most to the accuracy of device recognition.

5 Discussion

The proposed DL-SORT algorithm shows significant advantages in performance comparison experiments. From multiple dimensions such as target detection accuracy, average recognition accuracy and average response time, DL-SORT hybrid algorithm is superior to SSD, CSR-DCF, YOLOv4-tiny, FC-YOLO and FPN target detection algorithms. The five baseline algorithms were chosen for comparison because they offer fast computation speed and low resource consumption, which meet the high realtime requirements and the scenarios with limited hardware resources. However, the traditional Faster Region-based Convolutional Neural Network (Faster R-CNN) object detection algorithm has high computational cost and cannot meet the realtime requirements, so it is not selected to compare with the proposed algorithm. In the test of target detection and recognition accuracy and average accuracy, DL-SORT hybrid algorithm conducts target detection for five types of images including people, bicycles, electric vehicles, trains and trucks, and the accuracy is above 90%. However, the accuracy of YOLOv4-tiny object detection algorithm for these five types of images is below 90%, and the overall accuracy of SSD and CSR-DCF object detection algorithms is not high. Moreover, the mAP value of the proposed model is higher than that of SSD, CSR-DCF and YOLOv4-tiny. The reason is that the DL-SORT hybrid algorithm in the research and construction model deeply integrates DL and SORT, and improves the target tracking ability. Moreover, the applied SORT predicts the moving trajectory through Kalman filter, which is more stable for tracking the uniformly moving target. However, YOLOv4-tiny, SSD and CSR-DCF lack a tracking module and rely on post-processing based on multi-frame association. Their feature extraction capabilities are also relatively weak, so they are not superior to the model detection devices proposed in the research. In addition, when DL-SORT, SSD, CSR-DCF, YOLOv4-tiny, FC-YOLO and FPN were tested in VOC data set, the average image processing response time was 15 ms, 20 ms, 35 ms, 32 ms, 19 ms and 37 ms, respectively. The research algorithm has the fastest processing time. This is due to the excellent dynamic parallel computing capability of DL-SORT hybrid algorithm for target detection and tracking. Although the DL-SORT algorithm excels in target tracking, the hybrid model may underperform in highly dynamic scenes, such as car racing or drone-captured football matches captured by drones. When DL and SORT are processed jointly in this model, the frame rate will be affected by the high-speed target, and the computation extension will gradually increase.

6 Conclusion

To address the precision issues in component detection, this study developed a DL-SORT-based hybrid model for automatic recognition and grasping of electronic components. In the construction of the model, an improved visual feature recognition method, combining recursive labeling and BRIEF, was used to optimize image acquisition and enhance the target recognition capability of the electronic component model. Experimental results showed that the DL-SORT hybrid algorithm achieved a recognition accuracy of 92.6% for human detection, while the SSD, CSR-DCF, and YOLOv4-tiny target detection algorithms had recognition accuracies of 90.0%, 78.2%, and 84.5%, respectively, all of which were lower than the proposed algorithm. Furthermore, empirical analysis of the constructed electronic component automatic recognition and grasping hybrid model revealed an MAE of 2.12 when the target count reached 70, meeting the accuracy requirements for electronic device grasping. While models constructed using SSD, CSR-DCF, and YOLOv4-tiny target detection algorithms had MAE values of 3.12, 2.82, and 3.01, respectively, all of which were higher than that of DL-SORT. In conclusion, the electronic component automatic recognition and grasping hybrid model based on the DL-SORT multi-target visual detection algorithm enhances the intelligent recognition capability of electronic components and significantly improves production efficiency. However, the current experiments still have limitations, such as computational resource consumption and data recognition constraints. Therefore, future work could expand the target recognition range and enhance the experimental validity.

References

[1] Yin W, He K, Xu D, Yue Y, Luo Y. Adaptive low light visual enhancement and high-significant target detection for infrared and visible image fusion. The Visual Computer, 2023, 39(12):6723-6742. https://doi.org/10.1007/s00371-022-02759-w

[2] Tang D, Yu W, Lv X, Wang, G., Shen W. Research progress on pole-climbing robots: a review. Recent Patents on Engineering, 2024, 18(8):32-59. https://doi.org/10.2174/1872212118666230914104239

[3] Ruan D, Zhang W, Qian D. Feature-based autonomous target recognition and grasping of industrial robots. Personal and Ubiquitous Computing, 2023, 27(3):1355-1367. https://doi.org/10.1007/s00779-021-01589-2

[4] Liu J, Liu Z. The vision-based target recognition, localization, and control for harvesting robots: a review. International Journal of Precision Engineering and Manufacturing, 2024, 25(2): 409-428. https://doi.org/10.1007/s12541-023-00911-7

[5] Tan Y. Highway visibility detection using hough circle detection and incremental probabilistic neural networks. Informatica, 2024, 48(23): 91-105. https://doi.org/10.31449/inf.v48i23.6539

[6] Sun X, Liu K, Chen L, Cai Y, Wang H. LLTH-YOLOv5: a real-time traffic sign detection algorithm for low-light scenes. Automotive Innovation, 2024, 7(1):121-137. https://doi.org/10.1007/s42154-023-00249-w

[7] Zhang X Y, Fan R, Liu W M, Xue, J F, Liu Q C. Optimization method for robot moving object recognition and grasping strategy based on binocular vision. Journal of Computers, 2024, 35(1):207-215. https://doi.org/10.53106/199115992024023501016

[8] Lin S, Zeng C, Yang C. Robot grasping based on object shape approximation and LightGBM. Multimedia Tools and Applications, 2024, 83(3):9103-9119. https://doi.org/10.1007/s11042-023-15547-y

[9] Wang Y, Guo Z, Chen Y, Guo C, Xia M, Qi T. A robot grasping detection network based on flexible selection of multi-modal feature fusion structure. Applied Intelligence, 2024, 54(6):5044-5061. https://doi.org/10.1007/s10489-024-05427-9

[10] Liu W, Xu M, Jiang H. Design, integration, and experiment of transplanting robot for early plug tray seedling in a plant factory. AgriEngineering, 2024, 6(1):678-697. https://doi.org/10.3390/agriengineering6010040

[11] Yan S, Zhang L. CR-Net: Robot grasping detection method integrating convolutional block attention module and residual module. IET Computer Vision, 2024, 18(3):420-433. https://doi.org/10.1049/cvi2.12252

[12] Li F, Xiao K, Hu Z, Zhang G. Fabric defect detection algorithm based on improved YOLOv5. The Visual Computer, 2024, 40(4):2309-2324. https://doi.org/10.1007/s00371-023-02918-7

[13] Wang H, Fu L, Wang L. Detection algorithm of aircraft skin defects based on improved YOLOv8n. Signal, Image and Video Processing, 2024, 18(4):3877-3891. https://doi.org/10.1007/s11760-024-03049-9

[14] Ji C, Zhang F, Huang X, Song Z, Hou W, Wang B, Chen G. STAE-YOLO: Intelligent detection algorithm for risk management of construction machinery intrusion on transmission lines based on visual perception. IET Generation, Transmission & Distribution, 2024, 18(3):542-567. https://doi.org/10.1049/gtd2.13093

[15] Xiong J, Wu J, Tang M, Xiong P, Huang Y, Guo H. Combining YOLO and background subtraction for small dynamic target detection. The Visual Computer, 2025, 41(1): 481-490. https://doi.org/10.1007/s00371-024-03342-1

[16] Boudraa M, Bennour A, Mekhaznia T, Alqarafi A, Marie R R, Al-Sarem M, Dogra A. Revolutionizing historical manuscript analysis: a deep learning approach with intelligent feature extraction for script classification. Acta Informatica Pragensia, 2024, 13(2):251-272. https://doi.org/10.18267/j.aip.239

[17] He B, Qian S, Niu Y. Visual recognition and location algorithm based on optimized YOLOv3 detector and RGB depth camera. The Visual Computer, 2024, 40(3):1965-1981. https://doi.org/10.1007/s00371-023-02895-x

[18] Yadav S P, Jindal M, Rani P, de Albuquerque V H C, dos Santos Nascimento C, Kumar M. An improved deep learning-based optimal object detection system from images. Multimedia Tools and Applications, 2024, 83(10):30045-30072. https://doi.org/10.1007/s11042-023-16736-5

[19] Chen G Y, Krzyzak A. Face recognition via selective denoising, filter faces and hog features. Signal, Image and Video Processing, 2024, 18(1): 369-378. https://doi.org/10.1007/s11760-023-02769-8

[20] Çakır M, Ekinci M, Kablan E B, Şahin M. AVD-YOLOv5: a new lightweight network architecture for high-speed aortic valve detection from a new and large echocardiography dataset. Medical & Biological Engineering & Computing, 2024, 62(8):2511-2528. https://doi.org/10.1007/s11517-024-03090-3

[21] Ma J. A High Performance computing web search engine based on big data and parallel distributed models. Informatica, 2024, 48(20): 27-38. https://doi.org/10.31449/inf.v48i20.6776

[22] Patnaik S. Speech emotion recognition by using complex MFCC and deep sequential model. Multimedia Tools and Applications, 2023, 82(8):11897-11922. https://doi.org/10.1007/s11042-022-13725-y

[23] Zhong M, Zhou Z. 3D reconstruction of grassland landforms using intelligent robot vision and numerical simulation. Informatica, 2024, 48(15):179-190. https://doi.org/10.31449/inf.v48il5.6294