Journal of Electronic Research and Application, 2026, Volume 10, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Traffic Police Gesture Recognition Based on an Improved YOLOv11-Pose Algorithm

Shijie Jia, Haoxiang Zhang*

Dalian Jiaotong University, Dalian 116028, Liaoning, China

**Author to whom correspondence should be addressed.*

Abstract: To address challenges in feature extraction and real-time processing during traffic police pose estimation, this paper proposes an improved YOLOv11-pose network for traffic police gesture recognition. By replacing the C3K2 module in the backbone network with an enhanced C3K2-Star-CAA module, we achieve efficient extraction of traffic police posture features. A multi-branch star topology enables cross-level feature fusion and multi-scale information propagation, enhancing the model's perception of minute posture details and complex background interference. Embedding the CAA attention mechanism at the key feature layer models critical locations and their spatial contextual relationships through contextual anchors, effectively enhancing key-point feature representation while suppressing complex background interference. Experimental results demonstrate that the improved model achieves 78.6% mAP on the self-built dataset with a detection speed of 186.9 fps, outperforming comparison models in both accuracy and real-time performance. The findings indicate that this approach provides a robust and highly real-time practical solution for traffic police gesture recognition.

Keywords: Pose estimation; YOLOv11-pose; Traffic police gesture recognition; C3K2-Star-CAA; Real-time processing

Online publication: March 31, 2026

1. Introduction

With the rapid advancement of artificial intelligence technology, autonomous vehicles can now capture real-time information such as road signs and traffic signals within traffic scenarios to provide driving recommendations [1]. However, during peak traffic hours or at accident sites requiring urgent intervention, manual traffic police direction remains essential. In 2017, the Beijing Municipal Commission of Transport issued the "Guiding Opinions on Accelerating the Road Testing of Autonomous Vehicles" and the "Implementation Rules for the Management of Road Testing of Autonomous Vehicles." These documents stipulate that autonomous vehicles must be capable of detecting traffic police hand signals [2]. The objective of traffic police posture estimation is to reconstruct skeletal key-point information from captured images, generating a skeletal spatial vector map, a prerequisite for

gesture recognition [3]. Consequently, research on traffic police posture estimation holds significant importance in autonomous driving and driver assistance systems.

Traditional gesture recognition methods primarily rely on image processing techniques, such as skin tone detection to isolate hand regions and track hand movements [4,5]. However, these methods are only effective in simple and controlled environments. To address this limitation, researchers have explored using various sensor types to directly capture motion data [6–8] . While this approach achieves higher accuracy, its costly hardware installation and complex real-time communication requirements hinder practical implementation.

In recent years, deep learning-based gesture recognition methods have made significant progress. Zhou *et al.* proposed a traffic police gesture recognition framework based on Faster R-CNN for image feature extraction combined with fuzzy matching to capture gesture features from complex scene images and achieve accurate recognition [9]. Xiao *et al.* proposed a lightweight traffic police gesture recognition framework for intelligent transportation applications [10]. This method fuses CNN and LSTM to learn the spatio-temporal features of gestures, enhancing model real-time performance and environmental robustness while maintaining high recognition accuracy.

Although the aforementioned methods can handle dynamic changes in gestures to some extent, they typically rely on feature extraction from the entire image, neglecting the spatial structural information of key joints in gestures. This makes them susceptible to background interference, occlusion, and variations in lighting and viewing angles. Consequently, skeleton-based gesture recognition methods have gained prominence in recent years by explicitly modeling key joints and their spatio-temporal relationships, thereby enhancing recognition robustness and accuracy. Xu *et al.* proposed a traffic scene action recognition framework for autonomous driving systems [11]. This method acquires skeletal key-points of road participants via 3D human pose estimation and combines a Dynamic Adaptive Graph Convolutional Network (Dynamic Adaptive GCN) to learn action features across spatial and temporal dimensions, enabling efficient recognition of traffic police and other road user actions. Fu *et al.* proposed a three-dimensional traffic police gesture recognition method based on the Spatio-Temporal Adaptive Graph Convolutional Network (FA-STGCN) [12]. This approach normalizes key skeleton points of traffic police officers, then learns dynamic relationships between skeleton nodes via an adaptive spatial graph convolutional module. An adaptive temporal graph convolutional module captures temporal features of action sequences, enabling recognition of gestures with varying amplitudes and frequencies. Ma *et al.* proposed a multi-view temporal Hypergraph Neural Network (HNN) [13]. By constructing hypergraphs to represent high-order relationships among multiple nodes and views, it enhances global spatio-temporal modeling capabilities for action recognition in complex environments. Guo *et al.* introduced the Motion-Guided Graph Convolutional Transformer (MG-GCT) method [14]. This approach extracts structural features through skeleton keypoint extraction, highlights key frames via a motion-guided mechanism, and combines graph convolutions with Transformer self-attention modules for spatio-temporal feature modeling, achieving high-precision traffic police gesture recognition in complex backgrounds.

Given the aforementioned research, this paper addresses the issues of insufficient accuracy and poor real-time performance in existing methods for traffic police pose recognition in complex traffic scenes. We propose an improved YOLOv11-pose pose estimation method with enhanced robustness and high real-time capability. This paper modifies the original network backbone into a feature extraction network based on C3K2-Star-CAA. It integrates the C3K2 module with StarBlock and embeds the CAA attention mechanism at key layers, enhancing feature expression capabilities for key points and limb connections while maintaining model lightweightness to

reduce redundant parameters and accelerate feature extraction. Traffic police images undergo flipping, rotation, and lighting enhancement to improve pose feature distinguishability. Through multiple traffic police pose estimation experiments, the model's performance in detection accuracy and speed is evaluated, validating the proposed method's accuracy and real-time capability.

2. Methodology

2.1. Overview of methodology

The backbone network of YOLOv11-pose aims to capture multi-scale semantic information of the human body by deeply stacking feature extraction units. Its core component, the C3k2 module, optimizes gradient flow propagation while maintaining a large receptive field by introducing cross-stage local connections [15]. However, when handling minute human key-points in complex backgrounds, the traditional C3k2 module often faces challenges balancing parameter redundancy and computational efficiency, with room for improvement in nonlinear mapping capabilities across feature channels.

To address these limitations, this paper introduces the Star-CAA Block to optimize C3k2, constructing a novel C3k2-Star-CAA architecture. StarBlock employs a unique star operation to achieve high-order interactions among feature channels through element-wise multiplication, simulating global representation capabilities akin to Transformers with minimal computational overhead [16]. Integrating StarBlock into the residual branch of C3k2 not only enhances the backbone network's accuracy in modeling key-point spatial topology but also suppresses background noise through more efficient nonlinear transformations. The introduction of the CAA attention mechanism within StarBlock enables the model to adaptively capture dependencies between key-points and their surrounding context. This enhances the integration of local details and global semantic information, further improving the accuracy of human skeleton feature extraction under occlusion, complex backgrounds, and multi-scale transformations. The model achieves more robust and refined pose estimation while maintaining real-time inference speed. The network architecture of the improved YOLOv11-pose is shown in **Figure 1**.

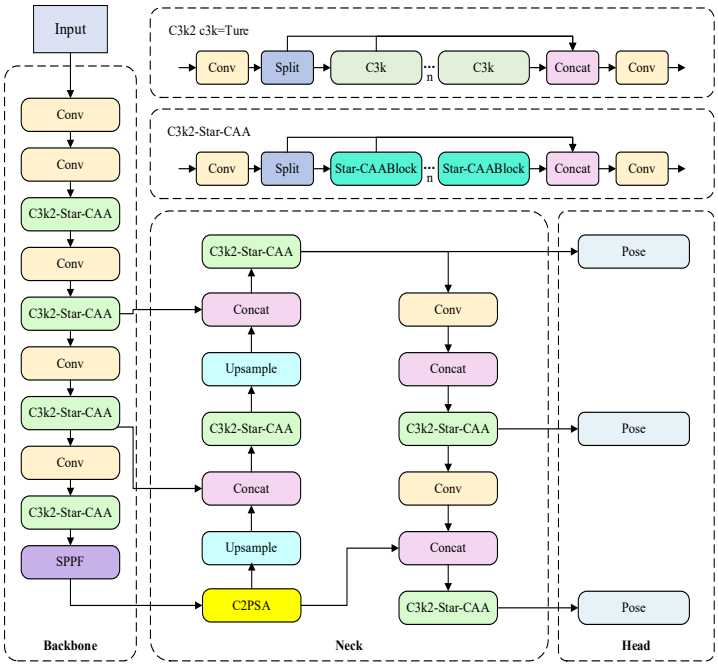


Figure 1. Improved network structure of YOLOv11-pose.

2.2. Star-CAABlock

StarNet is an efficient network based on element-wise multiplication proposed by Microsoft in 2024 [17]. StarNet fuses features from different subspaces through element-wise multiplication, achieving outstanding performance and low latency with its compact network structure and reduced computational load. Its core module, StarBlock, employs star-shaped operations to map inputs into high-dimensional nonlinear feature spaces, equivalent to a polynomial kernel mapping, each layer exponentially increases the complexity of the latent dimension. When stacked, star-shaped operations achieve extremely high-dimensional representation within a compact space. StarNet demonstrates a simple structure with outstanding results. Its powerful expressiveness stems from efficiently leveraging the implicit high-dimensional space, rather than enhancing performance through deepening or widening the network. Consequently, this method significantly reduces computational overhead while improving spatial feature extraction capabilities and increasing computational resource utilization. This paper introduces the CAA attention mechanism to the original StarBlock module, forming the Star-CAABlock structure, as shown in **Figure 2**.

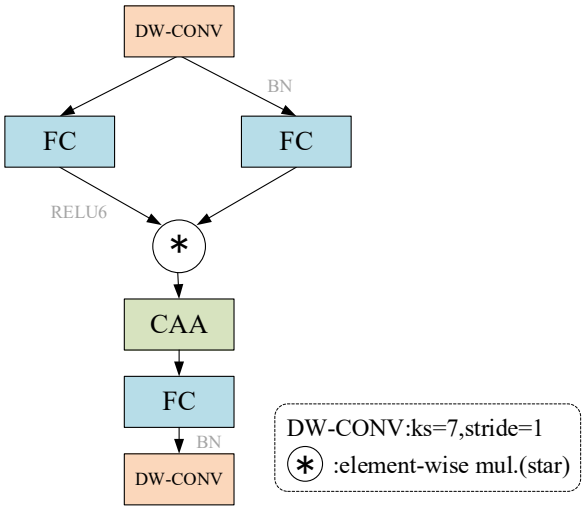


Figure 2. Architecture of Star-CAA Block.

CAA (Context Anchor Attention) utilizes a context anchoring mechanism to fuse global semantic information with local feature details, guiding the model to focus on critical regions within feature maps [18]. Unlike traditional attention modules, CAA partitions feature maps into multiple sub-feature groups along the channel dimension and reshapes them, thereby optimizing feature distribution and enhancing the perception of multi-scale features.

By cross-outputting features between branches, it effectively enriches feature representations and improves the network's ability to integrate multi-scale information. Its structure is shown in **Figure 3**.

CAA excels in achieving a balanced trade-off between computational efficiency and feature modeling capability. By focusing computations on key regions and fusing critical information across dimensions, it effectively enhances the model's ability to represent complex features while avoiding significant computational overhead. The implementation process of the CAA module is as follows:

$$F_{pool} = Conv_{1\times1}(P_{avg}(X)) \quad (1)$$

Here, X represents the input feature map, Pavgdenotes global average pooling, and $Conv_{1\times1}$ indicates a 1×1 convolution operation.

$$F_{strip}^{h} = DWConv_{1\times k_b}(F_{pool}) \quad (2)$$

$$F_{strip}^{v} = DWConv_{k_b\times 1}(F_{strip}^{h}) \quad (3)$$

Equation (2) and **Equation (3)** describe the extraction of long-range features in horizontal and vertical directions using two 1D strip convolutions, where kb denotes the kernel size.

$$A = \sigma(Conv_{1\times 1}(F_{strip}^{v})) \quad (4)$$

Equation (4) represents the fusion of the directional convolution output with the original features to generate attention weights, where σ denotes the Sigmoid activation function.

$$F_{attn} = A \odot X + X \quad (5)$$

Equation (5) represents the enhancement of the original features through a weighted operation, where ⊙ denotes element-wise multiplication.

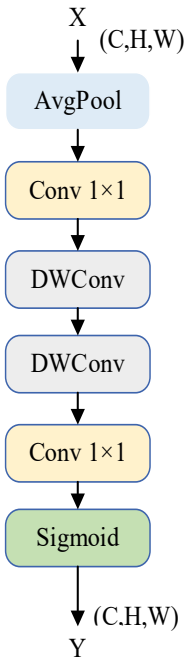


Figure 3. Architecture of CAA.

2.3. SIoU Loss Function

To enhance localization accuracy and training stability for traffic officer gesture detection in autonomous driving, this work employs S_{IoU} as the positional loss function in the bounding box regression branch of object detection, replacing the original C_{IoU} [19]. In traffic officer gesture recognition tasks for autonomous driving, detected targets typically manifest as slender limb structures with significant directional variations (e.g., raised arms, angled arms). Although S_{IoU} introduces constraints on center point distance and aspect ratio consistency beyond the *IoU* overlap metric, its modeling of target orientation remains implicit. This makes it difficult to effectively constrain the offset direction of predicted boxes when pose variations are large or directional shifts are pronounced, often leading to misalignment between localization boxes and key-points. In contrast, S_{IoU} explicitly incorporates angle and shape terms to jointly model both the center offset direction and aspect ratio discrepancy of predicted boxes. This enables

simultaneous refinement of bounding boxes across position, orientation, and scale dimensions relative to ground truth, significantly improving regression accuracy for non-rectangular objects and those with substantial pose variations. The S_{IoU} formula is as follows:

$$S_{IOU} = IoU - \left(\frac{d_{center}}{d_{max}} + \alpha \frac{d_{aspect}}{d_{aspect,max}}\right) \quad (6)$$

Here, *IoU* measures the overlap between the predicted and ground-truth bounding boxes, defined as the ratio of their intersection area to their union area. d_{max} represents the normalized maximum reference distance, typically defined as the diagonal length of the bounding rectangle that simultaneously encloses both the predicted and ground-truth boxes. α is a weighting coefficient used to balance the influence of different geometric constraint terms.

To further enhance the constraint capability of bounding box regression, S_{IoU} introduces the Euclidean distance between the center points of the predicted and ground truth boxes as a supplementary metric on top of *IoU*. Its calculation is as follows:

$$d_{center} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (7)$$

where (x_1, y_1) and (x_2, y_2) denote the center coordinates of the predicted and ground truth bounding boxes, respectively.

To further characterize shape consistency, S_{IoU} measures shape deviation through the aspect ratio difference d_{aspect} between predicted and ground truth boxes, calculated as:

$$d_{aspect} = |r_r - r_2| \quad (8)$$

r_1 and r_2 denote the width-to-height ratios of the two bounding boxes, respectively, while d_{aspect}, max represents the maximum aspect ratio difference.

3. Experiment

3.1. Datasets and experimental environment

Due to the limited availability of publicly accessible traffic police datasets, this study employs a combination of partially open datasets and a custom-built traffic police command gesture dataset for training. The custom dataset incorporates traffic command gestures sourced from literature, supplemented by traffic police command gesture videos extracted at 15 frames per second from online resources [20]. Data augmentation was then applied to the training images, as illustrated in **Figure 4**, through techniques such as rotation, cropping, and adjusting light intensity. This process yielded a training set of 2,400 traffic police samples, with 300 samples each for the test and validation sets. The dataset was partitioned in an 8:1:1 ratio for training, validation, and testing.

Figure 4. Examples of dataset augmentation.

Model training and performance evaluation were conducted on the AutoDL cloud platform. Settings included 500 epochs, a batch size of 32, and an input image resolution of 640×640. **Table 1** details the hardware and system

configurations used during experimentation.

Table 1. Hardware and system configuration

Hardware and system	Version
Processor	EPYC 9754 Processor (128-core)
GPU	Nvidia RTX 4090D
Memory	60GB
Operating system	Linux

3.2. Traffic gesture recognition method

Traffic police hand gesture recognition methods are based on human skeleton key-point detection technology. By extracting key points such as the elbow joint, wrist joint, and shoulder joint, the movement angles of four common traffic police command gestures are calculated. Let the coordinates of the left shoulder, left elbow, and left wrist be A(x_1, y_1), B(x_2, y_2), and C(x_3, y_3), and the coordinates of the right shoulder, right elbow, and right wrist as D(x_4, y_4), E(x_5, y_5), and F(x_6, y_6). The angle θ is defined as the angle between the line connecting the left shoulder, left elbow, and left wrist key points, and the angle α is defined as the angle between the line connecting the right shoulder, right elbow, and right wrist key points. Their expressions are as follows:

$$\theta = arccos\left(\frac{(x_1-x_2)-(x_3-x_2)+(y_1-y_2)(y_3-y_2)}{\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}\sqrt{(x_3-x_2)^2(y_3-y_2)^2}}\right) \tag{9}$$

$$\theta = arccos\left(\frac{(x_4-x_5)-(x_6-x_5)+(y_4-y_5)(y_6-y_5)}{\sqrt{(x_4-x_5)^2+(y_4-y_5)^2}\sqrt{(x_6-x_5)^2(y_6-y_5)^2}}\right) \tag{10}$$

The four common traffic police hand signals are shown in **Figure 5** and as follows:

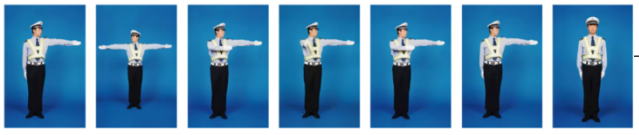
(1) Stop: When the vertical coordinate y_3 of the left wrist is less than y_2 of the left elbow, and the angle θ between the line connecting the left shoulder, left elbow, and left wrist key points satisfies $80° < \theta < 90°$, it is determined to be a stop gesture;

(2) Straight: When the horizontal coordinate x_3 of the left wrist is greater than x_2 of the left elbow, and the angle α formed by the line connecting the right shoulder, right elbow, and right wrist key points satisfies $80° < \alpha < 90°$,it is determined to be a straight gesture;

(3) Left turn: When the horizontal coordinate x_6 of the right wrist is less than x_5 of the right elbow, and the angle θ between the line connecting the left shoulder, left elbow, and left wrist key points and the horizontal plane satisfies $40° < \theta < 50°$, it is determined as a left turn gesture;

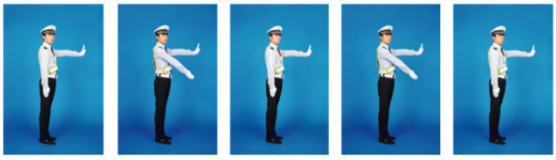
(4) Right turn: When the horizontal coordinate x_3 of the left wrist is greater than x_2 of the left elbow, and the angle α formed by the line connecting the right shoulder, right elbow, and right wrist key points with the horizontal plane satisfies $40° < \alpha < 50°$, it is determined as a right turn gesture.

(a) Stop signal



(b) Go signal

(c) Left turn signal



(d) Right turn signal

Figure 5. Four types of traffic police hand signals.

3.3. Evaluation criteria

To objectively evaluate the performance of the improved algorithm, we employ Precision (P), Recall (R), mean average precision (mAP), and Frames Per Second (FPS) as evaluation metrics for model performance. The formulas are as follows:

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

$$Precision = \frac{TP}{TP+FN} \tag{12}$$

TP denotes samples correctly classified as positive; *FP* denotes samples incorrectly classified as positive; *FN* denotes samples incorrectly classified as negative.

mAP represents the mean average precision in human key-point detection tasks, calculated as the average of *AP* (average precision) values across all categories. It is computed by averaging the *AP* values for each key-point category.

The formula is shown below:

$$AP = \int_0^1 Precision(Recall)dRecall \tag{13}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{14}$$

3.4. Comparative experiments

To validate the effectiveness of the improved model across different datasets, OpenPose, the original YOLOv11n-pose model, and the improved YOLOv11n-pose model were compared on COCO, MPII, and a self-built dataset [21,22]. The specific comparison results are shown in **Table 2**, the improved YOLOv11n-pose model achieves a 1.2 percentage point higher mAP score than the original YOLOv11n-pose model on the COCO dataset, with respective improvements of 1.1 and 1.4 percentage points on the MPII and self-built datasets.

Table 2. Comparison across different datasets

Dataset	Model	P/%	mAP@0.5%	FPS

COCO	OpenPose	85.4	75.4	41
	YOLOv11n-pose	86.7	77.8	162
	Improved YOLOv11n-pose	87.9	78.6	184.1
MPII	OpenPose	87.6	77.4	41
	YOLOv11n-pose	89.7	79.1	162
	Improved YOLOv11n-pose	90.2	80.2	185
Self-built dataset	OpenPose	73.8	69.2	43
	YOLOv11n-pose	83.8	77.2	164
	Improved YOLOv11n-pose	84.3	78.6	186.9

To further validate the detection performance of the proposed method in traffic police hand gesture applications, the algorithm was compared with several other algorithms. The results are shown in **Table 3**.

Table 3. Comparison experimental results

Methods	Parameters/M	FLOPS (G)	mAP50/%	mAP50-90 (%)
OpenPose	73.8	70.1	69.2	43
YOLOv5n-pose	75.5	73.3	75.3	150
YOLOv8n-pose	81.4	75.3	76.1	170.6
YOLOv10n-pose	79.2	76.4	74.5	155.1
YOLOv11n-pose	83.1	78.9	77.2	164
Improved YOLOv11n-pose	84.3	78.5	78.6	186.9

Table 3 experimental results demonstrate that compared to the original YOLOv11n-pose model, the improved model achieves a 1.2 percentage point increase in accuracy, a 1.4 percentage point improvement in map coverage, and a frame rate (FPS) boost from 164 to 186.9. Consequently, the proposed algorithm exhibits high detection accuracy while enhancing processing speed, enabling precise and rapid detection during autonomous vehicle operation.

3.5. Ablation experiment

To validate the effectiveness of the improvements to the network model in detecting traffic police hand signals, corresponding enhancements were sequentially applied to the C3K2-Star backbone network and CAA attention mechanism. The ablation test results are shown in **Table 4**, where "√" indicates the method was employed.

As seen from **Table 4**, after using the C3K2-Star to improve the original C3K2 module, although the mAP@0.5 slightly decreased, the number of parameters was reduced by 1.1×10^5, and the FPS increased by 23.2 frames per second. When the C3K2-Star was further improved to C3K2-Star-CAA and the SIOU loss function was replaced, the ability to extract global features was enhanced, the mAP@0.5 increased by 1.4%, and the FPS increased to 186.9 frames per second. This indicates that the network's ability to detect and recognize the skeletal points of traffic police during command improved after the modification. Compared with the original YOLOv11n-pose model, although the improved YOLOv11n-pose model has a slightly increased number of parameters, it has a faster detection speed and higher accuracy.

Table 4. Ablation experiment results

YOLOv11n-pose	C3K2-Star	C3K2-Star-CAA	SIOU	P/%	R/%	mAP@0.5%	Parameters/10^6	FPS
√				83.1	77.8	77.2	2.8	164
√	√			83.3	77.5	75.2	2.76	187.2
√	√	√		84.1	78.1	77.6	3.32	185.1
√	√	√	√	84.3	78.5	78.6	3.32	186.9

3.6. Visual analysis

To visually demonstrate the improvement of the proposed algorithm, a visual analysis is conducted using both the YOLOv11n-pose algorithm and the improved YOLOv11n-pose algorithm. These algorithms are applied to recognize traffic police hand gestures in different scenarios (**Figure 6**).

(a) YOLOv11n-pose

(b) Improved YOLOv11n-pose

Figure 6. Visual comparison.

The visual results reveal that the YOLOv11n-pose algorithm exhibits inaccuracies in detecting key points at the wrist and elbow for the first three traffic police hand gesture images. The fourth image shows missed detections for the right hip and knee key points. In contrast, the improved YOLOv11n-pose algorithm accurately detects key points in the images and successfully identifies key points even in cases of missed detections. The recognition performance for four common traffic police hand gestures is shown in **Figure 7**, demonstrating accurate identification of different hand gestures.

In summary, the improvements to the algorithm proposed in this paper are effective compared to the YOLOv11n-pose algorithm. Not only does it enhance the detection accuracy of the model, but it also achieves faster processing speeds, demonstrating good recognition performance across various scenarios.

Figure 7. Traffic police hand gesture recognition results.

4. Conclusion

To address the issues of low detection accuracy and poor robustness of traffic police hand gestures during vehicle driving, this paper proposes a traffic police gesture detection method based on YOLOv11-pose. On the basis of YOLOv11-pose, the C3K2 module in the backbone network is replaced with the improved C3K2-Star-CAA module, and the original CIOU loss function is replaced with the SIOU loss function. Through comparative experiments and analysis, the improved model shows good detection performance. Compared with YOLOv11n-pose, it not only increases the fps by 22.9 but also improves the detection accuracy by 3.7% and mAP@0.5 by 1.2%. In the future, we will continue to study the simplification of the network structure and its transplantation and deployment on edge devices.

Disclosure statement

The author declares no conflict of interest.

References

[1] Dong H, 2025, A Review of Vision-Based Multi-Task Perception Research Methods for Autonomous Vehicles. Sensors, 25(8): 2611.

[2] Xie Y, 2018, Analysis of Legal Regulations for Autonomous Vehicles in China. Journal of Beijing University of Technology (Social Sciences Edition), 18(6): 72–77.

[3] Cao Z, Simon T, Wei S, et al., 2017, Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1302–1310.

[4] Zheng M, Crouch M, Eggleston M, 2021, Surface Electromyography as a Natural Human-Machine Interface: A Review. IEEE Sensors Journal, 22(10): 1.

[5] Arsalan M, Kim D S, Owais M, et al., 2020, OR-Skip-Net: Outer Residual Skip Network for Skin Segmentation in Non-Ideal Situations. Expert Systems with Applications, 141: 112922.

[6] Hao Z, Dongzhi Z, Bao Z, et al., 2022, Wearable Pressure Sensor Array with Layer-by-Layer Assembled MXene Nanosheets/Ag Nanoflowers for Motion Monitoring and Human-Machine Interfaces. ACS Applied Materials & Interfaces, 14(43): 48907–48916.

[7] Bai H, Li S, Barreiros J, et al., 2020, Stretchable Distributed Fiber-Optic Sensors. Science, 370(6518): 848–852.

[8] Liu M, Hang C, Wu X, et al., 2022, Investigation of Stretchable Strain Sensor Based on CNT/AgNW Applied in Smart Wearable Devices. Nanotechnology, 33(25): 255501.

[9] Zhou Q, Wang S, Wang Y, et al., 2023, Traffic Police Gesture Recognition Based on Faster R-CNN and Fuzzy Matching Algorithm. Advances in Transportation Studies, 2023(60): 159–170.

[10] Xiao J, Li H, Zhao J, 2026, A Lightweight and Efficient Gesture Recognizer for Traffic Police Commands Using Spatiotemporal Feature Fusion. Scientific Reports, 15(2025): 18256.

[11] Xu F, Xu F, Xie J, et al., 2022, Action Recognition Framework in Traffic Scene for Autonomous Driving System. IEEE Transactions on Intelligent Transportation Systems, 23(11): 22301–22311.

[12] Fu Z, Chen J, Jiang K, et al., 2023, Traffic Police 3D Gesture Recognition Based on Spatial-Temporal Fully Adaptive Graph Convolutional Network. IEEE Transactions on Intelligent Transportation Systems, 2023(24): 9518–9531.

[13] Ma N, Wu Z, Feng Y, et al., 2024, Multi-View Time-Series Hypergraph Neural Network for Action Recognition. IEEE Transactions on Image Processing, 2024(33): 3301–3313.

[14] Guo X, Zhu Q, Wang Y, et al., 2024, MG-GCT: A Motion-Guided Graph Convolutional Transformer for Traffic

Gesture Recognition. IEEE Transactions on Intelligent Transportation Systems, 2024(25): 14031–14039.

[15] Khanam R, Hussain M, 2024, YOLOv11: An Overview of the Key Architectural Enhancements, arXiv, https://doi.org/10.48550/arXiv.2410.17725

[16] Vaswani A, Shazeer N, Parmar N, et al., 2017, Attention Is All You Need, arXiv, https://doi.org/10.48550/arXiv.2410.17725

[17] Rewrite the Stars, 2024, In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5694–5703.

[18] Cai X, Lai Q, Wang Y, et al., 2024, Poly Kernel Inception Network for Remote Sensing Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[19] Gevorgyan Z, 2022, SIoU Loss: More Powerful Learning for Bounding Box Regression, arXiv, https://doi.org/10.48550/arXiv.2205.12740

[20] He J, Liao J, Zhang C, et al., 2020, Visual Gesture Recognition Technology Based on Long Short-Term Memory and Deep Neural Networks. Journal of Graphics and Information Technology, 41(3): 372–381.

[21] Lin T, Maire M, Belongie S, et al., 2014, Microsoft COCO: Common Objects in Context. Springer International Publishing.

[22] Andriluka M, Pishchulin L, Gehler P, et al., 2014, Human Pose Estimation: New Benchmark and State of the Art Analysis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3686–3693.