# Research on gesture recognition based on YOLOv8

**Yang Yang** TIGER86YY@SINA.COM
*Leshan Vocational and Technical College, Leshan 614000, China*

**Editors:** Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

## Abstract

Recognizing gestures quickly and accurately has always been a research topic that has attracted much attention. However, existing gesture recognition algorithms still face two challenges. The computational complexity and parameters of gesture recognition deep learning models are often numerous, making them difficult to deploy on resource-limited embedded devices. Secondly, the deep learning model for dynamic gesture recognition is still insufficient in its ability to extract location spatial features. To solve the above problems, this paper proposes a gesture recognition algorithm based on an attention mechanism. First, You Only Look Once (YOLO) v8n lightweight object detection algorithm was selected to reduce parameters and calculations. Furthermore, the Multi-Head Self-Attention (MHSA) model was integrated into the YOLOv8n network to enhance the feature extraction capabilities from the position and spatial dimensions. Experimental results demonstrated that the proposed algorithm achieved 99.2% accuracy, surpassing by 1.1% compared to the original algorithm. Furthermore, it had a 233 FPS detection speed on the Nvidia RTX 3070 GPU.

**Keywords:** Gesture Recognition, YOLOv8, MHSA, Human-computer Interaction.

## 1. Introduction

The development of computer technology has led to the explosion of artificial intelligence applications. People's demand for technology to change their lives is becoming stronger and stronger, especially in smart homes. The innovation of artificial intelligence applications is inseparable from the innovation of human-computer interaction. Gesture recognition has become the focus of many researchers due to its intuitive, flexible, and diverse characteristics (Hui et al., 2024).

In recent years, the emergence of smart devices such as touch screens and iPads has pushed human-computer interaction technology to a new level. They use intuitive operation methods such as touch, slide, and voice control, which makes the information transmission more efficient and brings users more diversified and personalized entertainment methods. Human-computer interaction products are gradually developing towards intelligence and convenience (Sunuwar et al., 2020). Therefore, gestures, as an intuitive body language, play an important role in human-computer interaction.

With the development of artificial intelligence technology, gesture recognition based on deep learning has gradually become the mainstream of gesture recognition technology in human-computer interaction (Sharma et al., 2023). Deep learning-based algorithms can significantly improve recognition accuracy and demonstrate good performance when processing complex gestures.

However, the existing gesture recognition technology still faces some issues. The YOLOv5 algorithm was open-sourced in 2020, balancing detection accuracy and speed. However, with the development of deep learning, its performance can no longer meet the current needs of the industrial field. The MobileNet series algorithms for mobile devices have achieved fast detection speed, but the detection accuracy is unsatisfactory.

Therefore, this paper proposed an improved gesture recognition algorithm based on YOLOv8. The contributions of this paper are as follows: (1) The most lightweight YOLOv8n algorithm was selected to train the gesture recognition model. (2) The MHSA attention model was integrated into the YOLOv8n network to mitigate accuracy drops due to the network's lightweight.

## 2. Related Works

In 2025, Li and Yu (2025) proposed a gesture recognition algorithm for educational robots based on MobileNetv3 by combining the attention mechanism of time and space. The proposed algorithm achieved an average recognition accuracy of 98 31% and 98 48% on the two gesture datasets.

Cheng et al. (2024) proposed a gesture recognition algorithm based on YOLOv5 in 2024. First, the lightweight backbone network MobileNetV3 replaced the backbone of YOLOv5s, and then the Squeeze-Excitation (SE) attention mechanism was integrated into the new backbone. The parameters of the proposed model were reduced by 33%, the calculation amount was reduced by 54%, and the accuracy rate was 97.8%.

In 2025, Li et al. (2025) proposed a two-stage static gesture recognition algorithm based on background optimization. In the first stage, the YOLOv5s model was selected to detect the hand's location. In the second stage, Visual Geometry Group (VGG) 16 was used as the recognition network to recognize gestures. The accuracy of the proposed algorithm reached 97.9%, and the F1 value reached 92.3%.

Wang (2024) proposed an improved static and dynamic gesture recognition algorithm based on YOLOv5 in 2024. First, the feature pyramid structure in YOLOv5 is simplified to improve the accuracy of small target detection. Second, the Ghost module is introduced to reduce parameters and improve detection speed. Finally, the convolutional block attention module is introduced to improve the detection accuracy. The proposed algorithm achieves 94.55% accuracy and 52.9 frames per second speed.

In 2025, Wang et al. (2025) designed a dynamic gesture recognition system in the low-light environment based on the improved YOLOv7 network framework. The method of calculating frame location loss based on the auxiliary frame is integrated with the YOLOv7 network framework to improve the problem of low recognition accuracy and inaccurate positioning of YOLOv7 under low light conditions. The average accuracy of gesture recognition of the proposed algorithm is 99.4%.

Although the above methods based on YOLOv5 and YOLOv7 have achieved high accuracy, they have many model parameters and high computational complexity. YOLOv5s model has more than 7 million parameters, and YOLOv7 tiny also has more than 6 million. Therefore, deploying these algorithms to implement gesture recognition on embedded devices is a challenge. The number of algorithm parameters based on MobileNetv3 has decreased, but the accuracy has declined. The YOLOv8n model has more than 3 million parameters equivalent to the MobileNetv3 model while maintaining high accuracy.

## 3. Methodology

Like the previous YOLO series algorithm, YOLOv8 maintains the backbone, neck, and head structures. The role of the backbone is feature extraction. The attention mechanism improves the performance of the detection network by re-adjusting the weight of network parameters. Therefore, to enhance the feature extraction ability of the YOLOv8n network for gestures, this paper integrated

a lightweight MHSA attention module in the backbone to improve the model's accuracy without adding too many parameters.

### 3.1. MHSA

Fig. 1 shows the MHSA structure.

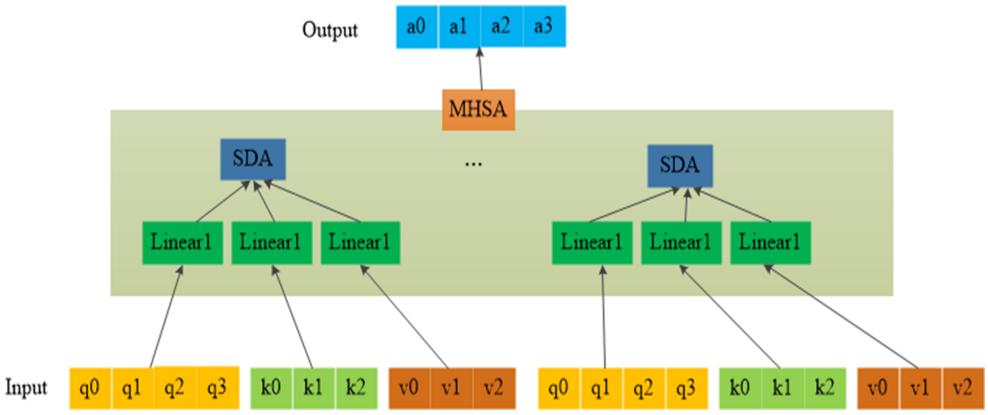



Figure 1: MHSA Structure.

MHSA comprises one or more one or more parallel unit structures, called one-head attention. It includes scaled dot-product attention (SDA) and three weight matrices (three parallel full connection layers). The input of each one-head attention is three sequences: query, key, and value. Multiple one-head attention output is concatenated to get the final output.

MHSA captures different relational patterns of input features through the parallel learning of multiple sets of attention weights. First, the input features are decomposed into three learnable weight matrices through linear transformation. Second, divide multiple heads based on feature dimensions. Third, one-head attention calculation is performed through similarity calculation and weighted aggregation. Finally, concatenate and merge the outputs of multiple heads to output a feature map.

### 3.2. Improving the YOLOv8n using MHSA

Fig. 2 shows the improvement of YOLOv8n using MHSA.

MHSA module can be integrated into any layer of the YOLOv8n network. Furthermore, multiple MHSA modules can be integrated. However, the increase in the number of modules will lead to a substantial increase in parameters. Therefore, this paper integrated an MHSA module into the YOLOv8n backbone, as shown in yellow areas.

## 4. Results

The YOLOv8n and YOLOv8n-MHSA models were trained using the customized gesture dataset. The validation dataset was used to assess the performance of these two models. This paper utilized mAP_0.5 and FPS to evaluate the model's detection accuracy and speed.

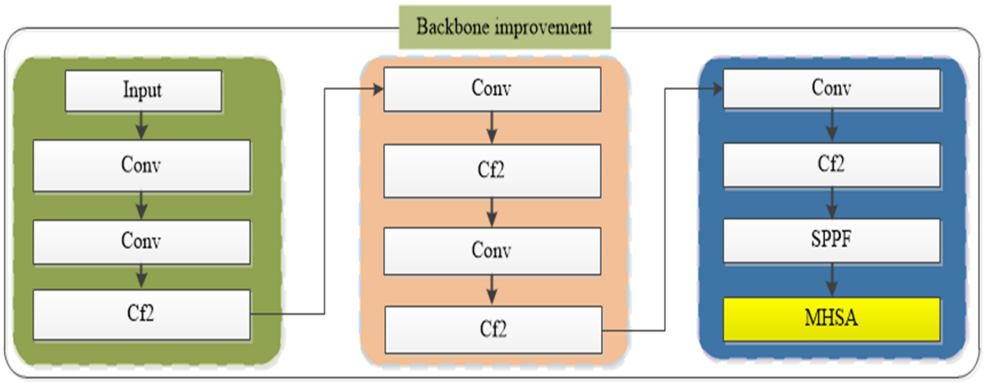



Figure 2: Improving the YOLOv8n using MHSA.

### 4.1. Experimental Environment and Dataset

The customized gesture dataset comprised 1,400 images, 1200 for training and 200 for validation. The experimental configuration is as follows: I5 13600KF CPU, 16G RAM, Nvidia RTX 3070 GPU, CUDA 11.8, torch 2.3.1, torchvision 0.18.1, YOLOv8 8.2.103.

### 4.2. Training Curves

Fig. 3 shows the training curves of the YOLOv8n model. The box_loss, cls_loss, and dfl_loss gradually dropped during the training period. However, the box_loss and dfl_loss changed frequently during the validation phase, while the cls_loss gradually decreased. The precision, recall, mAP50, and mAP50-95 metrics gradually increased during 200 training epochs.

Fig. 4 shows the training curves of the YOLOv8n-MHSA model.

The training curve of the YOLOv8n-MHSA model was similar to that of the YOLOv8n model. However, the box_loss and dfl_loss changed slightly compared to the YOLOv8n model.

### 4.3. Validation Results

Table 1 shows the YOLOv8n model's validation results. The YOLOv8n model achieved 98.1% mAP_0.5 for all classes. Furthermore, the accuracy of all classes exceeded 98%, except for class "I," which had an accuracy of 88.8%.

Table 2 displays the YOLOv8n-MHSA model's validation results.

The YOLOv8n-MHSA model achieved 99.2% mAP_0.5 for all classes. Furthermore, the accuracy of all classes exceeded 99%, except for classes "D" and "I," which had an accuracy of 89.7% and 97.8%, respectively.

### 4.4. Performance Comparison

Table 3 compares the mAP_0.5, parameters, latency and FPS. Latency represents the average inference time per image of the model on the validation dataset. FPS is the reciprocal of latency.

The YOLOv8n-MHSA model increased accuracy by 1.1% compared to the YOLOv8n model, with a 17,376 parameter increase. Furthermore, its detection speed was 233 FPS, which was slightly lower than the 238 FPS of the YOLOv8n model.

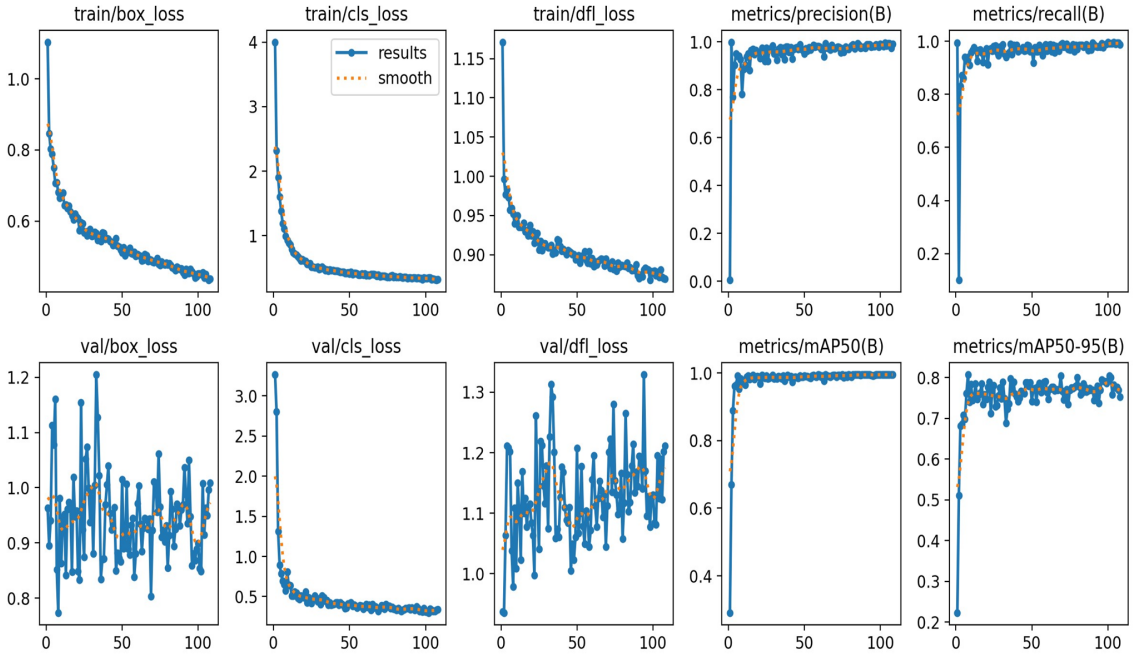



Figure 3: Training Curves of the YOLOv8n model.

Table 1: Validation results of the YOLOv8n model.

| Classs | Precision (%) | Recall (%) | mAP_0.5 (%) |
|---|---|---|---|
| all | 92.5 | 93.9 | 98.1 |
| “A” | 100.0 | 93.1 | 99.0 |
| “7” | 96.1 | 100.0 | 99.5 |
| “D” | 100.0 | 73.4 | 98.7 |
| “I” | 75.3 | 76.2 | 88.8 |
| “L” | 82.0 | 100.0 | 99.5 |
| “V” | 91.2 | 100 | 99.5 |
| “W” | 100.0 | 96.3 | 99.5 |
| ‘Y” | 87.6 | 100.0 | 99.5 |
| “LOVE” | 97.0 | 100.0 | 99.5 |
| “5” | 96.3 | 100.0 | 99.5 |

### 4.5. Detection Results

Fig. 5 shows the detection results of the YOLOv8n-MHSA model. Images containing each gesture were used to evaluate the model’s performance. Each result had a class and confidence value.

The YOLOv8n-MHSA model correctly detected all gestures of sample images. Furthermore, each result achieved a confidence value of more than 90%.

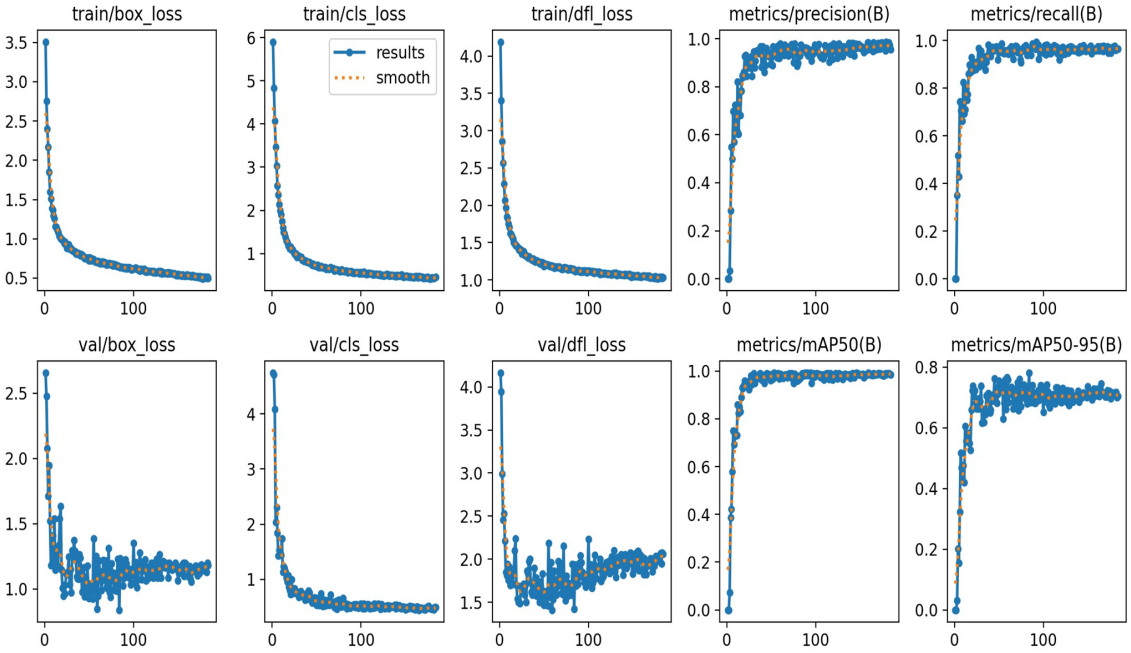


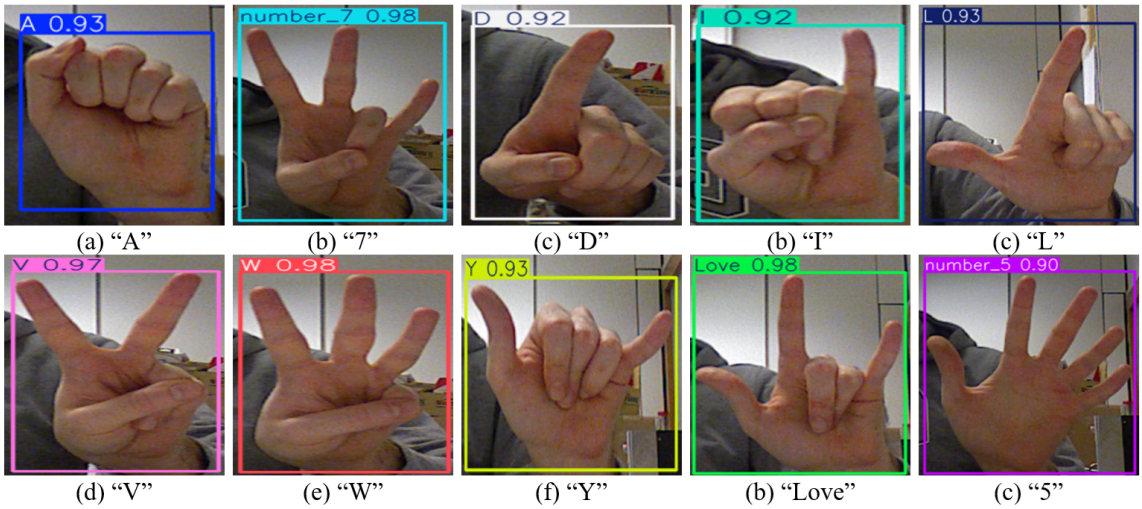

Figure 4: Training Curves of the YOLOv8n-MHSA model.

Table 2: Validation results of the YOLOv8n-MHSA model.

| Classs | Precision (%) | Recall (%) | mAP_0.5 (%) |
|---|---|---|---|
| all | 95.8 | 95.7 | 99.2 |
| “A” | 98.8 | 100.0 | 99.5 |
| “7” | 94.7 | 100.0 | 99.5 |
| “D” | 87.1 | 72.1 | 89.7 |
| “I” | 86.1 | 100.0 | 97.8 |
| “L” | 100.0 | 97.3 | 99.5 |
| “V” | 100.0 | 97.3 | 99.5 |
| “W” | 93.0 | 100.0 | 99.5 |
| “Y” | 100.0 | 99.9 | 99.5 |
| “LOVE” | 100.0 | 95.4 | 99.5 |
| “5” | 98.2 | 100.0 | 99.5 |

Table 3: Performance comparison of two models.

| Model | mAP_0.5 (%) | Parameters | Latency (ms) | FPS |
|---|---|---|---|---|
| YOLOv8n | 98.1 | 3,007,598 | 4.2 | 238 |
| YOLOv8n-MHSA | 99.2 (+1.1) | 3,024,974 (+17,376) | 4.3 | 233 |




(a) "A" (b) "7" (c) "D" (b) "I" (c) "L"

(d) "V" (e) "W" (f) "Y" (b) "Love" (c) "5"

Figure 5: Detection results of the YOLOv8n-MHSA model.

## 5. Conclusions

This paper proposed a gesture recognition algorithm based on YOLOv8n enhanced with a MHSA module, addressing challenges in computational complexity and spatial feature extraction. The YOLOv8n-MHSA model achieved 99.2% accuracy, a 1.1% improvement over YOLOv8n while maintaining lightweight parameters and a high detection speed of 233 FPS. Future work will focus on optimizing for low-power devices and expanding datasets for broader generalization.

## Acknowledgments

The author thanks the School of Artificial Intelligence and Leshan Vocational and Technical College for all their help and support in this research.

## References

Y. Cheng, J. Liang, and Y. Zou. Detection model for static gesture recognition based on yolov5. *Software Guide*, 23(11):181–186, 2024.

Z. Hui, W. Sheng, Z. Zhou, and H. Yuan. Research on gesture recognition technology. *Internet of Things Technologies*, 14(11):36–38+41, 2024.

N. Li and X. Yu. Research on educational robot gesture interaction based on attention improvement of mobilenet-v3. *Automation & Instrumentation*, (1):295–299, 2025.

S. Li, Y. Wang, X. Zhao, and Z. Zhong. Static hand gesture recognition algorithm based on deep learning and background optimization. *Journal of Shanxi University (Natural Science Edition)*, 48(1):180–191, 2025.

V. Sharma, H. Kolivand, S. Asadianfam, D. Al-Jumeily, and M. Jayabalan. Gesture recognition techsniques. In *15th International Conference on Developments in eSystems Engineering*, pages 244–249. IEEE Inc, Baghdad, 2023.

J. Sunuwar, S. Borah, and R. Pradhan. Gesture recognition approaches and its applicability: a study. In *4th International Conference on Electronics, Communication and Aerospace Technology*, pages 1458–1463. IEEE Inc, Coimbatore, 2020.

L. Wang. Research on static and dynamic gesture recognition algorithm based on improved yolov5s. *China CIO News*, (12):138–141, 2024.

Q. Wang, S. Wang, and M. Hu. Dynamic gesture recognition based on improved yolov7. *Heilongjiang Science*, 16(4):94–96+100, 2025.