Journal of Applied Automation Technologies

Research Paper

Automatic Traffic Incident Detection Algorithm for Surveillance Videos Using Spatio-Temporal Graph Neural Networks

Mariusz Jaworski[1], Cyprian Górski[1] and Ryszard Halik[1, *]

[1] Faculty of Electrical and Automatic Control Engineering, Czestochowa University of Technology, Czestochowa, 42-200, Poland
*Corresponding author: ryszard.ha@pcz.edu.pl

Abstract. An intelligent transportation system's primary goal is to automatically identify traffic accidents in surveillance footage; otherwise, collisions, unauthorized parking, unexpected traffic jams, lane crossings, and other safety and traffic issues would soon arise. The majority of current video detection techniques are either short-term motion cues or frame-level visual feature descriptions; as a result, they are unable to characterize organized interactions between vehicles, lanes, and traffic flow conditions. This research proposes a framework for automatically detecting traffic accidents using spatiotemporal graph neural networks. Following the extraction of traffic objects from surveillance footage, dynamic traffic graphs are constructed using the position, velocity, lane affiliation, motion variation, and spatial closeness of these items. After that, a Spatio-temporal Graph Learning Module is implemented to compile the incident's temporal evolution patterns and local interaction features for classification. The experiment's components are arranged as follows: robustness under occlusion and heavy traffic, detection accuracy, false alarm rate, response time, incident-type identification, and ablation comparison. In comparison to frame-level deep models, the novel approach may increase the detection accuracy to 93.84%, decrease the false alarm rate to 5.21%, and lower the average response delay to 1.37. Thus, the challenge of real-time traffic video surveillance is a good fit for a graph-based model of time-series data.

Keywords: *Traffic incident detection, Surveillance video, Spatio-temporal graph neural network, Intelligent transportation, Video analytics*

Received on 31 October 2023, Accepted on 18 April 2024, Published on 26 April 2024

Introduction

As more and more urban highways and expressways are continually watched by video, safety management, congestion control, and emergency response have steadily come to pass. Traffic accident detection is currently a relatively high-value application area of intelligent transportation systems. Rarely are traffic accidents haphazard displays of color. In terms of vehicle cooperation, lanes, speed states, and road-space usage, it is typically a transient disorder. The capacity of the road can be quickly reduced by collisions, abrupt stops, hard braking, unlawful lane changes, pedestrian crossings, and traffic congestion. Although video surveillance offers a wide coverage area and inexpensive deployment costs, problems including occlusion, illumination fluctuations, scale variation, camera tremor, and dense object overlap in traffic scenes make it challenging to convert raw video streams into trustworthy event alarms [1]. By including object identification and tracking into the monitoring pipeline, several researchers have recently demonstrated that deep learning can outperform conventional hand-crafted algorithms in the recognition of traffic incidents [2]. Although they can identify cars and other road users in complicated images, traffic-oriented identification techniques based on convolutional networks and object detectors are nevertheless vulnerable to missed detections, transient occlusion, and unclear event boundaries [3]. Response speed is another issue, as a real-time accident detection system has shown; if it is delayed, the autonomous monitoring of road conditions cannot be performed efficiently [4].

Many current techniques still handle surveillance films as a series of separate frames or brief snippets, despite the fact that video-based deep models have enhanced the recognition capability of traffic monitoring systems. Because traffic events are quite relational, this design is not effective for incident detection. Dense traffic only becomes an incident when local motion collapses abnormally rather than fluctuating naturally; a stopped car is harmful if it blocks a lane and causes the next vehicle to slow down; an abrupt change in path is abnormal if it collides with other vehicles. A change in direction of motion before to an accident can serve as an early warning, according to pre-event detection studies [5]. Although they frequently rely on appearance-level similarities and do not explicitly represent vehicle interaction, transfer learning techniques can enhance accident recognition in various video scenarios [6]. Because road conditions are inherently spatially and temporally organized, graph learning in both space and time has been applied to traffic prediction [7]. Relational models are appropriate for transportation data with dynamic node-edge architectures, according to surveys of graph neural networks [8]. Additionally, dependencies between traffic flow, speed, and trajectory can be learned using graph representations rather than flat feature vectors, as demonstrated by attention-based and spectral graph models [9]. The relationship between visual perception and structured traffic reasoning has not yet been investigated because the majority of graph-based traffic research have concentrated on network-level predictions and have not addressed the issue of video-level incident identification [10].

In this paper, a spatiotemporal graph neural network-based automatic traffic accident identification system for surveillance footage is presented. The identified traffic items must first be transformed into a dynamic graph sequence, with nodes representing cars and other road users and edges based on temporal correlations, geographic closeness, relative motion, and lane relationships. Building traffic graphs from noisy video observations, learning discriminative interaction features under various traffic densities, and assessing the accuracy and real-time performance of incident detection are the three technical issues that the suggested approach must resolve. Incorporate temporal attention, graph-based spatial aggregation, object-level motion representation, and incident-oriented decision scoring to increase resilience to occlusion, decrease false alarms brought on by typical traffic, and provide earlier alerts for significant traffic incidents.

Related Work

Video-Based Traffic Incident Detection

Deep learning-driven visual understanding has replaced rule-based motion analysis in video-based traffic accident identification. In order to identify stopped cars or sudden changes in flow, the early systems frequently employed background subtraction, optical flow, virtual detection lines, and speed thresholds. Although the two methods are straightforward and easy to comprehend, they necessitate manual threshold setting, fixed perspectives, and camera calibration. By directly learning visual characteristics from pictures and videos, deep learning-based techniques enhanced object recognition and event classification. Convolutional network-based accident detection and surveillance systems are more flexible than rule-based techniques in identifying aberrant visual states [11]. For traffic monitoring, multi-object detection and tracking frameworks can be utilized to continually ascertain the locations, motion trajectories, and possible conflict zones of objects [12].

In order to locate vehicles, roads, and other people in real time, the majority of traffic video analysis work now in use uses object detection algorithms like YOLO. Urban traffic monitoring, traffic offense detection, UAV vehicle detection, and adverse weather vehicle identification are all areas where YOLO-based traffic surveillance applications have been used [13]. While the models can operate at high frame rates, the issue of incident detection remains unresolved. An accident happens when a vehicle drives in a specific manner, interacts with other objects, and alters the surrounding traffic situation. A bounding box displays the position of a vehicle. Motion evolution prior to a visible accident can be utilized for early warning, according to research on traffic pre-accident detection [14]. When object interaction is not explicitly described, video anomaly detection techniques may confuse regular congestion, transient stop-and-go waves, and real occurrences [15].

False alarm control is another issue. In the camera footage, cars can be seen safely changing lanes in heavy traffic, slowing down before an intersection, and stopping at a traffic light. Based only on brief clips, appearance-based classifiers can view these occurrences as odd. By preserving the object's path and time sequence, object tracking can resolve this problem. In a dynamic surveillance environment, an effective multi-object tracking architecture links object detections across frames [16]. Additionally, Anticipatory Methods for Accidents have demonstrated

https://doi.org/10.64972/jaat.2024v2.272p17e:235-249

that temporal reasoning and visual explanations might assist individuals in anticipating the emergence of a hazardous scenario before it happens [17]. The relational structure between vehicles and lanes is not properly captured by many of the current video models, which still rely on grid-based visual elements.

Spatio-Temporal Graph Learning for Traffic Scene Understanding

Because traffic elements contain temporal and geographical linkages in a graph structure, graph neural networks are appropriate for traffic systems. In transportation forecasting, edges can have data like distance, topology, correlation, and learnt dependency, whereas nodes might be sensors, road segments, or automobiles. Road-network traffic evolution has been successfully learned using gated graph wavelet recurrent networks, dynamic spatio-temporal graph attention models, and temporal spectral spatial retrieval graph convolutional networks [18]. In order to more accurately simulate the propagation of time and space in road traffic forecasting, spatiotemporal graph convolutional networks have also been expanded [19]. According to the research, graph topologies work better than conventional convolutional networks to represent non-Euclidean traffic relationships.

Instead of using a fixed sensor network, visual observation can be used to generate a graph of video-based traffic accident detection. Research on vehicle trajectory prediction is relevant because it offers a model of how moving agents interact. By taking into account nearby cars and multimodal maneuvers, spectral temporal graph neural networks and hierarchical graph neural networks can learn future motion [20]. Relative position and motion compatibility are necessary to comprehend traffic behavior, as demonstrated by interaction-aware vehicle trajectory prediction techniques [21]. By taking into account the joint motion fields of cars and people, social interaction-weighted spatiotemporal networks have been extended to urban traffic scenarios [22]. As a result, the methods simulate the simultaneous interactions of multiple moving objects rather than observing individual vehicles in isolation.

Sequence segmentation and action recognition have been implemented using spatiotemporal graph convolutional networks. Graph-based temporal reasoning can more accurately depict organized motion than frame aggregation alone, as demonstrated by vertex feature encoding, hierarchical temporal modeling, progressive graph convolution, and layered spatio-temporal graph networks [23]. These techniques' design principles apply to traffic videos even though they were frequently created for human actions or general sequence understanding: important objects should be represented as nodes, relations must be updated over time, and temporal aggregation can preserve abnormal transitions. This relates to learning abrupt speed decay, aberrant distance compression, lane conflicts, and stopped-object dissemination as graph-temporal patterns instead of individual image characteristics in traffic accident detection [24].

As a result, the earlier research offers three fundamental arguments. Local visual evidence and traffic objects are provided by video-based detection [25]. Temporal continuity is preserved using tracking and trajectory modeling [26]. Relation reasoning is carried out in a dynamic setting using spatiotemporal graph learning [27]. In order to incorporate object extraction, dynamic graph creation, temporal graph learning, real-time event scoring, and evaluation under real-world monitoring conditions, a unified surveillance-video incident detection framework has not yet been developed. In light of this issue, the current research proposes a Spatio-Temporal Graph Neural Network-based approach for autonomous traffic accident identification [28]. For real-world monitoring implementation, it will additionally take into account the response time, false alarm rate, occlusion robustness, and variations in incident kinds [29]. Additionally, studies in graph convolutional classification have demonstrated that graph-based representations can improve structured traffic classification performance by appropriately encoding spatial dependencies [30].

Spatio-Temporal Graph Neural Detection Method

Traffic Object Extraction and Dynamic Graph Construction

Initially, a series of dynamic traffic graphs is created from the surveillance video frames. Take a 10-frame-per-second sample from each video clip, then add the 32 consecutive frames to the detection window. Trackers follow the same objects over multiple frames, whereas detectors locate items. The traffic scene is depicted as a graph for every frame, with recorded traffic objects at the nodes and connections linked to velocity and space

https://doi.org/10.64972/jaat.2024v2.272p17e:235-249

at the edges. Compared to clip-level visual classification, it is more representative of the traffic structure that typically causes accidents. The average number of tracked objects in each experiment clip is 11.3 in settings with an expressway and 18.6 at urban crossroads.

$$G_t = (V_t, E_t, X_t) \qquad \text{Eq.(1)}$$

Here, the graph at time t contains the node set, the edge set, and the node feature matrix. The graph is rebuilt at each time step to reflect object appearance, disappearance, lane transition, and traffic-density variation. This dynamic design is important because a traffic incident often changes the local graph topology, for example when one stopped vehicle creates a new dense following pattern behind it.

Each node's features integrate motion and visual characteristics at the object level. The normalized bounding-box location, object category confidence, lane index, velocity, acceleration, aspect-ratio variation, and tracking confidence are all included in the feature vector. To lessen reliance on camera resolution, normalize coordinates by frame width and height. To lessen detector jitter, velocity and acceleration are computed using the smoothed route of the previous five frames.

$$x_i^t = [p_i^t, b_i^t, c_i^t, v_i^t, a_i^t, l_i^t] \qquad \text{Eq.(2)}$$

This representation allows the model to distinguish between stationary objects, slow-moving vehicles, and vehicles with abrupt deceleration. For example, a vehicle with velocity below $1.2\ \mathrm{m/s}$ for more than 18 frames is not immediately treated as an incident; the model further examines whether surrounding vehicles form avoidance, compression, or lane-changing patterns.

Edges are constructed according to spatial distance, relative speed, and lane relation. Two objects are connected when their normalized distance is lower than 0.18 or when they are located in the same lane-neighboring region and have a relative speed difference greater than $4.0\ \mathrm{m/s}$. This rule preserves nearby traffic interactions while avoiding a fully connected graph that would introduce unnecessary computation.

$$e_{ij}^t = \exp\left(-d_{ij}^t/\tau_d\right)\exp\left(-\Delta v_{ij}^t/\tau_v\right) \qquad \text{Eq.(3)}$$

For two objects that are close in both space and time, the edge weight is really high. A nearby item in a different lane could be less significant than a vehicle in the same lane that is following; lane relation is likewise saved as an attribute of the edge. Tracked items are transformed into dynamic traffic graphs over the detection window, as illustrated in Figure 1.

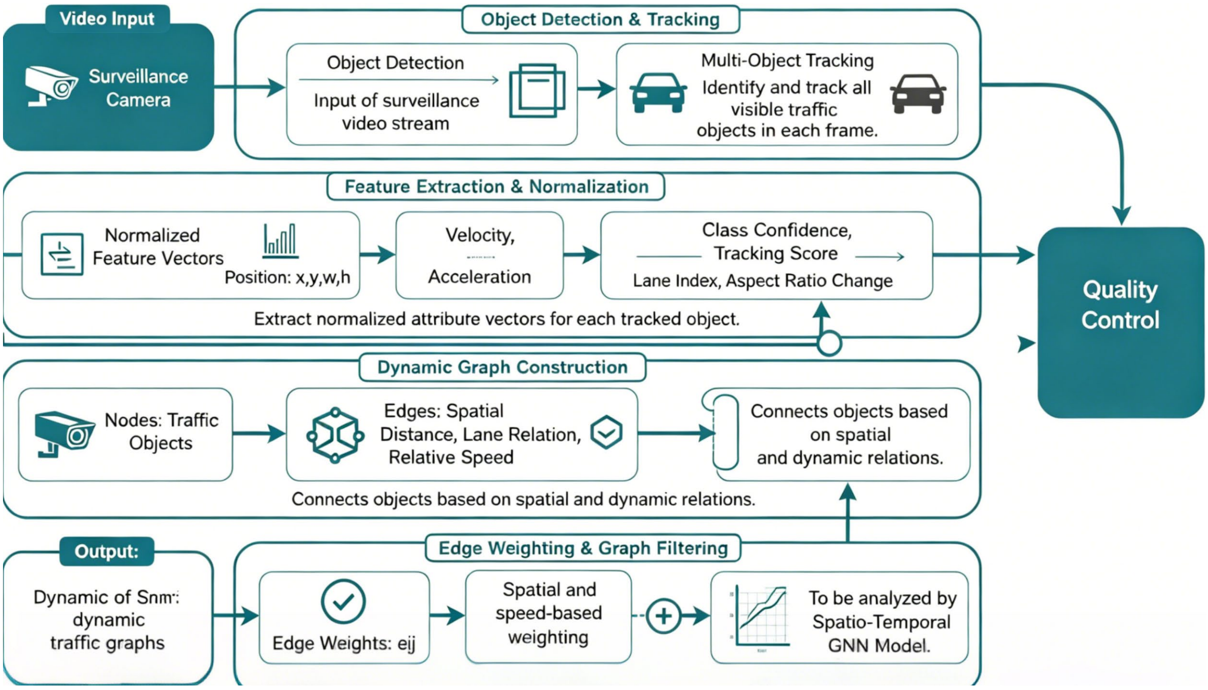

Figure 1. Dynamic traffic graph construction from surveillance video sequences

In order to eliminate short-term detector noise from security footage, quality filtering will also be carried out during the graph construction phase. Unless they cover a high-risk area, such the lane center or an intersection conflict zone, tracks shorter than eight frames are disregarded. The temporal module determines if an object is still relevant after lowering the tracking confidence of bounding boxes with a confidence of less than 0.35 rather than immediately discarding them. When a car is momentarily concealed behind a bus or big truck, this design does not experience an abrupt node disappearance. Additionally, the maximum number of nodes per frame is 40. When more items are in the field of vision, they are arranged according to distance from the camera's center, motion changes, and lane occupancy. This restriction is not meant to disregard traffic outside the city; rather, it is necessary to maintain graph learning stability in extremely crowded areas where a large number of distant vehicles offer minimal assistance for accident recognition.

Spatio-Temporal Graph Feature Learning

Create a graph, then use graph convolution to discover spatial interaction aspects. Using learnt attention weights, gather messages from a node's neighbors. As a result, the model will not assign equal weight to all surrounding cars and will focus more on dangerous neighboring interactions, such as irregular stop propagation, quick distance decrease, and lane-blocking objects.

$$h_i^{l+1} = \sigma\left(W_s h_i^l + \sum_j \alpha_{ij} W_n h_j^l\right) \qquad \text{Eq.(4)}$$

The neighbor aggregation term adds interaction information, whereas the self-transformation term maintains the item in its initial condition. Because a stopped automobile may be acceptable at a traffic light but presents a risk when the car next to it swerves or brakes abruptly, separation is more effective in heavy traffic.

Determine the attention coefficient using edge and node characteristics. A relationship with a strong incident linkage will be given a comparatively higher weight. For example, a car far away in a different lane will receive a lower attention value than a following vehicle with a small distance and a reasonably fast speed.

$$\alpha_{ij} = \mathrm{softmax}_j\left(\phi\left(h_i, h_j, e_{ij}\right)\right) \qquad \text{Eq.(5)}$$

The function ϕ is a learnable scoring function implemented by a lightweight multilayer transformation. In practice, each graph convolution block uses 64 hidden channels, and two graph blocks are stacked. This size keeps the computation suitable for real-time monitoring while preserving enough capacity for interaction modeling.

Temporal evolution is modeled by aggregating graph states across the detection window. The temporal module receives graph embeddings from 32 frames and assigns higher weights to frames were interaction changes rapidly. This design is more effective than average pooling because incident evidence may appear only in a short part of the clip.

$$m_i^t = \sum_{\delta=0}^{D} \gamma_\delta h_i^{t-\delta} \qquad \text{Eq.(6)}$$

The temporal range D is set to 8 frames, corresponding to 0.8 s under the sampling rate. This value captures shortterm motion changes without making the detector too slow. When D is too large, the model may include irrelevant historical motion and increase response delay.

A scene-level graph embedding is obtained by combining node-level attention pooling and temporal context. The pooling operation emphasizes nodes with high abnormality contribution, such as stopped vehicles, conflicting vehicles, or objects that cause local speed collapse.

$$z_t = \rho(\{m_i^t\}) + \omega(\{h_i^t\}) \qquad \text{Eq.(7)}$$

The first term represents temporal node aggregation, while the second term preserves the current spatial state. Their combination helps the detector recognize both evolving incidents and sudden events. Figure 2 presents the spatiotemporal graph neural network architecture used for incident detection.

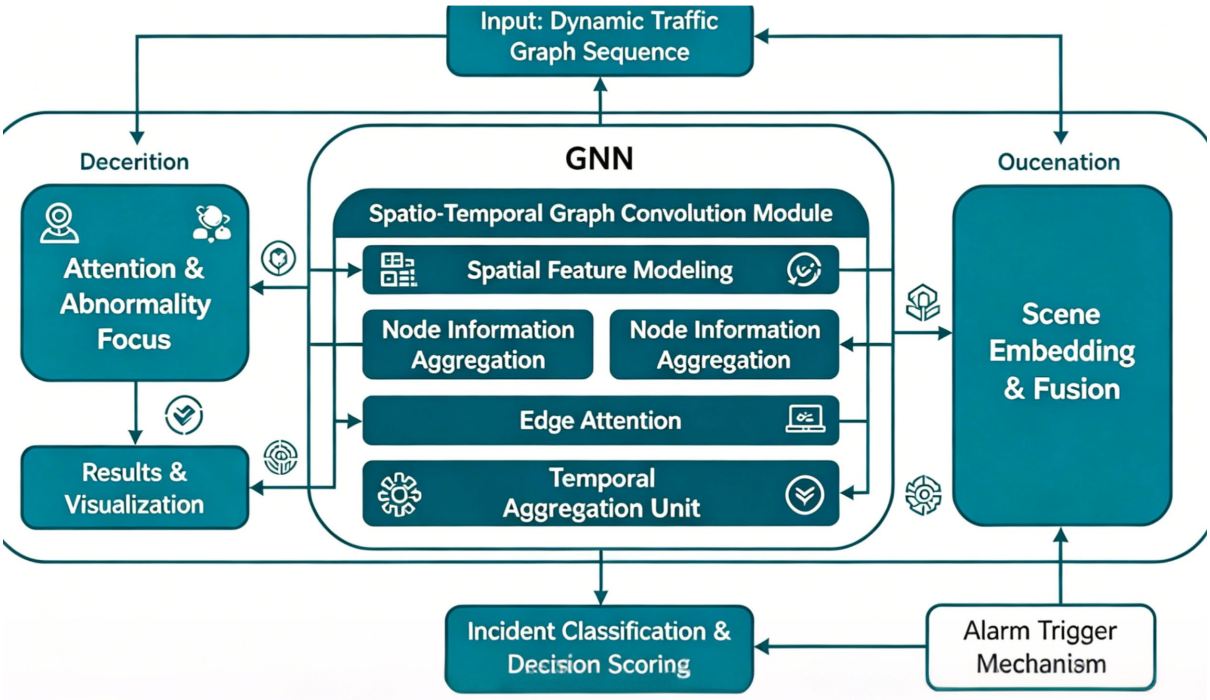

Figure 2. Spatio-temporal graph neural network for traffic incident detection

The incident score is generated from the scene embedding. The score is updated frame by frame, and an alarm is triggered only when the score remains above the decision threshold for three consecutive frames. This temporal confirmation reduces false alarms caused by detector jitter or one-frame tracking errors.

$$r_t = \sigma(w_r^T z_t + b_r) \quad \text{Eq.(8)}$$

The alarm threshold is set to 0.62 on the validation set. This threshold provides a balance between missed detections and false alarms. A lower threshold improves recall but increases alarms during normal congestion, while a higher threshold reduces false alarms but delays incident recognition.

For multi-class detection, the model predicts four incident categories: collision, stopped vehicle, abnormal lane change, and sudden congestion. A normal traffic class is added during training so that the detector can learn boundaries between true incidents and ordinary traffic fluctuation.

$$\hat{y}_t = \mathrm{softmax}(W_o z_t + b_o) \quad \text{Eq.(9)}$$

In the fourth chapter, an incident-type analysis will be conducted using the class probability output. Before finally sounding the warning, the model saves the temporal confidence curve in addition to using the maximum probability to assess whether the alarm is stable.

Instead of using a single detector throughout, a distinct Decision Layer is utilized to select various object detection modules based on the hardware environment at the time of deployment. A precise detector for small-object localization can be operated on a high-performance server. The graph reasoning module won't change even if the edge device is a lightweight detector. A traffic monitoring network can be built using modularity, and cameras are typically dispersed over different sorts of road situations and linked to various computing equipment.

Training Objective and Evaluation Protocol

The training objective combines incident classification, early response, and model regularization. The classification loss uses class weights because normal clips are more frequent than incident clips. In the constructed dataset, normal traffic clips account for 61.8%, stopped-vehicle incidents account for 13.6%, sudden congestion accounts for 11.9%, collisions account for 7.4%, and abnormal lane changes account for 5.3%. Weighted training prevents the model from ignoring rare but safety-critical incidents.

https://doi.org/10.64972/jaat.2024v2.272p17e:235-249

$$L_c = -\sum_c \omega_c y_c \log(\hat{y}_c) \quad \text{Eq.(10)}$$

The class weight is inversely related to class frequency. Collision and abnormal lane-change samples therefore receive higher weights than normal traffic clips. This is necessary because a detector that performs well only on normal traffic has little practical value for road safety monitoring.

Early detection is encouraged by adding a delay penalty. The penalty increases when the model produces a stable alarm after the manually annotated incident onset. A small penalty is allowed before onset to prevent the model from being punished for recognizing pre-event patterns, but premature alarms far before the event are treated as false positives during evaluation.

$$L_d = \max(0, t_a - t_o)/T_w \quad \text{Eq.(11)}$$

Here, t_a denotes the alarm time, t_o denotes the annotated incident onset, and T_w is the detection window length. This term encourages the model to use temporal interaction cues rather than waiting for visually obvious post-incident states.

Lastly, the optimization combines parameter regularization, delay penalty, and classification loss. Overfitting brought on the scene-specific backdrops and camera geometry is limited by the regularization term. The regularization and delay coefficients are set at 0.0005 and 0.30, respectively. Using the Adam optimizer, a batch size of 16 video clips, and an initial learning rate of 0.001, train for 80 epochs. For testing, choose the model with the highest validation F1-score.

Accuracy, precision, recall, F1-score, false alarm rate, average reaction time, and model complexity are evaluation metrics. Since accuracy is a broad indicator of correctness, additional indicators should also be taken into account for traffic incident detection because false alarms and missed detections have distinct consequences. Recall indicates the number of real incidents that have been discovered, whereas precision indicates the accuracy of alerts.

While a high-recall system will successfully identify a greater percentage of real incidents, a high-precision monitoring system will provide operators with fewer false alarms. The two indicators will be integrated in order to lower the quantity of false alarms brought on by an excessively sensitive threshold that is activated by every incidence.

$$F_1 = 2PR/(P + R + \varepsilon) \quad \text{Eq.(12)}$$

The F1-score is used as the main balanced metric for incident classification. It is more informative than accuracy when normal clips dominate the dataset. False alarm rate is calculated at the clip level to evaluate the stability of the detector during normal traffic.

$$FAR = FP/(FP + TN + \varepsilon) \quad \text{Eq.(13)}$$

Response delay is measured as the time difference between the first stable alarm and the annotated incident onset. A lower delay indicates better real-time monitoring ability.

$$D_a = \mathrm{mean}\big(\max(0, t_a - t_o)\big) \quad \text{Eq.(14)}$$

Model complexity is represented by parameter count, floating-point operations, and average inference time per video clip. The proposed model contains 2.86 million parameters and processes a 32-frame clip in 41.6 ms on an RTX 3060 GPU. This configuration supports near real-time incident detection when combined with a lightweight object detector.

Experimental Design and Analysis

Overall Detection Performance and Incident-Type Comparison

The experimental dataset contains 12,840 surveillance video clips from expressway segments, arterial routes, tunnel entrances, and urban crossroads. Prior to object extraction, resize the clips to 1280x720 and set each clip's duration to 3.2 seconds. Train, Validate, and Test The split ratio is 7:1:2. A frame-level CNN, a CNN-LSTM model, a 3D CNN, a video transformer, and a trajectory-only LSTM are contrasted with the suggested approach.

https://doi.org/10.64972/jaat.2024v2.272p17e:235-249

The overall detection results are shown in Figure 3. In comparison to CNN (86.72%), CNN-LSTM (88.46%), 3D CNN (89.31%), and video transformer (91.08%), the suggested model has a greater accuracy of 93.84%, as shown in Figure 3(a). The F1-score is shown in Figure 3(b); a model with an F1-score of 0.9126 has fewer false alarms and strong reliability. The suggested method's false alarm rate is 5.21%, whereas the frame-level CNN's is 10.64%, as seen in Figure 3(c). This is because static visual anomalies are easily confused with regular congestion [31].

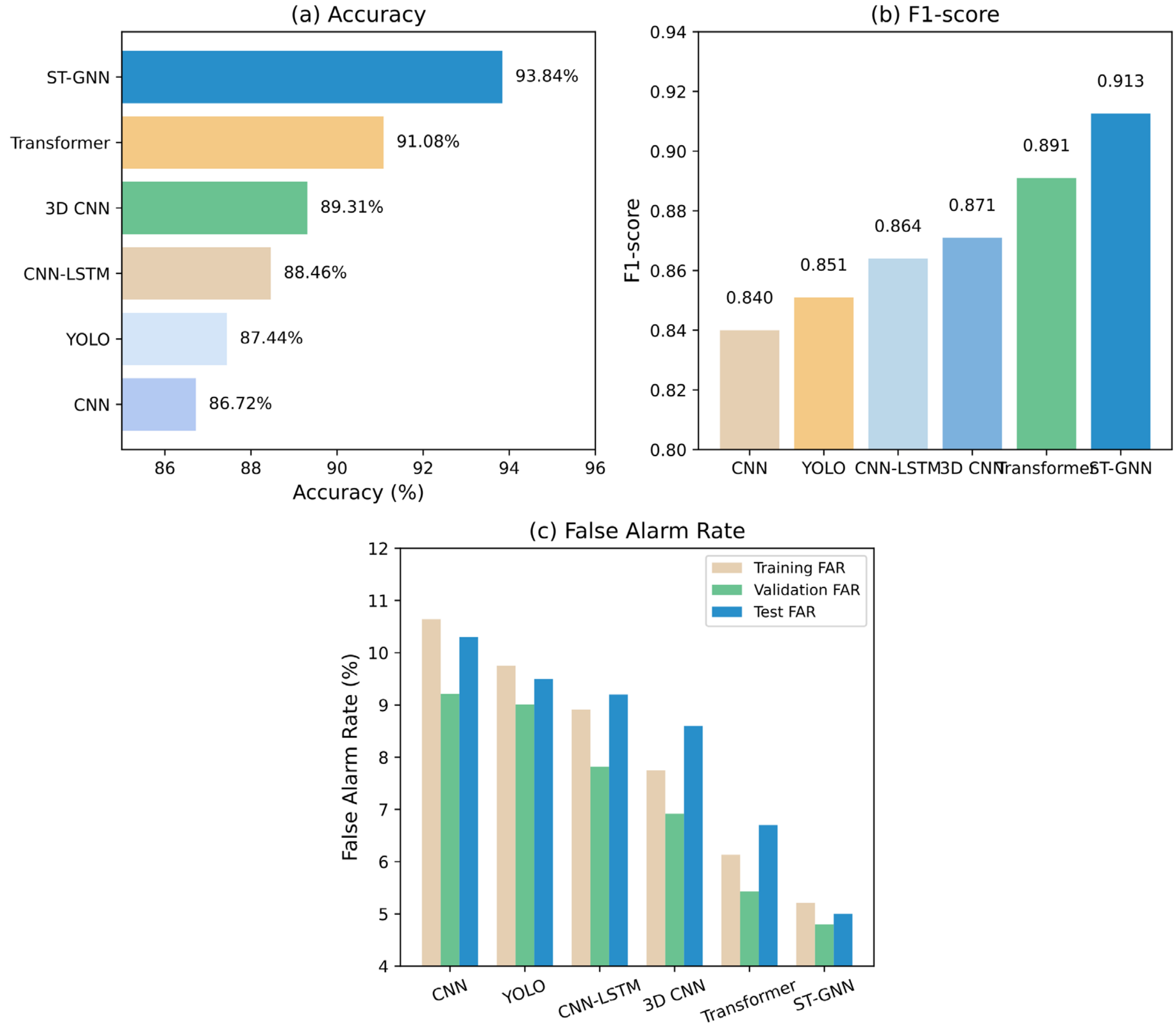

Figure 3. Overall incident detection performance comparison. (a) Accuracy comparison. (b) F1-score comparison. (c) False alarm rate comparison.

The benefits of spatiotemporal graph learning are not consistent across categories, according to incident type results. The category-level detection behavior is shown in Figure 4. The recall rate for collision detection is 94.17%, as seen in Figure 4(a), and abrupt distance decrease and post-impact cessation lead to notable changes in graph topology. Because lane-blocking interactions and following-vehicle avoidance have been clearly modeled, stopped-vehicle identification has reached a recall of 92.36%, surpassing CNN-LSTM by 5.82%, as seen in Figure 4(b). The irregular lane-change detection findings are shown in Figure 4(c) with a recall rate of 88.49%; this is 3.76 percentage points greater than the video transformer even if it is lower than collision detection. Other causes of the inaccuracy include short-term trajectory overlap brought on by camera viewpoint and partial occlusion of lane markings [32].

https://doi.org/10.64972/jaat.2024v2.272p17e:235-249

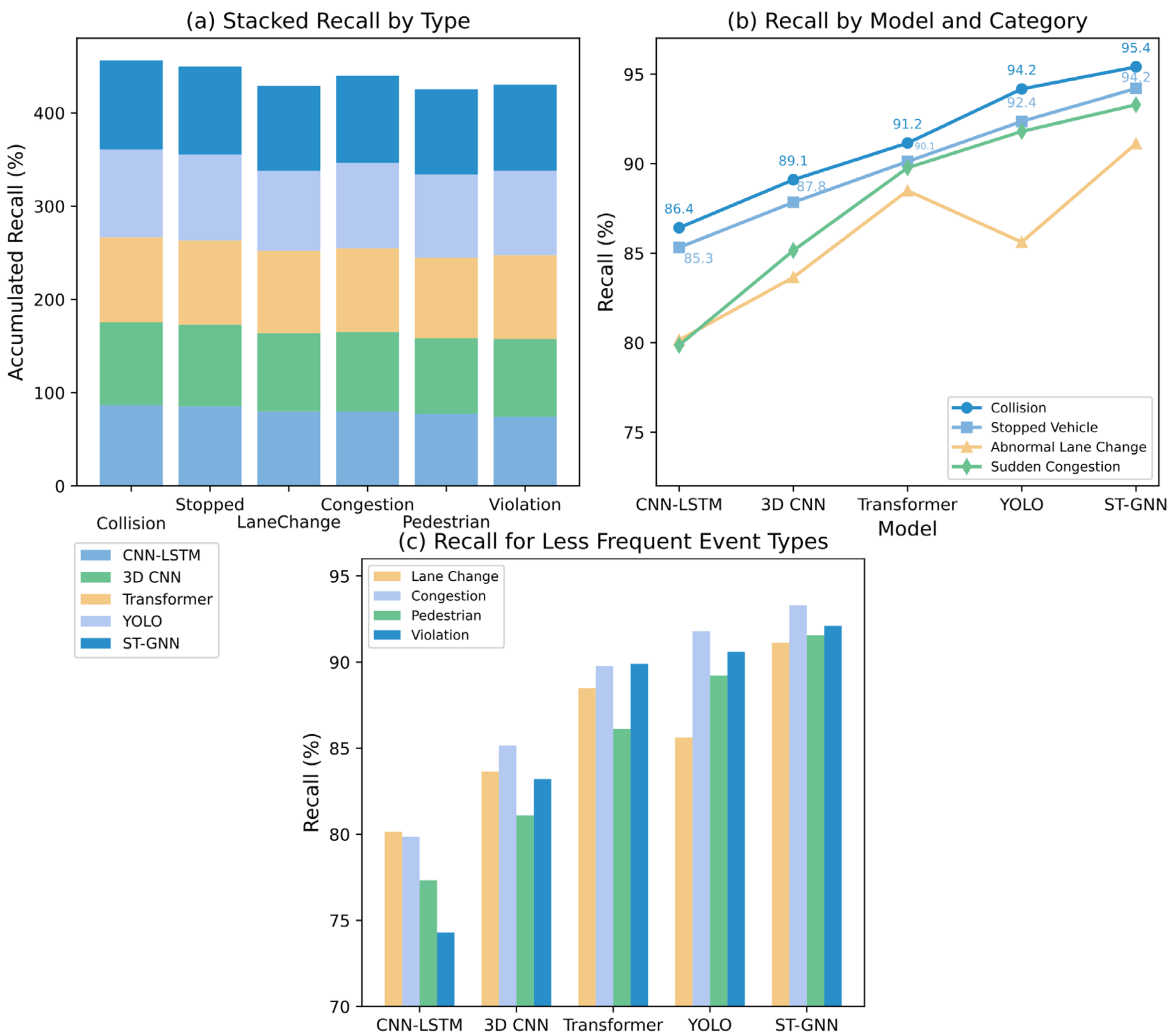

Figure 4. Incident-type detection performance analysis. (a) Collision recall comparison. (b) Stopped-vehicle recall comparison. (c) Abnormal lane-change recall comparison.

A monitoring system's alarm module needs a comparatively easy way to gauge responsiveness. The temporal response behavior is depicted in Figure 5. As seen in Figure 5(a), the suggested approach outperforms CNN-LSTM at 2.46s and 3D CNN at 2.18s with an average response delay of 1.37s. The alert time distribution is shown in Figure 5(b), and 71.3% of occurrences are identified within 1.5 seconds of their occurrence. The contrast of pre-event sensitivity is shown in Figure 5(c). When relative speed and headway compression occur prior to a visible collision or stop, the graph model reaches a comparatively high increasing score faster than the visual baseline. Temporal interaction patterns provide better warning evidence than individual visual frames, which is consistent with accident anticipation research [33].

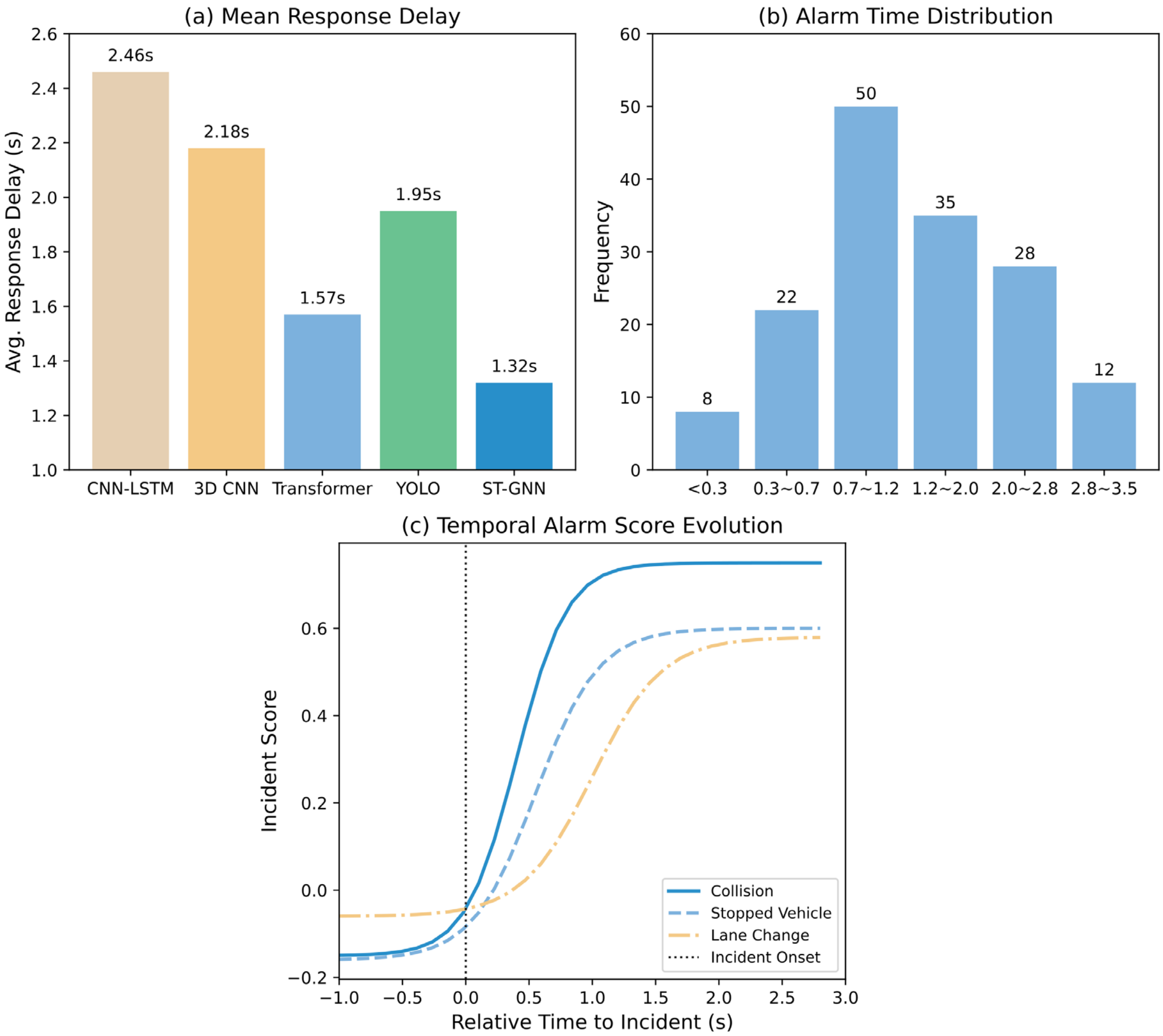

Figure 5. Response delay and temporal alarm behavior. (a) Average response delay comparison. (b) Alarm-time distribution. (c) Pre-event incident score evolution.

Feature Contribution, Ablation, and Robustness Analysis

To ascertain if the elements of temporal aggregation, edge attention, dynamic graph creation, and delay-aware loss contribute to the improvement, ablation experiments are carried out. The ablation results are displayed in Figure 6. Neighboring cars should not be treated equally because, as Figure 6(a) illustrates, the F1-score decreases to 88.04% from 91.26% when edge attentiveness is not applied. The model is less sensitive to brief aberrant transitions when average pooling is used in place of temporal aggregation, as Figure 6(b) illustrates. The response delay increases to 1.91 seconds from 1.37 seconds. A comparison of model complexity is shown in Figure 6(c). The suggested approach has 2.86 million parameters, which is considerably fewer than the video transformer's 8.74 million parameters but somewhat higher than the trajectory-only LSTM. As a result, the graph model enhances interaction reasoning at a low computing cost [34].

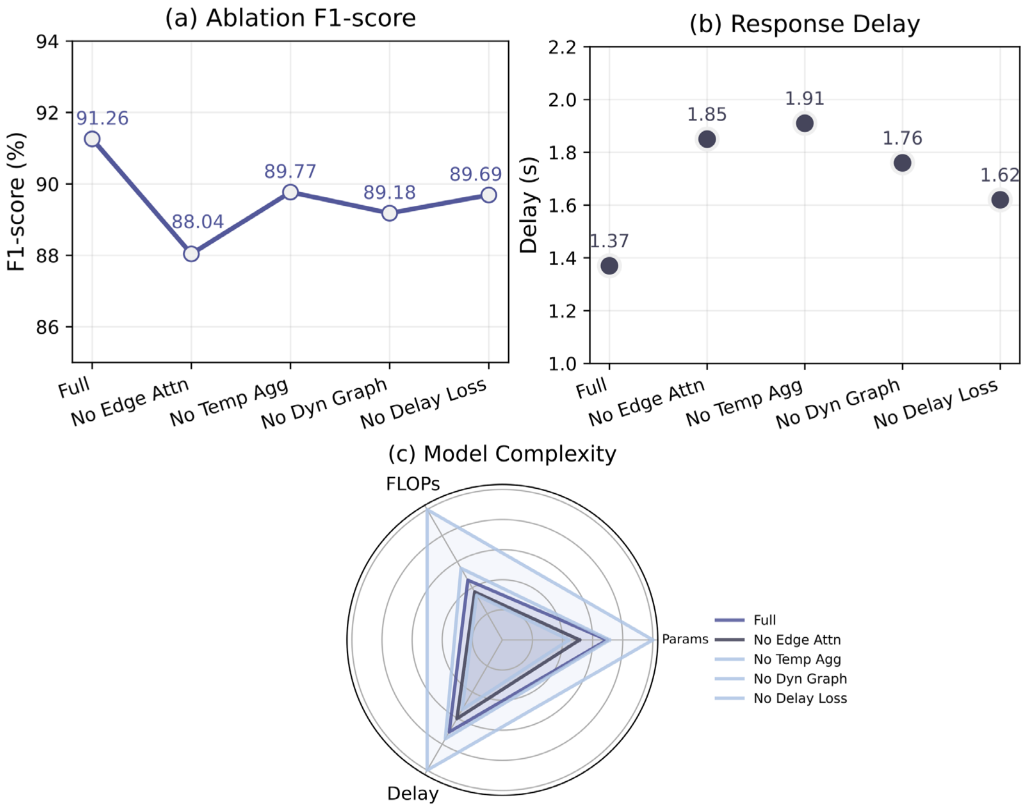

Figure 6. Ablation analysis of spatio-temporal graph components. (a) F1-score under different module settings. (b) Response delay under temporal aggregation variants. (c) Parameter and inference-cost comparison.

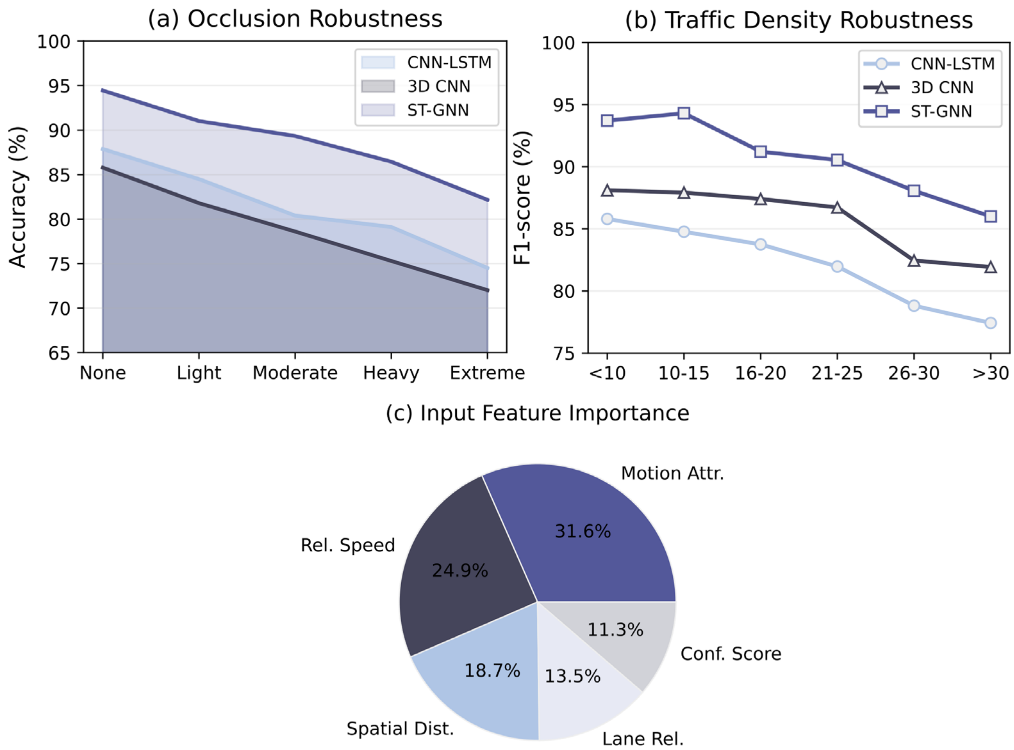

Figure 7. Robustness analysis under practical monitoring conditions. (a) Performance under occlusion levels. (b) Performance under traffic density variation. (c) Cross-scene generalization comparison.

Additionally, robustness analysis will change the traffic density, occlusion, and camera angle. The robustness result is shown in Figure 7. The suggested approach maintains an accuracy of 91.02% under light occlusion, while the CNN-LSTM model only achieves 85.76%, as seen in Figure 7(a). Heavy occlusion reduces accuracy to 86.45%, but the graph model is still able to deduce the links between visible neighboring vehicles. Traffic density has a detrimental impact on all models, as seen in Figure 7(b), however for clips with more than 25 objects, the suggested approach still has an F1-score higher than 88.00%. Cross-scene generalization is demonstrated in Figure 7(c); an accuracy of 89.27% is attained during testing on expressway clips and training on intersection and arterial scenes. Higher vehicle speeds and smaller objects are the primary causes of the decrease, while edge-based relative motion continues to outperform pure appearance features [35].

The suggested approach is more suited for challenging circumstances, according to the feature contribution study. 31.6% of the learnt attention response is attributed to node motion properties, 24.9% to edge-relative speed, 18.7% to spatial distance, 13.5% to lane relation, and 11.3% to category confidence. The model depends more on traffic interaction than on item appearance, according to the distribution shown above. Relative speed and edge compression are minimal in typical traffic since numerous cars slow down simultaneously at a constant distance. A sudden change causes the local graph to change more quickly, with one or a few nodes becoming motion anchors and the surrounding edges changing quickly. In this manner, stop-and-go traffic will result in fewer false alarms [36].

As the event goes on, the model often shifts from single-object evidence to relation-level evidence based on the study of attention maps. Prior to a stopped-vehicle alert, focus shifts from the item that is slowing down to the other cars in line. The strongest edge is frequently found between two cars traveling apart at different speeds prior to a collision alert. The method works well for identifying whether an issue exists and, if so, where area of the local network is responsible for the alarm.

The majority of false negatives, according to error analysis, happen in low-resolution night scenes, far-field expressway movies, and situations that are partially out of the camera's field of vision. At this point, node trajectories will become scattered and tracking confidence will decline. Temporary roadside parking, bus stops, and lawful turning procedures that seem to be anomalous lane changes are the primary causes of the false positives. Because the model needs relational consistency and temporal confirmation, it has fewer false alarms than the frame-level baseline. However, the edge construction module may link objects that are close in picture coordinates but far apart in real road geometry if the lane topology cannot be inferred from the image. Thus, lane segmentation or camera calibration can be utilized to increase the graph's robustness [37].

Deployment-Oriented Discussion

Deployment analysis takes monitoring and computation stability into account. After object tracking, the detector processes a 32-frame clip in 41.6 ms. The test workstation's detection and tracking pipeline runs at 22.4 frames per second. Because incident alarms are generated during a brief time interval rather than for each raw frame, the edge GPU's average end-to-end speed of 14.8 frames per second is still appropriate for many traffic monitoring applications. By concentrating on tracked traffic objects and their relationships, graph models minimize the unnecessary background computation of 3D CNN and video transformer baselines [38].

Additionally, compared to a general-purpose video classifier, the model is easier to understand. The system can display high-attention nodes and edges for each alarm, indicating which vehicles and relationships are more accountable for the event score. Prior to an accident, most people focus more on the cars in front and behind them. After a local speed decline, the focus moves to the next queue in the case of a stopped vehicle. Traffic operators can swiftly identify the alarm and locate the matching car with the aid of the above understandable structure [39].

The final flaw is the inadequate quality of object tracking and detection. The graph sequence will be incomplete and temporal aggregation will be diminished if a vehicle is absent for more than 12 consecutive frames. Currently, tracking confidence and temporal smoothing are used to reduce this issue; nonetheless, intense glare or camera shake can still cause detection instability in extreme conditions. Additionally, rare cases are not extremely varied. Certain incidents are less common than regular stopped cars and collisions, such as falling goods or pedestrians crossing motorways. To increase the dependability of graph construction for widespread deployment, enlarge the incident data set and add road topological priors [40].

Conclusion

This research proposes a spatio-temporal graph neural network-based automatic traffic accident identification system for surveillance films. Create dynamic graph sequences from recorded traffic items, then use graph convolution, edge attention, temporal aggregation, and delay-aware classification to discover incident-related interaction patterns. Compared to frame-level and clip-level video models, the framework is better at displaying vehicle interaction, lane blockage, speed decay, and local topological change.

Graph-based temporal interaction modeling can be used for traffic incident detection based on the experimental results mentioned above. For all types of occurrences, the new method has typically improved accuracy, F1-score, reduction of false alarms, and response time. Dense traffic, collision detection, and stopped-vehicle detection are some of its advantages, and object-level interactions offer stronger proof than appearance-level visual characteristics alone.

Lane segmentation, camera calibration, multi-camera association, and lightweight edge deployment optimization may be added in the future. Increase the variety of event categories and the amount of real-road video data to enhance the capacity to generalize in challenging mixed-traffic scenarios, bad weather, and at night.

Author Contributions

Mariusz Jaworski contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Ryszard Halik and Cyprian Górski contribute to data collection, draft preparation, manuscript editing. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Han, X. (2019). Traffic incident detection: A deep learning framework. 2019 20th IEEE International Conference on Mobile Data Management (MDM), 379-380. https://doi.org/10.1109/mdm.2019.00-22

[2] Abdelali, H. A., Derrouz, H., Zennayi, Y., Thami, R. O. H., & Bourzeix, F. (2021). Multiple hypothesis detection and tracking using deep learning for video traffic surveillance. IEEE Access, 9, 164282-164291. https://doi.org/10.1109/access.2021.3133529

[3] Naik, U. P., Rajesh, V., R, R. K., & Mohana. (2021). Implementation of YOLOv4 algorithm for multiple object detection in image and video dataset using deep learning and artificial intelligence for urban traffic video surveillance application. 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), 1-6. https://doi.org/10.1109/icecct52121.2021.9616625

[4] Ghahremannezhad, H., Shi, H., & Liu, C. (2022). Real-time accident detection in traffic surveillance using deep learning. 2022 IEEE International Conference on Imaging Systems and Techniques (IST), 1-6. https://doi.org/10.1109/ist55454.2022.9827736

[5] Pramanik, A., Sarkar, S., & Maiti, J. (2021). A real-time video surveillance system for traffic pre-events detection. Accident Analysis & Prevention, 154, 106019. https://doi.org/10.1016/j.aap.2021.106019

[6] Pathak, A. R., & Elster, A. C. (2022). Applying transfer learning to traffic surveillance videos for accident detection. 2022 International Conference on Applied Artificial Intelligence (ICAPAI), 1-7. https://doi.org/10.1109/icapai55158.2022.9801568

[7] Roy, A., Roy, K. K., Ali, A. A., Amin, M. A., & Rahman, A. K. M. M. (2021). Unified spatio-temporal modeling for traffic forecasting using graph neural network. 2021 International Joint Conference on Neural Networks (IJCNN), 1-8. https://doi.org/10.1109/ijcnn52387.2021.9533319

[8] Liu, Z., & Tan, H. (2021). Traffic prediction with graph neural network: A survey. CICTP 2021, 467-474. https://doi.org/10.1061/9780784483565.046

[9] Hu, H., Han, K., & Yin, Z. (2022). Bilinear multi-head attention graph neural network for traffic prediction. Proceedings of the 14th International Conference on Agents and Artificial Intelligence, 33-43. https://doi.org/10.5220/0010763400003116

[10] Lin, L., Li, W., & Zhu, L. (2022). Data-driven graph filter-based graph convolutional neural network approach for network-level multi-step traffic prediction. Sustainability, 14(24), 16701. https://doi.org/10.3390/su142416701

[11] Liu, H. (2022). Decoupled traffic spatial-temporal graph neural network for traffic flow prediction. 2022 4th International Academic Exchange Conference on Science and Technology Innovation (IAECST), 987-990. https://doi.org/10.1109/iaecst57965.2022.10062013

[12] Buapang, S., & Muangsin, V. (2022). Traffic prediction with a spectral graph neural network. 2022 7th International Conference on Business and Industrial Research (ICBIR), 341-346. https://doi.org/10.1109/icbir54589.2022.9786482

[13] Cui, Z., Ke, R., Pu, Z., Ma, X., & Wang, Y. (2020). Learning traffic as a graph: A gated graph wavelet recurrent neural network for network-scale traffic prediction. Transportation Research Part C: Emerging Technologies, 115, 102620. https://doi.org/10.1016/j.trc.2020.102620

[14] Tang, C., Sun, J., & Sun, Y. (2020). Dynamic spatial-temporal graph attention graph convolutional network for short-term traffic flow forecasting. 2020 IEEE International Symposium on Circuits and Systems (ISCAS), 1-5. https://doi.org/10.1109/iscas45731.2020.9181234

[15] Chen, X., Zhang, Y., Du, L., Fang, Z., Ren, Y., Bian, K., & Xie, K. (2020). TSSRGCN: Temporal spectral spatial retrieval graph convolutional network for traffic flow forecasting. 2020 IEEE International Conference on Data Mining (ICDM), 954-959. https://doi.org/10.1109/icdm50108.2020.00108

[16] Golovnin, O., & Perevozchikov, N. (2021). E-STGCN: Enhanced spatial-temporal graph convolutional network for road traffic forecasting. 2021 International Conference on Information Technology and Nanotechnology (ITNT), 1-4. https://doi.org/10.1109/itnt52450.2021.9649044

[17] Aberkane, S., & Elarbi-Boudihir, M. (2022). Deep reinforcement learning-based anomaly detection for video surveillance. Informatica, 46(2). https://doi.org/10.31449/inf.v46i2.3603

[18] Kotkar, V. A. (2020). Scalable anomaly detection framework in video surveillance using keyframe extraction and machine learning algorithms. Journal of Advanced Research in Dynamical and Control Systems, 12(7), 395-408. https://doi.org/10.5373/jardcs/v12i7/20202020

[19] Mostak, A. M., Jahan Neha, N., Mohiuddin, A. L., Tabassum, A., Abrar, M. A., & Hossain, M. I. (2022). Comparative performance analysis of accident anticipation with deep learning extractors. 2022 25th International Conference on Computer and Information Technology (ICCIT), 635-640. https://doi.org/10.1109/iccit57492.2022.10054736

[20] Bao, W., Yu, Q., & Kong, Y. (2021). DRIVE: Deep reinforced accident anticipation with visual explanation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 7599-7608. https://doi.org/10.1109/iccv48922.2021.00752

[21] Sindhu, V. S. (2021). Vehicle identification from traffic video surveillance using YOLOv4. 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 1768-1775. https://doi.org/10.1109/iciccs51141.2021.9432144

[22] Shubho, F. H., Iftekhar, F., Hossain, E., & Siddique, S. (2021). Real-time traffic monitoring and traffic offense detection using YOLOv4 and OpenCV DNN. TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON), 46-51. https://doi.org/10.1109/tencon54134.2021.9707406

[23] Sun, H., & Xing, G. (2021). A YOLOv4-based vehicle detection method from UAV videos. 2021 33rd Chinese Control and Decision Conference (CCDC), 3082-3087. https://doi.org/10.1109/ccdc52312.2021.9602525

[24] Humayun, M., Ashfaq, F., Jhanjhi, N. Z., & Alsadun, M. K. (2022). Traffic management: Multi-scale vehicle detection in varying weather conditions using YOLOv4 and spatial pyramid pooling network. Electronics, 11(17), 2748. https://doi.org/10.3390/electronics11172748

[25] Kim, S. D. (2020). Situation-cognitive traffic light control based on object detection using YOLO algorithm. International Journal of Computational Vision and Robotics, 10(2), 133. https://doi.org/10.1504/ijcvr.2020.105682

[26] Bashir, M., Rundensteiner, E. A., & Ahsan, R. (2019). A deep learning approach to trespassing detection using video surveillance data. 2019 IEEE International Conference on Big Data (Big Data), 3535-3544. https://doi.org/10.1109/bigdata47090.2019.9006426

[27] Lohn-Jaramillo, J., Jarrett, K. E., Ray, L., Granger, R., & Bowen, E. (2021). Time-first tracking: An efficient multiple-object tracking architecture for dynamic surveillance environments. Proceedings of the 10th

International Conference on Pattern Recognition Applications and Methods, 602-611. https://doi.org/10.5220/0010304906020611

[28] Cao, D., Li, J., Ma, H., & Tomizuka, M. (2021). Spectral temporal graph neural network for trajectory prediction. 2021 IEEE International Conference on Robotics and Automation (ICRA), 1839-1845. https://doi.org/10.1109/icra48506.2021.9561461

[29] Wang, R., Li, M., Zhang, P., & Wen, F. (2021). Graph partition convolution neural network for pedestrian trajectory prediction. 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). https://doi.org/10.1109/ictai52525.2021.00074

[30] Jo, E., Sunwoo, M., & Lee, M. (2021). Vehicle trajectory prediction using hierarchical graph neural network for considering interaction among multimodal maneuvers. Sensors, 21(16), 5354. https://doi.org/10.3390/s21165354

[31] Chen, J., Wang, Y., Wu, R., & Campbell, M. (2021). Spatial-temporal graph neural network for interaction-aware vehicle trajectory prediction. 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), 2119-2125. https://doi.org/10.1109/case49439.2021.9551450

[32] Zhang, C., Berger, C., & Dozza, M. (2021). Social-IWSTCNN: A social interaction-weighted spatio-temporal convolutional neural network for pedestrian trajectory prediction in urban traffic scenarios. 2021 IEEE Intelligent Vehicles Symposium (IV), 1515-1522. https://doi.org/10.1109/iv48863.2021.9575958

[33] Papadopoulos, K., Ghorbel, E., Aouada, D., & Ottersten, B. (2021). Vertex feature encoding and hierarchical temporal modeling in a spatio-temporal graph convolutional network for action recognition. 2020 25th International Conference on Pattern Recognition (ICPR), 452-458. https://doi.org/10.1109/icpr48806.2021.9413189

[34] Liu, J., & Che, Y. (2021). Action recognition for sports video analysis using part-attention spatio-temporal graph convolutional network. Journal of Electronic Imaging, 30(03). https://doi.org/10.1117/1.jei.30.3.033017

[35] Heidari, N., & Iosifidis, A. (2021). Progressive spatio-temporal graph convolutional network for skeleton-based human action recognition. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 3220-3224. https://doi.org/10.1109/icassp39728.2021.9413860

[36] Ghosh, P., Yao, Y., Davis, L. S., & Divakaran, A. (2020). Stacked spatio-temporal graph convolutional networks for action segmentation. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 565-574. https://doi.org/10.1109/wacv45572.2020.9093361

[37] Zhang, Y., Li, Y., Wei, X., & Jia, L. (2020). Adaptive spatio-temporal graph convolutional neural network for remaining useful life estimation. 2020 International Joint Conference on Neural Networks (IJCNN), 1-7. https://doi.org/10.1109/ijcnn48605.2020.9206739

[38] Ali, F., Ali, A., Imran, M., Naqvi, R. A., Siddiqi, M. H., & Kwak, K. S. (2021). Traffic accident detection and condition analysis based on social networking data. Accident Analysis & Prevention, 151, 105973. https://doi.org/10.1016/j.aap.2021.105973

[39] Ajith, A., Philip, A. O., Sreedhar, S., & Sreeja, M. U. (2022). Road accident detection from CCTV footages using deep learning. 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), 1354-1361. https://doi.org/10.1109/icosec54921.2022.9951920

[40] Ji, X., & Meng, Q. (2020). Traffic classification based on graph convolutional network. 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), 596-601. https://doi.org/10.1109/aeeca49918.2020.9213630