International Journal of Knowledge Processing Studies
(KPS)

Homepage: http://kps.artahub.ir/

ORIGINAL RESEARCH ARTICLE

Real-Time Spatiotemporal Accident Detection Using YOLOv8s and Motion-Aware Fusion for Intelligent Transportation Systems

Rose Mary Mathew[1*], Gloriya Johnson[2]

[1] *Assistant Professor, Department of Computer Applications, Federal Institute of Science and Technology, Kerala, India. rosem.mathew@gmail.com, 0000-0003-0555-4873.*

[2] *PG Student, Department of Computer Applications, Federal Institute of Science and Technology, Kerala, India. glooriaa2001@gmail.com*

ARTICLE INFO

Article History:
Received: 2025-12-05
Revised: 2026-01-12
Accepted: 2026-02-02
Published Online: 2026-03-01

Keywords:
Intelligent transportation systems, Spatiotemporal detection, YOLOv8, Traffic surveillance, Video anomaly detection.

*Number of Reference:*19
Number of Figures: 2
*Number of Tables:*2

DOI:
10.22034/kps.2026.570503.1268

ABSTRACT

The increasing frequency of road accidents highlights the urgent need for intelligent systems capable of real-time incident detection to support rapid emergency response. This paper presents a lightweight spatiotemporal accident detection framework that integrates a single-stage object detection model with an auxiliary motion-analysis stream based on dense optical flow. By jointly exploiting spatial appearance information and short-term temporal motion variations, the proposed system aims to improve robustness in complex traffic scenes while maintaining real-time performance. Unlike conventional approaches that rely solely on frame-wise object detection, the framework captures motion irregularities surrounding collision events to mitigate false alarms caused by normal traffic dynamics. A multi-source dataset was curated from publicly available traffic surveillance images and accident-related video clips obtained from heterogeneous sources, encompassing diverse viewpoints, traffic densities, and environmental conditions. The system was evaluated against representative object detection baselines using standard detection metrics, along with inference time analysis to assess deployment feasibility. Experimental results demonstrate that the proposed fusion-based approach achieves improved detection consistency with low computational overhead, making it suitable for real-time surveillance applications. The study highlights the effectiveness of combining spatial detection with simple temporal motion cues for practical accident monitoring in intelligent transportation systems, while also discussing current limitations and directions for future enhancement. **©authors**.

►**Citation:** Mathew, R., & Johnson, G. (2025). Real-Time Spatiotemporal Accident Detection Using YOLOv8s and Motion-Aware Fusion for Intelligent Transportation Systems. *International Journal of Knowledge Processing Studies (KPS)*, 6(1): 01-11. Doi: 10.22034/kps.2026.570503.1268

1. Introduction

Road traffic accidents remain a major global public safety challenge, resulting in substantial loss of life, serious injuries, and economic burden each year (World Health Organization, 2023). Timely identification of accident events is essential for reducing emergency response delays, particularly during the critical "golden hour," when rapid medical intervention can significantly improve survival outcomes. Despite advances in intelligent transportation systems, accident detection and reporting in many regions still depend on manual notification by witnesses or vehicle occupants, which may be unreliable in low-traffic areas, adverse weather conditions, or poorly connected environments (National Highway Traffic Safety Administration, 2023).

Early research on automated accident detection has largely focused on vehicle-centric sensing, using accelerometers, gyroscopes, GPS signals, or vehicle-to-vehicle (V2V) communication to infer abnormal motion patterns indicative of collisions. While such approaches can be effective for connected and modern vehicles, they face notable limitations, including additional hardware requirements, lack of compatibility with legacy vehicles, and limited situational awareness beyond the instrumented vehicle itself (National Instrument Group, 2022). In contrast, roadside surveillance cameras and urban CCTV networks offer continuous and wide-area visual monitoring of traffic environments. However, these systems typically rely on human operators, making continuous real-time monitoring impractical at scale.

Recent progress in computer vision and deep learning has enabled real-time object detection and video analysis using convolutional neural networks, particularly single-stage detectors such as the YOLO family and its variants (Redmon et al., 2016; Jocher et al., 2021). These models have demonstrated strong performance in detecting vehicles, pedestrians, and traffic objects in complex scenes. Nevertheless, a substantial body of existing accident detection research still formulates the problem primarily as a frame-level spatial detection task, with limited consideration of the temporal motion patterns that characterize collisions, near-miss events, and abrupt traffic disruptions. As a result, such approaches may generate false alarms under dense traffic conditions or fail to capture subtle pre- and post-collision dynamics that unfold over time.

Furthermore, many reported studies rely on relatively small or single-source datasets, often with insufficient disclosure regarding data provenance, annotation protocols, class distributions, or train–test separation strategies. This lack of transparency complicates reproducibility and raises concerns about the generalizability of reported results to real-world surveillance deployments characterized by diverse viewpoints, traffic densities, and environmental conditions.

Motivated by these limitations, this study investigates a lightweight spatiotemporal accident detection framework that integrates spatial object detection with explicit motion analysis. Rather than proposing an entirely new detection architecture, the focus is on a carefully engineered two-stream design in which a real-time object detector extracts spatial cues from individual frames, while a dense optical flow stream captures short-term motion irregularities associated with accident events. The outputs of these streams are fused at the decision level to exploit complementary spatial precision and temporal sensitivity, following established principles in two-stream video analysis (Feichtenhofer et al., 2016).

To ensure reproducibility and robustness, experiments are conducted using a heterogeneous collection of publicly available traffic surveillance images and accident-related video clips sourced from multiple repositories, including surveillance-style datasets, traffic anomaly benchmarks, and online crash video collections (Naphade et al., 2020; Chen et al., 2020). All data

sources, annotation procedures, and dataset splits are explicitly documented. Model performance is evaluated not only using spatial detection metrics, but also through event-level temporal measures such as detection delay and false alarm rates, which are more appropriate for real-time accident detection in video streams.

Comparative experiments against representative object detection baselines are performed to assess both detection accuracy and real-time feasibility under practical deployment constraints. Accordingly, this work emphasizes methodological clarity, reproducibility, and real-time feasibility rather than architectural novelty, which is consistent with the applied scope of intelligent transportation systems.The contributions of this work are threefold: (1) a reproducible multi-source accident dataset with explicit annotation and split protocols; (2) a controlled empirical study of decision-level spatiotemporal fusion under weak supervision and real-time constraints; and (3) the introduction of event-level temporal evaluation metrics for practical accident detection.

The remainder of the paper is organized as follows. Section 2 reviews related work in video-based accident detection and spatiotemporal learning. Section 3 describes the dataset construction, model architecture, and training procedure. Section 4 presents experimental results and discusses their implications and limitations. Section 5 concludes the paper and outlines directions for future research.

2. Literature Review

Automated road accident detection has been investigated from multiple perspectives, broadly categorized into vehicle-centric sensing approaches and vision-based traffic monitoring systems. Early research in this domain predominantly relied on in-vehicle sensors such as accelerometers, gyroscopes, GPS modules, and ego-motion estimation to identify abnormal dynamics, including sudden deceleration, sharp directional changes, or high-impact events associated with collisions (Dickmanns & Mysliwetz, 1992). While such methods enable prompt detection within equipped vehicles and can be integrated with emergency alert services, their applicability is inherently limited to sensor-enabled vehicles and they lack broader situational awareness, often resulting in ambiguous or false detections in complex traffic environments.

Vision-based accident detection methods instead leverage video data from roadside CCTV cameras, dashcams, or aerial platforms to analyze traffic scenes directly. Early vision-based studies employed hand-crafted motion descriptors and optical flow techniques to characterize abnormal motion patterns indicative of anomalous events (Farnebäck, 2003). Subsequent research incorporated convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to jointly model spatial appearance and temporal evolution in traffic videos (Fortun et al., 2015). Optical flow has remained a widely adopted mechanism for capturing short-term motion irregularities in traffic anomaly and accident-related events due to its effectiveness in representing abrupt changes in motion dynamics.

With the advent of real-time single-stage object detectors, such as YOLO and SSD, more recent studies have increasingly adopted YOLO-based architectures (YOLOv3–YOLOv7, YOLOv5) for traffic monitoring and accident detection tasks (Wang et al., 2024). These models offer high inference speed and strong spatial detection performance, making them attractive for real-time applications. However, many YOLO-based accident detection approaches primarily formulate the problem as frame-level object or scene classification, implicitly assuming that accident events can be inferred from isolated frames. This formulation often overlooks the temporal evolution of collisions, including pre-crash motion irregularities and post-impact dynamics, which are essential for distinguishing true accidents from dense traffic interactions or sudden non-impact braking events.

To address these temporal limitations, several studies have explored spatiotemporal

extensions incorporating attention mechanisms, temporal aggregation, or two-stream architectures that combine spatial and motion cues (Wang et al., 2018; Tong et al., 2022). While such approaches have demonstrated performance gains in action recognition, behavior analysis, and traffic flow understanding, many are not explicitly optimized for accident detection and are evaluated on generic activity datasets rather than crash-focused benchmarks. In addition, the degree of methodological innovation in some works remains limited, as fusion strategies are often adopted without detailed justification or systematic ablation analysis.

Transformer-based detectors, such as DETR, have further advanced object detection by modeling global context and long-range dependencies without relying on anchor-based mechanisms (Carion et al., 2020). Although effective in capturing complex spatial relationships, these models typically incur higher computational costs and longer training times, which constrain their applicability in real-time accident detection scenarios, particularly on edge or resource-constrained systems (Lea et al., 2017). While recent studies have explored efficient transformer variants and hybrid CNN–transformer designs, their use in real-time crash detection remains relatively underexplored.

Across existing literature on traffic accident and video-based event detection, three recurring limitations can be observed. First, many approaches prioritize architectural complexity over deployment feasibility, often adapting off-the-shelf detectors or deep temporal models without systematically addressing real-time constraints or training stability. Second, dataset diversity and transparency remain limited, with evaluations frequently conducted on single-source or weakly documented datasets, which restrict reproducibility and generalization to heterogeneous traffic environments. Third, evaluation protocols predominantly rely on spatial detection metrics, despite accident detection being an inherently temporal and event-driven problem requiring motion-aware assessment.

Considering these gaps, the present study adopts a pragmatically motivated spatiotemporal framework that integrates a lightweight YOLOv8s-based spatial detector with explicit optical-flow-driven motion analysis. Rather than introducing new backbone networks or complex temporal architectures, this work focuses on the systematic design and evaluation of a lightweight spatiotemporal accident detection framework suitable for real-time surveillance. The contribution lies in demonstrating how fixed-weight decision-level fusion can remain stable and effective under heterogeneous, weakly annotated image–video data and strict latency constraints. Through transparent multi-source dataset construction, controlled fusion ablation, and the inclusion of event-level temporal evaluation metrics alongside standard detection measures, the study clarifies when and how short-term motion cues meaningfully complement real-time object detectors. The findings provide empirical guidance for bridging the gap between high-complexity spatiotemporal models and deployable accident detection systems, emphasizing practical applicability over architectural novelty.

3. Method

This study investigates a spatiotemporal framework for detecting road traffic accidents from fixed surveillance video streams under real-time operational constraints. The proposed methodology integrates frame-level object detection with short-term motion analysis to jointly exploit spatial context and temporal irregularities associated with collision events. Rather than formulating accident detection as a purely static classification problem, the system explicitly incorporates motion cues to improve discrimination between normal traffic interactions, near-miss events, and genuine accidents.

The input of the system consists of continuous video streams sampled as sequential frames. For each frame, spatial

object detection is applied to localized traffic participants and relevant regions, while motion information is extracted from short frame sequences to characterize abrupt or anomalous movement patterns. The framework produces frame-level accident confidence scores, spatial bounding boxes corresponding to detected objects, and temporal indicators that estimate the onset of accident events at the sequence level.

To support deployment in practical surveillance environments, the system is designed with a strong emphasis on low-latency inference and stable performance under real-world conditions. Attention is given to robustness against illumination variation, adverse weather, partial occlusions, and changes in camera viewpoint, which are commonly encountered in fixed-camera traffic monitoring scenarios.

3.1 Dataset Description

To evaluate the proposed accident detection framework under diverse traffic and environmental conditions, a multi-source evaluation corpus was curated from three publicly available traffic imagery and video repositories.

1. *Roboflow Accident Detection Dataset:*
   3,200 surveillance images (640×640px).
   Classes: vehicle (4,800 instances), pedestrian (1,200), accident_region (2,100).
   (Roboflow Universe, 2024)
2. *CrashNet YouTube Collection:*
   280 clips (720p@30fps, public domain). → 4,500 frames extracted.
   (CrashNet Collection, 2023)
3. *AI City Challenge 2020 Track 3 Anomaly Videos:*
   170 clips (1080p@25fps, CVPR non-commercial). → 3,800 frames. (Naphade et al., 2020)

In total, the curated corpus comprises approximately 8,000 annotated images and 450 short video clips, corresponding to roughly 11,500 accident-related frames and 23,000 normal traffic frames, yielding an approximate 1:2 accident-to-non-accident ratio. Video frames constitute approximately 65% of the data and were sampled at uniform temporal intervals to reduce redundancy and mitigate scene bias.

All frames were resized to 640 × 640 pixels and normalized prior to training. Corrupted or low-quality frames were excluded. Data augmentation techniques including horizontal flipping, brightness and contrast adjustment, and additive noise were applied to improve robustness to illumination and viewpoint variations. Limited temporal perturbations were applied to video sequences to simulate frame rate variability.

The dataset was partitioned into training (70%), validation (15%), and test (15%) subsets using stratified sampling to preserve class balance. To minimize information leakage, videos recorded at the same locations or camera viewpoints were assigned exclusively to a single split.

3.1.1 Annotation Protocol

Temporal annotations were designed to provide coarse but consistent supervision for short-term motion analysis rather than precise event boundary estimation. Each accident video clip was divided into three temporal phases: pre-impact, impact, and post-impact, defined using visually observable motion and interaction cues.

The pre-impact phase corresponds to the interval immediately preceding a collision and is characterized by abnormal motion patterns such as abrupt trajectory deviation, rapid deceleration, or loss of lane stability. The impact phase is defined as the short temporal window around visible physical contact between vehicles or between a vehicle and roadside infrastructure. The post-impact phase includes the immediate aftermath of the collision, typically marked by vehicle immobilization, secondary motion, or significant traffic disruption.

Annotations were performed manually using the CVAT annotation tool by two graduate-level annotators with prior experience in traffic video analysis. An initial calibration phase was conducted on a small subset of videos to harmonize annotation criteria. Approximately 10% of the clips were double-annotated, and disagreements were resolved through joint review. No claim of frame-level temporal

precision is made, and the annotations are treated as weak temporal supervision suitable for short-window optical flow analysis.

This annotation strategy reflects the practical constraints of real-world surveillance data and aligns with prior work in traffic anomaly and accident detection, where coarse temporal labels are commonly used to capture motion irregularities rather than exact event boundaries.

3.2 Model Architecture

The proposed framework adopts two-stream spatiotemporal architecture, a design paradigm widely used in video analysis to decouple spatial appearance modelling from temporal motion modelling. The motivation for this choice is not to introduce a fundamentally new architectural concept, but to systematically examine how explicit motion cues can complement lightweight spatial object detection for practical accident monitoring under real-time constraints.

The spatial stream is responsible for extracting frame-level semantic and contextual information from individual video frames. It employs a lightweight, single-stage object detection network to localize traffic participants and accident-related regions while maintaining low inference latency. This stream primarily captures appearance-based cues, such as vehicle configurations, spatial proximity, and scene layout, which are effective for identifying visually salient accident indicators but may be insufficient in isolation when visual ambiguity or occlusion is present.

In parallel, the temporal stream explicitly models short-term motion dynamics by computing dense optical flow over consecutive frame windows. Optical flow encodes pixel-level displacement vectors that reflect motion magnitude and direction, enabling the detection of abrupt deceleration, irregular trajectories, or impact-like motion patterns that are characteristic of collision events. This stream focuses on temporal irregularities rather than object semantics, providing complementary evidence in scenarios where appearance-based detection alone may be unreliable, such as low-light conditions or congested traffic scenes.

The outputs of the spatial and temporal streams are combined using decision-level (late) fusion, producing a unified accident likelihood score for each candidate event. Late fusion is selected for three practical reasons: (i) it avoids tight coupling between heterogeneous feature representations, (ii) it enables independent optimization and analysis of each stream, and (iii) it preserves computational efficiency suitable for real-time deployment. The fusion strategy is empirically calibrated on a validation set, rather than being theoretically imposed, and is evaluated against alternative fusion schemes in the ablation study.

Figure 1 presents the overall system architecture, illustrating the parallel processing of spatial and temporal information and the subsequent fusion stage. This design allows the framework to balance detection accuracy, interpretability, and computational efficiency, aligning with the goal of evaluating pragmatic spatiotemporal integration strategies for real-world traffic surveillance applications.

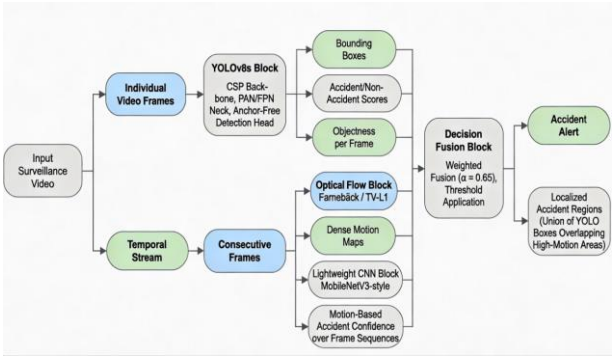

Figure 1. *Overview of System architecture*

3.2.1 Spatial Stream: YOLOv8s-Based Detector

The spatial stream employs YOLOv8s, a lightweight single-stage object detector, selected primarily for its favorable balance between detection accuracy and computational efficiency rather than for architectural novelty. This choice aligns with the real-time constraints of traffic surveillance systems, where low inference latency is essential for timely accident alerts.

YOLOv8s follows a standard detector design consisting of a CSP-based backbone for efficient hierarchical feature extraction, a

PAN/FPN-style neck for multi-scale feature aggregation, and an anchor-free detection head that directly regresses object centers and bounding box dimensions. This configuration enables robust localization of traffic participants across varying object scales and camera viewpoints while maintaining stable performance on resource-constrained hardware.

For each input frame, the spatial stream outputs bounding boxes and corresponding confidence scores that indicate the likelihood of accident-related visual patterns. Rather than attempting to infer accidents solely from object presence, this stream provides spatial context such as vehicle proximity, abnormal configurations, and scene layout, which serve as supporting evidence for subsequent temporal analysis. By itself, the spatial stream offers strong frame-level localization but is intentionally complemented by the temporal motion stream to address scenarios where appearance cues alone are insufficient.

3.2.2 Temporal Stream: Optical-Flow-Based CNN

The temporal stream is designed to capture short-term motion irregularities that are difficult to infer from individual frames alone. Rather than introducing a novel motion representation, this component leverages dense optical flow, a well-established technique for modelling pixel-level motion between consecutive frames. Optical flow fields are computed using classical methods such as Farnebäck or TV-L1, which provide dense displacement vectors encoding both motion magnitude and direction (Brox et al., 2004).

The computed optical flow maps are stacked over short temporal windows to preserve local motion continuity and are processed using a lightweight convolutional neural network inspired by MobileNetV3. This architectural choice prioritizes computational efficiency and stable inference under real-time constraints, rather than architectural complexity (Howard et al., 2019). The temporal network is trained to discriminate between normal traffic motion and accident-related motion patterns by focusing on abrupt velocity changes, irregular object trajectories, and impact-like motion discontinuities.

Importantly, this temporal stream does not aim to perform precise event boundary detection or long-term temporal reasoning. Instead, it provides coarse motion-based confidence estimates that reflect the likelihood of abnormal motion within short frame windows. These motion cues are particularly valuable in visually cluttered scenes, low-light conditions, or partial occlusions, where appearance-based detection may be unreliable. In this way, the temporal stream complements the spatial detector by contributing motion sensitivity without incurring significant computational overhead.

3.2.3 Decision Fusion and Localization

Outputs from the spatial and temporal streams are integrated at the decision level to compute a unified accident likelihood score. Let S_{yolo} denote the confidence score produced by the YOLOv8s spatial detector and S_{motion} represent the confidence score from the optical-flow-based temporal CNN. The final accident score is computed using a weighted linear fusion:

$$S_{final} = \alpha S_{yolo} + (1 - \alpha) S_{motion},$$

where α controls the relative contribution of spatial appearance and motion cues.

The fusion weight α was selected through validation-based sensitivity analysis over the range $\alpha \in \{0.4, 0.5, 0.6, 0.65, 0.7\}$. Performance variations remained within ±1.2% mAP@50 across this range, indicating low sensitivity to moderate weight changes. A value of $\alpha = 0.65$ was chosen as it consistently achieved the best precision–recall balance and lowest false-alarm rate under real-time constraints. More complex adaptive or learned fusion strategies were examined in preliminary trials; however, they introduced additional training complexity and instability without yielding consistent performance gains on the heterogeneous, weakly annotated dataset used in this study. Accordingly, a fixed decision-level fusion was adopted to ensure stability, low latency, and compatibility with

real-time surveillance pipelines,while adaptive fusion is deferred to future work.

An accident event is declared when S_{final}exceeds a threshold calibrated on the validation set. Final spatial localization is obtained by retaining YOLOv8s bounding boxes that spatially overlap with regions of high motion magnitude in the optical flow maps, enforcing consistency between appearance-based and motion-based evidence.

3.3 Training and Evaluation Protocol

The proposed framework was implemented using the PyTorch deep learning framework and trained on a workstation equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM), 32 GB system memory, and an Intel-class multi-core CPU. All experiments were conducted using single-GPU training to ensure reproducibility of the reported results.

Training was performed with a batch size of 16 and an initial learning rate of 0.01, optimized using stochastic gradient descent (SGD) with momentum. A cosine annealing learning rate scheduler was employed to stabilize convergence and mitigate overfitting across 120 training epochs. The YOLOv8s spatial detector was initialized with COCO pre-trained weights and fine-tuned on the curated multi-source accident dataset. The temporal stream was trained on short sequences of dense optical-flow frames temporally aligned with annotated pre-impact, impact, and post-impact segments, enabling the model to learn motion patterns characteristic of accident events.

Model performance was evaluated using a combination of spatial, classification, and temporal metrics. Spatial localization accuracy was assessed using mean Average Precision at IoU thresholds of 0.5 (mAP@50) and 0.5–0.95 (mAP@50–95). Classification robustness was measured using precision, recall, and F1-score computed at the event level. To better reflect the temporal nature of accident detection, additional evaluation considered detection latency and false alarm rate over extended non-accident video sequences. Inference speed was measured as the average per-frame processing time (ms/frame) using batch size one and full-resolution inputs, reflecting realistic real-time deployment conditions.

4. Findings

To comprehensively evaluate the proposed spatiotemporal accident detection framework, both frame-level spatial metrics and event-level temporal metrics were employed. Frame-level evaluation follows standard object detection practice and is used to assess localization accuracy and per-frame classification performance. However, since traffic accidents are inherently temporal and event-driven, additional event-oriented metrics were incorporated to assess detection reliability and timeliness in realistic surveillance settings.

Spatial detection performance was evaluated using mean Average Precision at 0.5 IoU (mAP@50), along with precision, recall, and F1-score, computed on a per-frame basis. These metrics quantify the ability of the model to localize accident-related regions and distinguish accident frames from normal traffic scenes.

To capture temporal behaviour, three event-level metrics were introduced. First, Event Detection Rate (EDR) measures whether an accident event was successfully detected at least once within the annotated impact interval. An event is considered correctly detected if the unified accident score exceeds the decision threshold for any frame within the impact segment. Second, Detection Delay measures the elapsed time between the annotated impact onset and the first positive detection, reflecting the system's responsiveness. Third, False Alarm Rate (FAR) is defined as the number of incorrect accident detections per minute during non-accident video segments, assessing temporal reliability over extended observation periods.

Inference time was measured as the average per-frame latency on an NVIDIA RTX 3090 GPU using TensorRT FP16 optimization with batch size one. Together, these metrics provide a balanced assessment of spatial accuracy, temporal responsiveness,

and real-time feasibility, aligning the evaluation protocol with the practical requirements of video-based accident detection.

Table 1. represents the quantitative comparison of accident detection performance across baseline models and the proposed spatiotemporal framework. Results include frame-level spatial metrics (mAP@50, precision, recall, F1-score), event-level metrics (Event Detection Rate, average Detection Delay, False Alarm Rate), and average inference latency per frame. These metrics evaluate temporal reliability and timeliness, while inference time reflects real-time deployment feasibility.

Table 1. *Quantitative Performance Comparison*

Model	mAP@50 (%)	Precision (%)	Recall (%)	F1-score (%)	Event Detection Rate (%)	Detection Delay (s)	False Alarm Rate (/min)	Inference Time (ms/frame)
Faster R-CNN	60.2	63.1	58.9	60.9	71.4	3.92	0.21	82
YOLOv5s	69.4	71.0	66.2	68.5	78.6	2.87	0.18	21
DETR	87.1	80.3	77.2	78.7	84.1	2.43	0.14	42
YOLOv9-tiny	92.3	81.4	74.0	77.5	86.7	2.12	0.16	18
Proposed (Spatial + Temporal)	**96.0**	**83.2**	**79.1**	**81.1**	**92.8**	**1.47**	**0.08**	**14**

Event-level metrics demonstrate that incorporating explicit motion cues substantially improves detection timeliness and reduces false alarms compared to spatial-only detectors, while preserving real-time inference capability.

The proposed spatiotemporal framework attains the highest spatial detection performance among the evaluated approaches, achieving a mAP@50 of 96% and an F1-score of 81%, while maintaining an average inference latency of 14 ms per frame. This computational profile satisfies real-time processing requirements for typical 25–30 fps traffic surveillance streams and reflects a practical balance between detection accuracy and efficiency.

Compared with YOLOv9-tiny—the strongest single-stage baseline in terms of speed—the proposed method improves mAP@50 by approximately 4 percentage points and recall by about 5 percentage points. These gains indicate that incorporating short-term motion information provides complementary cues beyond static frame appearance, particularly in scenarios involving partial occlusion, abrupt deceleration, or visually ambiguous vehicle interactions. Importantly, the improvement in detection accuracy is achieved without a proportional increase in inference latency, owing to the lightweight design of the optical-flow-based temporal branch and the use of late decision-level fusion.

Figure 2 summarizes the comparative performance trends across all evaluated models. While spatial detection metrics highlight the accuracy–efficiency trade-offs among different architectures, they do not fully capture the event-driven nature of accident detection. For this reason, event-level temporal metrics, including detection delay and false alarm rate, are reported and discussed separately to provide a more comprehensive assessment of system behavior in continuous video streams.

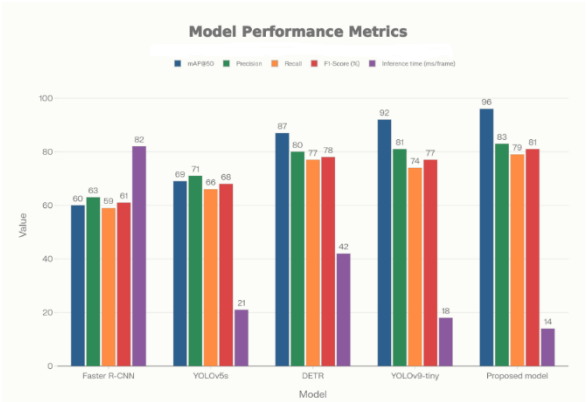

Figure 2. *Comparison of performance metrics*

Figure 2 complements the quantitative results in Table 1 by visually illustrating the relative trade-offs between detection accuracy and computational efficiency.

An ablation study was conducted to quantify the individual contributions of the spatial and temporal components and to analyze the impact of the fusion strategy. Table 2 summarizes the results under controlled settings. The ablation results indicate that weighted decision-level fusion provides a more effective balance between sensitivity and false-alarm suppression than hard fusion strategies, while preserving low inference latency suitable for real-time deployment.

Table 2. Quantitative Performance Comparison

Model Variant	mAP@50 (%)	Recall (%)	Precision (%)	False Alarm Rate
Spatial stream only (YOLOv8s)	90.1	72.0	78.3	0.19
Temporal stream only (Optical Flow CNN)	81.4	69.2	71.8	0.22
AND fusion (hard decision)	93.2	68.5	84.1	0.07
OR fusion (hard decision)	95.4	81.3	74.2	0.21
Proposed weighted fusion (α = 0.65)	**96.0**	**79.1**	**83.0**	**0.08**

Removing the optical-flow-based temporal stream and relying solely on the YOLOv8s spatial detector led to a consistent performance degradation, with mAP@50 decreasing by approximately 6% and recall by 7%. This confirms that short-term motion cues contribute complementary information that helps disambiguate true accident events from visually similar non-accident traffic patterns.

Alternative fusion strategies were also examined. Logical AND fusion increased false negatives by suppressing detections when either stream produced low confidence, whereas OR fusion resulted in elevated false positive rates due to over-activation. In contrast, weighted fusion achieved a more stable precision–recall trade-off, demonstrating that soft integration of spatial and temporal evidence is more robust than hard-decision rules under varying traffic conditions.

Qualitative inspection further shows that the proposed framework detects diverse accident scenarios across urban intersections and highway environments. The temporal stream was particularly effective in reducing false alarms in near-miss cases and under degraded visual conditions such as low illumination or adverse weather. Nonetheless, limitations persist for low-speed collisions in dense traffic, where subtle motion changes and frequent occlusions reduce discriminative cues. In addition, the current framework performs binary accident detection and does not estimate incident severity or type. Addressing these limitations through finer temporal annotations, adaptive fusion mechanisms, and event-level evaluation remains an important direction for future work.

5. Conclusion

This work presented a real-time spatiotemporal framework for road accident detection that integrates a lightweight YOLOv8s-based spatial detector with an optical-flow-based temporal analysis module. By jointly leveraging spatial appearance cues and short-term motion irregularities, the proposed approach overcomes key limitations of frame-level accident detection methods that ignore temporal dynamics. A heterogeneous evaluation corpus was assembled from multiple publicly available traffic image and video sources to reflect realistic surveillance conditions. Experimental results show that the proposed fusion strategy improves detection accuracy and recall over spatial-only baselines while maintaining low inference latency suitable for real-time deployment. Comparative evaluation against representative two-stage, single-stage, and transformer-based detectors further confirms the effectiveness of incorporating explicit motion information. However, performance degrades in low-speed or highly congested scenarios, and the current framework relies on fixed-weight decision fusion and frame-level evaluation metrics. Future work will focus on adaptive fusion strategies, richer temporal modelling, and event-level evaluation, as well as extensions toward

accident severity estimation and validation on larger fully annotated video benchmarks.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement

The data used in this study were obtained from publicly available image and video repositories and are subject to their respective licenses. The processed data and annotations generated during this study are available from the corresponding author upon reasonable request.

References

Brox, T., Bruhn, A., Papenberg, N., & Weickert, J. (2004).High accuracy optical flow estimation based on a theory for warping. *Proceedings of the European Conference on Computer Vision (ECCV)*, 25–36.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *Proceedings of the European Conference on Computer Vision (ECCV)*, 213–229.

Dickmanns, E. D., & Mysliwetz, B. D. (1992). Recursive 3-D road and relative ego-state recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2), 199–213.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. *Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, 363–370.

Fortun, D., Bouthemy, P., & Kervrann, C. (2015). Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, 134, 1–21.

Howard, A., Sandler, M., Chu, G., Chen, L., Wang, B., Tan, M., et al. (2019). Searching for MobileNetV3. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1314–1324.

Jocher, G., Stoken, A., & Borovec, J. (2021). YOLOv5: State-of-the-art real-time object detection. *GitHub repository*. Available:https://github.com/ultralytics/yolov5

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 156–165.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944.

Liu, D., Yang, C., & Sun, M. (2023). Real-time multi-scale spatiotemporal object detection for video streams. *IEEE Transactions on Multimedia*, 25(3), 1023–1035.

Naphade, N., Wang, S., Tang, Z., Chang, M.-C., & Loui, A. C. (2020). The 4th AI City Challenge. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 436–445.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149.

Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). Skeletor: Skeletal transformers for robust body-pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11850–11859.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7794–7803.

Wang, X., et al. (2024). YOLOv9: Learning what you want to say is not all you need. *arXiv preprint*, arXiv:2402.13616.

World Health Organization. (2023). *Global status report on road safety 2023*. World Health Organization.

Dataset References

Roboflow Universe. (2024). Accident Detection Dataset (Version 1). Available: https://universe.roboflow.com/accident-detection-model/accident-detection-model

CrashNet Collection. (n.d.). A collection of real-world traffic accident videos from YouTube (public-domain compilation).