RESEARCH ON PERSON AND VEHICLE DETECTION AND COUNTING METHOD BASED ON IMPROVED YOLOV5

Zhidan Yuan, Jialu Sun, Yinglu Wei and Yuqing Zhang

School of Information Engineering, Jiangsu Maritime Institute, Nanjing, China

ABSTRACT

With the rapid advancement of autonomous driving technologies, Person and vehicle detection and counting tasks in urban road environments are confronted with significant challenges, including object occlusion, background interference, and insufficiently discriminative feature representations. These factors directly degrade detection accuracy and counting reliability. To address these issues, this study proposes an improved algorithm that integrates YOLOv5 with the SE attention mechanism. The SE module adaptively recalibrates channel-wise feature responses by modeling inter-channel dependencies, thereby enhancing the network's focus on salient target features while suppressing irrelevant background noise and interference. Experiments were conducted on the BDD100K dataset to evaluate the effectiveness of the proposed approach. The results demonstrate that the improved YOLOv5 model outperforms the baseline YOLOv5 in Person and vehicle detection and counting tasks, achieving a performance improvement of 6.63%. These findings indicate that the proposed method enhances both detection accuracy and robustness, and thus exhibits greater efficiency and reliability for practical deployment in autonomous driving systems.

KEYWORDS

Object Detection, YOLOv5, Attention Mechanism, Human and Vehicle Detection and Counting

1. INTRODUCTION

With the profound integration of Intelligent Transportation Systems (ITS) and autonomous driving technologies, real-time and precise environmental perception has emerged as the cornerstone for ensuring vehicular safety and decision-making reliability. Within complex urban road scenarios, Persons and vehicles constitute the primary traffic participants; thus, their detection precision and counting accuracy directly dictate the efficacy of path planning, collision warning, and microscopic traffic flow control[1]. Particularly in the context of Vehicle-Infrastructure Cooperation (VIC) and smart city development, perception systems are required not only to possess ultra-fast real-time response speeds but also to provide stable semantic information support across diversified traffic conditions. Consequently, developing an environmental perception algorithm that balances detection efficiency with recognition precision holds significant research value for enhancing the engineering utility of autonomous driving systems.

To address the limitations of YOLOv5's perception capabilities in complex scenarios, this paper proposes an improved YOLOv5[2] algorithm integrated with the Squeeze and Excitation (SE) attention mechanism[3], specifically designed to tackle the challenges of Person and vehicle detection and counting. By modeling explicit interdependencies between feature channels, the SE module adaptively recalibrates channel-wise feature responses, thereby enhancing the efficiency of feature representation. This attention mechanism emulates the human visual system by

implementing differential focusing on specific feature channels or spatial regions, effectively bolstering the model's capacity for detection and counting. Particularly in cases of target occlusion or small object detection, the SE mechanism enhances the expression of critical features, significantly improving detection precision for occluded entities and distant small targets. Furthermore, by suppressing background noise and irrelevant features, the SE mechanism enables the model to concentrate on salient objects, further augmenting both detection accuracy and counting reliability. The primary contributions of this work include:

- **Architectural Enhancement:** An SE attention module is embedded into the YOLOv5 backbone for feature extraction. Through global average pooling and non-linear mapping, the module achieves adaptive adjustment of feature channel weights, thereby suppressing background noise and reinforcing the feature representation of occluded targets.
- **Experimental Validation:** Comprehensive scene-level validation was conducted on the large-scale autonomous driving dataset, BDD100K. The experimental results demonstrate that the improved algorithm enhances counting accuracy while maintaining real-time performance, verifying the robust generalization capability of the proposed method.

The remainder of this paper is organized as follows: Section 2 reviews related work regarding the improved YOLOv5 algorithm, focusing on the fundamental YOLOv5 framework and its applications in Person and vehicle detection. Section 3 elaborates on the detailed methodology and workflow of the proposed approach. Section 4 describes the experimental design, including the dataset characteristics, evaluation metrics, and baseline methods. Section 5 presents the experimental results and provides an in-depth analysis of the improved algorithm's performance. Finally, Section 6 concludes the paper and discusses potential directions for future work.

2. RELATED WORK

2.1. Overview of the YOLOv5 Network

Developed as an evolution of the YOLO series[4], YOLOv5 is a state-of-the-art one-stage object detection model implemented by the Ultralytics team based on the PyTorch framework. It inherits the end-to-end regression-based detection paradigm, significantly enhancing detection precision and model stability while preserving high-speed inference. Compared to its predecessors, such as YOLOv3[5] and YOLOv4[6], YOLOv5 undergoes systematic optimizations in network architecture and training strategies. By refining feature extraction and multi-scale fusion mechanisms, it bolsters the model’s adaptability to targets across diverse scales. Furthermore, YOLOv5 excels in lightweight design, offering multiple scaled versions that facilitate flexible deployment across varying computational constraints, thereby achieving an effective equilibrium between detection accuracy and inference latency.

In addition, YOLOv5 introduces enhancements in data augmentation strategies, automated anchor box clustering, and training pipeline optimization, which collectively strengthen the model’s generalization capability and convergence stability. In complex environments, YOLOv5 exhibits superior robustness and real-time performance compared to earlier models. Unlike two-stage detection algorithms, it bypasses the region proposal generation process, resulting in higher inference efficiency tailored for real-time tasks. When compared to conventional YOLO iterations, it demonstrates a higher degree of engineering maturity and ease of deployment, possessing substantial practical utility.

Owing to its superior performance and seamless engineering adaptability, YOLOv5 has been extensively applied in fields such as autonomous driving, intelligent surveillance, traffic

monitoring, and industrial inspection. In tasks involving Person and vehicle detection, object counting, and dynamic scene perception, YOLOv5 provides high detection precision while ensuring real-time responsiveness, establishing itself as one of the most mature and stable detection frameworks in contemporary engineering applications. Consequently, conducting research on structural improvements and performance optimizations based on YOLOv5 holds significant theoretical importance and practical engineering value.

2.2. Person and Vehicle Detection and Counting

Person and vehicle detection and counting constitute a pivotal foundation within Intelligent Transportation Systems and autonomous perception technologies[7]. These tasks are of paramount importance for road status assessment, traffic flow analysis, congestion forecasting, and refined urban management. With the continuous advancement of smart city initiatives, achieving real-time detection and statistics of Persons and vehicles via video surveillance and onboard sensing equipment has become a vital technical means to enhance traffic safety and operational efficiency. Consequently, constructing high-precision and robust detection and counting models has emerged as a prominent research focus in the fields of computer vision and intelligent transportation.

Historically, early methods for Person and vehicle detection primarily relied on traditional machine learning and handcrafted features, such as HOG[8] and SIFT[9], coupled with Support Vector Machines (SVM)[10] for classification and localization. While effective in simplistic scenarios, these methods exhibit limited adaptability to complex illumination variations, scale fluctuations, and occlusions. With the evolution of machine learning[11-13], convolutional neural network (CNN)-based algorithms[14] have become the mainstream. Two-stage detectors, exemplified by the R-CNN series[15], achieved significant gains in accuracy but suffered from high computational complexity, making it difficult to satisfy real-time requirements. Subsequently, one-stage detectors represented by YOLO and SSD[16] significantly accelerated inference speeds through end-to-end regression, leading to widespread application in real-world road scenarios. Recently, research incorporating multi-scale feature fusion[17], Feature Pyramid Networks[18], and optimized loss functions has further bolstered detection precision and stability.

Despite these advancements, Person and vehicle detection and counting still face substantial challenges in practical urban environments[19]. For instance, frequent occlusions and overlaps occur between targets against complex backgrounds; furthermore, significant scale variations in dense scenes lead to insufficient feature representation for small objects. Additionally, cluttered background interference often results in elevated false alarm and miss rates, directly compromising counting accuracy. Existing models often lack the capacity for selective attention to critical information during feature extraction, struggling to suppress irrelevant background features effectively. Therefore, it is imperative to introduce attention mechanisms to enhance feature representation capabilities. The SE mechanism adaptively models inter-channel dependencies to reinforce salient features and suppress redundant information, thereby improving detection and counting performance in complex road scenes. Building upon this, this paper incorporates the SE attention mechanism into the existing detection framework to further enhance detection accuracy and model robustness.

3. APPROACH

3.1. Workflow of the Improved YOLOv5 Algorithm

The overall workflow of the improved YOLOv5 algorithm is illustrated in Figure 1. Adopting an end-to-end design philosophy, the process encompasses the entire processing pipeline, ranging from raw data input to the final object counting output. Initially, the input video streams or image data undergo preprocessing before being fed into the CSPDarknet53[20] backbone of YOLOv5 for multi-scale feature extraction. Through successive layers of downsampling, the backbone generates feature maps at three distinct scales—P3, P4, and P5 to capture both the semantic information and spatial structural details of targets at various sizes. To enhance the model's focus on critical objects, SE attention modules are embedded at each scale’s feature output terminal. Upon entering the module, the feature maps undergo "Squeeze-and-Excitation" operations, achieving adaptive feature recalibration in the channel dimension. This process reinforces the discriminative features of key targets, such as pedestrians and vehicles, while simultaneously suppressing interference from complex backgrounds, thereby improving the overall quality of feature representation.

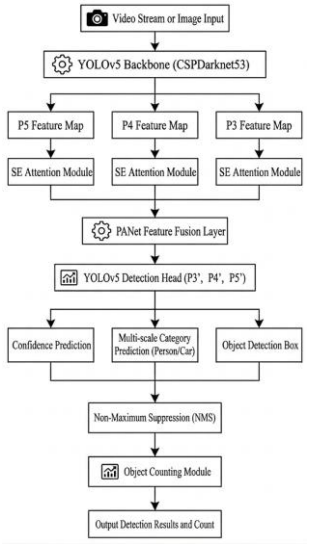

Figure 1. Overall Workflow of Improved YOLOv5 Algorithm

Subsequently, the three-way feature maps enhanced by the SE modules are fed into the PANet feature fusion layer. At this stage, feature interaction and fusion are executed via bidirectional paths: a top-down path transmits high-level semantic information to bolster the semantic representation of low-level features, while a bottom-up path aggregates shallow positional information to refine the spatial localization precision of high-level features. The fused features then generate three output feature layers, denoted as P3', P4', and P5', which are subsequently passed to the detection heads for prediction. The detection head is responsible for mapping abstract features into concrete detection results, employing a parallel branch structure to output three categories of critical information: bounding box regression (defining the geometric localization of targets), multi-scale category prediction (distinguishing between classes such as "pedestrian" and "vehicle"), and objectness confidence prediction (assessing the reliability of the detection results). Through this multi-scale prediction mechanism, the model achieves a robust balance in detection accuracy for both small and large-scale targets.

At the final stage of the pipeline, all prediction results are subjected to Non-Maximum Suppression[21]. This step filters redundant overlapping detection boxes and retains the optimal bounding box with the highest confidence score, thereby minimizing false positives and repeated detections. Ultimately, the filtered valid target information is passed to the object counting module, which independently enumerates the targets categorized as "pedestrian" and "vehicle." During the output phase, the system overlays the detection boxes and classified counting results in real-time, achieving a synchronized visual representation of target detection and statistical quantity estimation.

3.2. Squeeze and Excitation Attention Mechanism

The SE attention mechanism is designed to achieve adaptive feature recalibration by explicitly modeling the inter-dependencies between feature channels. The structural schematic of the SE mechanism is illustrated in Figure 2. The operational workflow of the SE attention mechanism primarily comprises three distinct stages: Squeeze, Excitation, and Scale.

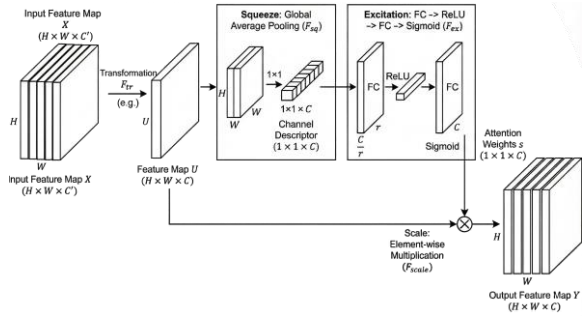

Figure 2. Structural schematic of the SE attention mechanism

During the Squeeze phase, the primary objective of the SE attention mechanism is to transform 2D spatial features into global channel descriptors. Specifically, the module receives the feature map $U \in R^{H \times W \times C}$ generated by the transformation F_{tr} and performs a global average pooling operation on each channel. In this process, the feature values at all spatial coordinates (i, j) within each channel are averaged, thereby compressing the two-dimensional feature maps which initially contain spatial distribution information into a single scalar z_c. The mathematical expression for z_c is defined as shown in Eq. (1):

$$z_c = F_{sq}(u_c) = 1/(H \times W) \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \quad (1)$$

In Eq. (1), z_c denotes the output feature of the c-th channel, and u_c represents the c-th channel of the feature map U. Following this operation, the original feature map is transformed into a channel descriptor with a dimension of $1 \times 1 \times C$, effectively achieving global information aggregation from the spatial dimension to the channel dimension. This phase integrates global contextual information, establishing a foundation for learning inter-channel dependencies and generating adaptive weights in the subsequent Excitation stage.

During the Excitation phase, the core task of the SE attention mechanism is to model the non-linear dependencies between channels and generate adaptive weights for each channel. This phase takes the channel descriptor $z \in R^{1 \times 1 \times C}$ obtained from the Squeeze stage as input and implements feature recalibration through a bottleneck structure consisting of two fully connected (FC) layers. Initially, the first FC layer reduces the channel dimensionality from C to C/r (where r denotes the reduction ratio) to decrease the parameter count and enhance the model's

generalization capability, while incorporating the ReLU activation function δ to bolster non-linear representation. In our work, we follow the original configuration of the SE attention mechanism and set the reduction ratio r to 16, consistent with the setting adopted in the original study[3]. Subsequently, the second FC layer restores the feature dimension back to C, completing the reconstruction of channel information. Finally, a Sigmoid activation function σ maps the output to the [0, 1] interval, generating the channel attention weight vector s. The calculation process is as follows:

$$s=F_{ex}(z,W)=\sigma(g(z,W))=\sigma(W_2\delta(W_1z)) \quad (2)$$

In Eq.(2), W_1 and W_2 represent the weight matrices of the two fully connected layers, respectively. Through this nonlinear mapping process, the model is capable of adaptively learning the relative importance of each channel, thereby providing a precise basis for weight allocation in the subsequent feature recalibration stage.

During the Scale phase, the channel attention weight vector s, generated in the excitation stage, is reapplied to the original feature map U to adjust the channel-wise features via element-wise multiplication. Specifically, for the c-th channel, the scaling calculation process is defined as follows:

$$\tilde{x}_c=F_{scale}(u_c,s_c)=s_c \cdot u_c \quad (3)$$

In Eq.(3), s_c represents the corresponding attention weight coefficient. The term $\tilde{x}_c$ refers to the c-th output channel feature map after weight recalibration. Through this linear recalibration approach, the model adaptively amplifies the response intensity of significant channels while suppressing redundant or interfering information, thereby enhancing the discriminative power of feature representation. The resulting output feature maps achieve dynamic optimization in the channel dimension while preserving the original spatial structure. Within the context of pedestrian and vehicle detection and counting tasks, this mechanism enables the model to focus more precisely on the critical features of targets, ultimately improving the accuracy of both detection and statistical counting.

4. EXPERIMENTAL SETUP

4.1. Experimental Dataset

Experimental validation was conducted using the BDD100K dataset[22], a large-scale benchmark for autonomous driving perception. Renowned for its exceptional scene coverage and environmental diversity, this dataset systematically integrates real-world driving imagery across heterogeneous geographic regions, complex meteorological conditions, and full-spectrum lighting variations. The core library of BDD100K consists of 100,000 high-definition video sequences, totaling 1,100 hours of footage. Each individual video is recorded at a resolution of 1280×720 with a frame rate of 30 fps and a duration of approximately 40 seconds. To strike an optimal balance between annotation cost and computational efficiency, the dataset extracts the keyframe at the 10th second of each sequence for multi-dimensional, fine-grained annotation. Its rich array of meteorological variants (e.g., sunny, overcast, and rainy) and all-weather sampling characteristics provide a reliable benchmark for evaluating the generalization performance of models in dynamic and complex environments.

Table 1. Statistical Distribution of the BDD100K Dataset Subset.

S. No.	Dimension	Category/ Time Interval	Quantity & Ratio	Scenario Characteristics

1	Illumination	Daytime	35,644 (60.6%)	High visibility, natural lighting
		Nighttime	23,195 (39.4%)	Complex artificial light, low visibility
2	Temporal	Peak Hours (06:00-09:00, 17:00-20:00)	–	Saturated traffic flow, high density
		Off-peak (11:00-14:00)	–	Stable and typical traffic flow
3	Night Conditions	Urban Artificial Lighting	14,381 (62% of Night)	Dominant street lighting, uniform
		Unlit Highway	5,335 (23% of Night)	Vehicle-light dependent, low ambient
		Mixed Lighting	3,479 (15% of Night)	Strong glare, interlocking shadows
4	Complex Scenes	Extreme Weather (Foggy, Rainy)	–	Mainstream adverse sensing conditions
		Long-tail Roads (Work Zone, Ramp)	–	Edge cases with complex topology
Total	Keyframes	Full Scenario Coverage	58,839 (100%)	Multi-dimensional high-fidelity set

Regarding experimental sample construction, this study utilizes a refined subset of 58,839 high-resolution keyframe images, officially curated from the 2024 Belt and Road and BRICS Skills Development and Technology Innovation Competition (Artificial Intelligence Algorithm Design and Application track)[1]. Detailed specifications are presented in Table 1. The subset comprises 35,644 daytime samples and 23,195 nighttime samples, exhibiting significant representational advantages across both temporal and spatial dimensions. Temporally, it precisely covers typical traffic peak periods, including morning and evening rushes (06:00-09:00, 17:00-20:00) and the midday off-peak period (11:00-14:00). In terms of illumination, it specifically addresses critical perception challenges by incorporating three typical nighttime conditions: urban artificial lighting (62%), unlit highways (23%), and mixed lighting (15%). Furthermore, through a stratified sampling strategy, the dataset deeply integrates extreme meteorological conditions, such as fog and rain, while accounting for "long-tail" scenarios including urban expressways, construction zones, and ramps. This high-fidelity, multi-dimensional sample distribution establishes a robust empirical foundation for validating the environmental adaptability and detection robustness of the proposed algorithm.

[1] Official website: http://aicontest.occupationedu.com

4.2. Evaluation Metrics

The weighted utility evaluation score, based on the absolute deviation of counting and officially formulated by the 2024 Belt and Road and BRICS Skills Development and Technology Innovation Competition (Artificial Intelligence Algorithm Design and Application track), is employed as the evaluation metric for the model. Let d denote the absolute deviation between the predicted count and the ground truth for a specific category in a single image, defined as $d = |y_{pred} - y_{gt}|$. For the two target categories—pedestrians and vehicles, nonlinear score decay functions, denoted as $f_{p(d)}$ and $f_{v(d)}$, are respectively defined. Given that pedestrian features are fine-grained and highly susceptible to occlusion, the corresponding scoring function exhibits higher sensitivity to counting deviations. Conversely, the vehicle scoring function is designed to provide a degree of error tolerance for moderate deviations while ensuring overall precision. The specific piecewise scoring criteria are defined as follows:

$$f_p(d)=\begin{cases}0.100, & d=0\\ 0.090, & d=1\\ 0.075, & d=2\\ 0.055, & d=3\\ 0.030, & 3<d\le 5\\ 0, & d>5\end{cases}, \quad f_v(d)=\begin{cases}0.100, & d=0\\ 0.095, & d=1\\ 0.090, & d=2\\ 0.085, & d=3\\ 0.055, & 3<d\le 5\\ 0.030, & 5<d\le 7\\ 0, & d>7\end{cases} \quad (4)$$

The comprehensive score for a single image, denoted as Si, is calculated using a multi-task weighted summation approach. Based on the relative importance and perception difficulty of targets within traffic scenarios, the weights are assigned as $w_p=0.4$ for pedestrians and $w_v=0.6$ for vehicles. The calculation formula is defined as follows:

$$S_i=w_p \cdot f_p(d_{p,i})+w_v \cdot f_v(d_{v,i}) \quad (5)$$

In Eq.(5), $d_{p,i}$ and $d_{v,i}$ denote the counting deviations for pedestrians and vehicles in the i-th image, respectively. Finally, the overall accuracy (OA) metric is obtained by aggregating the scores across all N images within the test set. The calculation is formulated as follows:

$$OA=\sum_{i=1}^{N} S_i \quad (6)$$

By quantifying perception performance across varied error-tolerance ranges, this metric effectively evaluates the detection robustness of the algorithm within real-world long-tail scenarios.

4.3. Experimental Setup

An evaluation of perception performance was conducted for object detection algorithms within road scenarios. The experimental platform was established on the Windows 11 (64-bit) operating system, with a hardware configuration centered on an Intel Core i5-12500H processor and an NVIDIA GeForce RTX 3050 Laptop GPU. On the software side, the experiments were implemented using the Python 3.10 programming language and the PyTorch 2.0.0 deep learning framework, integrated with CUDA 11.8 and CUDNN 8.7.0 acceleration libraries to optimize computational efficiency. During the data preprocessing stage, the BDD100K keyframes and their associated annotations underwent rigorous verification and normalization. To validate the generalization performance of the algorithm, a random sampling strategy was employed to partition the dataset and its corresponding label files into training, validation, and test sets

according to an 8:1:1 ratio, ensuring a balanced distribution across feature dimensions such as meteorology, time periods, and road topologies.

5. EXPERIMENTAL RESULTS

Figure 3 illustrates the loss function curves for the improved YOLOv5 model during the training and validation phases. As demonstrated by the curves in Figure 3, the improved model exhibits favorable convergence characteristics in both stages. With the increase in training epochs, the train/box_loss, train/obj_loss, and train/cls_loss components show a rapid initial decline followed by a gradual stabilization. This trend indicates that the model efficiently captures key target features during the early stages of training before entering a stable optimization phase where parameters progressively approach their optimal solutions. Simultaneously, the corresponding validation curves (val/box_loss, val/obj_loss, and val/cls_loss) remain highly consistent with the training trends, showing no significant oscillations or divergence. This consistency suggests the absence of pronounced overfitting and confirms the model's robust generalization capability. A detailed analysis of individual loss metrics reveals that the continuous decrease in box_loss signifies ongoing improvements in localization precision, while the steady convergence of obj_loss and cls_loss reflects enhanced discriminative ability regarding target existence and category classification, respectively. Throughout the training process, the smooth and stable convergence of all loss curves validates the effectiveness of the proposed mechanisms in enhancing feature representation and information filtering.

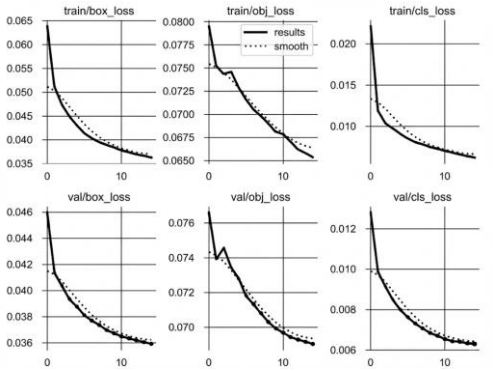

Figure 3. Convergence curves of the loss functions for the improved YOLOv5 architecture.

A comparative analysis of the detection performance between the baseline and the improved YOLOv5 models was further conducted. Experimental results indicate that the baseline YOLOv5 model achieves an overall score of 87.5, whereas the improved model reaches 93.3, representing a performance gain of 6.63%. This improvement underscores the efficacy of the proposed strategies in enhancing detection precision. Further analysis reveals that environmental factors significantly influence model performance. Under low-light conditions, the overall image contrast decreases, which restricts target feature representation and leads to increased false positives and missed detections. In scenarios with high target density, occlusion issues are exacerbated, diminishing the model's ability to recognize partially occluded targets and thereby negatively impacting the recall rate. Furthermore, the detection accuracy for small-scale targets is constrained by insufficient effective feature information. Although the baseline YOLOv5 model demonstrates a certain level of stability with its 87.5 score, it still exhibits missed detections in dense pedestrian scenes during morning and evening peaks. This limitation is attributed to complex occlusion patterns and inadequate data augmentation strategies during training, which constrain the model's generalization capability. In contrast, the improved YOLOv5 model shows simultaneous enhancements in both mean precision and mean recall, with the comprehensive

score rising to 93.3, signifying a substantial boost in overall detection performance. Notably, in scenes characterized by complex illumination and high target density, both false positive and false negative rates are reduced. This suggests that the incorporated mechanisms effectively strengthen feature representation and target discrimination, thereby enhancing the model's robustness and adaptability in complex environments. In summary, the proposed improvements demonstrate significant practical value for real-world pedestrian and vehicle detection and counting tasks.

6. CONCLUSION AND FUTURE WORK

This study addresses the challenges of pedestrian and vehicle detection and counting within urban autonomous driving scenarios by proposing an improved algorithm that integrates YOLOv5 with the SE attention mechanism. To tackle issues such as severe target occlusion, complex background interference, and insufficient feature representation in intricate traffic environments, an attention module is embedded into the YOLOv5 backbone. This integration enables adaptive feature recalibration across channels, thereby strengthening the model's feature extraction capabilities for critical target regions, enhancing the representation of effective information, and suppressing irrelevant background noise. Experimental results demonstrate that the improved model achieves significant gains in precision and counting accuracy compared to the baseline, with overall performance markedly surpassing traditional detection methods. These findings validate the effectiveness and stability of the proposed approach in complex urban road environments. Furthermore, the research indicates that the organic fusion of attention mechanisms with object detection networks substantially bolsters model adaptability and robustness in real-world traffic scenarios, providing a valuable reference for related technological advancements in the field of autonomous driving.

Future research can be further extended across the following dimensions. First, regarding structural optimization, more efficient attention mechanisms and refined network architectures could be introduced to simultaneously enhance detection precision and computational efficiency. Second, the scope of experimental scenarios should be expanded to include systematic validation under extreme weather, complex lighting, and ultra-high-density traffic conditions, thereby providing a comprehensive assessment of the model's generalization capability and environmental adaptability. Third, multi-modal data fusion strategies[23] could be explored by integrating visual information with sensor data such as LiDAR to bolster the integrity and reliability of object perception. Furthermore, aligned with the requirements of Intelligent Transportation Systems, applied research in areas such as traffic flow monitoring and anomalous event early-warning systems can be conducted, providing robust technical support for smart transportation and urban intelligent infrastructure development.

ACKNOWLEDGEMENTS

This work is supported in part by Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No.23KJD580001)

REFERENCES

[1] Wang, Z., Zhan, J., Duan, C., Guan, X., Lu, P., & Yang, K. (2022). A review of vehicle detection techniques for intelligent vehicles. IEEE Transactions on Neural Networks and Learning Systems, 34(8), 3811-3831.

[2] Kim, J. H., Kim, N., Park, Y. W., & Won, C. S. (2022). Object detection and classification based on YOLO-V5 with improved maritime dataset. Journal of Marine Science and Engineering, 10(3), 377.

[3] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).
[4] Xu, J. H., Li, J. P., Zhou, Z. R., Lv, Q., & Luo, J. (2024, December). A survey of the yolo series of object detection algorithms. In 2024 21st International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) (pp. 1-6). IEEE.
[5] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.
[6] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.
[7] Kamkar, S., & Safabakhsh, R. (2016). Vehicle detection, counting and classification in various conditions. IET Intelligent Transport Systems, 10(6), 406-413.
[8] Pang, Y., Yuan, Y., Li, X., & Pan, J. (2011). Efficient HOG human detection. Signal processing, 91(4), 773-781.
[9] Wu, J., Cui, Z., Sheng, V. S., Zhao, P., Su, D., & Gong, S. (2013). A Comparative Study of SIFT and its Variants. Measurement science review, 13(3), 122-131.
[10] Chandra, M. A., & Bedi, S. S. (2021). Survey on SVM and their application in image classification. International Journal of Information Technology, 13(5), 1-11.
[11] Yuan, Z., Zhu, M., Qian, H., Lv, T., & Huang, T. (2023, December). Multi-class failure prediction for distributed systems based on kpi data and feature selection. In 2023 6th International Conference on Software Engineering and Computer Science (CSECS) (pp. 01-06). IEEE.
[12] Yuan, Z., Wang, Z., Wu, E., Huang, T., & Chen, Y. (2024, September). Empirical Studies on Failure Prediction for Distributed Systems Based on Feature Selection. In Proceeding of the 2024 5th Asia Service Sciences and Software Engineering Conference (pp. 43-52).
[13] Kipruto, J., Pranggono, B., & Shukla, R. (2026). Scalable Hybrid Online Machine Learning for Real-Time Intrusion Detection in Electric Vehicle Charging Systems. IEEE Access.
[14] Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. Artificial Intelligence Review, 57(4), 99.
[15] Sumit, S. B., Joshi, S., & Rana, U. (2024). Comprehensive review of R-CNN and its variant architectures. Int. Res. J. Adv. Eng. Hub IRJAEH, 2(04), 959-966.
[16] Chen, Z., Guo, H., Yang, J., Jiao, H., Feng, Z., Chen, L., & Gao, T. (2022). Fast vehicle detection algorithm in traffic scene based on improved SSD. Measurement, 201, 111655
[17] Ma, C., Fu, Y., Wang, D., Guo, R., Zhao, X., & Fang, J. (2023). YOLO-UAV: Object detection method of unmanned aerial vehicle imagery based on efficient multi-scale feature fusion. IEEE Access, 11, 126857-126878.
[18] Zhu, L., Lee, F., Cai, J., Yu, H., & Chen, Q. (2022). An improved feature pyramid network for object detection. Neurocomputing, 483, 127-139.
[19] Iftikhar, S., Zhang, Z., Asim, M., Muthanna, A., Koucheryavy, A., & Abd El-Latif, A. A. (2022). Deep learning-based pedestrian detection in autonomous vehicles: Substantial issues and challenges. Electronics, 11(21), 3551.
[20] Wang, J., Zhang, Z., Dai, B., Zhao, K., Shen, W., Yin, Y., & Li, Y. (2024). Cow-YOLO: Automatic cow mounting detection based on non-local CSPDarknet53 and multiscale neck. International Journal of Agricultural and Biological Engineering, 17(3), 193-202.
[21] Symeonidis, C., Mademlis, I., Pitas, I., & Nikolaidis, N. (2023). Neural attention-driven non-maximum suppression for person detection. IEEE transactions on image processing, 32, 2454-2467.
[22] Höhne, M. O., Menke, M., & Bieshaar, M. (2025, June). Enhancing Data Efficiency for Training Object Detectors. In 2025 IEEE Intelligent Vehicles Symposium (IV) (pp. 285-292). IEEE.
[23] Liu, W., Fan, S., & Weng, G. (2025). Multi-modal deep learning framework for early Alzheimer's disease detection using MRI neuroimaging and clinical data fusion. Annals of Applied Sciences, 6(1).