applied sciences

MDPI

Article

Gesture Detection and Recognition Based on Object Detection in Complex Background

Renxiang Chen and Xia Tian *

Chongqing Engineering Laboratory for Transportation Engineering Application Robot, Chongqing Jiaotong University, Chongqing 400074, China
* Correspondence: xiatian_cqjtu@163.com

Abstract: In practical human–computer interaction, a hand gesture recognition method based on improved YOLOv5 is proposed to address the problem of low recognition accuracy and slow speed with complex backgrounds. By replacing the CSP1_x module in the YOLOv5 backbone network with an efficient layer aggregation network, a richer combination of gradient paths can be obtained to improve the network's learning and expressive capabilities and enhance recognition speed. The CBAM attention mechanism is introduced to filtering gesture features in channel and spatial dimensions, reducing various types of interference in complex background gesture images and enhancing the network's robustness against complex backgrounds. Experimental verification was conducted on two complex background gesture datasets, EgoHands and TinyHGR, with recognition accuracies of mAP0.5:0.95 at 75.6% and 66.8%, respectively, and a recognition speed of 64 FPS for 640 × 640 input images. The results show that the proposed method can recognize gestures quickly and accurately with complex backgrounds, and has higher recognition accuracy and stronger robustness compared to YOLOv5l, YOLOv7, and other comparative algorithms.

Keywords: object detection; hand gesture recognition; human–computer interaction (HCI); YOLOv5; complex background

Citation: Chen, R.; Tian, X. Gesture Detection and Recognition Based on Object Detection in Complex Background. *Appl. Sci.* **2023**, *13*, 4480. https://doi.org/10.3390/app13074480

Academic Editors: Wendong Xiao, Jin Guo and Wei Su

Received: 4 March 2023
Revised: 16 March 2023
Accepted: 28 March 2023
Published: 31 March 2023

1. Introduction

As the most flexible part of human body, the use of hands for communication is highly expressive, and the use of gestures can be seen in various interaction scenarios [1–3]. During interaction, machines need to correctly understand human gestures and react accordingly [4]. This interaction should be natural and metaphorical, allowing for more natural and intuitive communication with various systems [5,6]. Machine vision-based gesture recognition, however, does not require physical contact with the user, nor the use of additional sensors or specialized equipment. Instead, it only needs to be trained with a large amount of image data to obtain a highly accurate recognition model and achieve accurate gesture recognition [7–10]. Early gesture recognition methods relied on manually extracting gesture features (e.g., shape, color, texture, etc.) for gesture recognition [11–14], but such methods relied on expert knowledge and could not solve the problem of complex images with skin-like backgrounds [15]. With the development of computer and deep learning techniques, many scholars have used advanced deep learning models for gesture recognition [16–20]. Compared with early gesture recognition methods, using deep learning methods for gesture recognition has the characteristics self-extraction of features, strong adaptability and robustness, high accuracy, and greater efficiency to handle large-scale data [21–23]. However, when using deep learning methods for gesture recognition, the recognition accuracy will be significantly decreased if there are complex background interferences, such as skin colors, in the gesture image [24]. With the development of object detection algorithms, represented by SSD (single-shot detector) [25] and YOLO (you only look once) series algorithms [26–29], many scholars have transformed the gesture classification problem into a gesture object detection problem. According to whether

there is an explicit candidate region extraction stage during gesture detection, gesture detection algorithms can be divided into one-stage and two-stage algorithms [30]. One-stage gesture detection algorithms directly predict the position and classes of the gesture from the input image without an explicit candidate region extraction stage, which requires less computation and is faster. The two-stage gesture detection algorithms usually consists of a gesture candidate region extraction stage and a gesture classification stage. In the gesture candidate region extraction stage, some techniques (e.g., Selective Search [31] and RPN [32], etc.) are used to generate candidate regions, which are possible regions containing targets. Since explicit candidate region extraction is required, the algorithm has a higher computation cost and slower speed. Despite the improvement in recognition accuracy for gesture images with complex backgrounds, detecting and recognizing gestures using object detection algorithms still cannot provide a good trade-off between recognition speed and accuracy. To address these issues, this paper proposes a gesture recognition method based on improved YOLOv5l for complex backgrounds using the one-stage object detection network YOLOv5 as the benchmark network. An efficient layer aggregation network is also used to enhance the network's feature learning and expression ability while further improving its recognition speed. Moreover, we introduce the CBAM attention mechanism module to enhance the network's robustness to complex backgrounds. The main contributions of this paper are as follows:

- A gesture recognition algorithm based on improved YOLOv5 is proposed for complex background scenarios.
- By using an efficient layer aggregation network module, the benchmark network YOLOv5 is enhanced for complex background gesture feature extraction. It can also effectively reduce model complexity and improve network recognition speed.
- The CBAM attention mechanism module is introduced to filter out irrelevant features and improve network robustness to complex backgrounds.

2. Related Work

Gesture recognition is an important research direction in the field of human–computer interaction, which is widely used in gesture interaction, smart homes, virtual reality, and other fields. Many researchers are devoted to developing gesture recognition systems with higher recognition accuracy and better robustness using machine learning or deep learning algorithms. Chaudhary et al. [33] extracted features such as gesture histograms for training neural networks and used those networks for gesture recognition, but this method is inefficient and cumbersome. Yang et al. [34] extracted gesture feature vectors and matched them with selected similar gesture templates for recognition, but this algorithm depends on the main direction of the gesture. Ma jie et al. [35] proposed a new type of gesture recognition system based on a spatial transformation network and a dense convolutional network that can effectively reduce the impact of gesture deformation on classification accuracy. Xu et al. [36] used hybrid detection and generative adversarial networks (GAN [37]) to reconstruct hand appearance, which can reliably detect multiple hands from a single image. Soe et al. [38] performed gesture recognition based on the Faster R-CNN [39] algorithm, using the ZFNet [40] network as the backbone network to extract features. Although the network's structure is simple, it is difficult to fully extract gesture features. The accuracy on the NUS-II dataset [41] is only 90.08%, and the accuracy of gesture c is only 65.82%. Wang et al. [42] improved the YOLOv3-tiny algorithm for gesture recognition, achieving a recognition speed of 220 fps and mAP of 92.24%. Although the YOLOv3-tiny-T algorithm has a fast recognition speed, its mAP is relatively low. Xing et al. [43] used the ShuffleNetv2 [44] lightweight network model to replace Darknet-53 in the original YOLOv3 network backbone, reducing the model's computational complexity and adding the CBAM [45] (Convolutional Block Attention Module) attention mechanism module to enhance the network's attention to space and channels. The average recognition accuracy is 99.2%, the recognition speed is 44 FPS, and the inference time for a single 416×416 image on a GPU is 15 ms, but this method has low accuracy for complex

background gesture recognition. Lu Di et al. [46] used the YOLOv4-tiny [47] object detection network with the addition of 1 × 1 convolution and SPP (Spatial Pyramid Pooling [48]) modules to achieve a detection speed of 377 FPS, which can detect and recognize gestures quickly and accurately, but the proportion of gestures in the data used by this algorithm to recognize gestures was too large. Bao et al. [49] used a deep convolutional network model to recognize complex background gestures. The network model used a large number of convolution and pooling layers to extract deep features of complex background gesture images while using dropout regularization to reduce overfitting. This method achieves recognition accuracies of 97.1% and 85.3% on simple and complex background gesture images, respectively, but higher accuracy is needed for practical gesture recognition with complex backgrounds. Wang et al. [50] proposed a sewing gesture detection algorithm based on an improved YOLO deep convolutional neural network, which is also a complex background gesture detection and recognition algorithm. The algorithm achieved end-to-end sewing gesture detection by adding a dense connection layer at the deep YOLOv3 low-resolution network to enhance image feature transfer and reuse to improve network performance. The mAP of the algorithm for recognizing four complex sewing gestures as 94.45%, and the FPS was 43, but the memory usage FLOPs of the algorithm were too high, which was not conducive to practical applications. Peng Yuqing et al. [51] proposed an HGDR-Net algorithm based on YOLO detection and CNN classification recognition, which can have good detection and recognition accuracy with complex backgrounds, with an F1 mean value of 98.43% and a processing time of 0.065 s. However, the algorithm needs to use the YOLO algorithm to detect gestures first and then perform recognition classification, which is a more complex recognition process which cannot recognize images including multiple gestures. To address this problem, this paper proposes an end-to-end algorithm for complex background gesture detection and recognition based on improved YOLOv5, using both efficient layer aggregation networks and CBAM attention mechanisms to achieve robust, fast, and accurate detection and recognition of gesture images with complex backgrounds.

3. Methodology

As a one-stage object detection algorithm, YOLOv5 can transform the detection problem into a regression problem. Compared with other algorithms, YOLOv5 has faster detection speed and higher detection accuracy, which can meet the needs of real-time detection and recognition of complex background gestures. According to different usage requirements, the YOLOv5 network is scaled to obtain five different-sized network models. In this article, we balance the real-time performance and accuracy of complex background gesture recognition, with the YOLOv5l network selected as the base network model.

3.1. YOLOv5l

As one of the most commonly used networks in the field of object detection, YOLOv5 outperforms most previous models due to its faster detection speed and higher detection accuracy. The YOLOv5 network was scaled to obtain five different network models of different sizes according to different usage requirements. In this paper, the YOLOv5l network was chosen as the base network model to balance the speed and accuracy of gesture recognition in complex backgrounds.

The basic modules of YOLOv5l are CBS, SPPF, CSP1_x, and CSP2_x. The CBS module consists of a 3 × 3 convolution, a BN layer, and a SiLU activation function. This allows for the selection of a model with both high efficiency and accuracy. This module reduces the possibility of gradient dispersion and also retains more of the original information, to a certain extent, through feature reuse. The SPPF module improves classification accuracy by extracting and fusing high-level features and applying maximum pooling several times during the fusion process to extract as many high-level semantic features as possible.

The object detection process using the YOLOv5l algorithm is the same as the other network models in the YOLO series. The input image size is first adjusted so that all input

image sizes have the same fixed value of A × A(640 × 640 in this paper). The input images are extracted by the Backbone and Head parts of the network to obtain a feature map of size B × B × C, where the first two dimensions (B × B) represent the size of the extracted feature map, and the last dimension C = n × (5 + N), where n represents the number of bounding boxes predicted by each grid to be divided by the YOLO series algorithm, n is 3 in YOLOv5l, N represents the number of categories to be detected, and 5 is equal to 4 position coordinate information plus one confidence information.

3.2. ELAN

The use of more efficient, high-quality network architectures to enhance network representation is a key factor in improving the accuracy of complex background gesture recognition. Well-performing network architectures have richer gradient combinations, while the CSP1_x module in the YOLOv5l backbone network achieves model scaling through stacking of residual blocks. However, stacking of residual layers only increases the longest gradient paths, not the shortest ones. In contrast, the Efficient Layer Aggregation Network (ELAN) [52] makes the shortest gradient path of the entire network longer by using fewer transition layers. ELAN uses a new network design strategy, namely, a network architecture design based on gradient path analysis. By analyzing the shortest and longest gradient paths for each layer in the network, layer aggregation architectures with efficient gradient propagation paths are designed. The use of this design strategy enhances the expressiveness of the training model by improving the learning capability of the network.

The ELAN network structure is shown in Figure 1, where $F_{3\times3}$ and $F_{1\times1}$ denote 3 × 3 convolution and 1 × 1 convolution, respectively. In addition, 2 paths are obtained by using 2 times 1 × 1 convolution on the input feature map, 1 of which is again obtained by 2 times 3 × 3 convolution and 4 times 3 × 3 convolution to obtain 2 paths; finally, the feature maps extracted from these 4 paths are stitched together and scaled by 1 × 1 convolution to obtain the output feature map. The output feature map is obtained by 1 × 1 convolution.

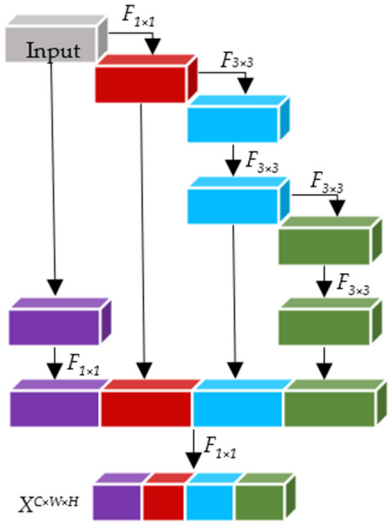


Figure 1. ELAN network structure.

3.3. CBAM Attention Mechanism

To allow the network to focus more on gesture feature information, the CBAM [45] attention mechanism module was added after the backbone network to enhance the network's focus on space and channels, as well as to improve the model's accuracy. CBAM is a simple but effective attention mechanism module for feed-forward convolutional neural networks that sequentially infer attention maps along two separate dimensions, spatial and channel, and then multiply the attention maps with the input feature maps for adaptive feature optimization.

The attention graph is then multiplied by the input feature map to perform adaptive feature optimization. This module is a lightweight general-purpose module a with negligible module overhead, and is seamlessly integrated into the YOLOv5l network architecture in this paper to complete end-to-end training with the backbone network. The CBAM attention mechanism module is shown in Figure 2, where AvgPool denotes the average pooling operation, MaxPool denotes the maximum pooling operation, MLP denotes the multilayer perceptron, and σ is the sigmoid function.

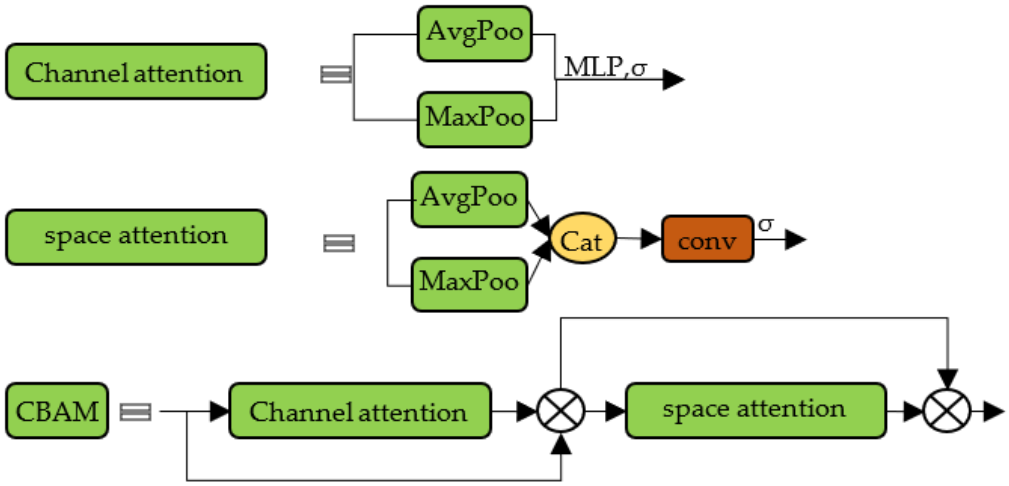


Figure 2. CBAM network structure.

3.4. Improved YOLOv5l

For complex background gesture recognition tasks, due to various interferences brought on by gesture images with complex backgrounds, the original YOLOv5l algorithm has an insufficient ability to extract gesture features, which easily leads to false positives or false negatives. This results in low accuracy when detecting gestures, and the network is too complex to guarantee real-time detection, making it unsuitable for practical applications. To address these issues, this paper proposes the use of an efficient layer aggregation network based on the YOLOv5l network to replace the CSP1_x module in the YOLOv5l network. This is intended to improve the network's ability to extract and express gesture features while reducing network parameters and improving the network's recognition speed. Specifically, the improvement consisted of using the efficient layer aggregation network to replace the original CSP1_x module in the third, fifth, seventh, and ninth layers of the YOLOv5l backbone network, while retaining the original downsampling CBS module. Then, the CBAM attention mechanism is introduced and embedded before the 10th layer of the SPPF module in the backbone network to focus the network on the most critical gesture features for the current task, reduce complex background interference, and further improve the network's gesture recognition accuracy and robustness. The improved algorithm's network structure is shown in Figure 3.

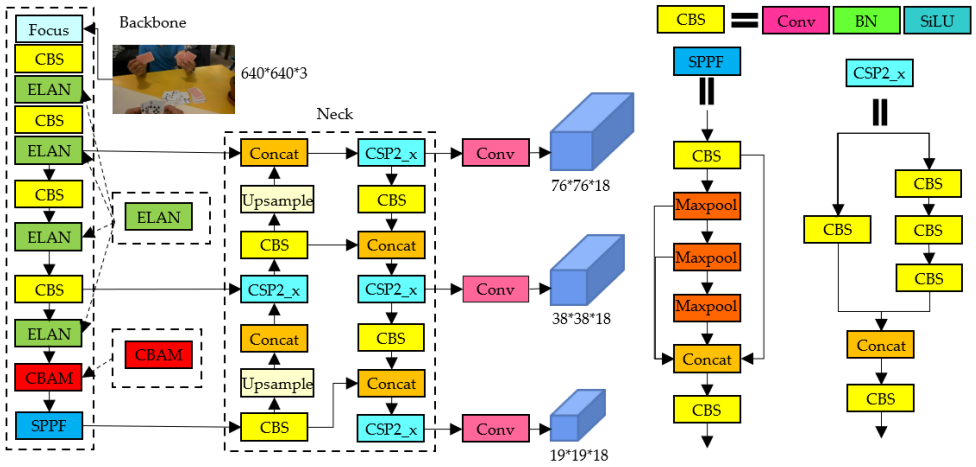


Figure 3. Network structure of the improved YOLOv5l algorithm.

4. Experimental Evaluation

In this section, we first introduce the dataset and evaluation metrics used for the experiments, and then conduct comparison experiments with other gesture recognition algorithms on the EgoHands complex background gesture dataset [53] to demonstrate the effectiveness and superiority of the proposed method. This is followed by ablation experiments to demonstrate the effectiveness of the proposed improvements, then experiments on the TinyHGR dataset [49] to demonstrate the generalization of the proposed method, and, finally, practical testing experiments to demonstrate the robustness of the proposed method.

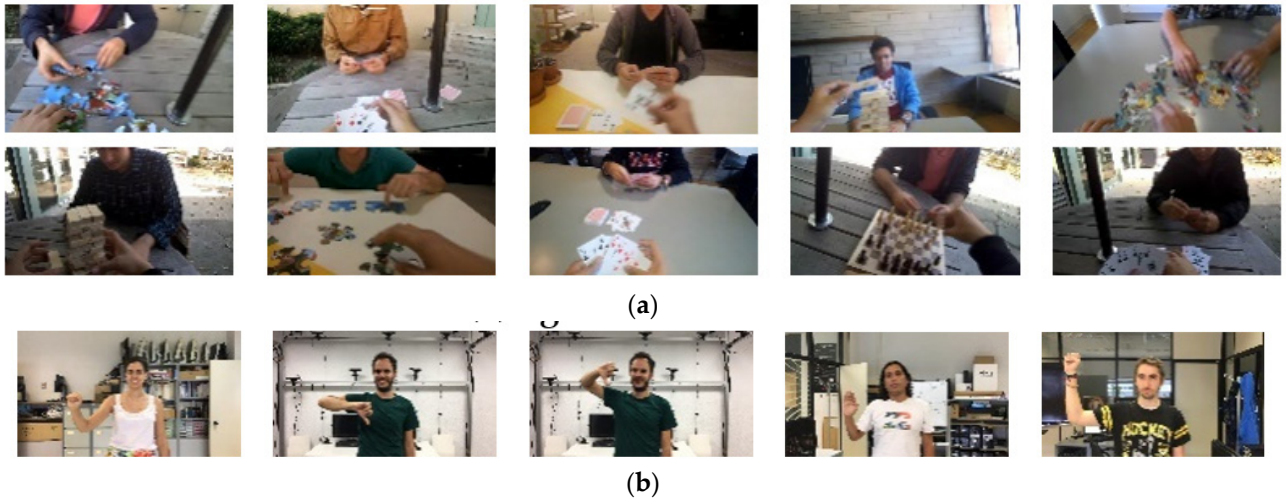
4.1. Experimental Environment and Datasets

All experiments in this paper were conducted in the same experimental environment, Ubuntu 16.04, with Inteli7-6700 CPU 3.40 GHz and NVIDIA GeForce GTX 1080Ti GPU. The Gradient Descent algorithm was used to optimize the loss function. The initial learning rate during training was 0.01, and the learning rate adjustment strategy was a cosine annealing algorithm. Two complex background public gesture datasets, EgoHands and TinyHGR, were used for the experimental data, the training and test sets were divided in a 9:1 ratio, and the network input image size was adjusted to 640 × 640.

The EgoHands dataset consisted of 48 videos, each recording a complex first-person interaction between two test subjects using Google Glass in different environments, such as playing cards, blocks, and chess. The TinyHGR dataset captured data from 40 people making 7 gestures, 20 of which were filmed against simple backgrounds and 20 against complex HCI backgrounds (highly cluttered backgrounds with large illumination variations, see Figure 4). Approximately 1400 frames of each gesture were taken consecutively per person, with sizes of 1920 × 1080, and the gestures appeared at different locations in the images. The largest proportion of each image in this dataset included the face and body, with the gestures only accounting for about 10% of the overall image size. For this paper, 1000 of these complex background gesture images were selected, and the gesture positions were labeled using Labelimg. Figure 4 shows the sample images from the two datasets.

4.2. Evaluation Indicators

The main objective of this paper was to provide a complex background gesture detection and recognition algorithm that could effectively weigh the accuracy and speed of the model. Therefore, this paper used three metrics, namely precision, recall, and mean average precision (mAP), to evaluate the model's accuracy. Frames per second (FPS), model size, and one billion floating point operations (GFLOPs) were used to evaluate the model's speed.



(**a**)

(**b**)

Figure 4. Datasets; (**a**) EgoHands dataset; (**b**) TinyHGR dataset.

For the calculation process of the above three metrics, the prediction frame and the real label frame needed to be compared to calculate their intersection over union (IOU) ratio, so as to determine the matching degree between the prediction frame and the real label frame. According to the matching degree, the prediction results were classified into true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The values of the required indicators were calculated accordingly.

The probability that a positive sample would be detected in all the detected targets was given by precision, which reflected the false detection rate of the model, i.e., False Detection Rate = 1-Precision, and could be obtained by calculating the proportion of positive samples in all the detected target frames, as:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall is a measure of the probability of detecting all positive samples. In this study, it reflected the rate at which the model missed positive samples, which was equal to 1 minus the recall. This measure could be obtained by calculating the ratio of the number of actual positive samples detected to the total number of positive samples. The calculation formula was:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

The mean average precision (mAP) takes into account both precision and recall, and provides a more comprehensive reflection of the model's performance. This measure was obtained by calculating the average precision (AP) for each gesture category at a fixed IOU threshold, and then taking the average of all AP values for all categories. The AP values were obtained by calculating the area under the precision–recall (P–R) curve for each category, where precision was plotted on the x-axis and recall on the y-axis. The formulas for calculating AP and mAP were:

$$AP = \int_0^1 p(r)dr \tag{3}$$

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{4}$$

Here, $p(r)$ represents the precision-recall curve and C represents the number of gesture categories.

To provide a more comprehensive and rigorous evaluation of the proposed algorithm, this paper used mAP@0.5:0.95 as the evaluation metric. This metric represents the average mAP at different IOU thresholds, ranging from 0.5 to 0.95, with a step size of 0.05 (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95). The mAP values used in this paper's experiments were all mAP@0.5:0.95.

4.3. Comparison with Other Gesture Recognition Methods

To verify the effectiveness of the proposed method in detecting gestures, representative lightweight networks ShuffleNetv2-YOLOv3 [43] and YOLOv3-tiny-T [42]; the complex background gesture recognition method used in the literature [50]; and YOLOv7 (the improved version of YOLOv5) [54], as well as the original YOLOv5l algorithm, were selected for comparison experiments with the proposed method. The EgoHands dataset was used for the experimental data, and the experimental results are shown in Table 1.

Table 1. Experimental results of different method on the EgoHands dataset.

Method	Precision/%	Recall/%	mAP/%	FPS	Model Size/MB	GFLOPs
YOLOv3-tiny-T	95.1	90.3	66.0	227	9.4	7.2
ShuffleNetv2-YOLOv3	92.8	78.8	61.6	69	49.1	42.5
YOLOv5l	96.2	92.2	73.5	54	92.8	108.2
YOLOv7	92.6	83.8	59.5	68	74.8	105.1
literature [50]	96.1	90.1	68.5	31	122.6	152.0
proposed algorithm	97.8	94.4	75.6	64	88.3	97.2

For the YOLOv3-tiny-T algorithm, the method used 1×1 convolution after each of the 6 maximum pooling layers in the YOLOv3-tiny network to further halve the output channels for fast gesture detection. As this method reduced the network width, resulting in poorer feature extraction and difficulty in extracting gesture features in complex backgrounds, it achieved only 66% mAP, a 7.5% reduction compared to the YOLOv5l algorithm, despite its good model size, GFLOPs, and FPS.

The ShuffleNetv2-YOLOv3 algorithm used a lightweight ShuffleNetv2 network to replace the backbone in the original YOLOv3 network, and used a CBAM attention mechanism after the backbone network to improve the recognition accuracy and speed. Although the method achieved a recognition speed of 69 FPS, it was difficult to achieve better recognition results using this method due to the presence of a large amount of interference in forms such as occlusion in the complex background gesture images. Its accuracy, completeness, and mAP values were lower than those of the other comparison methods, except the YOLOv7 algorithm.

The YOLOv7 algorithm proposed the use of the E-Elan extended efficient layer aggregation network to extract image features. It also used the structural reparameterization method to merge multi-branch network structures to improve the recognition speed, which had a certain effect, as can be seen from the fact that this method had the highest model size and GFLOPs, but its FPS reached 68. However, the mAP value of the method was only 59.6, which was lower than all of the comparison methods, showing that the use of extended efficient layer aggregation networks does not overcome the challenges present in complex background gesture detection and recognition. This can be demonstrated in subsequent experiments using different backbone networks.

The method used in the literature [50] improved network performance by adding a densely connected layer to the deeper network at the lower resolution of YOLOv3 to enhance image feature transfer and reuse for end-to-end complex background gesture recognition. Although the mAP value of this method reached 68.5%, it was still lower than the 75.6% achieved by the proposed method. Moreover, the memory footprint of this method was as high as 152 GFLOPs, the model size reached 122.6 MB, and the FPS was only 31. This method is not conducive to practical deployment applications.

Although the YOLOv5l algorithm achieved a mAP value of 73.5, the network was too complex, the feature reuse rate was not high, and the recognition speed was only 54 FPS, values which do not represent a good trade-off between recognition accuracy and speed.

The proposed method combines an efficient layer aggregation network with a higher feature reuse rate and better feature extraction capability, redesigns the YOLOv5l backbone network, and uses the CBAM attention mechanism to cause the network to focus on gesture features. Compared with the original YOLOv5l algorithm, the detection accuracy is improved by 1.6%, the detection completion rate is improved by 2.2%, the mAP is improved by 2.1%, the FPS is improved from 54 to 64, and the model size and GFLOPs are reduced, thus achieving a good trade-off between recognition accuracy and speed.

4.4. Ablation Experiments

In order to compare the effects of different backbone networks on recognition accuracy and speed, and to demonstrate the advantages of the backbone network used in this paper, experiments were conducted using different backbone networks based on the YOLOv5l algorithm, and the experimental results are shown in the first five columns of Table 2. It can be seen that, although using the ShuffleNetv2 network as the backbone network greatly reduced the model size and GFLOPs, its FPS did not increase, but rather decreased to 31, and its three indexes, namely the accuracy rate, the completion rate and the mAP, all decreased to different degrees. The accuracy metrics, such as mAP, when using the VoVNetv2 network [55] as a backbone network were not much different from the original YOLOv5l algorithm. However, the FPS using this method was the lowest, at 27, which cannot meet the real-time requirements, and the model size was also much larger. When using E-ELAN [54] as the backbone network, although speed metrics such as FPS were significantly improved, the accuracy metrics were reduced to varying degrees compared to the original YOLOv5l algorithm. Experiments showed that using ELAN as the backbone network not only improved the accuracy of the original YOLOv5 algorithm, but also significantly increased the speed of the model, reduced memory usage, and met the demand for real-time gesture recognition.

Table 2. Experimental results of different modules on the EgoHands dataset.

Method	Precision/%	Recall/%	mAP/%	FPS	Model Size/MB	GFLOPs
YOLOv5l	96.2	92.2	73.5	54	92.8	108.2
YOLOv5l + ShuffleNetv2	92.2	89.4	69.3	31	47	42.1
YOLOv5l + E-ELAN	97.0	90.5	73.5	70	73.5	76.8
YOLOv5l + VoVNetv2	96.8	92.3	73.6	27	103.2	107.7
YOLOv5l + ELAN	97.3	93.5	74.8	71	84.1	95.1
YOLOv5l + ELAN + sSE	97.2	89.8	72.7	58	88.5	96.9
YOLOv5l + ELAN + CBAM	97.8	94.4	75.6	64	88.3	97.2

To demonstrate the effectiveness of the utilized attention mechanism, the proposed method was used. The YOLOv5l+ ELAN + sSE was employed, with the sSE attention mechanism module [55] as a comparison, and the experimental results are shown in the last 2 columns of Table 2. The sSE attention mechanism module reduced the FC layer in the original SE attention mechanism module from 2 to 1 to avoid the loss of channel information due to a reduction in dimensionality. Table 3 shows that adding the sSE attention mechanism module did not achieve the desired effect, but instead reduced the value of each metric. On the other hand, adding the CBAM attention mechanism module improved the mAP by 0.8%, and also improved both the check-all rate and the check-accuracy rate, achieving a better trade-off between accuracy and speed despite the small difference in model size, FPS, etc. The experiments demonstrated that the addition of the CBAM attention mechanism module was able to improve the model's detection and recognition performance.

Table 3. Experimental results of different methods with the TinyHGR dataset.

Method	Precision/%	Recall/%	mAP/%
YOLOv3-tiny-T	94.1	97.6	65.1
YOLOv5l	97.4	98.3	65.3
YOLOv7	96.2	98.6	64.5
Proposed method	97.7	99.3	66.8

4.5. Validation on Other Dataset

In order to fully verify the generalization and robustness of the proposed method, 1000 complex background gesture images were selected for manual annotation in the public dataset TinyHGR and used as experimental data. The proposed method was compared with the original YOLOv5l algorithm, the YOLOv7 algorithm, and the YOLOv3-tiny-T algorithm for experimental experiments, and the results are shown in Tables 3 and 4.

Table 4. Each gesture recognition result with the TinyHGR dataset.

Hand Classes	Precision/%	Recall/%	mAP/%
fist	97.6	99.3	66.8
l	100	100	58.3
OK	96.5	95.2	70.9
palm	96.5	100	69.7
pointer	96.7	100	56.2
thumb down	99.1	100	69.9
thumb up	97.4	100	70.0

Table 3 shows a comparison of the recognition accuracies of the compared methods. It can be seen that all of the methods achieved good results on this dataset, which is due to the relatively low difficulty of recognition in this dataset, which included dark backgrounds, motion blur, and incomplete gestures. However, the proposed method still achieved the best results for each metric, which is good evidence that the proposed method has better recognition performance and robustness for all types of complex backgrounds.

Table 4 shows a comparison of the proposed method's gesture recognition results. It can be seen that the proposed method showed good recognition accuracy for all types of gestures, with only OK and 1 gestures having a full rate of less than 99%, although they reached 95.2%. For all gestures, an accuracy rate of more than 96.5% was reached. In these two gesture categories, despite their reduced AP values, 58.3% and 56.2% were reached, respectively.

4.6. Comparison of Actual Test Results

Figure 5 shows a visualization of the different algorithms' detection processes using the Egohands dataset. Comparing column 1, where the backgrounds are darker and there are incomplete gestures present, all gestures were correctly identified by the comparison algorithms. However, the YOLOv7 algorithm's recognition accuracy dropped to 80% for the incomplete gesture in the bottom right corner, and the identification of the rest of the gestures was also less accurate than with other comparison methods. In column 2, where the pictures have skin-like backgrounds and incomplete gestures, the YOLOv7 algorithm missed a detection. In column 3, we can see that the left gestures are blurred by motion. Despite this, all of the comparison methods accurately recognized the blurred gesture, except for the YOLOv7 algorithm, which had a slightly lower recognition accuracy. This extreme case was not recognized by any of the other comparison methods.

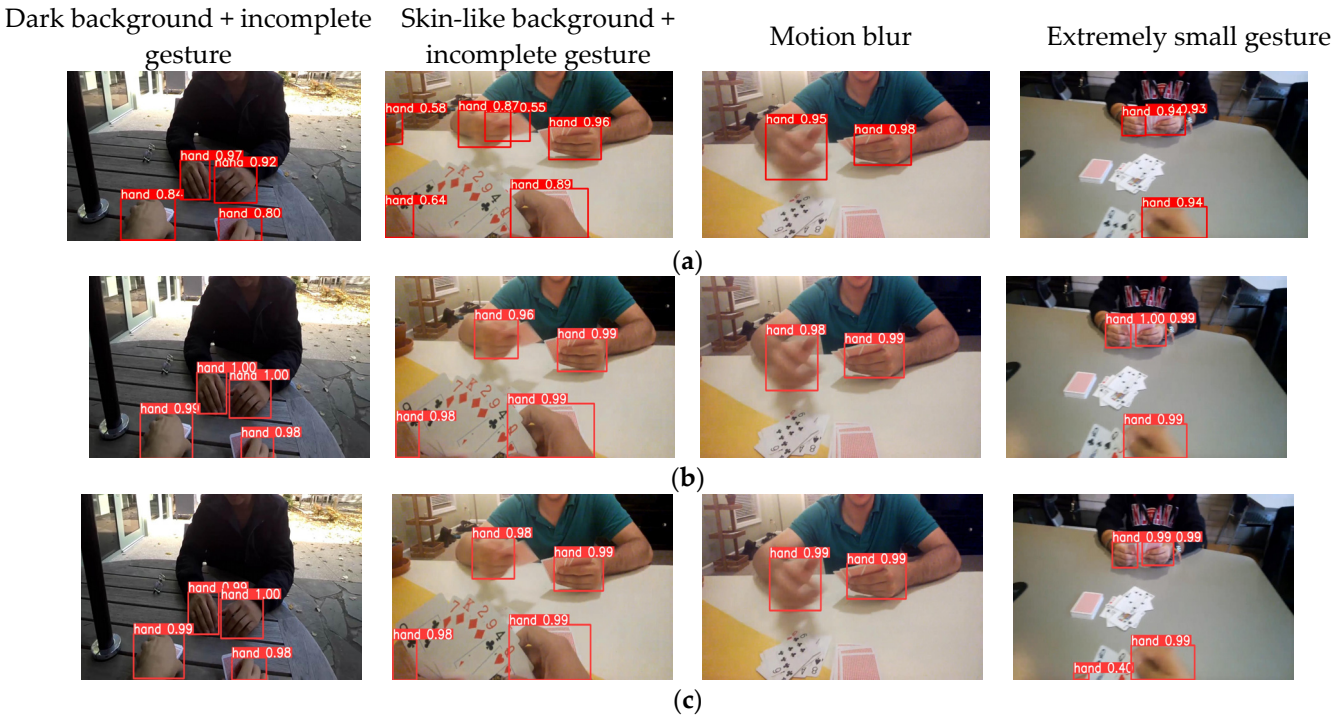


Figure 5. Comparison of detection results of different methods; (**a**) Results of YOLOv7; (**b**) Results of YOLOv5l; (**c**) Results of proposed method.

5. Conclusions

In this study, we proposed a gesture detection and recognition algorithm based on an improved YOLOv5 method, which enables the fast, accurate, and robust detection and recognition of gestures with complex backgrounds. The algorithm utilizes both the efficient layer aggregation network module and the CBAM attention mechanism module, which reduces the model's parameters while enhancing the feature extraction ability of the baseline network and the robustness in detecting complex backgrounds. Compared with existing gesture detection and recognition methods, our method achieved better performance, and its effectiveness and superiority were verified through extensive experiments. In the future, we will continue to research how to achieve high recognition accuracy with fewer samples, and we will attempt to enhance the network's adaptability by removing anchor boxes without changing the recognition accuracy. We will also explore the application of this model to detection and recognition in other complex scenarios.

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

Funding: This work was supported by the Postgraduate Research Innovation Project of Chongqing Jiaotong University under Grant No. YYK202209.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Guo, L.; Lu, Z.; Yao, L. Human-machine interaction sensing technology based on hand gesture recognition: A review. *IEEE Trans. Hum.-Mach. Syst.* **2021**, *51*, 300–309. [CrossRef]
2. Ahmed, S.; Kallu, K.D.; Ahmed, S.; Cho, S.H. Hand Gestures Recognition Using Radar Sensors for Human-Computer-Interaction: A Review. *Remote Sens.* **2021**, *13*, 527. [CrossRef]
3. Serrano, R.; Morillo, P.; Casas, S.; Cruz-Neira, C. An empirical evaluation of two natural hand interaction systems in augmented reality. *Multimedia Tools Appl.* **2022**, *81*, 31657–31683. [CrossRef]
4. Tsai, T.H.; Huang, C.C.; Zhang, K.L. Design of hand gesture recognition system for human-computer interaction. *Multimed. Tools Appl.* **2020**, *79*, 5989–6007. [CrossRef]
5. Gao, Q.; Chen, Y.; Ju, Z.; Liang, Y. Dynamic Hand Gesture Recognition Based on 3D Hand Pose Estimation for Human–Robot Interaction. *IEEE Sens. J.* **2021**, *22*, 17421–17430. [CrossRef]
6. Liao, S.; Li, G.; Wu, H.; Jiang, D.; Liu, Y.; Yun, J.; Liu, Y.; Zhou, D. Occlusion gesture recognition based on improved SSD. *Concurr. Comput. Pract. Exp.* **2021**, *33*, e6063. [CrossRef]
7. Sharma, S.; Singh, S. Vision-based hand gesture recognition using deep learning for the interpretation of sign language. *Expert Syst. Appl.* **2021**, *182*, 115657. [CrossRef]
8. Parvathy, P.; Subramaniam, K.; Prasanna Venkatesan, G.K.D.; Karthikaikumar, P.; Varghese, J.; Jayasankar, T. Development of hand gesture recognition system using machine learning. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *12*, 6793–6800. [CrossRef]
9. Yadav, K.S.; Kirupakaran, A.M.; Laskar, R.H.; Bhuyan, M.K.; Khan, T. Design and development of a vision-based system for detection, tracking and recognition of isolated dynamic bare hand gesticulated characters. *Expert Syst.* **2022**, *39*, e12970. [CrossRef]
10. Chen, G.; Xu, Z.; Li, Z.; Tang, H.; Qu, S.; Ren, K.; Knoll, A. A Novel Illumination-Robust Hand Gesture Recognition System with Event-Based Neuromorphic Vision Sensor. *IEEE Trans. Autom. Sci. Eng.* **2021**, *18*, 508–520. [CrossRef]
11. Li, H.; Wu, L.; Wang, H.; Han, C.; Quan, W.; Zhao, J.P. Hand Gesture Recognition Enhancement Based on Spatial Fuzzy Matching in Leap Motion. *IEEE Trans. Ind. Inform.* **2019**, *16*, 1885–1894. [CrossRef]
12. Zhou, W.; Chen, K. A lightweight hand gesture recognition in complex backgrounds. *Displays* **2022**, *74*, 102226. [CrossRef]
13. Chung, Y.L.; Chung, H.Y.; Tsai, W.F. Hand gesture recognition via image processing techniques and deep CNN. *J. Intell. Fuzzy Syst.* **2020**, *39*, 4405–4418. [CrossRef]
14. Güler, O.; Yücedağ, İ. Hand gesture recognition from 2D images by using convolutional capsule neural networks. *Arab. J. Scie. Eng.* **2022**, *47*, 1211–1225. [CrossRef]
15. Li, J.; Li, C.; Han, J.; Shi, Y.; Bian, G.; Zhou, S. Robust Hand Gesture Recognition Using HOG-9ULBP Features and SVM Model. *Electronics* **2022**, *11*, 988. [CrossRef]
16. Jain, R.; Karsh, R.K.; Barbhuiya, A.A. Literature review of vision-based dynamic gesture recognition using deep learning techniques. *Concurrency and Computation: Pract. Exp.* **2022**, *34*, e7159. [CrossRef]
17. Hu, B.; Wang, J. Deep Learning Based Hand Gesture Recognition and UAV Flight Controls. *Int. J. Autom. Comput.* **2019**, *17*, 17–29. [CrossRef]
18. Dong, Y.; Liu, J.; Yan, W. Dynamic Hand Gesture Recognition Based on Signals from Specialized Data Glove and Deep Learning Algorithms. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–14. [CrossRef]
19. Al-Hammadi, M.; Muhammad, G.; Abdul, W.; Alsulaiman, M.; Bencherif, M.A.; Alrayes, T.S.; Mathkour, H.; Mekhtiche, M.A. Deep Learning-Based Approach for Sign Language Gesture Recognition with Efficient Hand Gesture Representation. *IEEE Access* **2020**, *8*, 192527–192542. [CrossRef]
20. Wang, X.; Zhu, Z. Vision-based hand signal recognition in construction: A feasibility study. *Autom. Constr.* **2021**, *125*, 103625. [CrossRef]
21. Mahmoud, R.; Belgacem, S.; Omri, M.N. Towards wide-scale continuous gesture recognition model for in-depth and grayscale input videos. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 1173–1189. [CrossRef]
22. Mahmoud, R.; Belgacem, S.; Omri, M.N. Deep signature-based isolated and large scale continuous gesture recognition approach. *J. King Saud Univ.—Comput. Inf. Sci.* **2020**, *34*, 1793–1807. [CrossRef]
23. Wan, J.; Lin, C.; Wen, L.; Li, Y.; Miao, Q.; Escalera, S.; Anbarjafari, G.; Guyon, I.; Guo, G.; Li, S.Z. ChaLearn Looking at People: IsoGD and ConGD Large-Scale RGB-D Gesture Recognition. *IEEE Trans. Cybern.* **2020**, *52*, 3422–3433. [CrossRef] [PubMed]
24. Deng, M. Robust human gesture recognition by leveraging multi-scale feature fusion. *Signal Process. Image Commun.* **2019**, *83*, 115768. [CrossRef]
25. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
26. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
27. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
28. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
29. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
30. Zhang, S.; Wen, L.; Lei, Z.; Li, S.Z. RefineDet++: Single-shot refinement neural network for object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 674–687. [CrossRef]

31. Fang, Z.; Cao, Z.; Xiao, Y.; Gong, K.; Yuan, J. MAT: Multianchor Visual Tracking with Selective Search Region. *IEEE Trans. Cybern.* **2020**, *52*, 7136–7150. [CrossRef]
32. Wang, R.; Jiao, L.; Xie, C.; Chen, P.; Du, J.; Li, R. S-RPN: Sampling-balanced region proposal network for small crop pest detection. *Comput. Electron. Agric.* **2021**, *187*, 106290. [CrossRef]
33. Chaudhary, A.; Raheja, J.L. Light invariant real-time robust hand gesture recognition. *Optik* **2018**, *159*, 283–294. [CrossRef]
34. Yang, X.W.; Feng, Z.Q.; Huang, Z.Z.; He, N.N. Gesture recognition by combining gesture principal direction and Hausdorff-like distance. *J. Comput.-Aided Des. Comput. Graph.* **2016**, *28*, 75–81.
35. Ma, J.; Zhang, X.D.; Yang, N.; Tian, Y. Gesture recognition method combining dense convolution and spatial transformation network. *J. Electron. Inf. Technol.* **2018**, *40*, 951–956.
36. Xu, C.; Cai, W.; Li, Y.; Zhou, J.; Wei, L. Accurate Hand Detection from Single-Color Images by Reconstructing Hand Appearances. *Sensors* **2019**, *20*, 192. [CrossRef]
37. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative Adversarial Networks: An Overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
38. Soe, H.M.; Naing, T.M. Real-time hand pose recognition using faster region-based convolutional neural network. In Proceedings of the First International Conference on Big Data Analysis and Deep Learning; Springer: Singapore, 2019; pp. 104–112.
39. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; p. 28.
40. Fu, L.; Feng, Y.; Majeed, Y.; Zhang, X.; Zhang, J.; Karkee, M.; Zhang, Q. Kiwifruit detection in field images using Faster R-CNN with ZFNet. *IFAC-PapersOnLine* **2018**, *51*, 45–50. [CrossRef]
41. Pisharady, P.K.; Vadakkepat, P.; Loh, A.P. Attention based detection and recognition of hand postures against complex backgrounds. *Int. J. Comput. Vis.* **2013**, *101*, 403–419. [CrossRef]
42. Wang, F.H.; Huang, C.; Zhao, B.; Zhang, Q. Gesture recognition based on YOLO algorithm. *Trans. Beijing Inst. Technol.* **2020**, *40*, 873–879.
43. Xin, W.B.; Hao, H.M.; Bu, M.L. Static gesture real-time recognition method based on ShuffleNetv2-YOLOv3 model. *J. Zhejiang Univ. (Eng. Sci.)* **2021**, *55*, 1815–1824+1846.
44. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
45. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
46. Lu, D.; Ma, W.Q. Gesture Recognition Based on Improved YOLOv4-tiny Algorithm. *J. Electron. Inf. Technol.* **2021**, *43*, 3257–3265.
47. Osipov, A.; Shumaev, V.; Ekielski, A.; Gataullin, T.; Suvorov, S.; Mishurov, S.; Gataullin, S. Identification and Classification of Mechanical Damage During Continuous Harvesting of Root Crops Using Computer Vision Methods. *IEEE Access* **2022**, *10*, 28885–28894. [CrossRef]
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
49. Bao, P.; Maqueda, A.I.; del-Blanco, C.R.; García, N. Tiny hand gesture recognition without localization via a deep convolutional network. *IEEE Trans. Consum. Electron.* **2017**, *63*, 251–257. [CrossRef]
50. Wang, W.; He, M.; Wang, X.; Yao, W. Sewing gesture recognition based on improved YOLO deep convolutional neural network. *J. Text. Res.* **2020**, *41*, 142–148.
51. Peng, Y.; Zhao, X.; Tao, H.; Liu, X.; Li, T. Hand Gesture Recognition against Complex Background Based on Deep Learning. *Robot* **2019**, *41*, 534–542.
52. Wang, C.Y.; Liao, H.Y.M.; Yeh, I.H. Designing Network Design Strategies Through Gradient Path Analysis. *arXiv* **2022**, arXiv:2211.04800.
53. Bambach, S.; Lee, S.; Crandall, D.J.; Yu, C. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In Proceedings of the International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1949–1957.
54. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
55. Lee, Y.; Park, J. Centermask: Real-time anchor-free instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13906–13915.