applied sciences

MDPI

Article

Research on an Eye Control Method Based on the Fusion of Facial Expression and Gaze Intention Recognition

Xiangyang Sun [1,2,*,†] and Zihan Cai [1,*,†]

[1] School of Electronics and Information, Changchun University, Changchun 130022, China
[2] Key Laboratory of Intelligent Rehabilitation and Barrier-Free for the Disabled, Changchun University, Changchun 130022, China
* Correspondence: sunxy@ccu.edu.cn (X.S.); c1079611702@163.com (Z.C.)
† These authors contributed equally to this work.

Abstract: With the deep integration of psychology and artificial intelligence technology and other related technologies, eye control technology has achieved certain results at the practical application level. However, it is found that the accuracy of the current single-modal eye control technology is still not high, which is mainly caused by the inaccurate eye movement detection caused by the high randomness of eye movements in the process of human–computer interaction. Therefore, this study will propose an intent recognition method that fuses facial expressions and eye movement information and expects to complete an eye control method based on the fusion of facial expression and eye movement information based on the multimodal intent recognition dataset, including facial expressions and eye movement information constructed in this study. Based on the self-attention fusion strategy, the fused features are calculated, and the multi-layer perceptron is used to classify the fused features, so as to realize the mutual attention between different features, and improve the accuracy of intention recognition by enhancing the weight of effective features in a targeted manner. In order to solve the problem of inaccurate eye movement detection, an improved YOLOv5 model was proposed, and the accuracy of the model detection was improved by adding two strategies: a small target layer and a CA attention mechanism. At the same time, the corresponding eye movement behavior discrimination algorithm was combined for each eye movement action to realize the output of eye behavior instructions. Finally, the experimental verification of the eye–computer interaction scheme combining the intention recognition model and the eye movement detection model showed that the accuracy of the eye-controlled manipulator to perform various tasks could reach more than 95 percent based on this scheme.

Keywords: eye–computer interaction; intent recognition; facial expression; attention mechanisms

Citation: Sun, X.; Cai, Z. Research on an Eye Control Method Based on the Fusion of Facial Expression and Gaze Intention Recognition. *Appl. Sci.* **2024**, *14*, 10520. https://doi.org/10.3390/app142210520

Academic Editors: Thomas Lindner and Antonio Fernández-Caballero

Received: 22 August 2024
Revised: 25 October 2024
Accepted: 6 November 2024
Published: 15 November 2024

1. Introductory

The development goal of human–computer interaction technology is to create systems that can seamlessly understand human intentions. Visual information, including eye movements and facial expressions, has become a key factor in achieving this goal. In recent years, significant progress has been made in eye movement and facial expression recognition technology through deep learning methods.

Although eye movement recognition technology has achieved some success in controlled environments, its performance is still challenged when dealing with complex dynamic environments. For example, Tatsumi et al. [1] improved the accuracy of eye movement recognition through deep learning algorithms, but in practical application scenarios such as changing lighting conditions or user fatigue, the performance of single-mode eye tracking systems declined dramatically. In addition, Duchowski et al. [2] found that eye tracking systems may not be accurate when the user is wearing glasses or contact lenses.

Moreover, facial expression recognition technology faces similar challenges. Lucey et al. [3] used convolutional neural networks to effectively recognize facial expressions,

but small expression amplitude or a lack of expression may cause the system to be unable to accurately recognize user intent [4]. Single-modal data may not provide enough information to understand complex user behavior; for example, in the absence of contextual information, a rapid eye movement may be misinterpreted as a spontaneous eye movement rather than an expression of intent [5].

Given the limitations of single-modal systems, researchers began experimenting with multiple-mode fusion data supplementation methods to improve the accuracy and reliability of interactions. By combining information from multiple modes, such as eye movements and facial expressions, a multimodal approach can capture the user's interaction intent more comprehensively. For example, Berndt and Hall [6] used a hidden Markov model to design an intention prediction mechanism that can identify a driver's intention more quickly and accurately. Plopski et al. [4] used eye tracking technology instead of a mouse on the computer, and the interactor could choose different objects to interact with.

In 2023, Apple's Vision Pro integrated eye tracking technology into the core interaction, using a multimodal interaction model of "eye tracking + gesture + voice", which raised the importance of eye movement interaction to an unprecedented level. This multimodal interaction mode uses the movement of the eyes to select the object of interaction and uses gestures to determine or switch, which significantly improves the naturalness and intuitiveness of human–computer interaction. However, multimodal fusion also brings new challenges, including hardware systems that require more efficient fusion algorithms and higher operating speeds [7].

In this paper, we introduce an innovative interaction scheme for the fusion of eye movement and facial expression data that can significantly enhance the system's ability to recognize user intent and effectively overcome the limitations of single-modal technologies, such as challenges in changing lighting conditions, user fatigue, or the use of glasses. We propose that the combination of multimodal information can provide a more comprehensive understanding of user interaction intent and enable more natural and intuitive human–computer interactions.

The significance of this research lies in the fact that eye tracking and facial expression recognition technologies are not limited to specific research fields but have broad applications in areas such as health monitoring, augmented reality, gaming, education, and accessibility technologies. The advancements in these technologies will enhance user experience and drive technological progress in related industries.

The experimental results validate that our multimodal fusion approach significantly improves the accuracy and naturalness of HCI. By combining facial expression and eye movement information, the interaction system achieved over 95% accuracy in various tasks, demonstrating that this method can effectively enhance the interaction experience and lay a solid technical foundation for future eye control applications.

Research Ideas

The precise interaction of eye control depends on the accurate analysis and execution of the eye movement intention and eye movement instruction of the interactor by the robot. The core problem is to obtain the optimal fusion algorithm on the basis of ensuring the recognition of facial expression intention and eye movement intention. The basic idea of this study is to preprocess the video data, use the RetinaFace and VGG-19 models to recognize facial expressions, and complete the extraction of face and expression information. Then, in the NTH frame, the eye movement information is extracted and six key eye movement features are selected from this information. The extracted features are used for intention detection by a SVM to determine whether the user has a specific interaction intention. Finally, facial expression features and eye movement features are integrated through feature fusion and attention mechanism, and weighted summation is used to obtain the final fused features, which can be used to accurately judge the user's interaction intention. The technical route studied in this paper is shown in Figure 1.

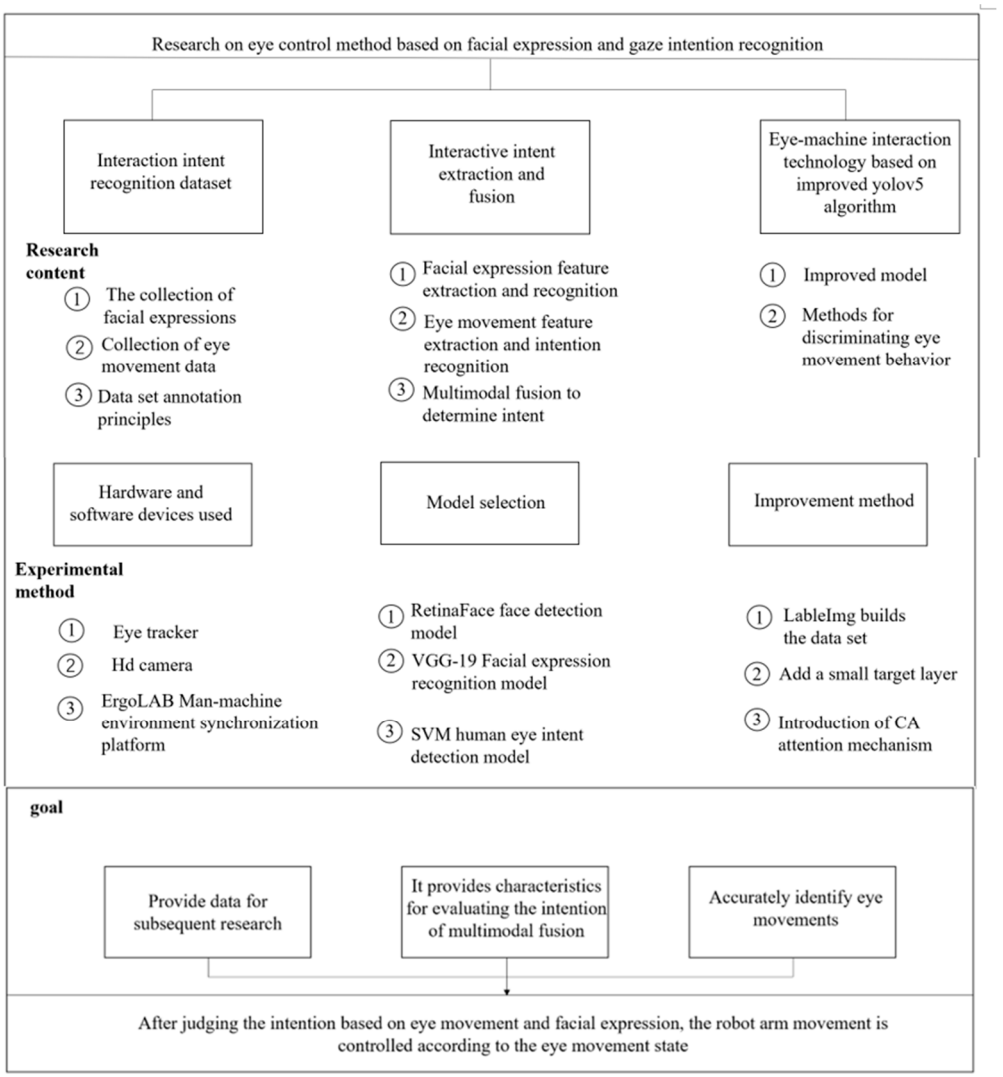


Figure 1. The technical route of this paper's research.

2. Research on the Key Technology of Eye Control Methods

2.1. Facial Expression Recognition

Facial expressions are characterized by a small spatial scale and being easily affected by light, angle, occlusion, etc. In order to enable the network to better learn the information of different characteristics, seven typical emotions such as calm, anger, sadness, happiness, fear, surprise and disgust are selected in this paper, and the features are extracted through the feature extraction network.

2.1.1. Datasets

In this article, we successfully constructed an interactive intent recognition dataset using a meticulously designed data collection scheme. This process involved capturing the facial expressions and eye movement information of 36 young participants while they were exposed to videos encompassing 7 distinct emotions, ensuring the dataset's diversity and representativeness. Prior to collecting eye tracking data, we performed calibration on the eye tracker to establish a mapping model and obtain accurate gaze position parameters. Throughout the collection process, we utilized a high-precision Tobii pro Fusion eye tracker (Tobii, Danderyd, Sweden) in conjunction with a webcam, instructing participants to sequentially fixate on target points distributed across the screen. The eye tracker recorded the coordinates of pupil center and target points, while the Tobii pro Fusion eye tracker

tracked pupils at a sampling rate of 30 Hz. Subsequently, preprocessed data were obtained through its software, which saved information regarding gaze duration and gaze position.

2.1.2. Face Image Preprocessing

Firstly, the video dataset was divided into picture datasets by extracting one frame every four frames. After cleaning, a total of 10,560 pictures were obtained. Because the image dataset obtained by direct decomposition contained a large amount of unnecessary environmental noise, it was necessary to eliminate the noisy data to increase the amount of effective information; for example, we could discard the image whose face confidence score was too low. After cleaning the image dataset, face detection was performed to determine the face location and save. In order to meet the network input requirements, the obtained face data were scaled uniformly. The picture example of the established face image dataset is shown in Figure 2.

Figure 2. Face image dataset example.

2.1.3. Example of Face Image Data Set

In order to improve the robustness and generalization ability of the facial expression recognition model, the Softmax loss function was adopted in the last fully connected layer of the VGG-19 network structure, which helped to accelerate the network convergence speed and effectively realize the seven classification tasks of facial expressions [5]. VGG-19 was chosen for this task because it performs well in image classification tasks, especially in capturing small spatial-scale features like facial expressions. Compared to other popular models like ResNet, VGG-19 strikes a balance between network depth and computational efficiency, making it effective at extracting local features from facial expressions while maintaining lower resource consumption. While ResNet excels in extracting deep features, its more complex structure results in additional computational overhead, which is not suitable for real-time applications. Additionally, although MobileNet offers advantages in lightweight architectures, its accuracy is insufficient for handling complex facial features. Therefore, VGG-19 provides a more balanced performance in terms of accuracy and resource usage, achieving higher classification accuracy in our experimental setup.

The feature vector of the Softmax layer in the expression recognition model based on VGG-19 was extracted and stored as the feature output at the decision level of facial expression information. Since the dimension of the features at the decision level was equal to the number of categories of the classification task, the seven-dimensional feature vectors of the Softmax layer could be saved to represent the confidence scores of seven different expression categories, respectively. The saving format is shown in Formula (1) [8].

$$\begin{bmatrix} N_0\ P_{0,0}\ P_{0,1}\ P_{0,2}\ P_{0,3}\ P_{0,4}\ P_{0,5}\ P_{0,6} \\ N_1\ P_{1,0}\ P_{1,1}\ P_{1,2}\ P_{1,3}\ P_{1,4}\ P_{1,5}\ P_{1,6} \\ \cdots \\ N_{(n-1)}\ P_{(n-1),0}\ P_{(n-1),1}\ \cdots\ P_{(n-1),5}\ P_{(n-1),6} \\ N_n\ P_{n,0}\ P_{n,1}\ P_{n,2}\ P_{n,3}\ P_{n,4}\ P_{n,5}\ P_{n,6} \end{bmatrix} \quad (1)$$

Among them, N represents the total number of pictures in the face image dataset, n represents the number of pictures, i represents the category of expressions represented by numbers, and represents the seven-dimensional decision-level features saved, that is, the confidence scores of various expressions.

2.2. Extraction and Recognition of Eye Movement Features

Eye movement information is a key factor in human–computer interaction and can reveal the user's focus of attention when processing information [9]. In order to effectively identify the user's interaction intent, we used an improved YOLOv5 model to detect and classify eye movements. YOLOv5 is a single-shot multi-frame detector that is capable of processing images in real time and detecting multiple objects in the image. We collected six eye movement characteristics, including the duration of fixation, fixation frequency, fixation interval, eye movement velocity, jump amplitude, and pupil diameter.

For classifying the fused features, a support vector machine (SVM) was selected due to its strong performance in handling high-dimensional feature spaces, which is essential for capturing and expressing complex user intentions. Compared to random forest (RF) and neural networks (NNs), an SVM demonstrates better generalization ability, particularly in small datasets. An SVM is able to segment data points into different categories by finding an optimal boundary in a high-dimensional space, making it very suitable for distinguishing between intentional and unintentional eye movement behavior [7]. While neural networks might perform well with large datasets, they carry a higher risk of overfitting and require extensive tuning and computational resources, making them less suitable for small datasets in this study. Random forest, while effective in certain classification tasks, may lead to increased complexity and computational costs when dealing with high-dimensional data. The experimental results showed that the SVM outperformed both RF and NN in terms of accuracy and F1 score, further validating its suitability and superiority in this study. The Eye movement intent detection flow chart is shown in Figure 3.

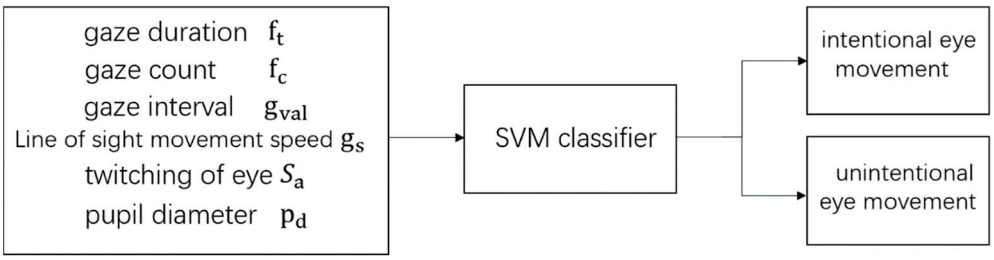


Figure 3. This eye movement intent detection flow chart describes the conversion of eye movement data to intent classification.

In this experiment, 1059 sets of eye movement feature data were collected and divided into training data and test data according to a ratio of 8:2. In order to evaluate the effectiveness of the intended eye movement detection model based on the SVM, the nearest

neighbor algorithm (KNN) and random forest algorithm (RF) were used in this study to construct a binary classification model for eye movement recognition. In the test stage, by comparing the accuracy rate, accuracy rate, recall rate, and F1 score of the eye movement recognition model based on the SVM, RF and KNN algorithms on the constructed dataset, the experimental results, as shown in Table 1, show that the SVM classifier has a good classification performance, which can be used in natural visual behaviors. The user's intentional eye movements are detected by the selected eye movement features.

Table 1. Accuracy, precision, recall, and F1 values for SVM, RF, and KNN algorithms.

Method	Accuracy	Precision	Recall	F1
SVM	90.8%	91.6%	92.7%	94%
RF	89.6%	88.4%	91.6%	91%
KNN	90.1%	87.3%	93.6%	89%

To quantitatively assess the contribution of individual features, namely facial expressions and eye movements, to the performance of the overall intent recognition system, we conducted an ablation study. The study involved training the model for each feature type independently and evaluating its performance separately. The results of this study are detailed in Table 2 below.

Table 2. Individual feature performance comparison.

Feature Type	Accuracy Rate	Precision	Recall Rate	F1 Score
Facial expression	87.6%	89.1%	92.6%	90.6%
Eye movement	90.8%	91.6%	93.7%	92.1%

In order to improve the accuracy of intention recognition through feature fusion, we compared various fusion techniques, including self-attention, joining, and bilinear pooling. Self-attention methods outperformed other methods by enabling the model to focus dynamically on the most relevant features. The results of comparative analysis are presented in Table 3.

Table 3. Comparison of fusion technologies.

Fusion Technique	Accuracy	Precision	Recall	F1 Score
Concatenation	92.0%	91.0%	93.0%	92.0%
Bilinear Pooling	92.5%	92.0%	93.5%	92.8%
Self-Attention	94.6%	93.8%	97.7%	94.5%

2.3. Interactive Intent Recognition Feature Fusion

Feature Fusion Strategy Based on Attention Mechanism

The attention mechanism fusion framework constructed in this paper is shown in Figure 4. Based on the two features extracted above, fusion was obtained after passing through an attention layer, and the support vector machine was used to classify the fused features to obtain the final intent recognition result.

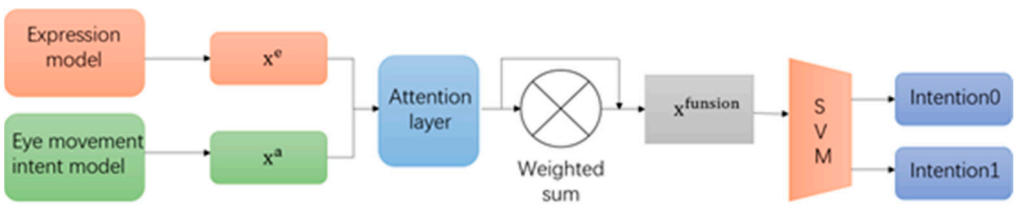


Figure 4. Integration framework based on attention mechanism.

The extracted facial expression features (marked as) and six kinds of eye movement features (marked as) judged with or without intention by the SVM were input into the classification network for feature fusion. Feature splicing is a standard feature-level fusion method widely used by many researchers. In this method, single features are connected in series to form new features, that is, the fused features are sent into the classification network for classification.

For the two features extracted in this paper, a self-attention fusion strategy was used to calculate the fused features, so that the information between different features could focus on each other. For the self-attention method, the linear projection operation was first carried out to map these features to the same vector space, and then the mapped features were spliced together, that is, the spliced features were projected into a new vector space through a multi-head bidirectional projection, expressed as the dimension of each projection space, and i = [1, 2] was the index of two different features. After two projection transformations of the features, the self-attention mechanism was used to explore the complementary relationship between the features and learn the common features between different features.

In multimodal fusion, eye movement features contribute more to accuracy compared to facial expression features, which can be explained from physiological and behavioral perspectives. Eye movements are a direct manifestation of human attention, closely linked to the user's gaze and information processing, making them more accurate in reflecting user intent. In contrast, facial expressions may be influenced by various factors such as emotional changes, lighting, and facial angles, which can lead to more indirect representations of user intent and potential recognition errors. Additionally, when facial expressions are subtle or less apparent, the system may struggle to accurately recognize user intent. By emphasizing the different roles of eye movements and facial expressions in intent recognition, the effectiveness of the multimodal fusion strategy can be better understood.

Using the fused features from the attention mechanism, the following is obtained:

$$\begin{cases} A_i = softmax\left(\frac{Z_i^1 \left(Z_i^2\right)^T}{\sqrt{D_s}}\right) Z_i^3 \\ \alpha_i = A_i q_i \\ \vartheta_i = \sum_{k=1}^{3} \alpha_i[k] A_i[k,:] \end{cases} \tag{2}$$

where $q_i \in D_s$ is the learning parameter of each feature; i is the third linear projection; all the projected features are spliced to obtain the fused feature, which is represented as $x\,fusion = [\vartheta_1, \vartheta_2, \ldots, \vartheta_p]$; and p is the number of applied projections. The hardware configuration of this paper runs on the Windows 10 operating system, the software environment runs on the PyCharm integrated development environment, and the programming language is Python. All the involved feature extraction network training and feature extraction operations are implemented in the PyTorch deep learning framework. In this paper, two comparative experiments are designed to compare the feature differences between single-feature and multi-feature fusion in intention recognition. The specific experimental 215 setup and test group configurations are detailed in Table 4.

Table 4. Feature comparison test groups' settings.

Test Group	Characteristic Information	Identification Method
Single-feature	Expression feature	VGG-19
	Eye movement feature	SVM
Multi-feature	Facial features + eye movement features + attention mechanisms	Concatenation + SVM

In this paper, three kinds of comparative experiments were used to evaluate intention. It can be seen in Figure 5 that the prediction accuracy is higher when the information combining facial expression and eye movement information is used.

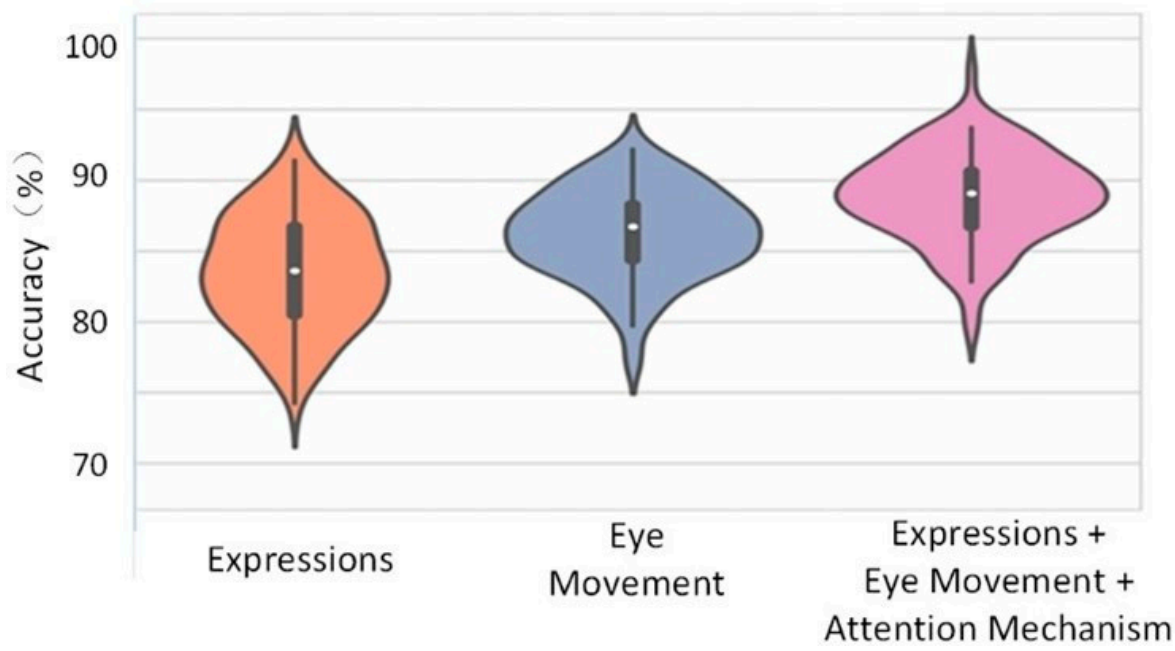


Figure 5. Comparison of performance in single-mode and multimodal prediction.

In order to ensure the consistency of the single-feature recognition method, we used the same network structure and parameters as the multi-feature model in this study. Thus, the only difference between the two models was how they handled the extracted features. Specifically, the single feature model relied only on a single feature to identify interaction intent, while the multi-feature model performed further secondary processing on the extracted single feature to improve the accuracy of interaction intent recognition by combining different features. In order to demonstrate the effectiveness of this method, we summarized the recognition results of facial expression features, eye movement features, and multimodal feature fusion, and we present these results in the form of graphs, as shown in Figure 6 and Table 5.

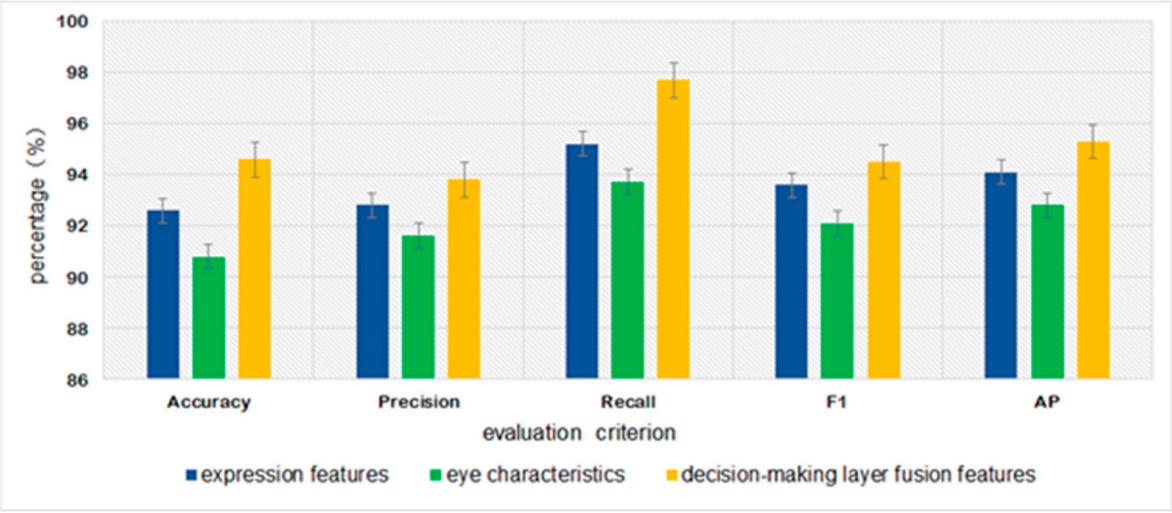


Figure 6. Line charts of five indicators of different models.

In this study, we also adopted a series of strategies to prevent overfitting problems caused by small datasets: First, we increased sample diversity by implementing data enhancement techniques, including random rotation, scaling, and color adjustment. Secondly, the early stop method was adopted to end the training process in advance by monitoring the loss of the validation set. The comprehensive application of these measures effectively improved the generalization ability of the model and reduced the risk of overfitting.

Table 5. Results of single-mode and multimodal intention recognition.

Model Class	Evaluation Index				
	Accuracy	Precision	Recall	F1	AP
Facial expression feature	87.6%	89.1%	92.6%	90.6%	89.8%
Eye movement feature	90.8%	91.6%	93.7%	92.1%	92.8%
Decision level fusion features	94.6%	93.8%	97.7%	94.5%	95.3%

The experimental results show that the accuracy of the interaction intention identification of the single-eye-movement feature using 1528 video datasets constructed in this paper was lower than that of the multi-feature fusion method, which proves that a single feature could not identify all interaction intentions completely and accurately, and there were obvious errors. It also shows that multi-feature fusion was very effective for improving the accuracy of interaction intention recognition. Table 5 reveals the significant advantages of multimodal features in improving recognition accuracy. The recognition model with multimodal decision level fusion had excellent performance in accuracy, precision, recall, F1 score, and average accuracy (AP), which were 94.6%, 93.8%, 97.7%, 94.5% and 95.3%, respectively. It is obviously better than the recognition result of the single mode. In addition, in order to balance the detection accuracy and computational efficiency of the model, a lightweight CA (Coordinate Attention) attention mechanism was introduced and compared with other advanced attention mechanisms. Table 6 reveals the potential of CA mechanisms to improve model classification accuracy, suggesting that CA mechanisms may provide an efficient optimization path for eye movement control techniques.

Table 6. Attention mechanism comparison.

Attention Mechanism	Accuracy	Precision	Recall	F1 Score
SE	93.1%	92.1%	94.0%	93.0%
CBAM	92.8%	91.5%	94.5%	93.0%
CA	96.3%	95.0%	97.5%	96.0%

3. Eye–Machine Interaction Technology Based on the YOLOv5 Network

3.1. Experimental Design of Eye–Machine Interaction Technology

3.1.1. Adaptive Anchor and Its Improvement

At present, the types of deep learning frameworks are extremely diverse. PyTorch, TensorFlow, Caffe, Keras, CNTK, MXNet, PaddlePaddle, and other mainstream frameworks have good performance in applications. Different frameworks have their unique advantages, and combined with the actual situation, there are different choices. Table 7 lists the hardware and software configurations for this study.

In the field of object detection, Anchor is a key component for feature mapping to expected bounding boxes. The adaptive Anchor method is based on the observation that different data sets require different sizes and proportions of Anchors to achieve optimal

detection results [10]. In order to improve the adaptability of the model to various size targets, we analyzed the size distribution of targets in the dataset. Based on the analysis results, we calculated a new set of Anchor sizes to better match the sizes of the targets in the dataset [11]. Therefore, we adopted an adaptive Anchor method, which could dynamically adjust the size and proportions of an Anchor according to the actual size of the target in the dataset. In order to improve the detection accuracy, the size and proportion of the Anchor should match the actual object, and the positive sample is usually taken as the IoU (intersection ratio) greater than 0.5. For datasets with intent recognition, YOLOv5's Anchor may need to be adjusted to accommodate objects of different sizes to reduce missed detection. In this paper, Anchor was improved by adding a small target layer to better adapt to detection targets. For adding the small target layer, the Anchor before and after improvement is shown in Table 8.

Table 7. Experimental environment configuration.

Experimental Environment	Version Model
Operating system	Window10
GPU	NVIDIA GeForce RTX3060 (NVIDIA, Santa Clara, CA, USA)
Compiled language	Python3.8
Compilation environment	Pycharm2021
Deep learning framework	Pytorch1.8.1

Table 8. Anchor before and after improvement.

Before Anchors:		After Anchors:	
-[10,13, 16,30, 33,23]	P3/8	-[5,6, 8,14, 15,11]	4
-[30,61, 62,45, 59,119]	P4/16	-[10,13, 16,30, 33,23]	P3/8
-[116,90, 156,198, 373,326]	P5/32	-[30,61, 62,45, 59,119]	P4/16
-[116,90, 156,198, 373,326]	P5/32		

In this study, we also adopted a series of strategies to prevent overfitting problems caused by small datasets: First, we increased sample diversity by implementing data enhancement techniques, including random rotation, scaling, and color adjustment. Secondly, the early stop method was adopted to end the training process in advance by monitoring the loss of the validation set. The comprehensive application of these measures effectively improved the generalization ability of the model and reduced the risk of overfitting.

According to the two Anchor calculation methods, the experiment was compared. The initial Anchor of YOLOv5 and the improved Anchor were, respectively, put into the model for 100 rounds of training, and the results were analyzed. The experimental results of training before and after the improvement are shown in Figure 7, which shows the change curve of the loss function before and after the improvement of the Anchor. The experimental results show that with the progress of training, the improved model shows a faster decline rate in terms of loss reduction, especially in the late training period, and the convergence of its loss function is more rapid. This observation intuitively proves that the addition of a small target layer has a positive effect on improving model training efficiency and fitting quality.

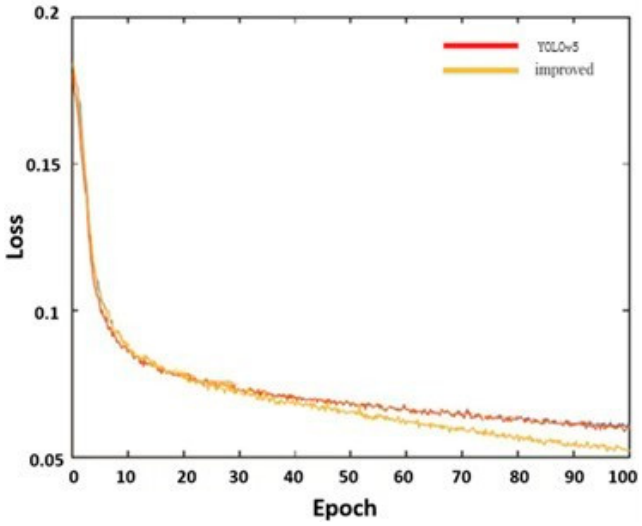


Figure 7. Loss function curve of Anchor method before and after improvement.

3.1.2. Adding an Attention Mechanism

In order to improve the performance of the YOLOv5 model in target detection tasks, this paper introduced the Coordinate Attention (CA) mechanism. The CA mechanism captures the features of the global sensitivity field by averaging pooling in the width and height direction of the feature map, and then it reduces dimensionality through a 1×1 convolution kernel and generates attention weights using the Sigmoid activation function, which is used to weight the original feature map, highlight important features, and suppress unimportant information [9]. The Structure diagram of the CA attention mechanism is shown in Figure 8.

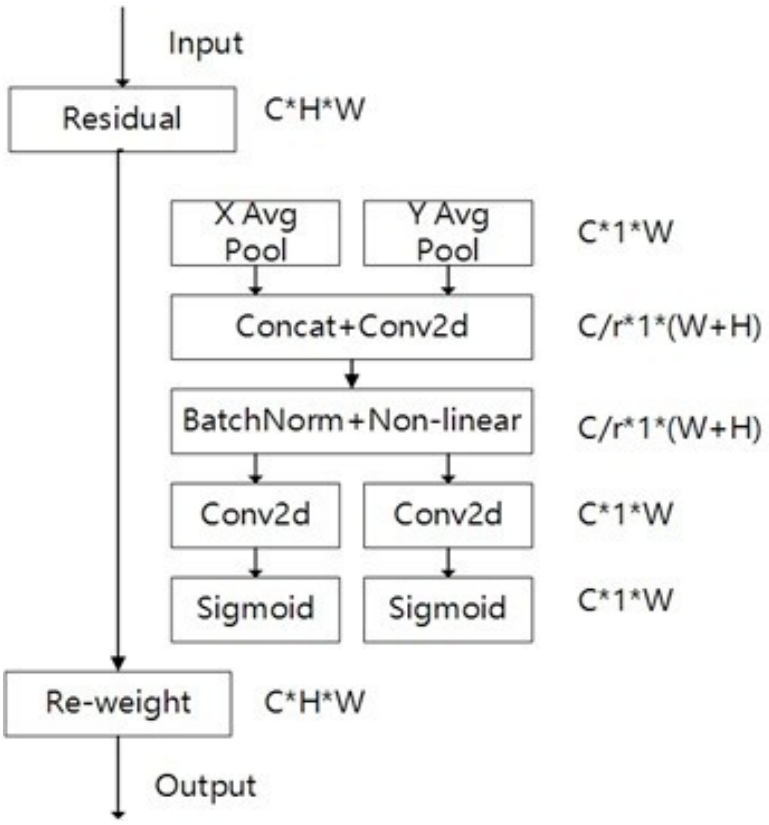
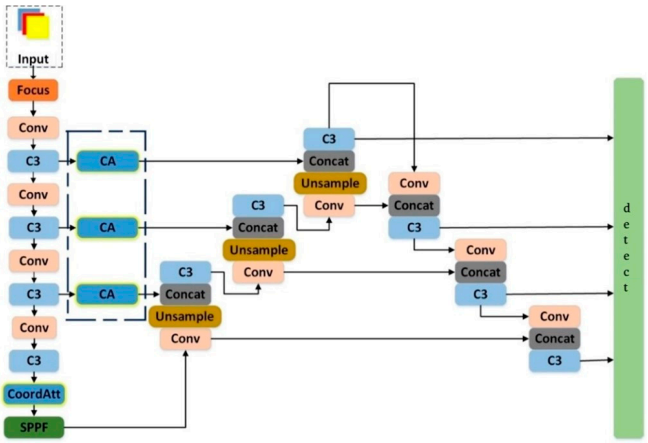


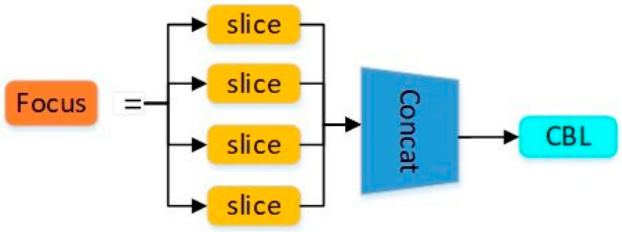
Figure 8. Structure diagram of the CA attention mechanism [9].

3.1.3. Analysis of Experimental Results of Improved Model Structure

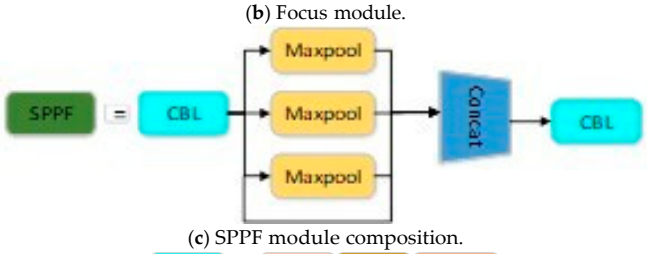
For the improvement of the Anchor and the introduction of attention mechanism optimization, the improved YOLOV5 model is shown in Figure 9. While the introduction of a small target layer and the CA (Coordinate Attention) mechanism significantly improve the model's accuracy in detecting small objects and enhancing feature representation, these enhancements come with additional computational costs. The small target layer increases computational demand, especially when processing higher resolution images, as it adds more detection layers to better handle smaller objects. Similarly, the CA mechanism requires more computational resources to capture global contextual features, increasing both the inference time and memory usage. Although these improvements boost accuracy, they may introduce challenges in resource-constrained environments, such as embedded devices or real-time applications. It is essential to balance these extra computational costs against the gains in detection precision, depending on the specific application needs.



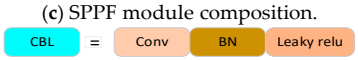
(**a**) Improved overall model structure.



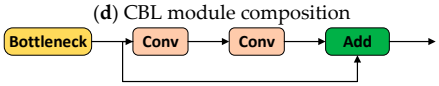
(**b**) Focus module.



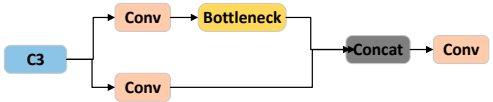
(**c**) SPPF module composition.



(**d**) CBL module composition



(**e**) A Bottleneck module is constructed

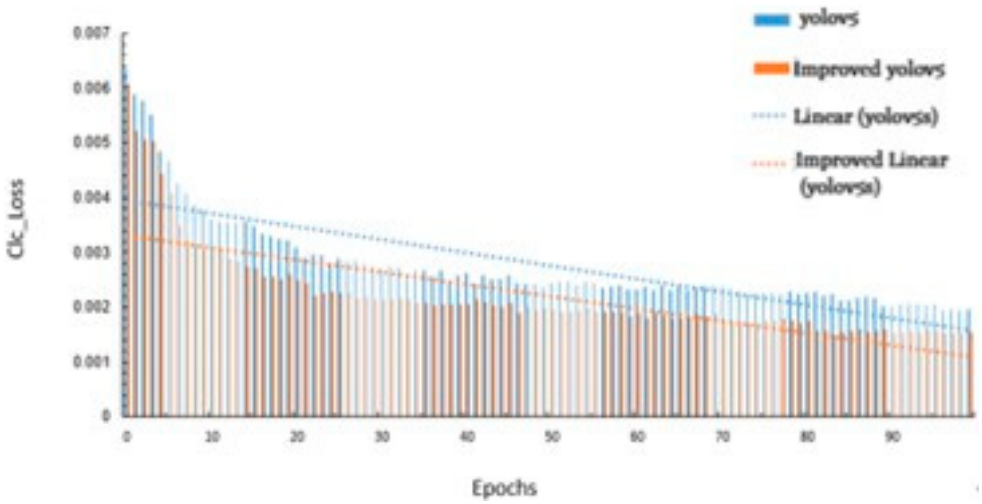


(**f**) C3 module

Figure 9. Improved YOLOv5 model structure.

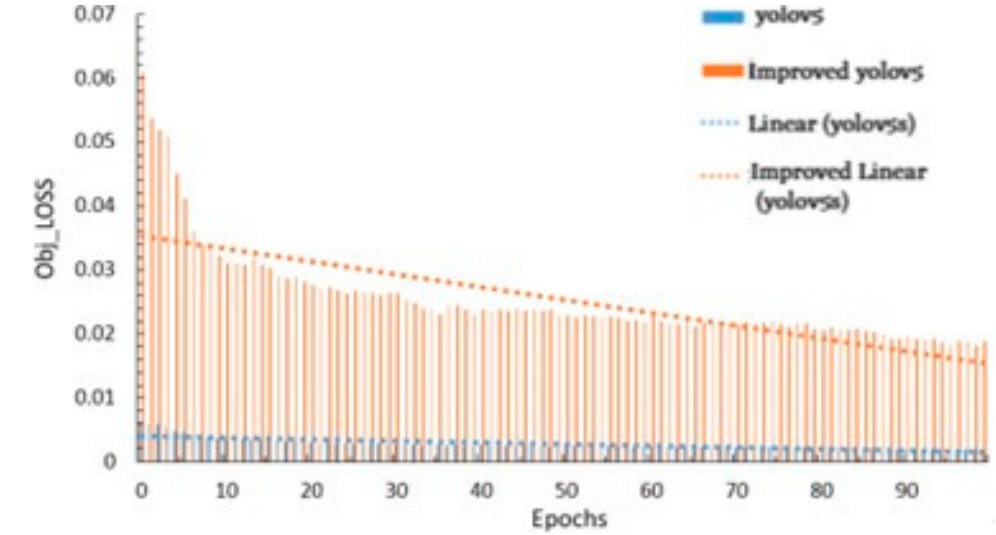
In Figure 9, (a) shows the overall improved model structure, where the part circled by the blue dashed line is the introduced CA attention mechanism. The backbone network includes Focus (Focus) and C3 modules. The structure of the Focus module is shown in (b). Operations such as slice, Concat, and the convolution (CBL) of the original feature map could reduce the calculation amount, improve the speed, and realize subsampling. The structure of SPPF module is shown in (c). After three MaxPool operations, multi-dimensional pooling and feature fusion of input features were carried out to improve the expression ability of feature maps. The composition of CBL, as shown in (d), consists of a series of convolution, batch normalization, and activation functions for better understanding and extracting features; the Bottleneck module, as shown in (e), is composed of a series of convolution layers, which were used to reduce the number of channels and spatial dimensions of feature graphs, reduce the number of parameters, and retain important feature information. The C3 module, as shown in (f), divides the input into two convolutions (CBLs) for output to enhance the model's ability to learn features at different levels. During the training of the YOLOv5 model and the improved YOLOv5 model, the same dataset and the same parameter settings were used to draw the training loss comparison curve of the two models as shown in Figure 7 according to the log files saved during the training process.

Figure 10 shows the performance of the improved YOLOv5 model in terms of target classification, location regression, and confidence loss. These loss indicators reflect, respectively, the model's ability to identify the category of the target in the image, determine the accuracy of the target location, and assess the probability of the presence of the target in the prediction box. With the progress of model training, we observed that all loss indexes showed a downward trend, indicating that the performance of the model was improving in all aspects. In addition, the steady decline in the loss function also suggests that the model was gradually converging, meaning that the network was learning and absorbing information from the training data. By comparing the loss curve between the improved model and the original model, we can evaluate the effectiveness of the introduced optimization measures, such as the new Anchor strategy and attention mechanism, in improving the model performance. Overall, the results in Figure 10 show that the model has significant learning progress and good convergence in the key aspects of object detection, which proves that the improvement measures we have taken were successful.

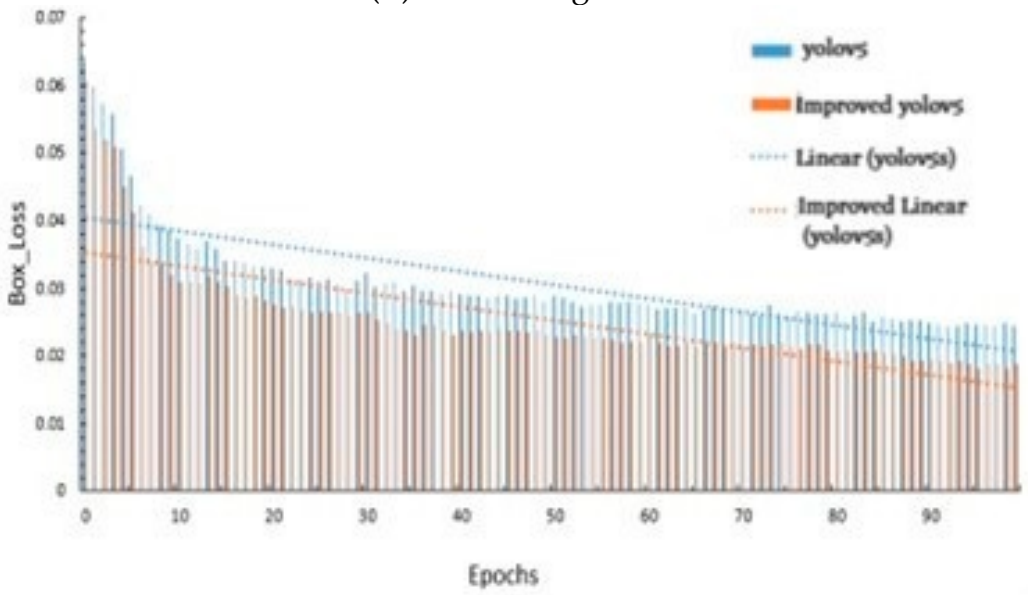


(**a**) Classification losses

Figure 10. *Cont.*



(**b**) Positioning loss



(**c**) Loss of confidence

Figure 10. Improved loss variation diagram for the YOLOv5 model.

Figure 11 provides a detailed view of the average precision (AP) of the improved YOLOv5 model in object detection tasks, which is a key metric for measuring overall model performance. The figure reveals the model's performance in recognizing different classes of objects, where each class's AP value reflects the balance between accuracy and recall of the model's predictions. Furthermore, by comparing the improved model with the original version, Figure 11 clearly demonstrates the performance improvement achieved through optimization measures such as introducing small object layers and attention mechanisms, thus confirming their effectiveness in enhancing object detection accuracy. The F1 score curve of the improved model is shown in Figure 12.

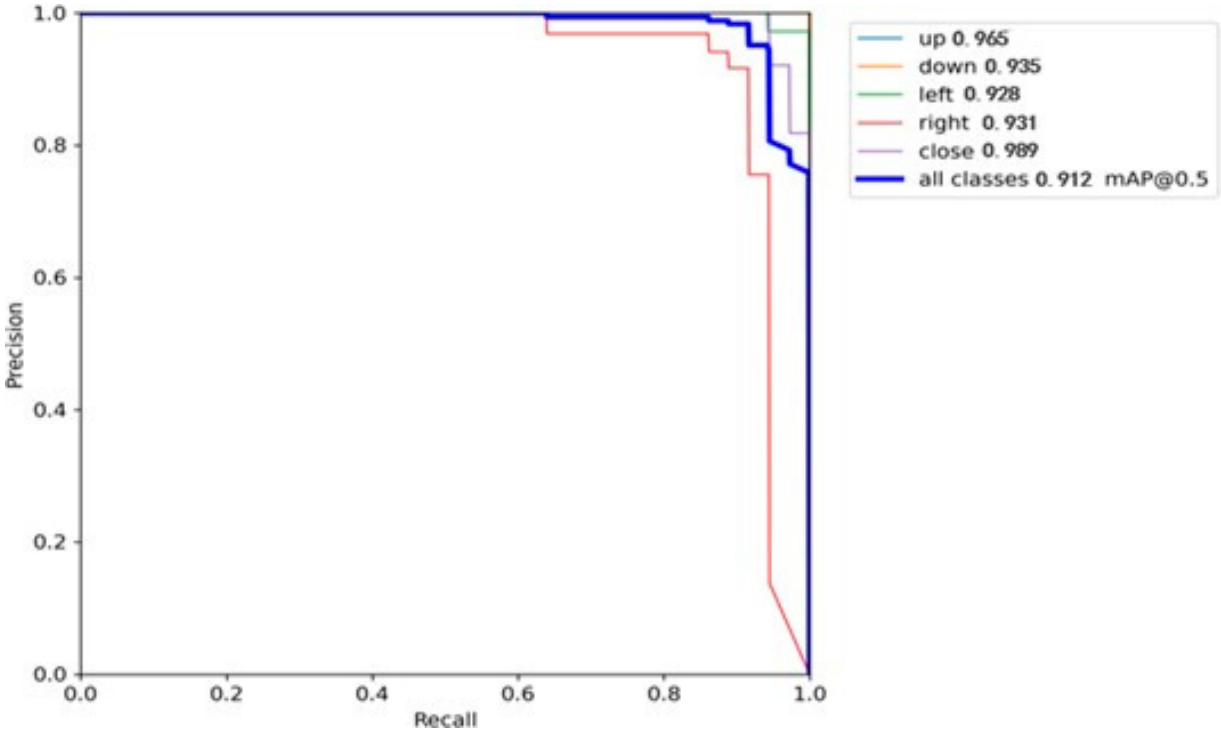


Figure 11. The average accuracy (AP) curve of the improved model.

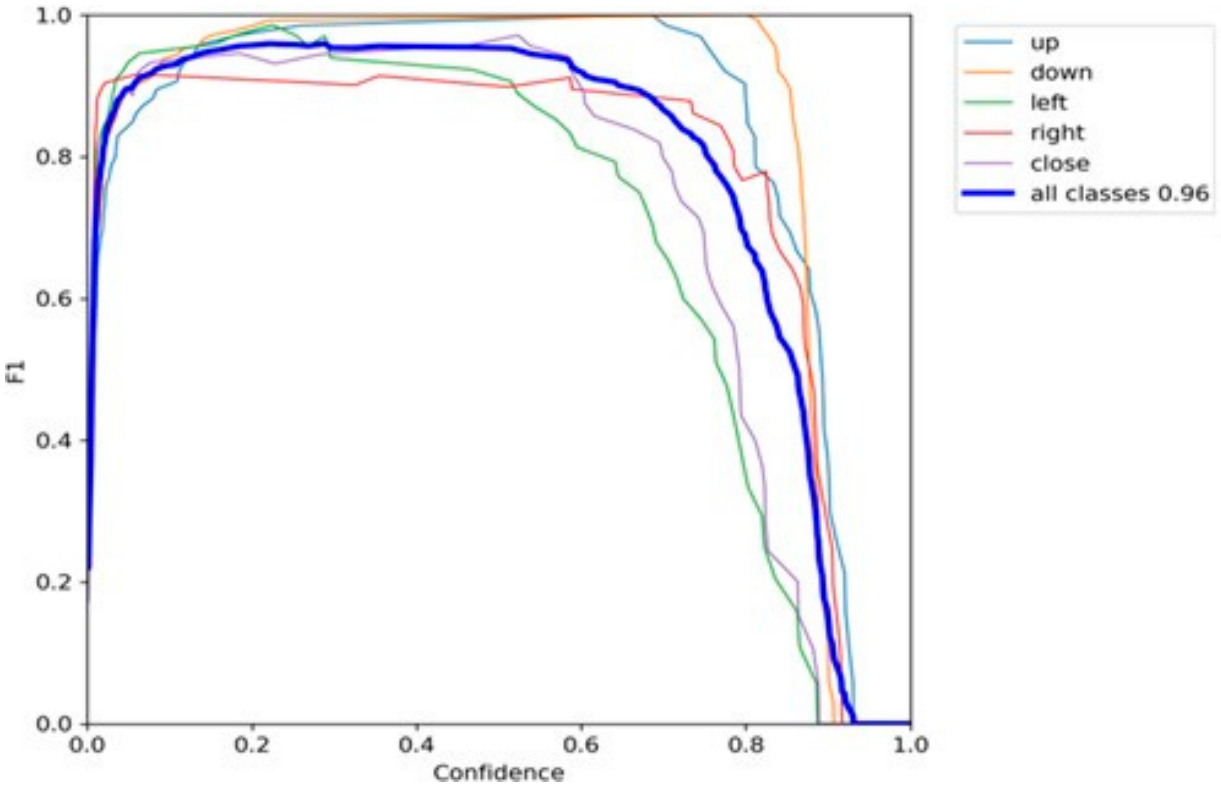


Figure 12. The F1 score curve of the improved model.

The improved YOLOv5 model, YOLOv5 model, YOLOv3-tiny model, Fast-Rcnn model, and SSD model were tested and compared, and the results are shown in Table 9. The accuracy, recall, F1 score, and average accuracy (mAP) together provide a comprehensive

assessment of the model's performance in object detection and classification tasks, revealing that the YOLOv5 model is more effective than several other models in identifying positive samples, balancing errors, and handling class imbalances.

Table 9. Comparison experiment.

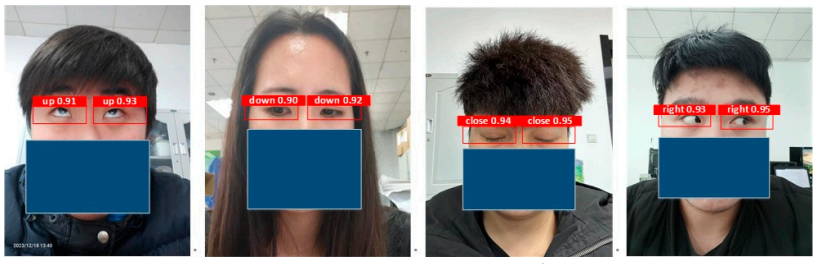
Model	Accurate/%	Recall Rate/%	F1 Score	Average Precision/%	mAP/%
SSD	87.3	89.6	88.5	86.4	43.3
YOLOv3-tiny	90.8	93.7	92.2	88.2	44.2
Fast-Rcnn	88.7	90.5	89.6	87.8	43.6
YOLOv5	92.2	95.6	93.9	88.9	44.6
Improvement of YOLOv5	96.8	97.6	96.2	91.2	48.5

In order to verify the effectiveness of the CA module used in this section and the addition of the small target layer, ablation experiments were conducted to verify the optimization effects of each improved module. The experimental results are shown in Table 10, where improvement 1 represents the addition of the small target detection layer and improvement 2 represents the addition of the attention mechanism. As can be seen from the data in the table, the average accuracy increased by 2.9 percentage points when these two improvements were combined into the model. The ability of the model to detect small targets was improved.

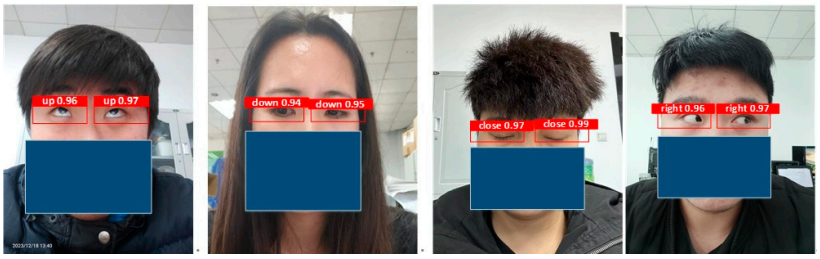
Table 10. Ablation Experiments.

Model	Join the Small Goal Tier	Incorporation of the Attention Mechanism	Average Precision/%	Mean Average Precision/%
YOLOv5s	×	×	87.5	44.6
Improvement 1	✓	×	88.5	46.3
Improvement 2	×	✓	89.3	46.5
Methodology of this paper	✓	✓	91.2	48.5

Some images were randomly selected from the dataset for testing. Some test results before and after improvement are shown in Figure 13. It can be seen that the improved YOLOv5 detection algorithm significantly exceeds the original version in target detection accuracy and can identify and locate targets more accurately.



(**a**) Pre-improvement test results



(**b**) Improved test results

Figure 13. Test results before and after improvement.

3.2. Discrimination of Eye Movement Behavior

In the stage of eye behavior discrimination, according to the criteria for eye behavior discrimination shown in Table 11, the collected eye images are classified into five eye behaviors: up and down, left and right, closed eye, single blink, and double blink. The three eye behaviors except single blink and double blink are all determined by monitoring the duration of the corresponding eye state. Single blink and double blink are used to distinguish eye behavior by monitoring whether the number of changes in "closed eye to open eye" in the eye image is 1 or 2 times in a continuous 1.5 s period.

Table 11. Criteria for eye behavior.

Ocular State	Criterion of Discrimination
Up, down, left, and right	Last more than 1.5 s
Close one's eyes	Last more than 1 s
Single wink	There was a continuous change in "closed eye—open eye" eye state within 1 s
Double wink	Two consecutive changes in "closed eye—open eye" appeared within 1.5 s

This paper proposes a set of criteria for eye movement behavior, which can be used in the eye-machine interaction technology of intelligent devices. First, for "up and down, left and right" eye movement behaviors, changes in eye state lasting more than 1.5 s are recognized as corresponding eye behavior instructions. Secondly, the criterion for judging the "closed eyes" behavior is that as long as the user maintains the "closed eyes" state for 1 s, it is determined as the "closed eyes" instruction, which effectively avoids the influence of subconscious blinking on the discrimination. Finally, for the "single blink" and "double blink" behavior, monitoring the number of changes in the "closed eye to open eye" state within 1.5 s is used to distinguish between behaviors. Since the normal blink cycle is about 0.3 to 0.4 s, taking into account individual differences, this paper sets the identification time window of "double blink" to 1.5 s and determines that two changes are detected as "double blink" instruction. These standards ensure the accuracy and practicability of the identification of eye movement behavior.

3.3. Construction of Experimental Platform

This experimental scheme aims to use the combination of facial expression and eye movement information to judge people's intentions, as shown in Figure 14. The platform mainly consists of the following parts: a six-degrees-of-freedom robot, an eye movement state capture camera, an intention recognition host, Raspberry PI, and robot–host computer.

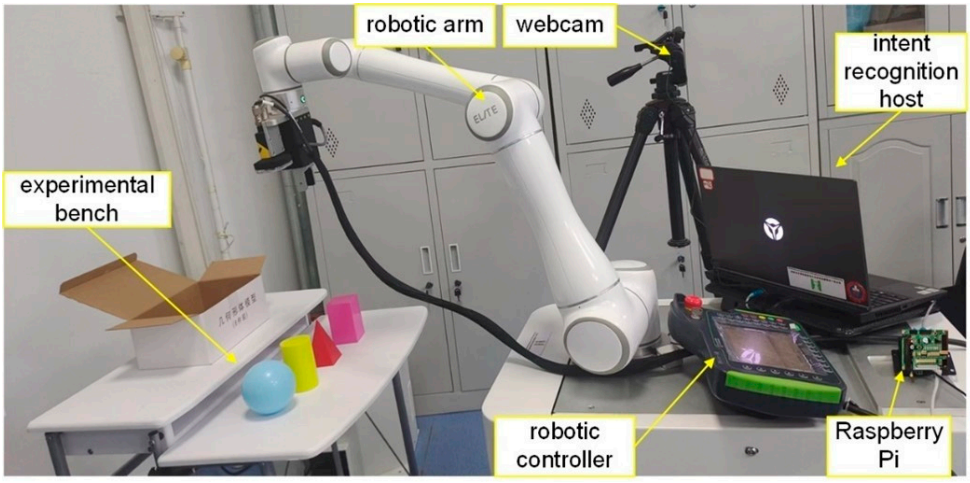


Figure 14. Human–computer interaction experiment platform.

First, the intent recognition results are sent via the PC side and sent to the Raspberry PI using the MQTT communication protocol. After the Raspberry PI receives the message, it runs the YOLOv5 model ported to the Raspberry PI for eye movement recognition. Finally, the identified eye movement state information is transmitted to the robot arm through the Modbus communication protocol for corresponding actions. The overall flow chart of the experiment is shown in Figure 15.

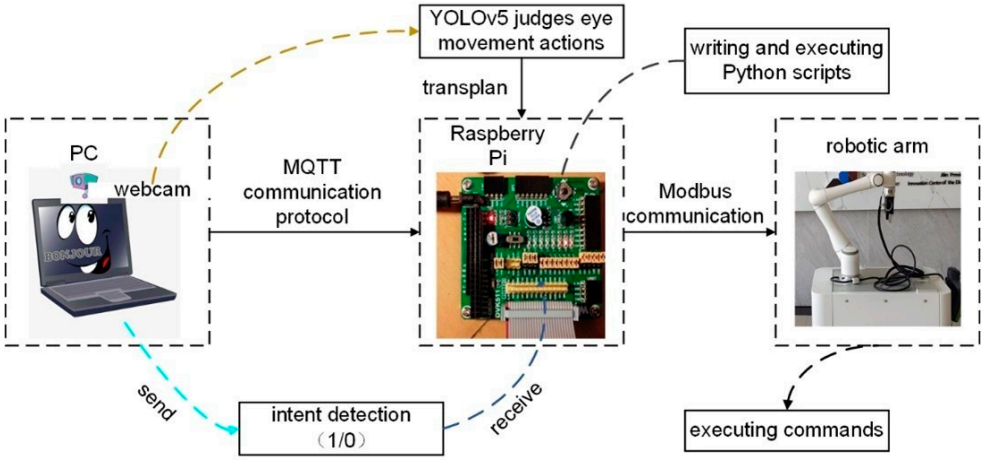


Figure 15. The overall flow chart of the experiment.

In this study, the improved YOLOv5 eye state detection algorithm is successfully transplanted to Raspberry PI 4B to achieve high-precision eye movement state judgment. Through the export.py tool, the YOLO model is converted to ONNX format for easy deployment on the Raspberry PI. Using MQTT protocol, the asynchronous communication between the computer's intention detection program and the Raspberry PI robot arm control program is realized. The Mosquitto server is installed on the Raspberry PI to ensure that the PC and Raspberry PI are in the same network segment. The weight and label are replaced by VNC Viewer to complete the model migration. In addition, the six-axis cooperative robot interacts with the Raspberry PI through the Modbus TCP protocol, and the user writes jbi scripts in the instructor to control the robot arm to perform specific actions. This integrated system provides an effective scheme for the deployment of eye control technology in practical applications.

We conduct a comprehensive evaluation of the computational efficiency of our models, especially for those deployed on resource-constrained devices such as the Raspberry PI. To measure the model's performance more accurately, we introduce three key computational efficiency metrics: processing time, memory usage, and power consumption. Table 12 shows the efficiency indicators of the YOLOv5 model before and after the improvement.

Table 12. Comparison of calculation efficiency indicators.

Model	Processing Time (ms)	Memory Usage (MB)	Power Dissipation (W)
The original YOLOv5	23.6	256	2.1
Improved YOLOv5	18.9	220	1.9

As shown in Figure 16, the improved YOLOv5 model outperforms the original model in terms of processing time, memory usage, and power consumption. This result shows that our improvement not only improves the accuracy of the model but also improves the computational efficiency of the model on the resource. The jbi script information is shown in Table 13.

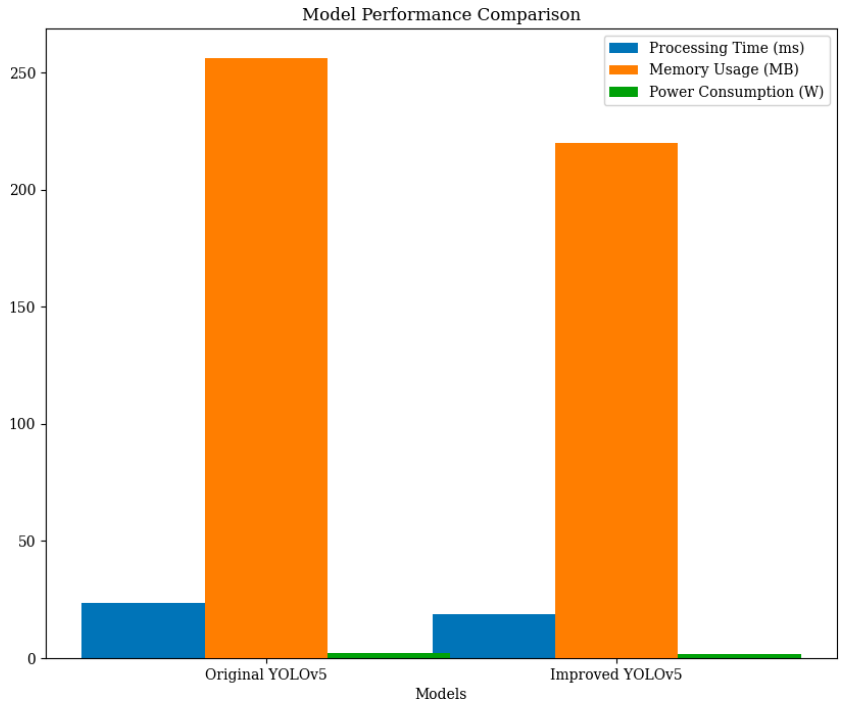


Figure 16. Comparison of calculation efficiency indicators.

Table 13. jbi script information.

Serial Number			
1	NOP	14	ELSEIF M(530 = 1)THEN
2	MOUT M(528)OFF	15	CALLJOB:leftmove
3	MOUT M(529)OFF	16	MOUT M(530)OFF
4	MOUT M(530)OFF	17	ELSEIF M(531 = 1)THEN
5	MOUT M(531)OFF	18	CALLJOB:rightmove
6	MOUT M(532)OFF	19	MOUT M(531)OFF
7	WHILE LB000=0 DO	20	ELSEIF M(532 = 1)THEN
8	IF M(528) = 1 THEN	21	CALLJOB:toolclose
9	CALLJOB:upmove	22	MOUT M(532)OFF
10	MOUT M(528)OFF	23	ENDIF
11	ELSEIF M(529 = 1)THEN	24	COUNTNUE
12	CALLJOB:downmove	25	ENDWHILE
13	MOUT M(529)OFF	26	END

3.4. Eye Movement Interactive Grasping Experiment

In this study, the user's intention is judged by combining the facial expression and gaze intention on the PC, and the eye movement state is captured by the camera to guide the robot to grab a specific target. In order to verify the effectiveness of the method, a human–computer interaction experiment is carried out. In the experiment, eight participants control the robot to grasp four different shapes of objects through eye movement, and each person repeats the operation four times. The experimental process is shown in Figure 17 to ensure the accuracy and reliability of the interactive system.

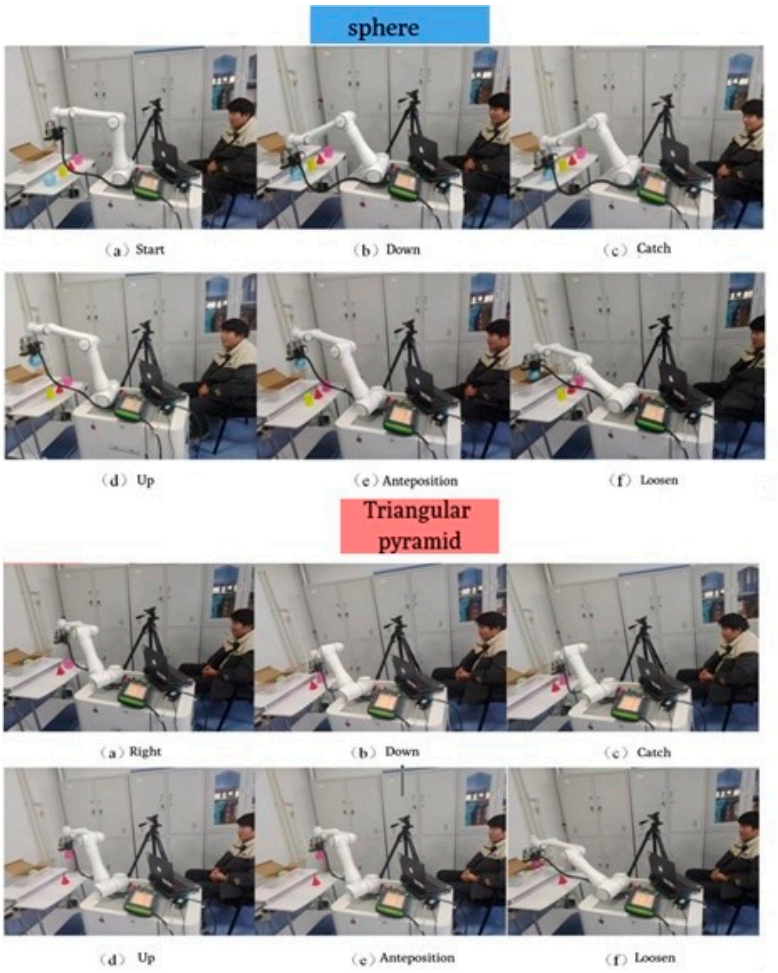


Figure 17. Complete human–computer interaction process.

3.5. Analysis of Eye–Machine Interaction Experiment Results

As shown in Figure 18, the performance verification results of the eye movement control technology in practical applications are critical for evaluating the model's performance in the real world. The chart includes accuracy, response time, and other key performance metrics for eight different testers using eye movements to control the arm, giving us a visual representation of the model's adaptability and reliability for different users and different tasks.

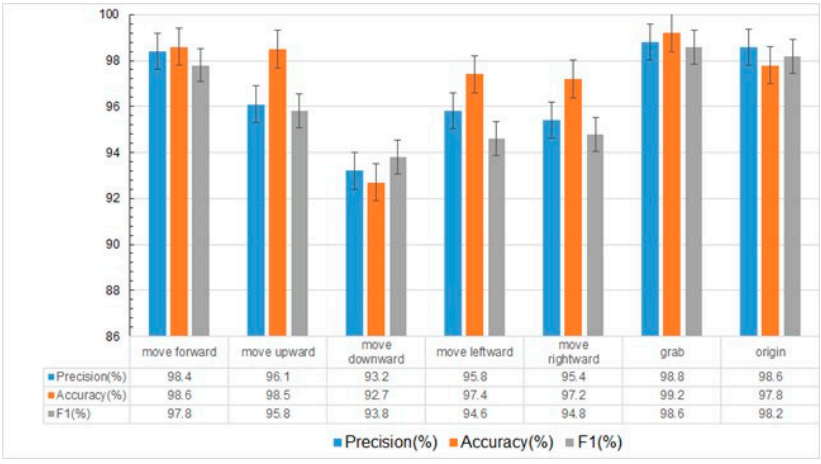


Figure 18. Test results.

Figure 19 offers a detailed analysis of the improved YOLOv5 model's performance during the eye movement-controlled grasping task. The experiment involved four distinct shapes—square, cylinder, pyramid, and sphere—with each shape being picked up 25 times. Tasks were labeled from 1 to 4 to represent different trials. According to the analysis in Figure 18, as the number of tests increased, the grasping accuracy of this method showed consistent improvement, stabilizing at around 95% after the eighth test. These results not only validate the effectiveness of the proposed method but also highlight its potential for practical applications in object recognition and grasping, making it a promising avenue for further research and development.

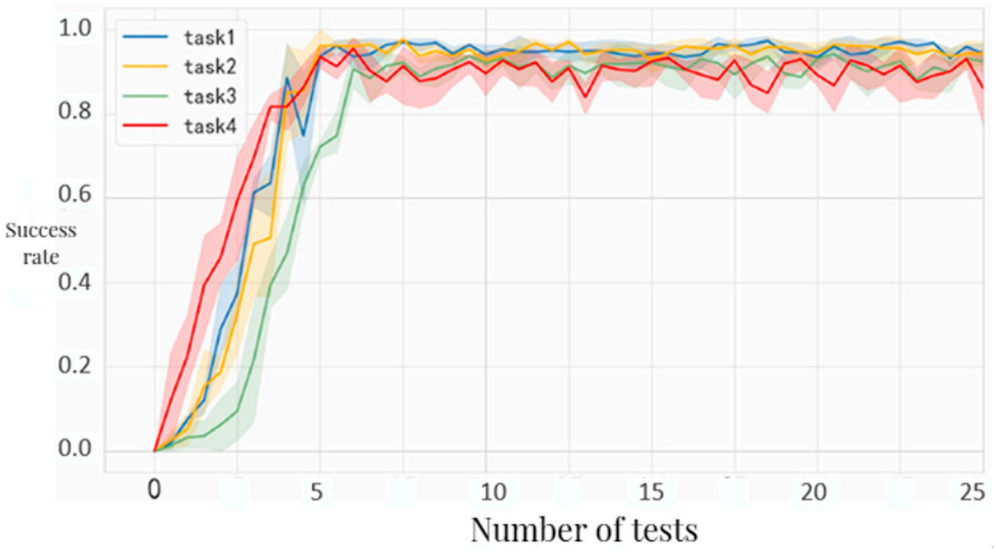


Figure 19. Test results for different tasks.

4. Summary of This Article

In this study, we successfully improved the YOLOv5 model to improve the accuracy of interaction intention judgment and implemented this model on the Raspberry PI platform, thus verifying the high performance of the eye movement control system based on this model. During the data acquisition phase, we conducted detailed experimental tests to ensure that the dataset we constructed could truly reflect the performance of eye movement control. By optimizing the network structure and adjusting the hyperparameters, we further improved the accuracy of eye movement recognition, making the system show excellent performance in adaptive movement and accurate grasping tasks and verifying its ability in high precision, fast response, and stable operation.

By combining facial expression and eye movement information, our multimodal fusion strategy not only enhanced the system's ability to recognize eye movement states but also significantly improves the overall accuracy and reliability of the system. Compared to existing studies, our approach was more accurate in detecting intentional eye movements in natural visual behaviors, promoting the naturalness and intuitiveness of interactions. In addition, our technology showed a wide range of application potentials in multiple industries such as gaming, healthcare and autonomous systems, which shows that our research is not only innovative in theory but also has broad application prospects in practical applications.

Nevertheless, we are aware of the limitations of this study. Our system was tested in a controlled laboratory environment, which may limit its ability to generalize in the real world. In future studies, we need to test the system in more diverse populations and more complex environments, such as different lighting conditions, facial features, and user states, to provide a more complete picture of the robustness and universality of the method. In addition, the research should explore how to further optimize the fusion algorithm to

adapt to the hardware system with higher running speed, so as to improve the real-time practicability of the system.

Overall, this study not only opens up a new research direction for the development of eye control technology but also greatly promotes the efficiency and convenience of human–computer interaction technology, and it provides strong technical support and a strong theoretical basis for future eye control applications. As the technology continues to evolve and improve, we expect this research to play an important role in the future application of eye control technology, significantly improving the user experience, enhancing operational efficiency, and driving technological progress in multiple areas.

Recent advancements in human–computer interaction highlight the importance of multimodal control systems in various fields, including healthcare and assistive technology. For instance, systems that integrate eye tracking with voice recognition have demonstrated improved user experience by more accurately understanding user intent. Additionally, the development of assistive technologies that incorporate both eye movements and facial expressions has shown great promise in enhancing user interactions. These trends underscore the relevance of our findings within the broader landscape of human–computer interaction.

Combining the emphasis on data privacy and user autonomy in this study with the ethical analysis measures that we adopted, we ensured that the principles of data minimization and encryption were strictly followed during the collection, processing, and storage of data. All participants signed informed consent forms before data collection and were informed of their rights, particularly their ability to withdraw consent and delete their data at any time. To protect participants' privacy to the greatest extent, we only collected the minimum data necessary to achieve the research objectives and immediately deleted redundant data after the experiment. Additionally, we employed the AES-256 encryption algorithm to encrypt all facial expression and eye movement data, ensuring the security of the data during transmission and storage. These privacy protection measures, along with the safeguarding of user autonomy, ensured the compliance of our research and laid a solid foundation for the ethical and societal standards of the technology. Through these measures, we not only enhanced the accuracy and naturalness of human–computer interaction technology but also ensured that the technology adheres to ethical standards in its application.

Author Contributions: Conceptualization, X.S. and Z.C.; methodology, X.S.; software, Z.C.; validation, X.S. and Z.C.; formal analysis, X.S.; investigation, Z.C.; resources, X.S.; data curation, X.S.; writing—original draft preparation, Z.C.; writing—review and editing, X.S. and Z.C.; visualization, Z.C.; supervision, X.S.; project administration, X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [Jilin Provincial Department of Science and Technology Project: Research on rehabilitation assistance system and key technologies for movement disorders based on visual tracking] grant number [20220201100GX] And The APC was funded by [Jilin Provincial Department of Science and Technology Project: Research on rehabilitation assistance system and key technologies for movement disorders based on visual tracking].

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of Changchun University (20220201100GX) on [20 August 2024].

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy and ethical restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tatsumi, E.; Yasumura, M.; Rashid, M. Deep Learning-Based Eye Movement Analysis for Predicting Landing Performance in Virtual Reality. *Appl. Ergon.* **2019**, *76*, 167–175.
2. Duchowski, A.T. *Eye Tracking Methodology: Theory and Practice*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2007.
3. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
4. Plopski, A.; Hirzle, T.; Norouzi, N. The Eye in Extended Reality: A Survey on Gaze Interaction and Eye Tracking in Head-Worn Extended Reality Interface. *J. Virtual Real.* **2022**, *55*, 1–39. [CrossRef]
5. Schindler, K.; Van Gool, L.; de Gelder, B. Recognizing Emotions Expressed by Body Pose: A Biologically Inspired Neural Model. *Neural Netw.* **2008**, *21*, 1238–1246. [CrossRef] [PubMed]
6. Berndt, E.K.; Hall, B.H. Hidden Markov Models for Economic Time Series Analysis. *Econometrica* **1963**, *31*, 63–84.
7. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Shanghai, China, 15–17 October 2021; pp. 13713–13722.
8. Liao, L.; Han, C.; He, C. Rice Disease Image Classification Method Based on VGG-19 Convolutional Neural Network and Transfer Learning. *Surv. Mapp.* **2023**, *46*, 153–157+181.
9. Liu, S.; Huang, J.; Wang, Z.; Wang, X.; Qiu, S.; Wen, S. Improved YOLOv5 for Real-Time Aerial Target Detection. *Remote Sens.* **2020**, *12*, 303.
10. Zhou, X.; Wang, D.; Kratz, L.; Lin, Y. Bottom-Up Attention: Fine-Grained Visual Question Answering with Bottom-Up Attention Flow. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
11. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; pp. 91–99.