Industry Science and Engineering Vol. 3 No. 9, 2026

An Intelligent Traffic Flow Monitoring Method Based on Deep Learning and Multi-Indicator Fusion

Yongyi Lin[1,*], Xiaoxu Wei[2]
[1]*School of Intelligent Media Engineering, Communication University of China Nanjing, Nanjing, China*
[2]*Department of Information and Intelligent Engineering, Shanghai Publishing and Printing College, Shanghai, China*
**Corresponding Author*

Abstract: In intelligent traffic video analysis, there is a semantic gap between low-level vehicle detection and tracking outputs and the macro-level traffic state. Besides, congestion assessment based on a single metric often fails to identify latent congestion. To address these problems, this paper formulates intersection traffic state assessment as a mapping from vehicle trajectories to discrete congestion levels and proposes a three-tier progressive monitoring method following the perception, measurement and analysis workflow. The perception layer combines a YOLO detector with a tracking-by-detection strategy to generate vehicle trajectories. The measurement layer performs deduplicated vehicle counting and computes three complementary indicators, namely vehicle density, speed loss ratio, and road space occupancy ratio. The analysis layer fuses these indicators into an interpretable weighted congestion index to determine the congestion level. Experiments on a self-built intersection dataset show that the proposed method achieves a congestion classification accuracy of 93.6% and a macro-averaged F1 score of 0.921, improving accuracy by 12.8 percentage points over the density-only baseline. Ablation experiments further confirm that the speed loss ratio and the road space occupancy ratio are key indicators for identifying latent congestion. The method supports multi-channel video monitoring on general-purpose computers without dedicated GPUs, achieving a good balance between classification accuracy and ease of engineering deployment.

Keywords: Deep Learning; Object Detection; Multi-Object Tracking; Multi-Indicator Fusion; Traffic Congestion Detection

1. Introduction

With continuing urbanization and the rapid growth of motor vehicle ownership, urban traffic pressure has been rising steadily. Traffic congestion has become a major factor limiting the efficiency of urban operations. Timely and accurate perception of traffic flow and congestion is a prerequisite for signal timing optimization and emergency management. It is also a fundamental component of Intelligent Transportation Systems (ITS) [1]. Among the various means of collecting traffic information, video surveillance stands out owing to its low deployment cost, wide coverage of the road network, and intuitive visual information. However, conventional video surveillance is mainly used for manual inspection and post-incident review. It cannot automatically interpret the semantic content of the footage and therefore cannot generate structured parameters such as traffic volume and congestion level in real time. In recent years, advances in deep learning-based object detection and multi-object tracking have laid a solid technical foundation for the automated analysis of traffic video [2]. Nevertheless, video analysis primarily generates instantaneous, microscale information such as vehicle bounding boxes and motion trajectories, while traffic management focuses on macroscale, aggregated traffic states, such as the congestion status of a road segment. A notable semantic gap therefore arises. How to robustly map detection and tracking results to traffic states consistent with human perception remains an open problem that requires in-depth study.

Existing congestion detection studies mainly rely on a single indicator, such as vehicle count, lane occupancy time, or average speed, to assess road conditions [3]. Although such methods are easy to implement, they are prone to misjudgment in complex traffic scenarios. Traffic congestion

results from the interplay of multiple factors, including the degree of vehicle aggregation, traffic efficiency, and spatial occupancy. A single indicator can hardly capture its underlying mechanism [4]. Consequently, fusing multidimensional traffic indicators extracted from detection and tracking outputs to construct an interpretable and robust congestion measurement model is critical for improving the accuracy of traffic state assessment. Moreover, city-scale management platforms demand considerable computational and network resources. They poorly adapt to small and medium scenarios, where low-cost, customizable, easily deployable monitoring solutions are scarce. To tackle these challenges, this paper formulates traffic state assessment as a mapping from vehicle spatiotemporal trajectories to discrete congestion levels. It proposes a three-stage progressive framework of perception, measurement, and analysis for intelligent traffic flow monitoring and conducts quantitative evaluations on a custom-built dataset.

2. Related Work

2.1 Vehicle Object Detection

Vehicle detection aims to locate vehicles in video frames and output their categories and confidence scores. It is the fundamental module of traffic video analysis. Early vehicle detection methods can be divided into two categories. The first extracts moving foreground regions via background modeling or frame differencing. The second adopts hand-crafted features including Histogram of Oriented Gradients (HOG) and Haar-like features, which are combined with classifiers to realize vehicle recognition. While easy to implement, these methods are vulnerable to illumination variations, occlusions and cluttered backgrounds, leading to limited generalization performance.

With the development of deep learning, object detectors based on convolutional neural networks have become mainstream and can be broadly divided into two-stage and single-stage approaches [2]. Two-stage detectors, represented by R-CNN [5], first generate region proposals for subsequent classification and bounding box regression. Such methods attain high detection accuracy yet suffer from low inference speed. Single-stage detectors, such as YOLO [6] and SSD [7], formulate detection as an end-to-end regression problem. They substantially boost inference speed while retaining competitive accuracy. YOLO achieves a favorable accuracy–speed trade-off by adopting backbone networks of different scales, making it suitable for deployment on resource-constrained devices. [8]. Since the YOLO family offers models with diverse scales for selection according to computational resources, it is well suited for deployment on resource-constrained devices. Therefore, this paper adopts YOLO as the vehicle detector for the perception layer.

2.2 Multi-Object Tracking

Multi-object tracking (MOT) establishes cross-frame identity associations for multiple targets and outputs stable trajectories, which is essential for deduplicated vehicle counting and speed estimation [2]. According to the coupling mode between trackers and detectors, existing MOT methods fall into two categories: tracking-by-detection and joint detection-and-tracking approaches. The former takes detection results as input and implements cross-frame matching via motion prediction and data association. Representative examples include SORT [9], DeepSORT [10], and ByteTrack [11]. SORT combines the Kalman filter with the Hungarian algorithm for association and features low computational overhead. DeepSORT introduces appearance features to improve association stability amid occlusions and crossing targets. ByteTrack exploits low-confidence detection boxes to increase recall, balancing real-time performance and trajectory completeness. The latter, exemplified by FairMOT [12], jointly learns detection and re-identification features within a single network. Recent tracking-by-detection variants, such as StrongSORT [13] and OC-SORT [14], further improve robustness via refined appearance modeling and observation-centric motion compensation. Owing to their clear structure, favorable adaptability and ease of engineering deployment, this work adopts the detect-first, track-later paradigm for the perception layer.

2.3 Traffic Condition Assessment

Classical traffic flow theory characterizes the macroscopic features of traffic flow using three fundamental parameters: flow, density, and speed. The linear speed–density relationship proposed by Greenshields provides a theoretical basis for congestion measurement. Early methods usually identified congestion based on a single indicator,

such as vehicle count or travel time, which is simple but prone to bias. In recent years, integrating multidimensional indicators and embedding traffic flow theory into data-driven models for congestion assessment has attracted increasing research attention [1]. However, most existing studies validate their models only on specific datasets and lack systematic analysis of the relative contribution of each indicator. Unlike these works, this paper selects three complementary indicators grounded in traffic flow theory: vehicle density, speed loss ratio, and road space occupancy ratio. It constructs an interpretable weighted fusion model and quantifies the contribution of each indicator via ablation experiments, achieving both interpretability and classification accuracy.

3. Proposed Method

3.1 General Framework

The proposed method consists of three layers: perception, measurement, and analysis. The perception layer performs vehicle detection and multi-object tracking on video frames to generate vehicle trajectories. The measurement layer performs deduplicated traffic counting and extracts congestion indicators from the obtained trajectories. The analysis layer fuses the indicators through weighted aggregation to classify the traffic state into different congestion levels. The system follows the principle of high cohesion and low coupling. The three layers communicate via standardized data structures (detection boxes, trajectories, and indicator vectors). This design not only facilitates modular development and component replacement but also provides clear boundaries for layer-wise performance analysis.

3.2 Perception Layer

The perception layer employs the YOLO detector to obtain vehicle detection sets. A confidence threshold s_{thr} is adopted to balance detection rate and false positive rate. Only detections with confidence scores no less than s_{thr} are preserved. To comply with traffic statistical standards, raw detection classes are semantically merged following predefined mapping rules: cars, buses and trucks are grouped into the motor vehicle category; motorcycles form an independent category; pedestrians, non-motor vehicles and other objects are regarded as non-statistical targets. The system supports two display modes: vehicles only and multiple categories. These modes are designed to satisfy diverse operational requirements. Nevertheless, traffic flow statistics are limited exclusively to motor vehicles, which guarantees consistent statistical criteria at the source. The detector scale and frame-skipping strategy jointly govern the accuracy–efficiency trade-off of the perception layer. For stable multi-channel monitoring on general-purpose computers without dedicated GPUs, the system samples video frames at a preset frame-skipping interval. This design reduces the per-unit-time inference load while exerting minimal influence on traffic statistics.

The tracker first predicts the position of each existing trajectory in the current frame using a motion model It then build a cost matrix based on the Intersection over Union (IoU) and appearance similarity between the predicted boxes and the detected boxes. The Hungarian algorithm is used to obtain the optimal assignment. Detections belonging to the same vehicle are linked across frames to form continuous spatiotemporal trajectories, each assigned a globally unique and stable track ID. Stable track IDs underpin subsequent unique vehicle counting and speed estimation.

3.3 Measurement and Analysis Layers

The measurement layer performs deduplicated traffic counting based on track IDs and supports two modes: line-crossing counting and region counting. In the line-crossing mode, the counting line is defined by endpoints P_1 and P_2. Let $p_{t\text{-}1}$ and p_t denote the center points of vehicle k in frames t-1 and t, respectively. The signed side function of the vehicle center relative to the counting line is defined as:

$$f(p) = \text{sign}((P_2 - P_1)(p - P_1)) \quad (1)$$

In Eq. (1), × denotes the cross product of two-dimensional vectors. When $f(p_{t\text{-}1})$ and $f(p_t)$ have opposite signs, the vehicle center crosses from one side of the counting line to the other. This event is recorded as a line-crossing occurrence, and the cumulative count for the corresponding vehicle category increases by one. In the region counting mode, a count is triggered once the vehicle center enters the counting region from outside. Both modes rely on track IDs to realize deduplicated counting. Accordingly, each vehicle is counted only once per passage, which guarantees reliable traffic statistics.

This paper leverages the three dimensions of traffic flow theory, namely density, speed, and

occupancy, to mitigate the bias arising from single-indicator congestion identification. Three complementary indicators computable from vehicle trajectories are selected for building the congestion measurement model. All indicators are normalized to the range [0, 1]. Larger values correspond to more severe congestion. Their definitions and theoretical rationale are presented as follows.

(1) Vehicle density index D_{ratio}. This index characterizes the degree of vehicle aggregation within the statistical region. It is defined as the ratio of the number of vehicles n in the region to the reference capacity N_{ref}. It reflects traffic conditions from the density perspective:

$$D_{ratio} = \min\left(\frac{n}{N_{ref}}, 1\right) \quad (2)$$

In Eq. (2), n denotes the number of motor vehicles in the statistical region in the current frame. N_{ref} denotes the reference vehicle capacity of the region. The truncation function min (·, 1) normalizes the index to [0, 1]. And a larger value indicates a higher degree of vehicle aggregation.

(2) Speed loss ratio S_{ratio}. This index characterizes the degradation of travel efficiency. It is calculated using the average vehicle speed within the statistical region and the free-flow speed. It reflects traffic conditions from the speed perspective:

$$S_{ratio} = 1 - \min\left(\frac{\overline{v}}{v_{free}}, 1\right) \quad (3)$$

In Eq. (3), $\overline{v}$ denotes the average speed of vehicles in the statistical region, and v_{free} denotes the free-flow speed of the intersection. This index quantifies congestion based on the reduction of actual speed relative to free-flow speed. The min (·, 1) operation truncates the output to the range [0, 1]. A larger value indicates more severe congestion.

(3) Road space occupancy ratio O_{ratio}. This index characterizes the degree of road space occupied by vehicles. It is defined as the ratio of the total area of all vehicle bounding boxes inside the statistical region to the region area A_{region}. It reflects traffic conditions from the spatial occupancy perspective and serves as a key indicator for identifying latent congestion such as low-density queues:

$$O_{ratio} = \min\left(\frac{\sum_i area(A_i)}{A_{region}}, 1\right) \quad (4)$$

In Eq. (4), $area(A_i)$ denotes the area of the bounding box of the i-th vehicle, and A_{region} denotes the area of the statistical region. The metric's value is constrained to the range [0, 1]. A larger value indicates a higher degree of road space occupancy.

These three indicators describe congestion from complementary perspectives. The density index represents the quantity of vehicles within the region. The speed loss ratio reflects traffic efficiency. The road space occupancy ratio characterizes road space utilization. The composite congestion index C is obtained by linearly weighting the three indicators:

$$C = w_1 D_{ratio} + w_2 S_{ratio} + w_3 O_{ratio} \quad (5)$$

In Eq. (5), the weights w_1, w_2, and w_3 denote the relative importance of the indicators at the target intersection. They can be set manually according to prior knowledge of the intersection or determined by constrained least-squares fitting on annotated data. Two thresholds τ_1 and τ_2 ($0 < \tau_1 < \tau_2 < 1$) are then used to map the composite index to three congestion levels: free-flowing ($C < \tau_1$), slow-moving ($\tau_1 \le C < \tau_2$), and congested ($C \ge \tau_2$).

4. Experiments

To quantitatively evaluate the proposed method, we constructed an intersection video dataset. The data were collected from surveillance videos of multiple intersections covering different time periods and traffic conditions. After frame extraction and manual screening, two annotators assigned labels to the samples across three levels: free-flowing, slow-moving, and congested. Annotation inconsistencies were reconciled through discussion. This process yielded 1,500 class-balanced samples. The three-class congestion classification task is evaluated in terms of Accuracy, Precision, Recall, and F1 score.

To verify the effectiveness of multi-indicator fusion, the proposed model is compared with three single-indicator baselines and one two-indicator combination. For each baseline, the optimal thresholds are selected on the same dataset for level classification. The results are shown in Table 1.

Table 1. Baseline Comparison

Methods	Accuracy	Precision	Recall	F1
D_{ratio}	80.8%	79.5%	78.6%	0.790
S_{ratio}	82.3%	81.0%	80.4%	0.807
O_{ratio}	81.5%	80.2%	79.1%	0.796
D_{ratio} + S_{ratio}	89.7%	88.9%	88.1%	0.885
Proposed Method	93.6%	92.8%	91.5%	0.921

As shown in Table 1, the three single-indicator baselines achieve similar yet limited accuracies between 80.8% and 82.3%, confirming that a

single indicator cannot fully characterize traffic congestion. Introducing a second indicator (D_{ratio} + S_{ratio}) markedly improves the accuracy to 89.7%. The proposed three-indicator fusion model achieves the best performance, with an accuracy of 93.6% and a F1 score of 0.921. Its accuracy is 11.3 percentage points higher than that of the best single-indicator baseline (S_{ratio}) and 12.8 percentage points higher than that of the density-only baseline (D_{ratio}). Results indicate that fusing the three complementary dimensions characterizes the traffic more comprehensively.

To further analyze the contribution of each indicator, we take the complete model as the reference, remove one indicator at a time, re-normalize and refit the weights of the remaining indicators, and observe the resulting performance changes. The results are shown in Table 2.

Table 2. Ablation Experiment

Configuration	Accuracy	F1	Changes
Complete Model	93.6	0.921	—
Remove O_{ratio}	89.7	0.885	−0.036
Remove S_{ratio}	87.4	0.861	−0.060
Remove D_{ratio}	90.2	0.892	−0.029

The ablation results show that removing any single indicator degrades model performance, indicating that all three indicators make positive contributions. Removing the speed loss ratio causes the largest drop in F1 score (−0.060), suggesting that speed is the key indicator for characterizing traffic efficiency and distinguishing slow-moving traffic from congestion. Removing the road space occupancy ratio leads to the second-largest drop (−0.036). Case analysis shows that without this indicator, latent congestion samples such as low-density queues are mainly misclassified, which confirms its critical role in identifying latent congestion. Removing the density index has the smallest impact, since its information can be partially compensated by the occupancy ratio. These results provide data-driven support for the design of complementary multi-indicator fusion.

In addition to the above quantitative evaluation, functional tests were conducted under both normal and abnormal conditions, covering user login, multi-channel monitoring, vehicle detection and counting, congestion detection, counting line configuration, and access control. In long-term continuous operation tests, the system exhibited no crashes, freezes, or memory leaks. When a single video stream failed, the fault was isolated without affecting the monitoring of other intersections. Monitoring data were correctly persisted and retrieved.

5. Conclusions

To address the semantic gap between traffic video analysis results and macro-level traffic states, as well as the misclassification of latent congestion by single-indicator methods, this paper formulates intersection traffic state assessment as a mapping from trajectories to congestion levels and constructs a three-tier progressive perception-measurement-analysis monitoring framework. The perception layer combines YOLO detection with a tracking-by-detection strategy to generate vehicle trajectories. The measurement layer performs deduplicated vehicle counting based on the trajectories and extracts three indicators, namely vehicle density, speed loss ratio, and road space occupancy ratio. The analysis layer fuses these indicators through an interpretable weighted model to output three-level congestion states. Experiments conducted on the self-built dataset show that the proposed method achieves a congestion classification accuracy of 93.6% and an F1-score of 0.921. It outperforms all single-indicator baselines. Ablation experiments verify the effectiveness of the speed loss ratio and road space occupancy ratio for identifying latent congestion. The system supports real-time multi-channel monitoring on ordinary computers without dedicated GPUs, balancing accuracy and deployment cost.

The proposed method still has several limitations. The weights and thresholds depend on prior knowledge or small-sample fitting and require recalibration for deployment at other intersections. Besides, the dataset does not adequately cover adverse conditions including nighttime, rain, and fog. In future work, we will optimize the tracking module to improve trajectory quality under occlusion, investigate adaptive learning of weights and thresholds to enhance generalization, and incorporate temporal modeling to enable short-term traffic forecasting. This development transforms the system into an integrated monitoring-and-forecasting platform that supports proactive traffic management.

References

[1] Amenah S. M. Thabit, Chaker Abdelaziz Kerrache, Carlos T. Calafate. A survey on monitoring and management techniques for road traffic congestion in vehicular networks. ICT Express, 2024, 10(6): 1186-1198.

[2] Zhengxia Zou, Keyan Chen, Zhenwei Shi, et al. Object detection in 20 years: A survey. Proceedings of the IEEE, 2023, 111(3): 257-276.
[3] Jahongir Azimjonov, Ahmet Özmen. A real-time vehicle detection and a novel vehicle tracking systems for estimating and monitoring traffic flow on highways. Advanced Engineering Informatics, 2021, 50: 101393.
[4] Shankha N. Chattopadhyay, Arvind K. Gupta. Do multivariate measures anticipate traffic congestion? Royal Society Open Science, 2026, 13(5): 252185.
[5] Ross Girshick, Jeff Donahue, Trevor Darrell, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, USA, 2014: 580-587.
[6] Joseph Redmon, Santosh Divvala, Ross Girshick, et al. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016: 779-788.
[7] Wei Liu, Dragomir Anguelov, Dumitru Erhan, et al. SSD: Single shot multibox detector. Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, Netherlands, 2016: 21-37.
[8] Chien-Yao Wang, I-Hau Yeh, Hong-Yuan Mark Liao. YOLOv9: Learning what you want to learn using programmable gradient information. Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 2024: 1-21.
[9] Alex Bewley, Zongyuan Ge, Lionel Ott, et al. Simple online and realtime tracking. Proceedings of the IEEE International Conference on Image Processing (ICIP), Phoenix, USA, 2016: 3464-3468.
[10] Nicolai Wojke, Alex Bewley, Dietrich Paulus. Simple online and realtime tracking with a deep association metric. Proceedings of the IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017: 3645-3649.
[11] Yifu Zhang, Peize Sun, Yi Jiang, et al. ByteTrack: Multi-object tracking by associating every detection box. Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 2022: 1-21.
[12] Yifu Zhang, Chunyu Wang, Xinggang Wang, et al. FairMOT: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision, 2021, 129(11): 3069-3087.
[13] Yunhao Du, Zhicheng Zhao, Yang Song, et al. StrongSORT: Make DeepSORT great again. IEEE Transactions on Multimedia, 2023, 25: 8725-8737.
[14] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, et al. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, 2023: 9686-9696.