TPGR-YOLO: Improving the Traffic Police Gesture Recognition Method of YOLOv11

Xuxing Qi[1], Cheng Xu[2]*, Yuxuan Liu[3], Nan Ma[4], Hongzhe Liu[5]
Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing, China[1, 2, 3, 5]
College of Information and Technology, Beijing University of Technology, Beijing, China[4]
Beijing Qiangqiang Yuanqi Technology Co., Ltd, Beijing, China[1, 2, 3]

***Abstract*—In open traffic scenarios, gesture recognition for traffic police faces significant challenges due to the small scale of the traffic police and the complex background. To address this, this paper proposes a gesture recognition network based on an improved YOLOv11. This method enhances feature extraction and multi-scale information retention by integrating RFCAConv and C2DA modules into the backbone network. In the Neck part of the network, an edge-enhanced multi-branch fusion strategy is introduced, incorporating target edge information and multi-scale information during the feature fusion phase. Additionally, the combination of WIoU and SlideLoss loss functions optimizes the positioning of bounding boxes and the allocation of sample weights. Experimental validation was conducted on multiple datasets, and the proposed method achieved varying degrees of improvement in all metrics. Experimental results demonstrate that this method can accurately perform the task of recognizing traffic police gestures and exhibits good generalization capabilities for small targets and complex backgrounds.**

***Keywords*—Traffic police gesture recognition; loss function; YOLO algorithm; multi-scale feature fusion**

I. INTRODUCTION

In the Road Traffic Law, it is stipulated that when traffic signals and signals issued by traffic control officers conflict, the latter shall prevail [1]. The quick and accurate identification of traffic control gestures is crucial for preventing traffic accidents, alleviating congestion, and improving road usage efficiency.

Traditional gesture recognition methods primarily rely on image processing techniques, separating hand regions through skin color detection [2, 3] and tracking hand movements. However, these methods are effective only in simple and controlled environments. To address this issue, researchers have explored the use of various types of sensors [4, 5, 6, 7, 8] to directly capture motion data. Although this approach achieves higher accuracy, its expensive hardware installation costs and complex real-time communication requirements limit practical implementation.

In recent years, multi-modal recognition technologies have gained prominence, with advancements such as those leveraging wireless signals [9, 10, 11] (e.g., Wi-Fi, infrared, or radar) and deep learning-based multi-modal recognition techniques. These methods alleviate problems caused by lighting variations and occlusions to some extent. However, in complex traffic scenarios, wireless signals are susceptible to interference, significantly impacting recognition accuracy.

With the rapid development of deep learning technologies, skeleton-based recognition methods [12, 13, 14, 15] have become a research focus. By extracting precise pose features, these methods have achieved notable results in gesture recognition. However, existing research mainly targets skeletal data at close distances, leaving gaps in recognition performance in open traffic scenarios with long-range, multi-target, and complex environments.

Effectively identifying traffic control gestures from a crowd under various lighting conditions and positions in open traffic environments remains a challenge. This study addresses the issue by optimizing methods for traffic control gesture recognition, eliminating the need for depth information or skeletal data, thereby significantly reducing processing time. The main contributions of this paper are as follows:

1) Based on publicly available traffic videos in complex scenarios, a new dataset was constructed. It contains eight types of traffic control gestures, characterized by small targets and complex backgrounds, providing a solid data foundation for future research.

2) To address challenges like significant scale variations and complex backgrounds in traffic control scenarios, we designed the C3k2RFCA and C2DA modules. These effectively preserve multi-scale information and detail features. Additionally, multi-scale deep convolution and the Sobel operator were integrated into the Neck part, achieving deep fusion of edge and scale information.

3) The combination of WIoU (Wise-IoU) and an improved SlideLoss was employed as the loss function. WIoU improves the precision of bounding box localization, while SlideLoss optimizes sample weight allocation. Their synergy significantly enhances the performance of traffic control gesture detection in open scenarios.

4) Comprehensive experiments on the custom dataset and two public datasets demonstrate the proposed method's high accuracy and strong generalization capability. It effectively handles the challenges of small targets and complex backgrounds.

The structure of this paper is as follows: Section II introduces the related work, Section III describes the improved YOLO method and its application algorithm for traffic police

*Corresponding Author

gesture recognition, Section IV reports the experimental results, and Section V provides the conclusions.

II. Related Work

A. YOLOv11 Algorithm

First, YOLOv11, as a groundbreaking advancement in the field of object detection, achieves significant improvements in detection efficiency and accuracy through a series of innovative modules and optimized designs. The core innovations of this method include the newly introduced C3k2 module and C2PSA module. C3k2 is an improved version of the CSP (Cross Stage Partial) bottleneck structure. By integrating a two-layer convolutional structure, it significantly enhances feature extraction capabilities while maintaining low computational overhead. This optimization makes the method particularly effective in handling complex scenarios, multi-scale targets, and detailed features. Meanwhile, the C2PSA module incorporates a position-sensitive attention mechanism that dynamically focuses on the critical spatial regions where targets are located, enhancing the capture and processing of spatial information. This is especially effective in tasks involving small object detection and precise localization.

Additionally, YOLOv11 introduces lightweight optimizations in the detection head. By simplifying the network structure and reducing unnecessary computational processes, the parameter count has been reduced by approximately 20% compared to the previous generation. This significantly lowers computational costs while maintaining high accuracy. This lightweight design not only enhances runtime speed but also supports the efficient deployment of the method on resource-constrained devices.

B. Gesture Recognition

For traffic police gestures, it has been observed that each gesture can be represented by key hand movements. Therefore, static gesture recognition becomes an effective implementation method. The development of object detection algorithms has provided new perspectives for this field. Zhang et al. [16] improved the YOLOv3 algorithm to achieve gesture recognition; Wang et al. [17] combined YOLO algorithm and Kalman filter to realize high-precision gesture recognition and tracking; Saxena et al. [18] utilized YOLOX and YOLOv5 methods to achieve high-accuracy and high-efficiency gesture recognition, facilitating communication for individuals with hearing and speech impairments; Helal et al. [19] combined CNN and YOLO techniques to recognize sign language alphabets, achieving high accuracy while optimizing training time.

The YOLO algorithm is widely applied due to its accuracy and speed, but it still requires adaptations for different scenarios. For instance, Yang et al. [20] proposed the YOLO-LRHG method, combining YOLO with attention mechanisms to address long-distance human-machine interaction gesture recognition problems. Zhou et al. [21] designed the PEA-YOLO lightweight network, improving recognition performance through adaptive spatial feature pyramids and multi-path feature fusion.

Most of the previous research methods have been limited to improving the accuracy of gesture recognition at close range and have not addressed distant scenes with increased interference. This paper builds upon the YOLOv11 network and introduces enhancements to tackle challenges such as the loss of detailed features and the presence of complex backgrounds in far-distance open traffic scenarios. The goal is to improve the detection of key gestures in complex environments, providing an efficient and accurate gesture recognition solution for intelligent traffic management and autonomous driving.

III. Model

In this study, the RFCAConv module was utilized for sampling operations, and the C3k2 module was restructured and optimized. Additionally, the original C2PSA module was replaced with the C2DA module, and an edge-enhanced multi-branch fusion strategy was introduced in the Neck section. To further improve the performance and efficiency of the method, WIoU [22] and Slide-Loss [23] were newly proposed. The TPGR-YOLO overall structure is shown in Fig. 1.

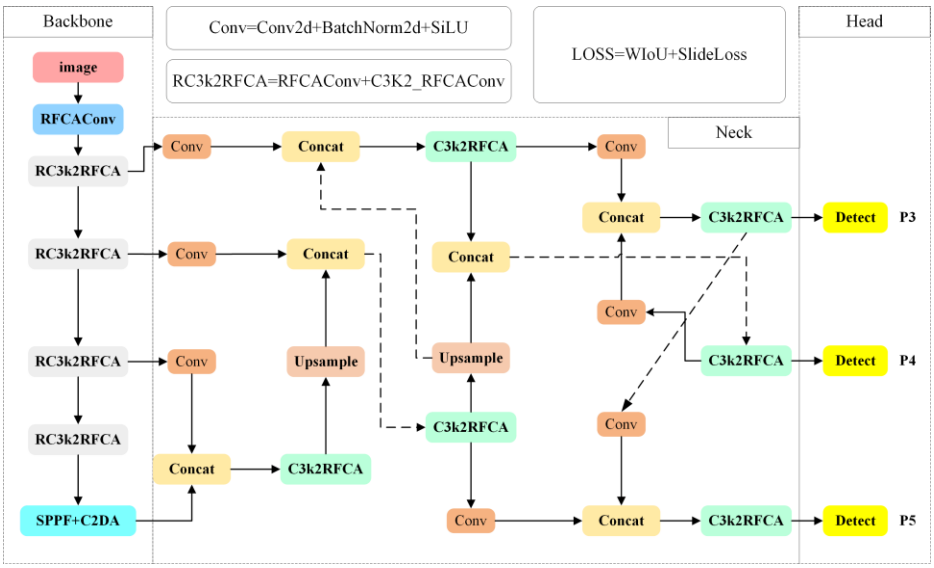

Fig. 1. TPGR-YOLO overall structure.

A. C3K2RFCA

RFCAConv [24] integrates receptive field generation convolution and coordinate attention mechanism (CA), effectively addressing the limitations of traditional convolution, where shared convolution kernels result in insufficient sensitivity to feature locations. It also overcomes the inadequacy of conventional spatial attention mechanisms in fully extracting regional features from contextual information. RFCAConv generates receptive field features, applies weight mapping to them, and then employs standard convolution operations to produce output features. This process combines the advantages of local feature enhancement and global positional attention, resulting in finer-grained and more effective attention weights. As shown in Fig. 2, its structure has been optimized and improved.

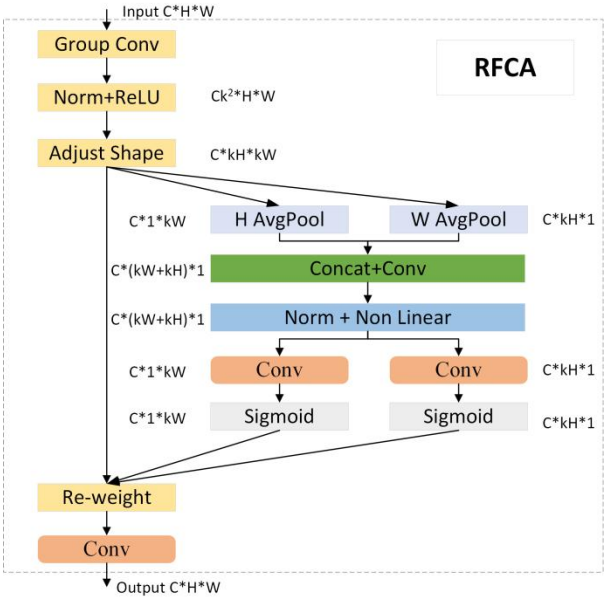

Fig. 2. RFCAConv structure.

For the input feature map $X \in R^{B\times C\times H\times W}$, the RFCAConv module first generates features $G \in R^{B\times(C\cdot k^2)\times H\times W}$ through the receptive field generation module. In the receptive field features G, the feature of each pixel is expanded into a high-dimensional feature map containing the local receptive field. Subsequently, the receptive field features G are unfolded into the spatial dimension.

$$G = \text{Conv}(X, \text{kernel\_size} = k, groups = C) \tag{1}$$

$$\begin{aligned} &G = \text{reshape}(G, (B, C, k^2, H, W)) \\ &\rightarrow G = rearrange(G, (H \cdot k, W \cdot k)) \end{aligned} \tag{2}$$

Next, the receptive field features G undergo global pooling along the horizontal and vertical directions, resulting in features X_h and X_w, which represent the global features in the horizontal and vertical directions, respectively. These two features are concatenated and processed through a 1×1 convolution, batch normalization, and the h-swish activation function to generate intermediate features Y. Y is then split into two parts: the horizontal features $X_h^{'}$ and the vertical features $X_w^{'}$. These parts are further processed through two separate 1×1 convolutions to generate attention weights in the two directions: horizontal weight A_h and vertical weight A_w, each normalized using a Sigmoid activation function.

$$X_h = Pool_h(G), X_w = Pool_w(G) \tag{3}$$

$$Y = \text{Act}(BN(Conv1(X_h \oplus X_w))) \tag{4}$$

$$X_h', X_w' = split(Y) \tag{5}$$

$$A_h = \sigma(Conv(X_h')), A_w = \sigma(Conv(X_w')) \tag{6}$$

Finally, the receptive field features G are pixel-wise weighted by multiplying A_h and A_w, resulting in enhanced feature representations. The weighted features are then passed through a convolutional layer to produce the final output feature map O. The principle is given in Eq. (1) – Eq. (6). This output feature map not only incorporates local receptive field information but also integrates global positional attention in the horizontal and vertical directions.

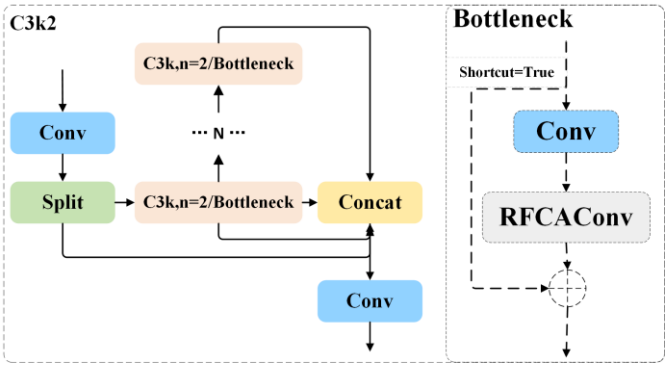

Fig. 3. C3k2RFCA structure.

At the same time, by introducing the global position sensitivity and local receptive field enhancement capabilities of RFCAConv, the C3k2 module has been structurally redesigned. This addresses the problem of reduced feature map resolution caused by multiple downsampling and stride convolutions in the C3k2 module, while compensating for the limited receptive field of the 3×3 convolution kernel, which hampers the ability to capture small objects and fine-grained features. The structure of the redesigned C3k2 module is shown in Fig. 3.

The input features first pass through RFCAConv, undergoing local receptive field enhancement and directional attention weighting to obtain feature representations with a stronger global receptive field. Subsequently, RFCAConv replaces traditional stride convolution for downsampling, further reducing resolution loss while enhancing focus on target regions. Finally, the output feature map retains more detailed information, providing stronger small-object detection capabilities and improved spatial perception.

B. C2DA

The C2PSA module, based on the global feature channel distribution attention mechanism, has limited local feature modeling capabilities, which can dilute spatial detail information and lead to reduced detection accuracy. Drawing inspiration from the concept of Deformable Attention [25], this paper proposes the C2DA module, introducing adaptive local feature learning based on deformable sampling to address the shortcomings of the C2PSA module. Its structure is shown in Fig. 4.

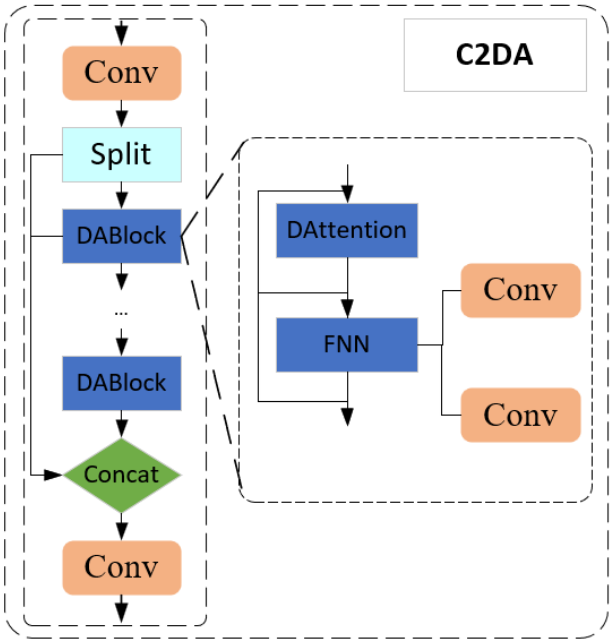

Fig. 4. C2DA structure.

First, the input features undergo dynamic offset sampling. The offset generation network in the C2DA module generates adaptive sampling positions based on the input features, enabling precise capture of critical features and detailed significant regions of small objects while avoiding interference from irrelevant areas. Second, a local sampling range is introduced by restricting dynamic sampling to a local area near the reference points. The offset magnitude is controlled using the tanh function and a scaling factor, ensuring the smoothness and local consistency of the sampling process. This approach enhances the modeling of object boundaries and details. In addition, C2DA retains the global characteristics of the attention mechanism. By leveraging query-key-value matching, it explores global relationships between pixels, allowing the capture of contextual information relevant to small objects and preventing their features from being overshadowed by complex backgrounds.

The generation of dynamic offsets begins by grouping the input feature map $X \in R^{B\times C\times H\times W}$. For each group of features $X_{group} \in R^{B\times \frac{C}{groups}\times H\times W}$, a convolutional network is used to generate the offset values $\Delta p = Conv_{offset}(X_{group})$. The offset values are $\Delta p \in R^{B\cdot g\times 2\times H_k\times W_k}$ then normalized and added to the reference points to obtain the offset positions p, where p represents the grid reference points. The principle is Eq. (7) – (8).

Using the offset positions p, the input feature map x is dynamically sampled through bilinear interpolation, producing sparsely sampled features $X_{sampled} \in R^{B\times C\times 1\times N}$.

$$p = \tanh(\Delta p)\cdot offset\_range + P_{ref} \tag{7}$$

$$X_{sampled} = GridSample(x, p) \tag{8}$$

The features obtained from dynamic offset sampling are then fed into the multi-head attention module. First, a convolutional layer is used to compute the values of Query(q), Key(k), and Value(v), where q, k, and v are the respective feature representations. Next, the similarity between the Query and Key is computed to obtain the attention weights A. The attention weights A are used to perform a weighted aggregation of v. The aggregated result z is then passed through the projection output layer W_{proj} to generate the final features. The input features x are added to the final features via a residual connection, resulting in the output y, where W_{proj} is the projection output layer, and X_{res} is914 the residual connection. The principle is given in Eq. (9) - Eq. (13).

$$L = H\cdot W, C_h = \frac{C}{heads} \tag{9}$$

$$q = W_q x, k = W_k x_{sampled}, v = W_v x_{sampled} \tag{10}$$

$$A = Soft\max(\frac{q\cdot k^T}{\sqrt{C_h}} + PE(q,k)) \tag{11}$$

$$z = A\cdot v \tag{12}$$

$$y = W_{proj}(z) + x_{res} \tag{13}$$

The C2DA module incorporates the powerful characteristics of the dynamic attention mechanism, leveraging deformable convolution and self-attention to enhance the ability to focus on critical spatial regions in visual tasks. It demonstrates strong adaptability while effectively balancing local details and global context.

C. SAFPN

Traffic police gesture recognition is often challenged by various complex background interferences, such as lighting conditions, vehicles, and pedestrians, which significantly increase the difficulty of gesture recognition. In scenarios with lighting variations or occlusions, edge features often provide more information than content features. However, the neck structure of the original network primarily focuses on high-level semantic features, lacking effective preservation of edge

and local details [26], which can easily lead to the loss of target edge information.

To address this issue, this paper proposes an edge-enhanced multi-branch fusion structure (SAFPN), which specifically includes the SSAF (Shallow Edge-Assisted Fusion) module and the AAF(Advanced Assisted Fusion) module. These modules are designed to strengthen edge information extraction and the comprehensive utilization of multi-scale features.

The SSAF module combines the capabilities of the Sobel operator and SAF, fusing the shallow extracted target edge information, high-resolution shallow features, information from the same level of the backbone network, and deep features. This approach significantly preserves critical details in shallow layers. The structure is illustrated in Fig. 5 of the paper.

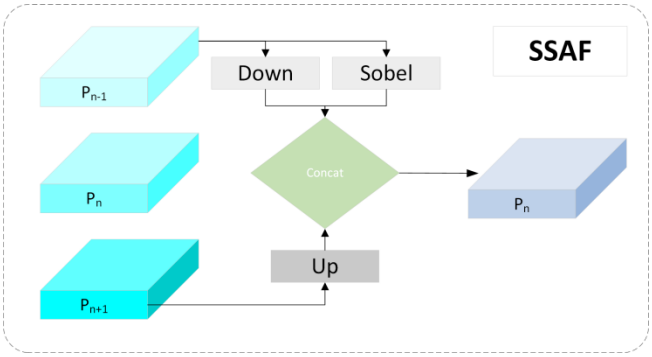

Fig. 5. SSAF structure.

SSAF demonstrates remarkable advantages in small object detection and handling complex backgrounds. The Sobel operator extracts the gradient information of images, effectively capturing the edge features of targets. It also suppresses interference from smooth areas in complex backgrounds, retaining only the edge regions with significant changes, thereby reducing the impact of background noise. Additionally, as a fixed edge extractor, the Sobel operator requires no additional training parameters, adds negligible computational complexity, and yet significantly enhances the network's sensitivity to target regions. The output results after applying SSAF are as follow [see Eq. (14)]

$$P'_n = concat\ (S(P_{n-2}), P_{n-1}, P_n, P'_{n+1}) \tag{14}$$

The AAF module aggregates shallow high-resolution features, shallow low-resolution features, information from sibling layers at the same level, and comprehensive information from the previous layer. This enables the transfer of more diverse gradient information to the network's output layer. As shown in Fig. 6, the final output layer combines information from multiple different layers, significantly enhancing the network's ability to detect targets at various scales. The output results after applying AAF are as follows. Additionally, the AAF module utilizes 1×1 convolutions to adjust the channel information of each layer, optimizing the contribution of features from each layer to the final result [see Eq. (15)]

$$P''_n = concat\ (P'_{n-1}, P''_{n-1}, P'_n, P'_{n+1}) \tag{15}$$

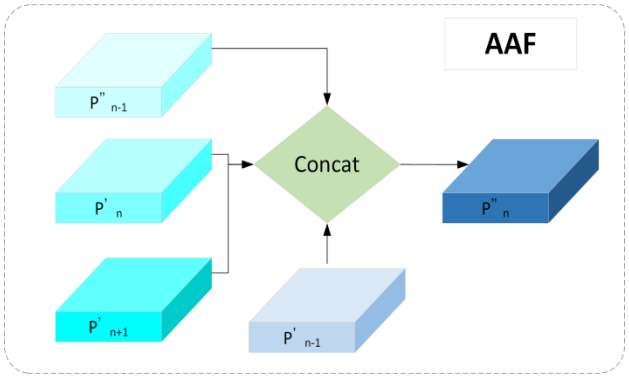

Fig. 6. AAF structure.

The SSAF module effectively reduces background noise interference, while the multi-scale fusion and dynamic weighting mechanism of the AAF module further integrate effective features. The overall design not only enhances the detection capability for small objects and prominent regions but also improves the method's robustness and detection accuracy in complex backgrounds.

D. Loss

In traffic police gesture recognition tasks in open scenarios, training data inevitably contains low-quality samples [27] due to issues with the images themselves or their annotations. Additionally, this task faces the challenge of imbalanced sample distribution, with an uneven distribution of hard and easy samples. These factors significantly limit the effectiveness of traditional loss functions in training.

To address this issue, we creatively introduce WIoU and Slide Loss, tackling the problem from two perspectives: target box optimization and sample weight allocation. Together, these approaches effectively mitigate the critical challenges in traffic police gesture recognition tasks.

1) WIoU: Since training data inevitably contains low-quality samples, geometric factors such as distance and aspect ratio exacerbate the penalties imposed on these samples, thereby reducing the method's generalization performance. When the anchor box aligns well with the target box, a good loss function should attenuate the penalties from geometric factors. Reduced training intervention can enhance the method's generalization capability. Based on this principle, Tong developed distance attention and introduced WIoU v1 with a two-layer attention mechanism, The principle is given in Eq. (16) – Eq. (18).

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \tag{16}$$

$$R_{WIoU} = \exp\left(\frac{(x-x_{gt})^2 + (y-y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \tag{17}$$

$$L_{IoU} = 1 - IoU = 1 - \frac{W_i \cdot H_i}{S_u} \tag{18}$$

Here, W_g and H_g represent the size of the minimum enclosing box, x and y are the center coordinates of the anchor box, and X_{gt} and Y_{gt} are the center coordinates of the ground truth box. The symbol $*$ indicates that W_g and H_g are detached from the computation graph to prevent gradients that hinder convergence. W_i and H_i denote the width and height of the predicted box, while S_u represents the union area.

WIoU v3 builds upon WIoU v1 by introducing a dynamic non-monotonic focusing mechanism, using outlier degree instead of IoU for anchor box quality evaluation. It also provides a more flexible gradient gain allocation strategy. This strategy reduces the competitiveness of high-quality anchor boxes while mitigating harmful gradients from low-quality samples, allowing the method to focus on optimizing medium-quality anchor boxes. Consequently, it enhances overall detection performance. The principle is given in Eq. (19) - Eq. (20)

$$L_{WIoUv3} = r \cdot L_{WIoUv1}, r = \exp(-\beta \cdot \delta) \tag{19}$$

$$\beta = \frac{L^*_{IoU}}{\overline{L}_{IoU}} \in [0, +\infty) \tag{20}$$

Here, β represents the outlier degree of the anchor box, and L^*_{IoU} is the exponentially weighted moving average of L_{IoU}. L^*_{IoU} is a dynamic value used to measure the quality of the anchor box. δ is a hyperparameter that controls the mapping relationship between the outlier degree β and the gradient gain r. When $\beta = \delta$, $r = 1$, meaning the anchor box receives the highest gradient gain.

In scenarios involving long distances and small targets, WIoU can provide more accurate target localization by applying weighted corrections to the shape, size, and offset position of the bounding box. It effectively alleviates the issue of bounding box displacement in complex backgrounds.

2) SlideLoss: Slide Loss introduces a dynamic modulation factor to segmentally adjust sample weights based on their IoU values. Additionally, Slide Loss dynamically adjusts the IoU threshold to ensure stability in loss calculation.

It categorizes samples into three IoU intervals: low, medium, and high. Different weights are assigned to each interval: low-IoU samples retain their original weights, medium-IoU samples are assigned amplified weights to enhance focus on hard-to-classify samples, and high-IoU sample weights are exponentially decayed to reduce interference from easily classifiable samples.

This mechanism enables the model to automatically assign higher optimization weights to challenging samples in gesture detection tasks under long-distance and complex background conditions. It addresses the shortcomings of Binary Cross-Entropy (BCE) loss, including imbalanced positive-negative sample weighting, insufficient optimization for hard-to-classify samples, and inadequate focus on low-confidence samples. The principle is given in Eq. (21) – Eq. (25).

$$L_{Slide} = L_{base} \cdot \alpha \tag{21}$$

$$\alpha_1 = 1.0, \text{ if true} \leq \beta - 0.1 \tag{22}$$

$$\alpha_2 = \exp(1.0 - \beta), \text{if } \beta - 0.1 < \text{true} < \beta \tag{23}$$

$$\alpha_3 = \exp(-(\text{true} - 1.0)), \text{ if true} \geq \beta \tag{24}$$

$$\text{auto\_iou} = \max(\beta, 0.2) \tag{25}$$

β represents the dynamic IoU threshold, constrained to a minimum value of 0.2. Low-IoU samples $true \leq \beta - 0.1$ retain their loss unchanged $\alpha_1 = 1.0$, medium-IoU samples $\beta - 0.1 < true < \beta$ have their loss amplified to $\exp(1.0 - \beta)$ to enhance focus on hard-to-classify cases, and high-IoU samples $true \geq \beta$ have their loss exponentially decayed to $\exp(-(\text{true} - 1.0))$, reducing the interference of easily classified samples on the overall loss. By assigning adaptive weights to samples at different IoU levels using the modulation factor α, Slide Loss optimizes hard-to-classify samples, preserves low-IoU samples, and minimizes the influence of high-IoU samples. This design makes Slide Loss particularly effective for traffic police gesture detection in long-distance and complex background scenarios.

WIoU enhances the geometric information modeling capability of bounding boxes through dynamic weighting, while Slide Loss adjusts sample weights based on IoU differences, prioritizing the optimization of rare and challenging samples with weaker gradients. The ingenious combination of these two methods effectively addresses the critical challenges in traffic police gesture recognition tasks, significantly improving the detection performance and generalization ability of the approach.

IV. Experience

To evaluate the effectiveness of the proposed method in traffic police gesture recognition tasks with long distances and complex backgrounds, this study conducted comprehensive testing and assessment across various scenarios. The proposed method was trained and validated on a custom-built traffic police gesture dataset and compared with several mainstream object detection methods. To further demonstrate the algorithm's generalization ability, performance comparisons were also made with leading algorithms on the WiderPerson and VOC datasets. Additionally, the study performed ablation experiments to analyze the contributions of different components of the method in detail, and visual methods were employed to conduct an in-depth discussion of the experimental results.

A. Setting

The improved YOLOv11 method was used for training and testing during the experiments. The experimental system was based on Ubuntu 22.04. Table I presents the experimental environment and the configurations of some hyperparameters.

TABLE I. EXPERIMENTAL ENVIRONMENT CONFIGURATION

Project Environment	Parameter Setting
Framework	Pytorch
CPU	Xeon(R) Platinum 8352V
GPU	3080*2
Batch Size	64
Epoch	300
Workers	8
CUDA	12.1
Optimizer	SGD

B. Datasets and Evaluation Metrics

The datasets used in this study include the self-collected CTPG dataset, the WiderPerson dataset, and the VOC dataset. Accuracy is evaluated using metrics such as Precision (P), Recall (R), and Mean Average Precision (mAP).

The self-collected CTPG dataset is sourced from Beijing University of Technology's public transportation control gestures [28]. It includes difficult samples from midday and evening scenarios, captured from three different viewing angles: left, center, and right. During the experiment, frames were randomly sampled from the provided video samples and annotated using the makesense platform, resulting in 5,857 training images, 1,200 validation images, and 1,200 test images. The gestures in the dataset represent real-world road scenarios, characterized by long distances, complex backgrounds, and occlusion. The gestures consist of eight categories: "Stop," "Straight Ahead," "Left Turn," "Left Turn and Wait," "Right Turn," "Lane Change," "Slow Down," and "Pull Over." Example samples are shown in Fig. 7.

The WiderPerson dataset [29] is a large-scale dataset for pedestrian detection, containing 13,382 images and over 400,000 high-quality annotated pedestrian instances. It covers complex scenarios such as occlusion, dense crowds, small targets, and various pedestrian poses. The dataset is sourced from real-world diverse scenes and provides precise bounding box annotations, supporting training and evaluation for pedestrian detection tasks. Its scene diversity and challenges make it an important benchmark dataset for pedestrian detection research.

The PASCAL VOC dataset [30] is one of the most widely used standard datasets in the field of computer vision, focusing on tasks such as object detection, image classification, and semantic segmentation. The dataset includes 20 common object categories and provides precise annotation information, including bounding boxes, category labels, and pixel-level segmentation masks. The VOC dataset features diverse scenes and object variations, and its clear evaluation metrics and rich challenges make it a key benchmark for object detection and image segmentation research.

C. Comparison Experiments

To demonstrate the superiority of the proposed method, this study compares several mainstream algorithms on the aforementioned datasets, including single-stage and two-stage conventional CNN algorithms, lightweight algorithms, and Transformer-based methods.

As shown in Table II, a comprehensive comparison is made between multiple object detection methods in terms of parameter count, computational cost, model size, performance metrics, and inference speed. The proposed TPGR-YOLO outperforms other methods across several key indicators. TPGR-YOLO has only 2.8 M parameters, a computational cost of 7.6 GFLOPs, and a model size of just 6 M. It achieves superior performance in accuracy (P = 94.4%), recall (R = 83.4%), mAP@50 (93.6%), mAP@50:95 (72.7%), and inference speed (FPS = 72.7), demonstrating its excellence in lightweight design and high-performance object detection. In contrast, traditional methods such as Faster R-CNN have a parameter count as high as 60 M, a computational cost of 255 GFLOPs, and while they show decent accuracy (P = 79.7%, mAP@50 = 86.2%), their inference speed is only 58.5 FPS, making them unsuitable for real-time applications. ASF-YOLO achieves a high FPS while achieving good accuracy due to its lightweight treatment; RT-DETR balances accuracy (P = 90.2%, R = 77.8%) and speed (FPS = 65.8), but its parameter count is 42 M, with a computational cost of 136 GFLOPs and a model size of 166 M, which is not suitable for resource-constrained scenarios. The YOLO series shows good performance in terms of both lightweight design and performance, but still lags behind TPGR-YOLO. For example, YOLOv11n has 2.6 M parameters and an accuracy of 90.1%, but its mAP and recall are lower than TPGR-YOLO. LD-YOLOv10 shows improvements in accuracy (P = 89.7%, R = 77.1%) and speed (FPS = 65.2), but its parameter count and computational cost are 8.1 M and 24.7 GFLOPs, respectively, making it less efficient than TPGR-YOLO.

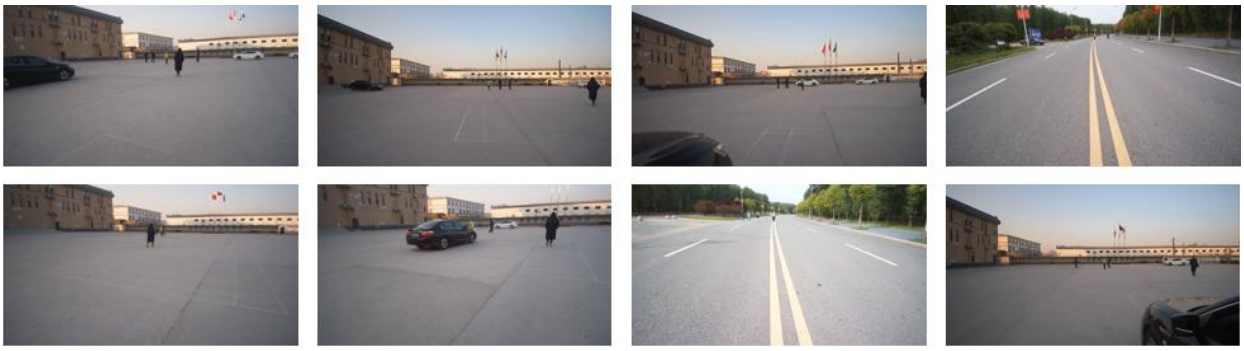

Fig. 7. Several types of action diagram.

TABLE II. COMPARISON OF MAINSTREAM ALGORITHMS ON THE CTPG DATASET

Models	Para(M)	GFLOPs(G)	Model size(M)	P	R	mAP@50	mAP@50:95	FPS
Faster R-CNN[31]	60	255	109	79.7	72.3	86.2	58.5	30
YOLOv5[32]	7.2	16.5	14	82.1	70.4	88.0	60.5	108
YOLOv7	36.9	104.7	72.5	85.3	73.2	86.3	63.4	70
ASF-YOLO[33]	15.9	79.1	25	88.5	74.5	86.0	64.7	137
YOLOv9s	9.7	37.6	15.2	87.6	75.2	85.6	63.4	100
LD-YOLOv10[34]	8.1	24.7	17	89.7	77.1	88.3	65.2	102
RT-DETR[35]	42	136	166	90.2	77.8	84.9	65.8	40
YOLOv11n	2.6	6.3	5.5	90.1	78.2	88.2	67.4	140
TPGR-YOLO	2.8	7.6	6.0	94.4	83.4	93.6	72.7	135

To evaluate the generalization ability of the algorithm, this study compares the proposed method with mainstream algorithms on the WiderPerson dataset. The results are shown in Table III. The experimental results demonstrate that the proposed TPGR-YOLO outperforms other object detection methods in multiple metrics, including precision, recall, and mAP. TPGR-YOLO achieves a precision of 94.5%, recall of 81.0%, mAP@50 of 89.2%, and mAP@50:95 of 69.6%, leading across all performance indicators.

In comparison, YOLOv5s and YOLOv7-tiny show weaker performance, with precisions of 82.2% and 83.2%, mAP@50 of 78.9% and 82.7%, and mAP@50:95 of only 59.7% and 61.7%. As the methods evolve, their performance gradually improves. When compared to the state-of-the-art YOLO detector, YOLOv11n, the proposed method surpasses its precision, achieving 91.8%, with an mAP@50 of 87.8%. TPGR-YOLO demonstrates significant advantages in balancing detection accuracy and performance, fully proving the superiority of the improved algorithm.

TABLE III. COMPARISON EXPERIMENTS OF THE WIDERPERSON DATASET

Models	P	R	mAP@50	mAP@50:95
YOLOv5s	82.2	66.7	78.9	59.7
YOLOv7-tiny	83.2	71.7	82.7	61.7
YOLOv8n[36]	86.4	76.4	85.3	65.9
Yolov9s	89.4	76.8	86.0	65.6
YOLOv10s	90.7	78.2	86.7	67.8
Yolov11n	91.8	79.3	87.8	68.4
Ours	94.5	81.0	89.2	69.6

The experimental results on the VOC dataset show that the proposed TPGR-YOLO performs the best across all key metrics. The results are shown in Table IV. TPGR-YOLO achieves a precision of 71.9%, recall of 64.8%, mAP@50 of 70.7%, and mAP@50:95 of 54.8%, significantly outperforming other methods.

In comparison, the weaker-performing YOLOv5s achieves only 55.7% precision and 53.5% mAP@50, while YOLOv7-tiny reaches precision and mAP@50 of 59.4% and 57.2%, respectively. As the methods evolve, performance improves. For example, YOLOv10s and YOLOv11n achieve mAP@50 values of 63.1% and 67.9%, respectively. However, TPGR-YOLO still leads across all metrics, particularly in mAP@50:95, where it surpasses YOLOv11n by approximately 3.1 percentage points. These results demonstrate that TPGR-YOLO has a clear advantage in balancing detection accuracy and performance, making it one of the best-performing methods on the VOC dataset.

TABLE IV. COMPARISON EXPERIMENTS OF THE VOC DATASET

Models	P	R	mAP@50	mAP@50:95
YOLOv5s	55.7	53.6	53.5	46.7
YOLOv7-tiny	59.4	55.2	57.2	45.0
YOLOv8n	62.7	56.3	61.8	46.0
Yolov9s	65.4	58.4	62.2	49.7
YOLOv10s	68.2	59.5	63.1	50.5
Yolov11n	71.2	61.7	67.9	51.7
Ours	71.9	64.8	70.7	54.8

The performance comparison between WIoU combined with Slide Loss and the original model loss (Table V) reveals that: The sliding window mechanism in Slide Loss refines the regression process of prediction boxes, significantly enhancing the model's adaptability to small targets and complex scenarios, demonstrating marked advantages over traditional BCE loss in small object detection and regression accuracy. While both WIoU and CIoU improve model performance when combined with Slide Loss, WIoU achieves superior comprehensive performance through its focus on medium-quality anchor boxes and suppression of low-quality samples interference, thereby forming complementary advantages with Slide Loss.

TABLE V. COMPARISON OF DIFFERENT LOSS FUNCTIONS

LOSS	P	R	mAP@50	mAP@50:95
CIoU+BCE	90.1	78.2	88.2	67.4
CIoU+SlideLoss	90.6	78.4	88.7	67.9
WIoU+BCE	90.7	78.3	88.6	68.2
WIoU+SlideLoss	91.4	78.9	89.5	68.5

D. Ablation Experiments

To validate the performance improvement contributed by each module, this study conducts ablation experiments by progressively introducing key modules and evaluating their impact on performance. The results are shown in Table VI.

TABLE VI. ABLATION ANALYSIS OF CTPG DATASET

Modules	Para/M	GFLOPs	Model size	P	R	mAP@50	mAP@50:95
YOLOv11n	2.6	6.3	5.5	90.1	78.2	88.2	67.4
+RFCAConv	2.7	6.9	5.8	90.7	79.4	89.2	68.1
+C3k2RFCA+C2DA	2.7	6.9	5.8	91.2	80.8	90.7	69.5
+C3k2RFCA+C2DA+SAFPN	2.8	7.6	5.8	92.5	81.5	91.4	70.2
+C3k2RFCA+C2DA+SAFPN+WIoU	2.8	7.6	6.0	93.9	82.6	92.8	71.1
TPGR-YOLO	2.8	7.6	6.0	94.4	83.4	93.6	72.7

The baseline method, YOLOv11n, has 2.6 M parameters, 6.3 GFLOPs, and a model size of 5.5 M, achieving a precision of 90.1%, recall of 78.2%, mAP@50 of 88.2%, and mAP@50:95 of 67.4%, with relatively limited performance. Upon introducing the RFCAConv module, the number of parameters slightly increases to 2.7 M, the computational cost rises to 6.9 GFLOPs, and the method's precision improves to 90.7%, recall increases to 79.4%, mAP@50 reaches 89.2%, and mAP@50:95 rises to 68.1%, indicating that RFCAConv effectively enhances feature extraction capabilities.

Further adding the C2DA module leads to continued performance improvement, with precision reaching 91.2%, recall rising to 80.8%, and mAP@50 and mAP@50:95 increasing to 90.7% and 69.5%, respectively, showcasing the significant potential of C2DA's dynamic offset and local feature modeling in object detection.

When combining C2DA and SAFPN modules, both precision and recall further improve to 92.5% and 81.5%, respectively, with mAP@50 reaching 91.4% and mAP@50:95 increasing to 70.2%, demonstrating the significant effect of these modules on multi-scale feature fusion and contextual modeling.

Finally, introducing the WIoU loss function to optimize the method's localization capability results in a precision increase to 93.9%, recall reaching 82.6%, and mAP@50 and mAP@50:95 achieving 92.8% and 71.1%, respectively, highlighting the importance of WIoU for optimizing small object localization strategies.

The complete TPGR-YOLO method, integrating all modules, has 2.8M parameters, 7.6 GFLOPs, and a model size of 6.0 M. Its final precision reaches 94.4%, recall is 83.4%, and mAP@50 and mAP@50:95 are 93.6% and 72.7%, respectively. The experimental results confirm that each module contributes positively to performance improvement, and TPGR-YOLO demonstrates exceptional performance in lightweight design and high-performance object detection.

E. Research Result

The study proposes an improved TPGR-YOLO for traffic police gesture recognition by introducing the RFCAConv module, the C2DA module, the edge-enhanced multi-branch fusion strategy (SAFPN), as well as the WIoU and SlideLoss loss functions. The proposed method demonstrates excellent performance in feature extraction, small object detection, and complex background handling. Experimental results show that TPGR-YOLO achieves high precision, recall, and mAP scores on the self-built CTPG dataset, the WiderPerson dataset, and the VOC dataset, while maintaining real-time inference speed. Furthermore, the ablation study validates the positive contributions of each improved module to the overall performance.

F. Visual Analysis

1) Dataset visualization: The label distribution after training on the CPRG dataset is shown in Fig. 8. From the figure, it can be observed that the "Straight Ahead" category has the largest number of samples, exceeding 1,000, while the "Pull Over" category has the fewest samples. It can also be noted that the labels are primarily distributed in the central region of the images, with most of them having a relatively uniform shape, exhibiting a clear concentration trend.

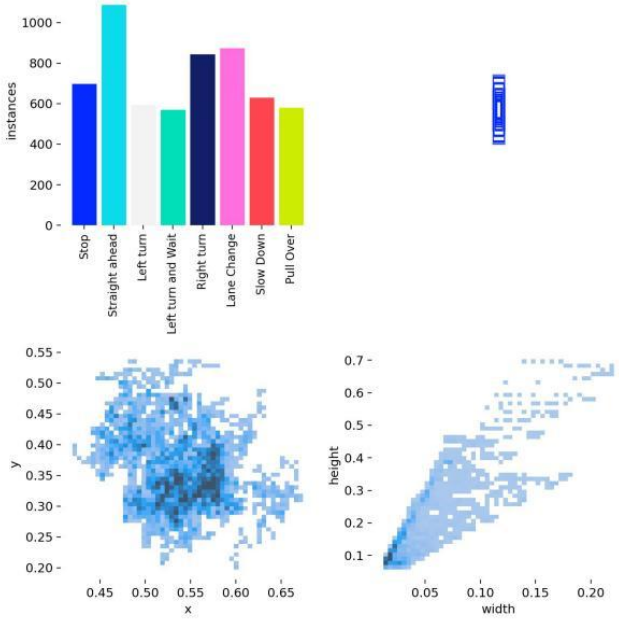

Fig. 8. Dataset visualization.

2) Results visualization: The visualization of the experimental results is shown in Fig. 9. The figure contains 8 rows, corresponding to the categories "Stop," "Straight Ahead," "Left Turn," "Left Turn and Wait," "Right Turn," "Lane Change," "Slow Down," and "Pull Over." The left column displays the original images, the middle column presents the visualization results after training with the original method, and the right column shows the visualization results after applying the improved method.

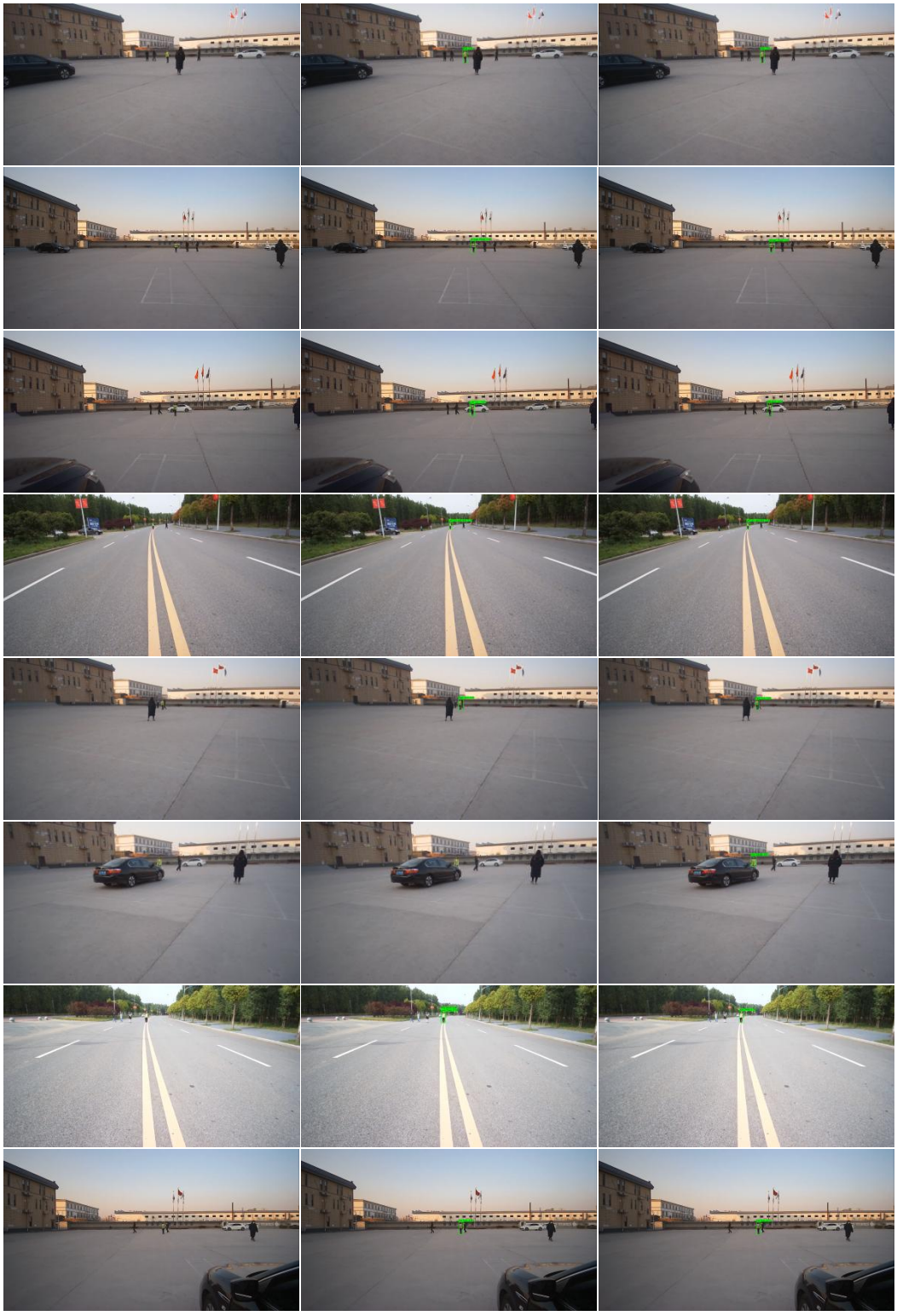

Fig. 9. Result visualization.

From the figure, it can be observed that in the "Line Change" scenario, the original model failed to correctly recognize the traffic officer's gesture due to the vehicle obstructing the lower part of the officer's body, whereas the improved model accurately identified the gesture. In the "Slow Down" and "Left Turn" scenarios, interference from a white vehicle in the background led the original model to incorrectly recognize two actions, while the improved model, by enhancing the recognition of target edges, demonstrated ideal performance in these scenes.

For the "Pull Over" and "Right Turn" gestures, the original method failed to accurately identify the edges of the body and gestures during inference due to background interference, resulting in larger bounding boxes or failure to properly frame the hand. This neglect of detailed information directly affected the inference accuracy. The improved method successfully overcame these issues, showing superior performance.

V. CONCLUSION

In order to address the challenges of small-scale traffic police gestures and complex backgrounds in open traffic scenes, this study improves the backbone network, neck network, and loss function of the YOLOv11 model. The modified algorithm demonstrates strong generalization ability across multiple public datasets. However, since the algorithm relies on single-frame image classification, misjudgments on boundary gestures are inevitable.

Although this research effectively addresses the issues of small-scale objects and complex backgrounds, future work should focus on overcoming the challenges of real-time deployment and exploring the integration of additional sensor data. Furthermore, the model should be improved to accommodate misjudgments of boundary gestures and non-standard traffic police signals. This will enhance the model's applicability and reliability in real-world applications.

ACKNOWLEDGMENT

This work was supported, the National Natural Science Foundation of China (Grant No. 62102033, U24A20331, 62371013, 61931012), the Beijing Natural Science Foundation (No. L247007), the R&D Program of Beijing Municipal Education Commission (Grant No. KZ202211417048), The Project of Construction and Support for high-level Innovative Teams of Beijing Municipal Institutions (Grant No. BPHR20220121), the Academic Research Projects of Beijing Union University (No. ZKZD202302, ZK20202403).

REFERENCES

[1] Mengying Chang, Huizhi Xu, Yuanming Zhang. Low light recognition of traffic police gestures based on lightweight extraction of skeleton features[J], Neurocomputing. 2025, 617: 129042-129042.

[2] Mingde Zheng, Michael S. Crouch, Michael S. Eggleston. Surface Electromyography as a Natural Human-Machine Interface: A Review[J], IEEE sensors journal, 2022, 22(10): 9198-9214.

[3] Muhammad A, Dong S K, Muhammad O, Kang R P, et al. OR-Skip-Net: Outer Residual Skip Network for Skin Segmentation in Non-Ideal Situations[J], Expert Systems with Applications, 2020, 141: 112922-112922.

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

[5] Hedan Bai et al. Stretchable distributed fiber-optic sensors. Science, 2020, 370, 848-852.

[6] LIUMY, HANGCZ, WUXY, et al. Investigation of stretchable strain sensor based on CNT/AgNW applied in smart wearable devices[J]. Nanotechnology, 2022, 33(25): 255501.

[7] LI XT, KOHKH, FARHANM, et al. An ultraflexible polyurethane yarn-based wearable strain sensor with a polydimethylsiloxane infiltrated multilayer sheath for smart textiles[J]. Nanoscale, 2020, 12(6): 4110-4118.

[8] Wenxuan M, Qingtian Z, Ge S, Minghao Z, et al. Design and Implementation of Traffic Police Hand Gesture Recognition System Based on Surface Electromyographic Signals[J], 2022 IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference(IMCEC), 2022, 5: 1888-1894.

[9] Wenxuan Ma;Ge Song;Qingtian Zeng;Hongxin Zhang;Minghao Zou;Ziqi Zhao. FFCSLT: A Deep Learning Model for Traffic Police Hand Gesture Recognition Using Surface Electromyographic Signals[J]. 2024, 15(8): 13640-13655.

[10] ZHANG Weidong, WANG Zexing, WU Xuangou. WiFi Signal-Based Gesture Recognition Using Federated Parameter-Matched Aggregation [J]. Sensors, 2022, 22(6): 2349.

[11] ZHOU Chengwei, GU Yujie, SHI Zhiguo, et al. Structured Nyquist correlation reconstruction for DOA estimation with sparse arrays[J]. IEEE Transactions on Signal Processing, 2023, 71: 1849-1862.

[12] Feiyi X, Feng X, Jiucheng X, Chi-Man P, Huimin L, Hao G, et al. Action Recognition Framework in Traffic Scene for Autonomous Driving System. [J], IEEE Transactions on Intelligent Transportation Systems, 2022, 23(11): 22301-22311.

[13] Zheng F, Junjie C, Kun J, Sijia W, Junze W, Mengmeng Y, Diange Y, et al. Traffic Police 3D Gesture Recognition Based on Spatial-Temporal Fully Adaptive Graph Convolutional Network[J], IEEE Transactions on Intelligent Transportation Systems, 2023, 24: 9518-9531.

[14] Nan M, Zhixuan W, Yifan F, Cheng W, Yue G, et al. Multi-View Time-Series Hypergraph Neural Network for Action Recognition[J], IEEE Transactions on Image Processing, 2024, 33: 3301-3313.

[15] Xiaofeng G, Qing Z, Yaonan W, Yang M, et al. MG-GCT: A Motion-Guided Graph Convolutional Transformer for Traffic Gesture Recognition[J], IEEE Transactions On Intelligent Transportation Systems, 2024. 25: 14031-14039.

[16] Ziwei Zhang, Bingbing Wu, Yulian Jiang. Gesture Recognition System Based on Improved YOLO V3[J], 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), 2022: 1540-1543.

[17] Shichao W, Chenxia G, Ruifeng Y, Qianchuang Z, Haoyu R, et al. A Lightweight Vision-Based Measurement for Hand Gesture Information Acquisition[J], IEEE sensors journal, 2023, 23(5): 4964-4973.

[18] Saloni S, Anushka P, Prachi J, Ahbaz M, Varsha N, et al. Hand Gesture Recognition Using YOLO Models for Hearing and Speech Impaired People[J], 2022 IEEE Students Conference on Engineering and Systems(SCES), 2022: 1-6.

[19] Maha H, Wesam S, Mohammed Z, Tariq K, et al. Hand Gesture Recognition Based on CNN and YOLO Techniques[J], Decision Science Letters, 2024, 13(4): 977-990.

[20] Yu-Yu Yang, Hsu-Han Yang, Jung-Kuei Yang. YOLO-LRHG: Long Range Hand Gesture Detection Using YOLO with Attention Mechanism[J], 2024 IEEE 7th International Conference on Electronic Information and Communication Technology(ICEICT), 2024: 985-990.

[21] Weina Zhou, Xile Li. PEA-YOLO: a Lightweight Network for Static Gesture Recognition Combining Multiscale and Attention Mechanisms[J], Signal Image and Video Processing, 2024, 18(1): 597-605.

[22] Jichi Liu, Wei Li, Houkun Lyu, Feng Qi. YOLO-based microglia activation state detection[J]. The Journal of Supercomputing. 2024: 24412-24434.

[23] Jian X, Chenglong Z, Jiaqiang M, Zibo C, Ying L, et al. Lightweight Detection Model for Safe Wear at Worksites Using GPD-YOLOv8 Algorithm[J], Scientific Reports, 2025, 15(1): 1-13.

[24] Zhaohui L, Wenshuai H, Wenjing C, Jiaxiu C, et al. The Algorithm for Foggy Weather Target Detection Based on YOLOv5 in Complex Scenes[J], Complex & Intelligent Systems, 2024, 11(1): 1-18.

[25] Renxiang Z, Guangyun Z, Rongting Z, Xiuping J, et al. A Deformable Attention Network for High-Resolution Remote Sensing Images Semantic Segmentation[J], IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-14.

[26] Tengfei Ma, Haitao Yin. MAFPN: a Mixed Local-Global Attention Feature Pyramid Network for Aerial Object Detection[J], Remote Sensing Letters, 2024, 15(9): 907-918.

[27] Xiaoqiang D, Hongchao C, Zenghong M, Wenwu L, Mengxiang W, Zhichao M, et al. DSW-YOLO: A Detection Method for Ground-Planted Strawberry Fruits under Different Occlusion Levels[J], Computers and Electronics in Agriculture, 2023, 108304.

[28] Nan M, Zhixuan W, Yifan F, Cheng W, Yue G, et al. Multi-View Time-Series Hypergraph Neural Network for Action Recognition[J], IEEE Transactions on Image Processing, 2024, 33: 3301-3313.

[29] Lihu P, Jianzhong D, Zhengkui W, Shouxin P, Cunhui Z, et al. HF-YOLO: Advanced Pedestrian Detection Model with Feature Fusion and Imbalance Resolution[J], Neural Processing Letters, 2024, 56(2).

[30] Jinsheng X, Haowen G, Jian Z, Tao Z, Qiuze Y, Yunhua C, Zhongyuan W, et al. Tiny Object Detection with Context Enhancement and Feature Purification[J], Expert Systems with Applications, 2023, 211: 118665.

[31] Kyungseo Min, Gun-Hee Lee, Seong-Whan Lee. Attentional Feature Pyramid Network for Small Object Detection. [J], Neural Networks, 2022, 155: 439-450.

[32] Xudong Dong, Shuai Yan, Chaoqun Duan. A Lightweight Vehicles Detection Network Model Based on YOLOv5[J], Engineering Applications of Artificial Intelligence, 2022, 113: 104914-104914.

[33] Shuai Yuan, Xiangjie Kong, Shuai Zhang. Research on Enhanced YOLOv8 Gesture Recognition Method for Complex Environments*[J], 2024 Wrc Symposium on Advanced Robotics and Automation, Wrc Sara, 2024: 141-146.

[34] Qiu, Xiaoyang; Chen, Yajun; Cai, Wenhao; Niu, Meiqi; Li, Jianying. LD-YOLOv10: A Lightweight Target Detection Algorithm for Drone Scenarios Based on YOLOv10. Electronics; Basel , 2024: 3269.

[35] Huanyu Y, Jun W, Yuming B, Jiacun W, et al. Istd-Detr: A Deep Learning Algorithm Based on Detr and Super-Resolution for Infrared Small Target Detection[J], Neurocomputing, 2025: 129289.

[36] Wong Min On, Nirase Fathima Abubacker. YOLO-Driven Lightweight Mobile Real-Time Pest Detection and Web-Based Monitoring for Sustainable[J]. International Journal of Advanced Computer Science and Applications, 2024, 15: 658-673.