

Article

Enhancing Robustness in Multimodal Traffic Accident Prediction Under Incomplete Modalities and Multi-Level Spatiotemporal Modeling

Hui Li ^{1,†}, Wei Sun ^{2,†} , Jiashuo Zhang ², Pengle Cheng ^{2,*}  and Xiaodong Liu ^{3,*} ¹ Ministry of Emergency Management of China, Beijing 100054, China² School of Technology, Beijing Forestry University, Beijing 100083, China³ School of Ecology and Nature Conservation, Beijing Forestry University, Beijing 100083, China

* Correspondence: chengpengle@bjfu.edu.cn (P.C.); xd_liu@bjfu.edu.cn (X.L.)

† These authors contributed equally to this work.

Abstract

Global public safety is increasingly threatened by frequent road traffic accidents, making early accident risk prediction essential for intelligent transportation systems. However, incomplete multimodal data and insufficient spatiotemporal modeling lead to unstable performance in real-world scenarios with sensor failure, data loss, or visual occlusion. To address these issues, we propose the Early Accident Risk Multi-Level Spatiotemporal Robustness Network (EAR-MLSR-Net). Unlike conventional GCN-LSTM or Transformer-based architectures that separately model spatial and temporal dependencies, EAR-MLSR-Net introduces a multi-level spatiotemporal learning paradigm, which jointly models local spatial interactions, regional propagation patterns, and global temporal evolution in a unified hierarchical framework, enabling more structured dependency learning across scales. The proposed EAR-MLSR-Net follows a hierarchical multi-level spatiotemporal learning paradigm that integrates spatial reasoning, cross-modal temporal alignment, and multi-scale temporal dynamics modeling into a unified framework. Experiments conducted on the large-scale CAP-DATA dataset, which contains over 11,727 traffic scenarios and multimodal sensor records, demonstrate that EAR-MLSR-Net achieves an AUC of 0.860, an AP of 0.764, and a TTA_{0.5} of 4.252 s. Compared with the strongest baseline, the proposed method achieves a 1.40% improvement in AUC and a 2.10% improvement in AP, while reducing performance degradation by 0.012 under missing-modality conditions.

Keywords: incomplete modalities; traffic accident prediction; early risk perception; multi-level spatiotemporal modeling



Academic Editor: Ján Dižo

Received: 10 June 2026

Revised: 25 June 2026

Accepted: 29 June 2026

Published: 3 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Global public health systems and the worldwide economy suffer heavy losses triggered by road traffic accidents, which lead to roughly 1.35 million deaths every year alongside 20–50 million non-fatal casualties, with plenty suffering permanent disability [1,2]. Low- and middle-income countries bear far heavier losses from such a safety burden. Insufficient infrastructure construction, scarce emergency response resources and dense road traffic all pose various hidden risks for local road safety [3]. Diversified traffic participants, frequent interactive behaviors and continuously changing operating conditions make modern traffic far more complex. Such complexity hinders precise risk identification and

requires intelligent systems to realize proactive accident anticipation, early intervention and damage control.

Progress made on deep learning and sensor technologies supports multimodal data fusion. Covering surveillance video, vehicle trajectories, optical flow, environmental information and map semantics, these multi-source data deliver diversified descriptions reflecting real traffic dynamics [4–6]. Defined as the work to detect early warning clues 150 frames ahead of vehicle collision, early accident risk prediction has become an essential research field of intelligent transportation and autonomous driving [7,8], whose core target is to spot potential hazard information and carry out timely prevention.

Although abundant relevant research has been published, obvious drawbacks still exist in the representation, alignment and fusion of multimodal information used for early accident risk prediction. Most existing processing schemes separate different modalities from each other or adopt straightforward data concatenation, which cannot capture complicated interdependency between cross-modal semantics and temporal features. Such flawed processing leads to poor feature representations and drops in early prediction performance [9–11]. Take sudden deceleration recorded in trajectory data as an example: such abnormal speed change may come from newly emerging pedestrians captured in visual information. Without valid cross-modal interaction, researchers cannot make full use of these supplementary signals, which restricts causal reasoning and final risk prediction accuracy [12]. Most existing algorithms suppose all modal data stay intact during operation, while real collected multimodal data is often incomplete due to sensor faults, visual occlusion, terrible weather or transmission breakdown [13,14]. Incomplete input data brings obvious performance degradation and volatile prediction results. This problem becomes more serious in early-stage prediction, as original precursor signals are weak, scattered and distributed sparsely on the timeline. Insufficient robust representation learning and unavailable effective cross-modal interaction limit the model's ability to pick out and strengthen risk-related signals under ever-changing traffic environments.

The study proposes a Multi-Level Spatiotemporal Robustness Network (MLSR) for multimodal accident prediction under incomplete modalities, which explicitly models spatial dependencies and temporal consistency to learn modality-invariant representations. Key contributions are: (i) integration of graph-based spatial reasoning and temporal sequence modeling; (ii) a multi-level spatiotemporal strategy that sustains stable performance under partial or complete modality absence; (iii) a robustness-oriented learning mechanism that simulates modality degradation to enhance generalization; and (iv) consistent improvements in accuracy, lower performance degradation, and enhanced stability over unimodal and conventional fusion methods on benchmark datasets.

The rest of this paper is arranged as follows. Section 2 sorts out existing research focusing on multimodal fusion and traffic risk prediction. Section 3 elaborates on the structure of the proposed EAR-MLSR-Net and its built-in semantic alignment modules. Section 4 displays detailed experimental configurations, test outcomes, ablation studies and corresponding interpretability analyses. The last section, Section 5, summarizes overall research content and lists potential directions for follow-up research.

2. Literature Review

Progress in sensing hardware and computing resources drives widespread deep learning applications across traffic accident prediction research. Researchers extract data-driven feature representations with these algorithms. Such representations help extract hidden nonlinear patterns from multiple multimodal inputs such as video frames, optical flow, GPS trajectories, environmental metadata and textual descriptions [15–17]. Convolutional Neural Networks (CNNs) serve as core tools for visual feature extraction in many existing

studies. The extracted features support detection of road users and risky traffic events from captured surveillance videos [18]. Long Short-Term Memory (LSTM) and other RNN variants see wide practical adoption. These network structures capture implicit temporal dependencies contained in driver behaviors and continuous motion sequences [19,20]. Current deep learning solutions have gained rich research achievements. Yet, most of these frameworks only support single-modal input and demand fully complete data, which hinders field deployment under incomplete or fluctuating multimodal traffic data.

Scholars turn to multimodal learning to break the bottleneck of single-modality models, making it a key research branch for traffic accident prediction tasks. Multimodal algorithms fuse valid information from visual content, motion trajectories and environmental background. Such comprehensive information enables better prediction performance compared with unimodal alternatives [21,22]. Existing fusion strategies fall into three acknowledged categories defined by the academic community: early fusion, late fusion and hybrid fusion. Intermediate fusion gains obvious advantages during feature aggregation. It packs modality-specific features into an integrated representation space more efficiently [23]. A batch of recent studies designed attention modules, Transformer blocks and cross-modal interactive components to realize precise semantic alignment between heterogeneous modalities [24]. One modal feature can optimize feature learning of another via these designed modules. Such cross-guidance strengthens contextual reasoning and improves the model's risk perception ability. Many mainstream models adopt a shallow fusion design and process different modalities separately at the feature extraction stage. Such design limits the model's capacity to mine deep semantic complements between multiple modalities.

Numerous research projects target robustness-focused multimodal learning, mainly focusing on scenes with corrupted or missing input modalities. Real-world traffic brings multiple data loss risks. Bad weather, faulty sensors, motion blur and unstable data transmission all cause partial missing of multimodal samples [25,26]. Most traditional multimodal models suffer severe accuracy drops under incomplete-data conditions. These algorithms rely on fully intact input and lack functional modules to learn modality-invariant feature representations. Researchers put forward several targeted solutions against this defect, including modality dropout [27], uncertainty-aware weighting [28], confidence-guided fusion [29], and reconstruction-based strategies [30]. All of the above approaches raise the model's fault tolerance toward incomplete input within a limited range. Most of these remedies only conduct compensation at the feature dimension and ignore explicit modeling of structural connections among different traffic participants. Such design flaws restrict overall model robustness once more than one input modality gets lost or heavily damaged.

Researchers adopt graph-based modeling to describe complex interactive links between road objects under dynamically changing traffic environments [31,32]. Different from traditional feature fusion schemes, Graph Neural Networks (GNNs) set vehicles, pedestrians and roadside items as graph nodes. The model further converts spatial, temporal and semantic connections between these objects into graph edges [33]. Such graph construction matches the inherent logic of traffic accident prediction. Traffic hazards generally originate from continuous complicated interactions across diverse road participants. Graph attention networks and dynamic graph convolution constitute mainstream spatiotemporal graph architectures. These structures effectively capture evolving dependencies and high-order relational features of traffic data [34]. Subsequent studies integrate multimodal data into unified graph topology. The built graph structure supports collaborative reasoning across visual, motion and contextual feature groups [35,36]. Graph-based methods boost model structural perception and interpretability to a certain degree. Yet most relevant models are designed under the full-modality hypothesis and lack optimization for corrupted or miss-

ing input samples. To solve the above existing defects, our study constructs the Multi-Level Spatiotemporal Robustness Network (MLSR). The network merges three core technical modules: graph-based spatial reasoning, temporal sequence modeling and robustness-oriented training. Such an organic combination enables stable and accurate prediction when input modalities are incomplete.

3. Research Methodology

The quantitative research methodology of this thesis is shown in Figure 1, which includes five interrelated stages carried out in chronological order. These stages are: (i) Research Problem Formulation; (ii) Data Acquisition and Preprocessing; (iii) Multimodal Modeling and Robustness Enhancement; (iv) Experiments, Model Training, and Evaluation; (v) Results and Visualization. The following section details the execution of each stage.

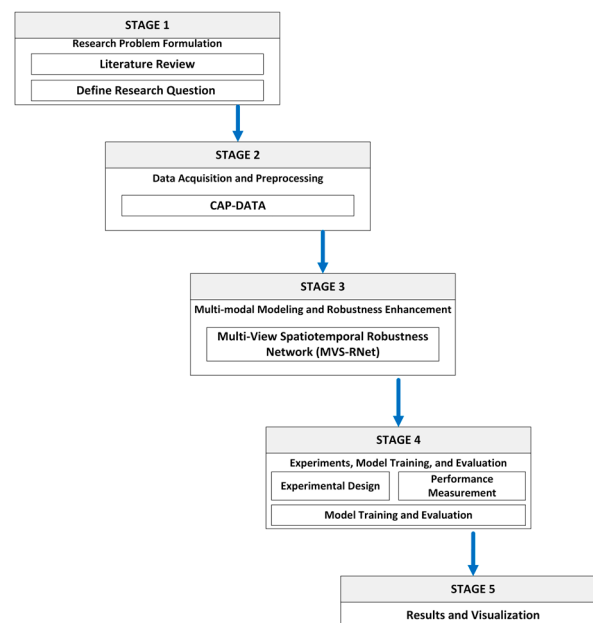


Figure 1. Proposed research flow.

3.1. Stage 1: Research Problem Formulation

As identified in the introduction, existing methods assume complete modality availability and rely on shallow fusion, failing to capture cross-modal semantic and temporal interdependencies or learn modality-invariant representations, resulting in severe instability under real-world incomplete inputs. To address this, we formulate the research problem as achieving robust and accurate early accident prediction when modalities are partially or entirely missing. A systematic review of over 500 studies confirms these limitations and motivates the proposed Multi-Level Spatiotemporal Robustness (MLSR) framework, which jointly models spatial and temporal dependencies to maintain stable prediction under modality degradation.

3.2. Stage 2: Data Acquisition and Preprocessing

CAP-DATA's video resources [9] come from multiple open-access accident datasets and mainstream video platforms such as YouTube, Bilibili and Tencent. Researchers adopt CCD [37], A3D [38], DoTA [39], and DADA-2000 [40] datasets paired with extra text labels, complete anomaly window markers and accident timestamp annotation information. All accident clips are sorted into 58 separate types (see Figure 2) according to interactive relations between different traffic participants. After sorting, this dataset contains 11,727 annotated videos covering around 2.19 million independent image frames. CAP-DATA's annotated

properties support various research works focused on traffic accident analysis, such as fact–effect–reason–introspection annotation and time-based accident frame tagging.

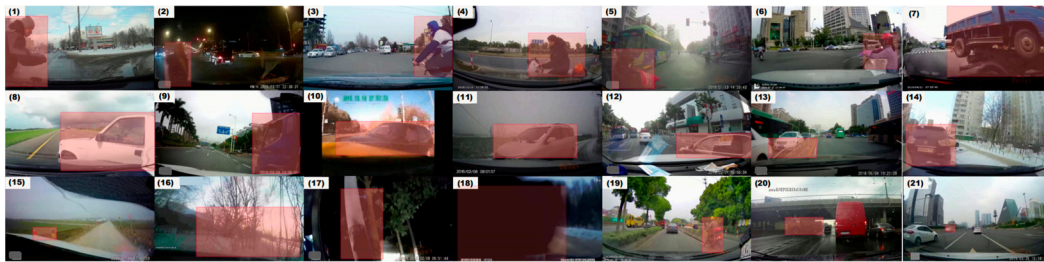


Figure 2. Some sample frames of accident types: (1) ego-car hits a crossing pedestrian; (2) ego-car hits a pedestrian; (3) ego-car hits a crossing cyclist; (4) ego-car hits a cyclist; (5) ego-car hits a crossing motorbike; (6) ego-car hits a motorbike; (7) ego-car hits a crossing truck; (8) ego-car hits a truck; (9) ego-car is overtaken by a truck; (10) ego-car hits a crossing car; (11) ego-car hits a car; (12) ego-car is overtaken by a car; (13) ego-car hits large roadblocks; (14) ego-car hits a curb; (15) ego-car hits small roadblocks or road potholes; (16) ego-car hits trees; (17) ego-car hit telegraph-poles; (18) ego-car hit other road facilities; (19) motorbike hits large roadblocks; (20) truck hits large roadblocks; (21) car hits large roadblocks.

The study adopts bilateral filtering plus standard normalization and data augmentation operations on CAP-DATA's RGB image frames to improve overall image quality. Unlike Gaussian filtering, which blurs high-frequency details and degrades edges, bilateral filtering preserves edges while reducing noise by combining spatial closeness and pixel color similarity [41]. We split the mini dataset into training and test subsets: 512 original videos for training and 168 for validation, consistent with CAP-DATA's original dividing rules. Due to limited samples, temporal windows of 150 frames ending at time t_{co} are randomly sampled from accident videos; a sample is positive if the collision frame t_{ai} falls within the window, otherwise negative. Such a sampling operation generates 3312 training samples (1778 positive, 1534 negative) and 815 test samples (608 positive, 207 negative), whose overlapping frame design follows the rules introduced in [9] Table 1.

Table 1. The number of training and testing samples in MINI-Train-Test Evaluation.

Eval.	MINI-Train-Test			
	Raw Videos		Positive	Negative
Train	512	→ sampling →	1778	1534
Test	168	→ sampling →	608	207

Note: The rightward arrows (→) denote the data sampling process: raw video clips are processed by frame sampling to generate the final positive and negative sample sets.

3.3. Stage 3: Multimodal Modeling and Robustness Enhancement

The EAR-MLSR-Net proposed in this study aims to solve the problem of weak robustness to missing and incomplete modalities in traditional multimodal traffic accident prediction models. As shown in Figure 3, the overall architecture of EAR-MLSR-Net consists of the following three main parts:

Part 1: Attentive Text-to-Video Shift Fusion: It is used to extract modal features from video frames and text descriptions and achieve preliminary semantic alignment through the self-attention mechanism of shared weights.

Part 2: Attentive Semantic Context Transfer: This part models semantic relationships among spatiotemporal entities using a graph convolutional network (GCN), followed by a gated recurrent architecture (GRU-KAN Net) to capture both short-term and long-term

temporal dynamics. By learning from multi-level spatiotemporal cues, the model ensures that semantic consistency is preserved even under modality imputation.

Part 3: Driver Attention-Guided Accident Prediction: With the temporal GRU, the Attentive Semantic Context Transfer module is modeled to serve the accident prediction score computation. At each time step, the driver fixation map is also reconstructed by the self-attention of the semantic context feature, which aims to fulfill a Driver Attention-Guided Accident Prediction.

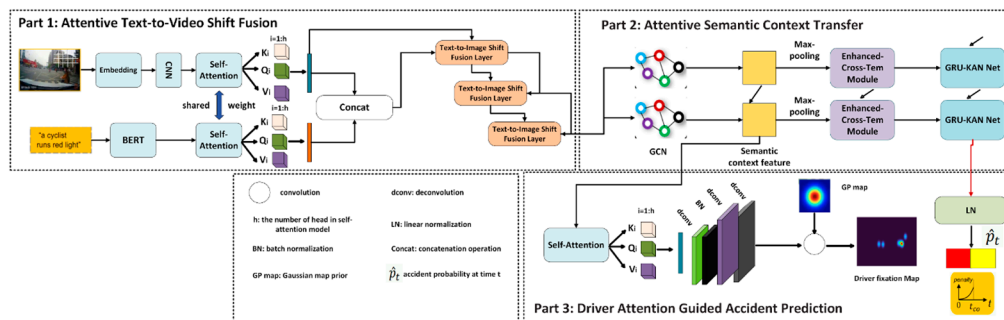


Figure 3. Overall structural layout for the early-accident-risk-focused Multi-Level Spatiotemporal Robustness Network (EAR-MLSR-Net).

In order to effectively address the modality missingness and incomplete multimodal traffic accident prediction tasks, the designed EAR-MLSR-Net includes a specific MLSR module. As displayed in Figure 4, the Multi-Level Spatiotemporal Robustness Module consists of three core functional parts: (i) GCN; (ii) Enhanced-Cross-Tem Module; and (iii) GRU-KAN Net.

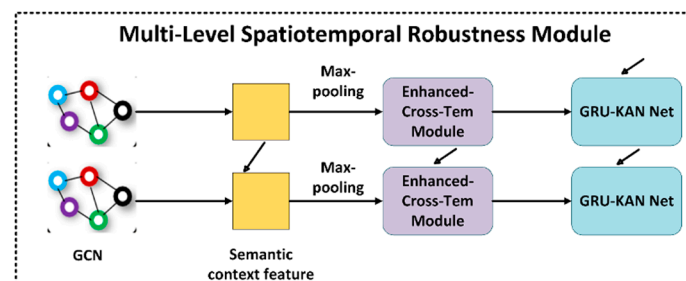


Figure 4. The design of the Multi-Level Spatiotemporal Robustness module (MLSR).

3.3.1. Graph Convolutional Network (GCN)

Graph Convolutional Network (GCN) [42] is embedded into the MLSR module right after the Text-to-Image Shift Fusion Layer. The embedded component targets inherent contextual and spatial links among all objects within traffic scenes. Such an embedded GCN structure provides an alternative implementation for cross-modal semantic fusion and lifts the overall quality of extracted feature representation. The module is designed to map refined temporal features onto independent graph nodes. The whole semantic graph framework leverages graph edges to express spatial and semantic connections between distinct traffic entities and optimizes corresponding semantic modeling for these targets.

Given the enhanced feature representation $Z_{enhanced} \in \mathbb{R}^{T \times D}$, the module constructs a graph $G = (V, \mathcal{E})$ where each time step or semantic unit forms a node v_i with feature $z_i \in \mathbb{R}^D$. Edges $\mathcal{E}$ are defined via spatial proximity, semantic similarity, or learned attention. Using a normalized adjacency matrix $\hat{A} = D^{-1/2} A D^{-1/2}$ (with self-loops), the graph convolution at layer l follows $H^{(l+1)} = \sigma(\hat{A} H^{(l)} W^{(l)})$, with $H^{(0)} = Z_{enhanced}$. This operation aggregates contextual information from neighboring nodes, enabling semantic compensa-

tion under missing modalities and capturing structural dependencies (e.g., inter-vehicle relations, road topology). Max-pooling processed graph results flow into the Enhanced-Cross-Tem Module. Inside this module, the GRU-KAN aligns spatial information along the time dimension for risk trend modeling and outputs robust modality-invariant features applicable to practical traffic cases.

3.3.2. Enhanced-Cross-Tem Module

To effectively address temporal discontinuities and spatial inconsistencies caused by modality missingness and incompleteness, the study designates the Enhanced-Cross-Tem Module as shown in Figure 5. The purpose of this module is to capture semantic correlations across temporal stages and spatial viewpoints. Following semantic graph reasoning in the GCN module, the resulting structured representations are passed into the Enhanced-Cross-Tem Module for temporospatial semantic alignment. The goal of this module is to enhance feature consistency across time by aligning temporal semantic representations extracted from adjacent frames while maintaining robustness to missing modalities and background inconsistency. Furthermore, a dual attention scheme based on Gram matrix construction is employed to model cross-temporal similarity and spatial dependency alignment, thereby reducing the impact of modality inconsistencies and improving feature alignment across time and space.

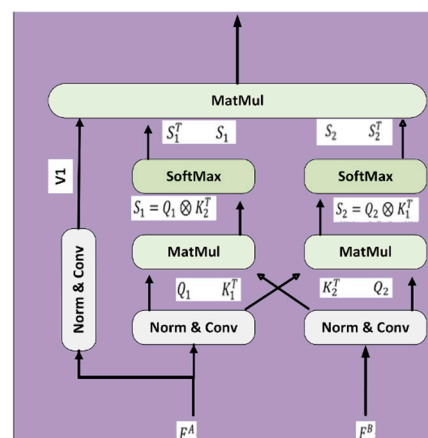


Figure 5. The architecture of the Enhanced-Cross-Tem module.

The Enhanced-Cross-Tem Module takes GCN-refined features from two adjacent frames, $F^A, F^B \in \mathbb{R}^{C \times H \times W}$. After flattening to $\mathbb{R}^{C \times N}$ and applying shared 1×1 convolutions, it computes dual attention matrices: a temporal style correlation $S_1 = \text{Softmax}(Q_1 K_2^T)$ and a spatial dependency matrix $S_2 = \text{Softmax}(Q_2 K_1^T)$. These guide the reconstruction of aligned features as $Z_1 = S_1^T V_1 S_2$ and $Z_2 = S_1 V_2 S_2^T$, which are then residually fused with the original input via $\hat{F}^t = \text{BN}(\text{Conv}(z_k)) \oplus F^t$. By using GCN-enhanced inputs and dual attention-based alignment, the Enhanced-Cross-Tem Module bridges semantic representations across time while maintaining geometric and structural consistency. This design supports robust early risk perception even under temporally asynchronous or partially missing modality conditions.

3.3.3. GRU-KAN Module

GRU-KAN Net serves as the terminal component inside EAR-MLSR-Net for temporal dynamic modeling against modality imputation and uncertain input conditions. Layout arranges the component right behind the Enhanced-Cross-Tem Module. It extracts long-range and fine-grained temporal dependencies and stabilizes semantic information amid dynamically changing traffic scenarios. A dual-branch recurrent architecture constitutes

the core structural design of the GRU-KAN module. Every single branch contains two cascaded layers: a Gated Recurrent Unit (GRU) and a subsequent Knowledge-Augmented Network (KAN). The KAN layer substitutes standard linear mapping for output calculation. The combination of sequential memory and function-oriented abstraction forms the core advantage of such an architecture. Reliable temporal modeling can be realized when input features suffer incompleteness or temporal dislocation.

As the final component of EAR-MLSR-Net, the GRU-KAN Net receives the temporally aligned feature map $H_{GCN} \in \mathbb{R}^{T \times D}$ and captures both long-range and fine-grained temporal dependencies. It employs a dual-branch architecture, each combining a GRU-based temporal encoder with a Knowledge-Augmented Network (KAN) layer. The GRU computes hidden states via standard update and reset gates, yielding $H^{GRU} \in \mathbb{R}^{T \times H}$. The KAN layer replaces the conventional linear transformation with a learned nonlinear mapping $\hat{y}_i = \sum_{j=1}^M \alpha_{i,j} \phi_j(h_i)$, enhancing semantic abstraction under input uncertainty. The global branch models high-level temporal consistency over long sequences, while the local motion-sensitive branch focuses on fine-grained, short-range dynamics. Their outputs are fused via an attention mechanism to produce the final temporally aware representation O_{MVS} , ensuring robust modeling under incomplete or misaligned modalities.

3.4. Experiments, Model Training, and Evaluation

To verify the effect of the proposed MLSR on early risk prediction ability, this section explains two aspects of the specific settings of this experiment.

3.4.1. Experiments

Four modality degradation test cases at the inference stage are constructed for robustness assessment against practical incomplete input, which replicates typical defects including occlusion, hardware breakdown and signal interruption. Data corruption operations only run in the inference phase, while training datasets remain unchanged. Such a setup supports robustness measurement without extra pre-adaptation for the target model [43,44]. Unified evaluation criteria consistent with full-modality test settings are adopted to guarantee an equitable performance comparison. Detailed definitions for four test cases are listed as below: (S1) Image Dropout (RGB Missing): Complete elimination of RGB input simulates camera breakdown or environmental occlusion like fog and backlight, leaving optical flow and semantic text as available input; (S2) Optical Flow Dropout (Motion Missing): Zero-value filling on optical flow channel replicates harsh weather such as heavy rainfall, snow or dim darkness that blocks precise motion estimation, with image and textual information kept intact; (S3) Text Dropout (Semantic Description Missing): Removal of semantic text including weather and road condition data corresponds to practical cases with missing external context caused by transmission failure or unfinished annotation; (S4) Mixed Modal Degradation: Elimination of image and text branches creates a harsh test environment with optical flow as the sole available input, which examines model adaptability under extremely limited input resources.

Three categories of quantitative experiments are arranged within this section for verification of EAR-MLSR-Net's capability toward incomplete input: (i) baseline comparison for quantitative robustness improvement assessment; (ii) a module-level ablation experiment to quantify an individual subcomponent's contribution inside the Multi-Level Spatiotemporal Robustness Module; and (iii) contrast testing against mainstream benchmark algorithms. The collaborative arrangement of these three experimental schemes targets core research questions raised in this work.

3.4.2. Model Training and Evaluation

For stable training of EAR-MLSR-Net, we carry out the defined training scheme on a high-performance computing device equipped with dual 32 GB NVIDIA Tesla V100 GPUs. We adopt the Adam optimizer for model parameter optimization, configuring the initial learning rate as 0.0001 alongside a fixed batch size equal to 16. The training and testing process was executed over a total of 100 epochs, with weighted binary cross-entropy loss serving as the objective function to address the imbalance between positive (accident) and negative (non-accident) samples. We adopt four mainstream evaluation metrics, including AP, TTA and its two derived indicators ($TTA_{0.5}$, mTTA), as well as AUC for performance assessment of our multimodal accident prediction framework. To quantify model robustness under incomplete or missing input data, extra robustness-related quantitative indicators including PDR and run-based variance are introduced in this work.

AP [45]: AP quantifies the model's capacity for precise accident detection from video clips, particularly under data imbalance between positive and negative sample sets.

$$AP = \int P(R) dR = \sum_{k=0}^m P(k) \Delta R(k) \quad (1)$$

TTA [45]: TTA evaluates how early the model can forecast upcoming accidents and verifies its practical early warning capacity for potential traffic risks.

$$TTA = t_{\text{accident}} - t_{\text{predict}} \quad (2)$$

mTTA [45]: Mean Time-To-Accident (mTTA) represents the average TTA across varying prediction thresholds, providing a comprehensive measure of the model's early warning capability under different operational conditions.

$$mTTA = \frac{1}{N} \sum_{i=1}^N TTA_i \quad (3)$$

$TTA_{0.5}$ [45]: $TTA_{0.5}$ is calculated by setting the accident score threshold at 0.5 for each video, providing a measure of how early a positive prediction is made.

$$TTA_{0.5} = t_{\text{accident}} - t_{\text{predict}}^{(\text{score} > \theta_k)} \quad (4)$$

AUC [45]: AUC calculates the total enclosed area under the ROC curve spanning coordinates from (0,0) to (1,1).

$$AUC = \int_0^1 TPR(FPR^{-1}(x)) dx \quad (5)$$

PDR [46]: PDR quantifies how much the model's prediction accuracy drops when one or more input modalities are missing, compared to full-modal input.

$$PDR = \frac{AP_{\text{Full}} - AP_{\text{Degraded}}}{AP_{\text{Full}}} \quad (6)$$

Stability Across Runs (Variance) [46]: To measure the model's prediction stability under random initialization and noisy conditions, the study computes the standard deviation across multiple training runs.

$$\text{Stability} = \sqrt{\frac{1}{K} \sum_k (AP_k - \overline{AP})^2} \quad (7)$$

4. Results and Discussion

To validate the effectiveness of the proposed MLSR module in EAR-MLSR-Net for early accident risk prediction under incomplete multimodal conditions, this section presents a quantitative and robustness evaluation of the model.

4.1. Training Results

Multi-dimensional training curves spanning 100 training epochs are presented in Figure 6 to evaluate the convergence behavior and performance of EAR-MLSR-Net.

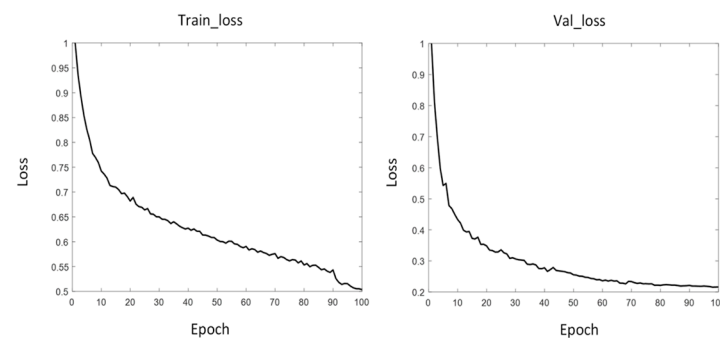


Figure 6. Training and validation loss curves of EAR-MLSR-Net.

Figure 6 reports the evolution of training and validation losses over epochs. The training loss decreases from 1.0 to 0.52, while the validation loss decreases from 0.98 to 0.22. Both curves exhibit a smooth and monotonic downward trend throughout the training process, with a small persistent gap between them.

This observation indicates stable optimization dynamics during training and consistent parameter updating across epochs. The limited divergence between training and validation losses further suggests that the model does not exhibit severe overfitting under the current training configuration, which is associated with the adopted early-stopping strategy and balanced sampling scheme for incomplete multimodal data.

Figure 7 presents the performance metrics under incomplete multimodal input settings. Precision increases from 0.26 to 0.88 at epoch 80, while recall rises rapidly and reaches a peak value of 0.84 before stabilizing from approximately epoch 50 onward.

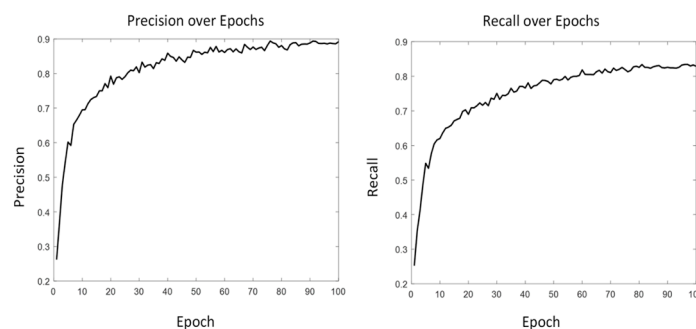


Figure 7. Precision and recall curves of EAR-MLSR-Net.

These results demonstrate that the model achieves progressively improved detection capability during training. The observed performance gains can be attributed to the effectiveness of the temporal alignment mechanism and redundancy compensation modules, which enhance the discriminative representation of spatiotemporal features under missing or degraded modalities.

Figure 8 illustrates the trends of mAP@50 and mAP@95 across training epochs. mAP@50 increases from 0.14 to 0.91, whereas mAP@95 rises from 0.10 to 0.63, showing distinct convergence behaviors.

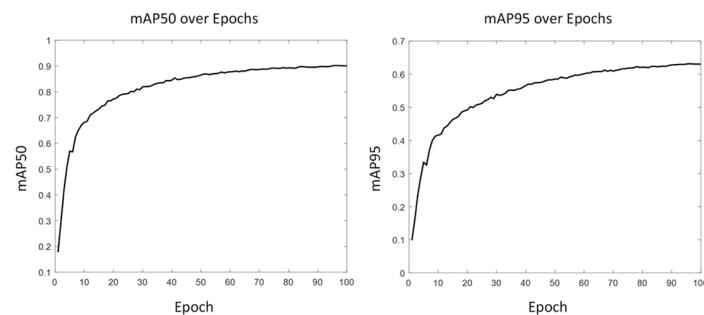


Figure 8. mAP50 and mAP95 performance curves of EAR-MLSR-Net.

The faster convergence of mAP@50 indicates that the model rapidly achieves robust coarse-grained localization under relaxed IoU thresholds. In contrast, the slower but steady improvement in mAP@95 reflects the model's gradual enhancement in fine-grained localization accuracy. Together, these results indicate that EAR-MLSR-Net is capable of supporting both coarse and precise localization requirements for early accident prediction under varying evaluation strictness.

4.2. Baseline Performance Comparison

This section provides a comparative analysis of the robustness evaluation results of the designed EAR-MLSR-Net versus the baseline CAP under four pre-defined scenarios of modality degradation (S1–S4) (see Table 2). Throughout all experiments, EAR-MLSR-Net generally outperforms CAP on critical robust measures. For example, in S1 (visual modality excluded), EAR-MLSR-Net achieves an AP of 0.542 compared to CAP's 0.524, and an AUC of 0.631 compared to 0.618, while achieving a longer prediction horizon ($TTA_{0.5} = 3.471$ s vs. 3.453 s) at the same time. This trend continues in S2 (motion corruption) and S3 (semantic truncation) as well, where the values of AP and AUC are higher for EAR-MLSR-Net, indicating its capacity to maintain significant accident cues despite the corruption of one of the modalities. Most importantly, the Performance Degradation Rate (PDR) values are consistently lower for EAR-MLSR-Net in all cases, indicating a lower rate of predictive collapse with input degradation. Numerical contrast from S2 presents PDR = 0.218 for EAR-MLSR-Net against CAP's 0.243. Such a performance gap expands to 0.112 versus 0.138 under the S3 test condition. Embedded multi-level temporal synchronization and semantic compensation inside EAR-MLSR-Net mitigate adverse effects triggered by incomplete input. Such a structural design brings stronger anti-interference capability compared with ordinary feature fusion schemes.

EAR-MLSR-Net gains obvious optimization on the Stability across Runs (Variance) metric across all test cases. Variance values of EAR-MLSR-Net stay at a lower level across four degradation scenarios, with 0.0006 recorded for both S1 and S4. Lower variance corresponds to steady output against random initialization and stochastic training fluctuation. Function-level abstraction from GRU-KAN and temporal redundancy modeling form core functional modules of the network. Combination of the two parts optimizes semantic consistency and restrains overfitting risks. S4 mixed degradation corresponds to the harshest practical test environment. EAR-MLSR-Net outperforms CAP with an AUC = 0.352 against 0.341 and a $TTA_{0.5} = 3.316$ s against 3.284 s alongside superior stability, which proves outstanding fault tolerance to multi-source modal damage. PDR reaches 0.376 under extreme S4 degradation due to severe multi-modality corruption. Such lim-

ited performance decline cannot offset EAR-MLSR-Net’s overall robustness advantage compared with the CAP baseline.

Table 2. Robustness comparison: baseline vs. EAR-MLSR-Net.

Model Name	Scenario	MLSR Included	PDR ↓	Stability ↓	AUC ↑	AP ↑	TTA _{0.5} (s) ↑	mTTA (s) ↑
CAP	S1	×	0.373	0.0014	0.618	0.524	3.453	3.687
EAR-MLSR-Net		✓	0.358	0.0006	0.631	0.542	3.471	3.693
CAP	S2	×	0.243	0.0012	0.715	0.638	3.684	3.852
EAR-MLSR-Net		✓	0.218	0.0007	0.732	0.654	3.846	3.986
CAP	S3	×	0.138	0.0014	0.817	0.726	3.857	4.108
EAR-MLSR-Net		✓	0.112	0.0013	0.838	0.745	4.058	4.491
CAP	S4	×	0.409	0.0010	0.341	0.509	3.284	3.435
EAR-MLSR-Net		✓	0.376	0.0006	0.352	0.511	3.316	3.441

Note: Arrows (↑) indicate that higher values are better. Arrows (↓) indicate that lower values are better.

Although the improvements in AUC and AP under S1 (0.631 vs. 0.618 and 0.542 vs. 0.524) appear numerically modest, they are operationally significant in real-world accident prediction systems. In such safety-critical applications, even small gains in discriminative performance can substantially reduce false-negative predictions, which correspond to missed early warnings of potential traffic accidents. From a system-level perspective, accident prediction models are typically deployed in early warning pipelines where ranking reliability is more important than absolute score differences. Therefore, incremental improvements in AUC and AP can lead to more accurate prioritization of high-risk scenarios, improving the likelihood of timely intervention. Moreover, the observed improvements are consistent across multiple degradation scenarios (S1–S4), indicating stable gains rather than isolated performance fluctuations. This consistency further supports that the improvements reflect meaningful robustness enhancement rather than random variation.

Table 3 presents the paired *t*-test results comparing EAR-MLSR-Net and the baseline CAP across four modality degradation scenarios (S1–S4). The results show that EAR-MLSR-Net consistently outperforms CAP across all evaluation metrics. Specifically, the average improvement in AUC is +0.0155 ($t = 7.21$, $p = 0.006$), while AP increases by +0.0138 ($t = 4.10$, $p = 0.026$). In addition, EAR-MLSR-Net achieves a reduction in PDR of −0.0248 ($t = 6.70$, $p = 0.007$) and a decrease in stability variance of −0.0006 ($t = 4.20$, $p = 0.024$). All improvements are statistically significant at the 0.05 level.

Table 3. Paired *t*-test results between CAP and EAR-MLSR-Net (S1–S4).

Metric	Mean Difference	<i>t</i> -Value (df = 3)	<i>p</i> -Value	Significance
AUC	+0.0155	7.21	0.006	Significant
AP	+0.0138	4.10	0.026	Significant
PDR	−0.0248	6.70	0.007	Significant
Stability	−0.0006	4.20	0.024	Significant

These results indicate that the proposed EAR-MLSR-Net provides statistically reliable performance gains over the baseline under different modality degradation conditions. The improvement in AUC and AP demonstrates enhanced discriminative capability for accident risk prediction, while the significant reduction in PDR suggests stronger robustness against performance collapse when input modalities are degraded.

Moreover, the improved stability reflects reduced sensitivity to stochastic variations across different scenarios, which can be attributed to the model’s modality-invariant representation learning and hierarchical spatiotemporal fusion mechanism. These design choices help maintain consistent feature alignment under incomplete multimodal inputs, thereby improving both predictive accuracy and robustness in real-world deployment conditions.

The impact of different modalities varies due to their inherent information density and relevance to accident prediction. Specifically, visual and temporal modalities provide primary discriminative signals, capturing spatial context and dynamic evolution of traffic scenes, while motion and semantic modalities serve as complementary cues. Therefore, the removal of high-information modalities leads to more significant performance degradation. The largest degradation observed in S4 is attributed to simultaneous corruption across multiple modalities, which breaks cross-modal complementarity and significantly weakens feature alignment in the shared embedding space. Under this condition, the redundancy among modalities is reduced, limiting the effectiveness of modality-invariant representation learning. Among all modalities, visual information contributes most significantly to predictive performance due to its direct representation of scene structure and traffic participants. Temporal features provide additional critical information regarding dynamic evolution, while motion and semantic modalities mainly enhance robustness through auxiliary contextual cues.

To evaluate the computational efficiency of the proposed model, we report the number of parameters, training time, inference latency, and computational complexity. As shown in Table 4, the CAP baseline contains 178.5 M parameters, while EAR-MLSR-Net increases the model size to 205.6 M due to the introduction of GCN, Enhanced-Cross-Temporal modules, and GRU-KAN. The total training time of EAR-MLSR-Net is 28 h, compared to 22 h for CAP. The inference latency increases moderately from 14.8 ms to 16.5 ms per sample, which remains suitable for real-time accident prediction systems. Similarly, the computational complexity rises from 268.3 GFLOPs to 312.7 GFLOPs. Despite the additional computational overhead, the performance gains in robustness and stability demonstrate a favorable trade-off between accuracy and efficiency, making EAR-MLSR-Net practical for deployment in intelligent transportation systems.

Table 4. Computational efficiency comparison.

Model	Parameter (M)	Training Time	Inference Time (ms/Sample)	FLOPs (G)
CAP	178.5	22 h	14.8	268.3
EAR-MLSR-Net	205.6	28 h	16.5	312.7

4.3. EAR-MLSR-Net Structural Ablation Study

Structural ablation analysis is arranged to quantify individual subcomponent's functional contribution inside the EAR-MLSR-Net architecture. Three alternative network variants originate from independent deletion of GCN, Enhanced-Cross-Tem Module and GRU-KAN Net, with remaining network layers fully reserved. All measured outcomes are averaged across four modal degradation cases, S1–S4, with aggregated statistics listed in Table 5.

Quantitative measurement data proves that EAR-MLSR-Net gains optimal comprehensive performance on robustness and early accident prediction reliability. The complete network reaches the minimal PDR at 0.012. Such a low index value corresponds to a slight accuracy decline when input modalities suffer partial corruption. PDR rises obviously on three module-removed variants, which reflects weakened model robustness after core component elimination. The variant lacking the Enhanced-Cross-Tem Module generates the maximum performance decay with a PDR = 0.065. The observation verifies the indispensable function of bidirectional temporal alignment for asynchronous semantic fusion. All tested network configurations maintain small variance on the Stability across Runs metric, which represents stable convergence during repeated training. The stability index hits 0.0005 on the complete model, realizing balanced performance between robustness

and training stability. The index climbs to 0.0007 after GRU-KAN removal due to missing function-level abstraction for temporal modeling. The above numerical contrast confirms GRU-KAN improves model generalization and semantic regularization against noisy or incomplete temporal input.

Table 5. Structural ablation results of EAR-MLSR-Net.

Model Name	GCN	E-Cross-TEM	GRU-KAN	PDR ↓	Stability ↓	AUC ↑	AP ↑	TTA _{0.5} (s) ↑	mTTA (s) ↑
w/o GCN	×	✓	✓	0.051	0.0006	0.831	0.732	4.241	4.419
w/o Enhanced-Cross-Tem Module	✓	×	✓	0.065	0.0005	0.824	0.738	4.151	4.173
w/o GRU-KAN	✓	✓	×	0.038	0.0007	0.839	0.743	4.181	4.285
EAR-MLSR-Net (Full)	✓	✓	✓	0.012	0.0005	0.860	0.764	4.252	4.687

Note: Arrows (↑) indicate that higher values are better. Arrows (↓) indicate that lower values are better. w/o represents without. E-Cross-TEM represents Enhanced-Cross-Tem Module.

AUC and AP quantification supplies extra evidence for the effectiveness of the multi-module integrated framework. Intact EAR-MLSR-Net acquires peak AUC = 0.860 and AP = 0.764. Higher numerical results indicate advanced classification precision and sensitive early-risk identification. AUC falls to 0.831 once the GCN component is removed from the network. Spatial and relational context extraction provided by GCN promotes feature distinction and supplements temporal reasoning logic. TTA_{0.5} and mTTA work as critical metrics to quantify the model's early accident prediction capacity with stable output. The complete network obtains optimal TTA_{0.5} = 4.252 s and mTTA = 4.687 s. Larger values enable longer available prediction intervals facing unpredictable practical input noise.

4.4. EAR-MLSR-Net Performance Comparison

The data listed in Table 6 demonstrate that EAR-MLSR-Net consistently outperforms all competing methods across both classification and temporal prediction metrics. In terms of classification performance, the proposed model achieves an AUC of 0.860 and an AP of 0.770, surpassing CAP (0.832/0.752) and DRIVE (0.691/0.723). For temporal prediction, EAR-MLSR-Net achieves a TTA_{0.5} = 4.318 s and a mTTA = 4.684 s, outperforming all baselines with clear margins. These improvements indicate stronger early warning capability and more reliable risk-ranking performance.

The superior performance can be attributed to the model's modality-invariant representation learning and multi-level spatiotemporal modeling, which enhance robustness under heterogeneous and incomplete input conditions. From a generalization perspective, the proposed framework is expected to transfer well to different traffic environments and countries, as it does not rely on dataset-specific heuristics but instead learns shared embedding representations across modalities. In addition, the model is designed to capture universal spatiotemporal patterns of accident risk, which are commonly observed across different geographic regions and road systems. For varying weather conditions, the modality-invariant design helps mitigate distribution shifts caused by environmental noise, thereby improving robustness under adverse conditions. However, since all experiments are conducted on the CAP-DATA dataset, cross-dataset validation on other accident benchmarks remains a limitation of this study and will be investigated in future work.

Table 6. The EAR-MLSR-Net performance comparison.

Baselines	AUC ↑	AP ↑	TTA _{0.5} ↑	mTTA ↑
CAP [9]	0.832	0.752	4.022	4.494
UncertaintyTA [37]	0.601	-	3.846	-
DRIVE [45]	0.691	0.723	3.655	4.294
DSA-RNN [47]	0.472	-	3.094	-
AdaLEA [48]	0.553	-	3.891	-
(EAR-MLSR-Net) proposed	0.860	0.770	4.318	4.684

Note: Arrows (↑) indicate that higher values are better.

5. Conclusions

This study proposes EAR-MLSR-Net, a robustness-oriented multimodal framework designed for early accident risk prediction under incomplete and corrupted modality conditions. The core contribution of this work lies in enabling stable and reliable prediction through multi-level spatiotemporal modeling, which integrates graph-based spatial reasoning, cross-temporal alignment, and knowledge-augmented temporal dynamics learning. Unlike conventional approaches that assume fully available multimodal inputs, the proposed framework is specifically designed to address real-world challenges such as sensor failures, camera occlusions, communication interruptions, and adverse environmental conditions. By learning modality-invariant and hierarchically structured spatiotemporal representations, EAR-MLSR-Net maintains robust predictive capability even when partial or noisy modalities are present.

Experimental results on CAP-DATA demonstrate that the proposed method consistently improves robustness, enhances stability, enables earlier accident risk identification, and significantly reduces performance degradation under missing-modality scenarios. Rather than focusing on isolated metric improvements, these results collectively indicate stronger resilience and reliability in complex and uncertain traffic environments. The proposed framework has important practical implications for intelligent transportation systems, advanced driver-assistance systems (ADAS), autonomous driving platforms, smart city traffic monitoring, and emergency response systems, where robust perception under incomplete sensor inputs is critical for safety assurance. Future work will focus on uncertainty-aware modality completion and online reliability estimation of heterogeneous data sources, aiming to further improve adaptability under dynamic and unpredictable sensing failures.

Author Contributions: Conceptualization, H.L., W.S. and P.C.; Methodology, H.L. and W.S.; Software, H.L. and W.S.; Writing—original draft, H.L. and W.S.; Writing—review & editing, J.Z.; Supervision, P.C. and X.L.; Project administration, P.C. and X.L.; Funding acquisition, P.C. and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The CAP-DATA dataset integrates multiple public traffic accident datasets with self-compiled multimodal annotations. The full integrated dataset established in this study is not publicly available due to internal research management constraints. Researchers with legitimate academic demands may contact the corresponding author to apply for data access.

Conflicts of Interest: The authors declare no conflict of interest.

References

- World Health Organization. Save Lives: A Road Safety Technical Package. 2017. Available online: <https://apps.who.int/iris/bitstream/handle/10665/255199/9789241511704-eng.pdf> (accessed on 6 May 2018).
- World Health Organization. *Global Status Report on Road Safety 2018*; World Health Organization: Geneva, Switzerland, 2019.
- Heydari, S.; Hickford, A.; McIlroy, R.; Turner, J.; Bachani, A.M. Road safety in low-income countries: State of knowledge and future directions. *Sustainability* **2019**, *11*, 6249. [\[CrossRef\]](#)
- Anik, B.T.H.; Islam, Z.; Abdel-Aty, M. A time-embedded attention-based transformer for crash likelihood prediction at intersections using connected vehicle data. *Transp. Res. Part C Emerg. Technol.* **2024**, *169*, 104831. [\[CrossRef\]](#)
- Buian, M.F.I.; Arde, R.A.; Billah, M.M.; Debnath, A.; Siddique, I.M. Advanced analytics for predicting traffic collision severity assessment. *World J. Adv. Res. Rev.* **2024**, *21*, 2007–2018. [\[CrossRef\]](#)
- Chen, J.; Tao, W.; Jing, Z.; Wang, P.; Jin, Y. Traffic accident duration prediction using multi-mode data and ensemble deep learning. *Heliyon* **2024**, *10*, e25957. [\[CrossRef\]](#) [\[PubMed\]](#)
- Elsayed, H.A.G.; Syed, L. An automatic early risk classification of hard coronary heart diseases using Framingham scoring model. In Proceedings of the Second International Conference on Internet of Things, Data and Cloud Computing, Cambridge, UK, 22–23 March 2017; pp. 1–8.
- Ma, J.; Jia, C.; Yang, X.; Cheng, X.; Li, W.; Zhang, C. A data-driven approach for collision risk early warning in vessel encounter situations using attention-BiLSTM. *IEEE Access* **2020**, *8*, 188771–188783. [\[CrossRef\]](#)
- Fang, J.; Li, L.L.; Yang, K.; Zheng, Z.; Xue, J.; Chua, T.S. Cognitive accident prediction in driving scenes: A multimodality benchmark. *arXiv* **2022**, arXiv:2212.09381.
- Fang, S.; Liu, J.; Ding, M.; Cui, Y.; Lv, C.; Hang, P.; Sun, J. Towards interactive and learnable cooperative driving automation: A large language model-driven decision-making framework. *arXiv* **2024**, arXiv:2409.12812. [\[CrossRef\]](#)
- Jamshidi, H.; Jazani, R.K.; Khani Jeihooni, A.; Alibabaei, A.; Alamdari, S.; Kalyani, M.N. Facilitators and barriers to collaboration between pre-hospital emergency and emergency department in traffic accidents: A qualitative study. *BMC Emerg. Med.* **2023**, *23*, 58. [\[CrossRef\]](#) [\[PubMed\]](#)
- Chen, W.; Jiang, Q.; Zhuang, X.; Ma, G. How deceleration onset time and vehicle speed shape pedestrian perceived safety and comfort in yielding autonomous vehicles. In *Proceedings of the International Conference on Human-Computer Interaction*; Springer: Cham, Switzerland, 2025; pp. 165–174. [\[CrossRef\]](#) [\[PubMed\]](#)
- Sadiq, T.; Omlin, C.W. Sensing in smart cities: A multimodal machine learning perspective. *Smart Cities* **2025**, *9*, 3. [\[CrossRef\]](#)
- Thomas, K.V.; Rattan, P.; Nair, R.R. An introduction to multimodal data representation. In *Multimodal Learning Using Heterogeneous Data*; Morgan Kaufmann: Cambridge, MA, USA, 2026; pp. 17–30.
- Adewopo, V.A.; Elsayed, N.; ElSayed, Z.; Ozer, M.; Abdelgawad, A.; Bayoumi, M. A review on action recognition for accident detection in smart city transportation systems. *J. Electr. Syst. Inf. Technol.* **2023**, *10*, 57. [\[CrossRef\]](#)
- Liu, C.; Xiao, Z.; Long, W.; Li, T.; Jiang, H.; Li, K. Vehicle trajectory data processing, analytics, and applications: A survey. *ACM Comput. Surv.* **2025**, *57*, 1–36. [\[CrossRef\]](#)
- Wang, L.; Xiao, M.; Lv, J.; Liu, J. Analysis of influencing factors of traffic accidents on urban ring road based on the SVM model optimized by Bayesian method. *PLoS ONE* **2024**, *19*, e0310044. [\[CrossRef\]](#) [\[PubMed\]](#)
- Khan, S.W.; Hafeez, Q.; Khalid, M.I.; Alroobaea, R.; Hussain, S.; Iqbal, J.; Almotiri, J.; Ullah, S.S. Anomaly detection in traffic surveillance videos using deep learning. *Sensors* **2022**, *22*, 6563. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hasan, F.; Huang, H. MALS-Net: A multi-head attention-based LSTM sequence-to-sequence network for socio-temporal interaction modelling and trajectory prediction. *Sensors* **2023**, *23*, 530. [\[CrossRef\]](#) [\[PubMed\]](#)
- Leon, F.; Gavrilescu, M. A review of tracking and trajectory prediction methods for autonomous driving. *Mathematics* **2021**, *9*, 660. [\[CrossRef\]](#)
- Yang, D.; Huang, S.; Xu, Z.; Li, Z.; Wang, S.; Li, M.; Wang, Y.; Liu, Y.; Yang, K.; Chen, Z.; et al. Aide: A vision-driven multi-view, multimodal, multi-tasking dataset for assistive driving perception. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; pp. 20459–20470.
- Tang, Q.; Liang, J.; Zhu, F. A comparative review on multimodal sensors fusion based on deep learning. *Signal Process.* **2023**, *213*, 109165. [\[CrossRef\]](#)
- Zheng, Y.; Xu, Z.; Wang, X. The fusion of deep learning and fuzzy systems: A state-of-the-art survey. *IEEE Trans. Fuzzy Syst.* **2021**, *30*, 2783–2799. [\[CrossRef\]](#)
- Marchetti, F.; Mordan, T.; Becattini, F.; Seidenari, L.; Bimbo, A.D.; Alahi, A. CrossFeat: Semantic cross-modal attention for pedestrian behavior forecasting. *IEEE Trans. Intell. Veh.* **2024**, early access. [\[CrossRef\]](#)
- Liao, C.; Lei, K.; Zheng, X.; Moon, J.; Wang, Z.; Wang, Y.; Paudel, D.P.; Van Gool, L.; Hu, X. Benchmarking multimodal semantic segmentation under sensor failures: Missing and noisy modality robustness. *arXiv* **2025**, arXiv:2503.18445. [\[CrossRef\]](#)
- Zhang, Q.; Wei, Y.; Han, Z.; Fu, H.; Peng, X.; Deng, C.; Hu, Q.; Xu, C.; Wen, J.; Hu, D.; et al. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv* **2024**, arXiv:2404.18947. [\[CrossRef\]](#)

27. Duan, M.; Sun, S.; Liu, M. A multimodal deep fusion framework for highway traffic anomaly detection. *Sci. Rep.* **2025**, *15*, 33573. [[CrossRef](#)] [[PubMed](#)]
28. Cui, Q.; Ning, E.; Du, M.; Lu, B.; Li, S.; Xiang, J.; He, S. Uncertainty-aware traffic prediction and routing for emergency vehicles: A multi-objective risk-aware framework. *Connect. Sci.* **2026**, *38*, 2658934. [[CrossRef](#)]
29. Zeqiao, L.; Yijing, W.; Haoyu, W.; Zheng, L.; Peng, L.; Chuan, H. Confidence-guided human-AI collaboration: Reinforcement learning with distributional proxy value propagation for autonomous driving. *arXiv* **2025**, arXiv:2506.03568.
30. Chen, Y.; Wang, W.; Hua, X.; Zhao, D. Survey of decomposition-reconstruction-based hybrid approaches for short-term traffic state forecasting. *Sensors* **2022**, *22*, 5263. [[CrossRef](#)] [[PubMed](#)]
31. Khan, T.A.; Qin, Y. Graph-based deep learning and multi-source data to provide safety-actionable insights for rural traffic management. *Vehicles* **2025**, *7*, 151. [[CrossRef](#)]
32. Mo, X.; Lou, B.; Mao, Z.; Huang, Q.; Sun, W.; Wang, Y.; Lv, C. Traffic scene representation and encoding with graph structure learning and exploration. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 16840–16853. [[CrossRef](#)]
33. Harsha, T.R.; Bhat, V.P.; Prabhanjan, K.; Vinay, S.K.; Thomas, L. Graph neural networks for scene understanding: A review and future directions. In *Graph Neural Networks: Essentials and Use Cases*; Springer: Cham, Switzerland, 2025; pp. 277–339.
34. Ba, Y.; Ji, S.; Pan, S.; He, S.; Ji, Y.; Guo, D. Urban road traffic prediction with dynamic spatial-temporal graph convolutional network based on attention mechanism. *IEEE Sensors J.* **2025**, *25*, 27187–27199. [[CrossRef](#)]
35. Liao, H.; Rao, B.; Sun, H.; Wang, C.; Chang, Q.; Li, S.E.; Xu, C.; Li, Z. Chain-of-thought guided multimodal large language models for scene-aware accident anticipation in autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 19371–19380. [[CrossRef](#)]
36. Liu, J.; Peng, L.; Yan, X.; Zhang, L.; Yang, C.; Tao, Y.; Tan, A.Y.X.; Xu, T.; Guo, S.; Wang, H.; et al. Enhancing autonomous vehicle safety with knowledge graphs and large language models: A comprehensive review. *Commun. Transp. Res.* **2026**, *6*, 100156. [[CrossRef](#)]
37. Bao, W.; Yu, Q.; Kong, Y. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, WA, USA, 12–16 October 2020; pp. 2682–2690. [[CrossRef](#)]
38. Yao, Y.; Xu, M.; Wang, Y.; Crandall, D.J.; Atkins, E.M. Unsupervised traffic accident detection in first-person videos. In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Macau, China, 3–8 November 2019; IEEE: New York, NY, USA, 2019; pp. 273–280. [[CrossRef](#)]
39. Yao, Y.; Wang, X.; Xu, M.; Pu, Z.; Wang, Y.; Atkins, E.; Crandall, D.J. DoTA: Unsupervised detection of traffic anomaly in driving videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 444–459. [[CrossRef](#)] [[PubMed](#)]
40. Fang, J.; Yan, D.; Qiao, J.; Xue, J.; Wang, H.; Li, S. Dada-2000: Can driving accident be predicted by driver attention? analyzed by a benchmark. In *Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, Auckland, New Zealand, 27–30 October 2019; IEEE: New York, NY, USA, 2019; pp. 4303–4309.
41. Sultana, S.; Ahmed, B. Robust nighttime road lane line detection using bilateral filter and SAGC under challenging conditions. In *Proceedings of the 2021 IEEE 13th International Conference on Computer Research and Development (ICCRD)*, Beijing, China, 5–7 January 2021; IEEE: New York, NY, USA, 2021; pp. 137–143. [[CrossRef](#)]
42. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
43. Krishna, G.; Dharur, S.; Rudovic, O.; Dighe, P.; Adya, S.; Abdelaziz, A.H.; Tewfik, A.H. Modality drop-out for multimodal device directed speech detection using verbal and non-verbal features. In *Proceedings of the ICASSP 2024-IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, Republic of Korea, 14–19 April 2024; IEEE: New York, NY, USA, 2024; pp. 8240–8244. [[CrossRef](#)]
44. Nezakati, N.; Reza, M.K.; Patil, A.; Solh, M.; Asif, M.S. MMP: Towards robust multimodal learning with masked modality projection. *arXiv* **2024**, arXiv:2410.03010.
45. Bao, W.; Yu, Q.; Kong, Y. Drive: Deep reinforced accident anticipation with visual explanation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 11–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 7619–7628. [[CrossRef](#)]
46. Neverova, N.; Wolf, C.; Taylor, G.; Nebout, F. Moddrop: Adaptive multimodal gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1692–1706. [[CrossRef](#)]
47. Chan, F.H.; Chen, Y.T.; Xiang, Y.; Sun, M. Anticipating accidents in dashcam videos. In *Proceedings of the Computer Vision-ACCV 2016*, Taipei, Taiwan, 20–24 November 2016; Springer: Cham, Switzerland, 2017; pp. 136–153. [[CrossRef](#)]
48. Suzuki, T.; Kataoka, H.; Aoki, Y.; Satoh, Y. Anticipating traffic accidents with adaptive loss and large-scale incident db. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: New York, NY, USA, 2018; pp. 3521–3529. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.