

# SUMO-enhanced traffic anomaly detection: enhancing spatiotemporal transferability through simulation-based data augmentation

Haonan Yang<sup>1,2</sup>, Chenlei Liao<sup>3</sup>, Zhiwu Dong<sup>3</sup>, Chuqiao Chen<sup>2\*</sup>, Dong Mo<sup>3,4</sup> and Wenhao Chen<sup>5</sup>

<sup>1</sup> School of Information and Communication Engineering, Hainan University, Haikou 570228 China

<sup>2</sup> School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310018, China

<sup>3</sup> Institute of Intelligent Transportation Systems, College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China

<sup>4</sup> College of Civil and Transportation Engineering, Shenzhen University, Shenzhen 518060, China

<sup>5</sup> School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China

\* Corresponding author, E-mail: [chuqiaochen@hdu.edu.cn](mailto:chuqiaochen@hdu.edu.cn)

## Abstract

Traffic anomaly detection is crucial for road safety and traffic efficiency, but the scarcity of accident data poses challenges for model training. Although current research has utilized deep neural networks for data augmentation in small-sample scenarios, it has largely overlooked real-world physical constraints and lacks exploration into zero-shot scenarios, areas that still require further development. The paper proposes a SUMO-based simulation-driven data augmentation approach that enhances detection performance under limited real-world data conditions. Synthetic accident data are generated under five physical constraints and integrated with real traffic data to augment the training set. This study evaluates the method using three deep learning models (MLP, CNN, and LSTM) on PeMS datasets from two California highways (I80-E and I10-E). Experimental results demonstrate that the simulation-augmented training significantly improves model performance. In the spatial transfer experiment, the MLP model's accuracy improved from 75.53% to 76.58%, and AUC increased from 53.62% to 56.74%. These findings indicate that this approach effectively enhances both temporal and spatial transferability for traffic anomaly detection in small-sample scenarios.

**Keywords:** Accident prediction and detection; Spatial and temporal transferability; SUMO simulation; Data augmentation; Freeway safety

**Citation:** Yang H, Liao C, Dong Z, Chen C, Mo D, et al. 2025. SUMO-enhanced traffic anomaly detection: enhancing spatiotemporal transferability through simulation-based data augmentation. *Digital Transportation and Safety* 4(4): 264–274 <https://doi.org/10.48130/dts-0025-0025>

## Introduction

Highways, as a critical component of modern transportation systems, have significantly promoted interconnectivity between regions and provided a solid foundation for national economic development. However, traffic accidents remain one of the primary public safety challenges worldwide, causing substantial fatalities, injuries, and economic losses annually. According to the World Health Organization (WHO), traffic accidents are the leading cause of death for individuals aged five to 29, accounting for approximately 1.19 million fatalities per year<sup>[1]</sup>. Additionally, millions of injuries resulting from these accidents impose a heavy burden on society. With the acceleration of urbanization and the increasing number of vehicles, the issue of frequent traffic accidents has become more pronounced, prompting academia and industry to urgently develop efficient technologies for rapid accident detection and response.

Currently, various accident prediction and detection techniques have been developed to improve traffic safety through risk identification. These techniques can be categorized into two types: statistical models and machine learning-based models<sup>[2]</sup>. Traditional statistical models rely on historical data and mathematical analysis to uncover patterns within the data. However, they exhibit limitations when addressing complex scenarios and large-scale datasets<sup>[3]</sup>. In contrast, with the rapid development of artificial intelligence (AI) technology<sup>[4,5]</sup>, AI-based models have shown greater potential in traffic accident prediction and detection. These methods extract complex features from large-scale data and achieve high-precision predictions. Nevertheless, AI-based solutions are highly dependent on high-quality data, and the traffic accident data often suffers from

insufficient quantity and quality. This issue limits the effectiveness of AI technologies in practical applications. Advanced data cleaning and preprocessing methods can address this challenge by improving data quality and enriching traffic event datasets with additional information.

Despite increasing interest in using deep learning for traffic accident prediction and detection, significant limitations remain. Most existing data augmentation strategies—such as those using Generative Adversarial Networks (GANs)—focus on statistical realism without incorporating real-world physical constraints. These approaches may produce synthetic data that lacks behavioral interpretability or alignment with actual traffic dynamics. Meanwhile, advanced traffic simulation platforms like SUMO have been extensively used for traffic modeling and control evaluation, but their potential as tools for generating physically grounded synthetic data to improve anomaly detection models remains underexplored. Furthermore, very few studies investigate the ability of augmented models to generalize in zero-shot or data-scarce scenarios, particularly across different temporal or spatial domains.

To address the aforementioned challenges, the paper proposes a SUMO (Simulation of Urban Mobility) -based simulation data augmentation method aimed at improving the performance of traffic accident detection and prediction by enhancing data quality and coverage. Specifically, the main contributions of the paper are as follows:

(1) Incorporation of physical constraints: by integrating traffic-related physical information, the paper introduces a physical-informed model that incorporates physical constraints into the data augmentation process, addressing the lack of physical constraints in

existing methods. These physical constraints include maximum acceleration, comfortable deceleration, minimum safe distance, acceleration index, and driver reaction time. While these five parameters capture key aspects of vehicle dynamics and are directly supported in SUMO, other influential real-world factors, such as weather conditions or driver behavior variability, were not included due to limited simulator support and the scope of this study. Future work will explore incorporating such factors to further enhance the realism of the augmented data.

(2) Temporal and spatial transfer modeling: using the SUMO software, the study conducts temporal transfer modeling for roads with limited accident data and spatial transfer modeling for roads without accident data, thereby extending the temporal and spatial applicability of the data.

(3) Deep learning-based accident detection: the study utilized three commonly applied neural network models for accident detection and validated the model's detection performance using real-world data and simulation data. Experimental results demonstrate that the model significantly improves detection performance and reliability.

The remainder of this paper is organized as follows: the 'related work' section reviews existing studies on traffic accident detection and prediction, and surveys relevant research on traffic flow data augmentation. The 'methodology' section introduces the SUMO simulation techniques and the accident prediction and detection models, including multilayer perceptron (MLP), convolutional neural network (CNN), and long short-term memory (LSTM). The 'data preparation' section details the data sources, selected roads, and data preprocessing methods. The 'results' section presents the simulation-generated normal traffic flows, temporal and spatial transfer modeling, and experimental results. Finally, the 'discussion and contributions' section summarizes the main contributions of the paper and highlights potential directions for future research.

## Related work

Accident prediction and detection have long been central topics in transportation research. With the rapid advancement of machine learning (ML), AI-based models have gradually become a core area of research, leading to numerous studies that utilize traffic accident data for analysis, prediction, and detection.

### Accident prediction

Research in accident prediction primarily falls into two categories: statistical analysis models and machine learning-based models. Statistical analysis, such as logistic regression<sup>[6–8]</sup>, log-linear models<sup>[9]</sup>, Bayesian logistic models<sup>[10]</sup>, and Bayesian dynamic logistic regression models<sup>[11]</sup>, provide valuable insights, however, they often struggle to capture the complex nonlinear relationships inherent in traffic flow data, leading to limited performance.

In contrast, ML models demonstrate greater potential due to their ability to handle high-dimensional and nonlinear data. Huang et al.<sup>[12]</sup> utilized deep learning techniques to predict crash risks, employing data from radar sensors on Interstate 235 in Des Moines, IA. Their model demonstrated notable improvements in detection accuracy. Pourroostaei et al.<sup>[13]</sup> explored road car accident data patterns and proposed a data preprocessing model, including missing data removal, attribute generalization, and outlier removal, to improve the accuracy of traffic accident prediction using machine learning techniques.

There are likewise many other studies that have investigated accident risk prediction by fusing traffic data and combining statistical and ML methods. For instance, Yu et al.<sup>[14]</sup> proposed a Deep Spatio-Temporal Graph Convolutional Network (DSTGCN) to predict traffic

accidents by effectively learning spatial, temporal, and external data correlations, and achieved superior performance on real-world datasets. Guo et al.<sup>[15]</sup> incorporated risky driving behavior and traffic flow data to predict traffic accident risks, achieving a prediction accuracy of 84.48%. Zhang et al.<sup>[16]</sup> developed a bidirectional LSTM model with convolutional layers to analyze roadside detector data and connected vehicle information, reaching a recall of 0.772 and an AUC of 0.857. Yuan et al.<sup>[17]</sup> combined statistical and ML models to explore the relationship between traffic conflicts and flow features using the HighD dataset, highlighting the significant influence of traffic flow on conflict probability.

### Accident detection

Early-stage methods for incident detection predominantly relied on comparative algorithms, such as the well-known California algorithm<sup>[18]</sup>, which compares traffic parameters (e.g., speed, occupancy) to predefined thresholds<sup>[19]</sup>. While effective, these methods depend heavily on fixed detector readings and lack adaptability.

With the development of computational intelligence, Artificial Intelligence algorithms apply AI and ML models to identify the normal and abnormal traffic patterns and then classify the given input data as either incident or normal conditions<sup>[20]</sup>. For example, Dogru et al.<sup>[21]</sup> proposed an intelligent traffic accident detection system based on VANETs and machine learning, using vehicle speed and position data for high-accuracy accident detection with the Random Forest algorithm. White et al.<sup>[22]</sup> proposed a smartphone-based traffic accident detection system that combines accelerometers, acoustic data, and context information to automatically notify emergency services, provide situational awareness, and reduce false positives. Ozbayoglu et al.<sup>[23]</sup> designed a real-time autonomous detection system using computational intelligence techniques to analyze Istanbul's traffic data, predicting accidents with nearest neighbor models and feed-forward neural networks. Gu et al.<sup>[24]</sup> explored social media as an alternative data source, mining Twitter for real-time traffic incident detection and classification. Mehrannia et al.<sup>[25]</sup> applied LSTM networks for freeway accident detection, improving class separability and reducing dimensionality.

### Data augmentation

The above AI-based solutions for incident detection have highlighted challenges, particularly regarding data quantity and quality. Traditional traffic datasets often lack sufficient volume and exhibit quality issues, limiting their applicability. To address these challenges, data enhancement techniques have emerged as a critical research focus.

Lin et al.<sup>[26]</sup> introduced a hybrid model using Generative Adversarial Networks (GANs) to expand sample size and balance datasets, complemented by Temporal and Spatially Stacked Autoencoders (TSSAE) for feature extraction. Xie et al.<sup>[27]</sup> proposed SASYNORF-RSKNN, a hybrid method tackling imbalanced datasets using over-sampling techniques and ensemble algorithms. Similarly, Jilani et al.<sup>[28]</sup> utilized GAN-based augmentation to train CNN models, achieving high classification accuracy. Other studies emphasize advanced augmentation strategies. Huang et al.<sup>[29]</sup> integrated GANs with TSSAE to address sample limitations and detect temporal and spatial traffic patterns. Dabboussi et al.<sup>[30]</sup> employed GANs to generate realistic traffic speed data, improving prediction for specific scenarios. Benabdallah Benarmas et al.<sup>[31]</sup> highlighted the role of Deep Generative Models (DGMs) in enhancing traffic speed predictions through synthetic dataset generation.

Summarily, these advancements underscore the significant progress in accident prediction and detection, particularly through the integration of deep learning and data augmentation. However, challenges remain in addressing the dependency on high-quality

data and balancing the scalability of models for diverse traffic scenarios are key. Based on these insights, this paper proposes a method that combines SUMO simulation data with real-world data. This method not only increases the diversity of the data but also effectively improves the performance of the model.

## Data preparation

This study collected and processed 160 accident and non-accident cases, with data sourced from the California Department of Transportation's Performance Measurement System (PeMS) and the California Highway Patrol (CHP). The study focuses on two major highways in California—Interstate 10 East (I10-E) and Interstate 80 East (I80-E)—covering the period from December 1 to June 30 in both 2016 and 2021, thereby capturing traffic conditions before and after the COVID-19 pandemic. The dataset encompasses traffic records from 8,041 mainline detector stations distributed across 5,360 miles of highway. Data from the year 2017 was used for model training, while 2022 data was utilized to evaluate temporal and spatial transferability.

Although PeMS performs automated imputation to address short-term missing data, the dataset remains vulnerable to residual anomalies resulting from intermittent hardware failures and communication disruptions at detector stations. To enhance data integrity and ensure the reliability of downstream modeling, a comprehensive preprocessing framework was implemented. Specifically, data entries were excluded under the following conditions: (1) a complete absence of recorded measurements from a detector station; (2) simultaneous zero values in average speed, flow, and occupancy, indicating likely sensor malfunction; (3) identical readings across multiple detectors, suggesting data duplication or transmission errors; and (4) repeated values at a given detector that exactly matched those recorded at the same timestamp on the previous day or week, indicating placeholder or default imputations.

### I10-E

I-10E<sup>[32]</sup> is the eastbound section of Interstate 10 in the United States, spanning multiple states and connecting major cities from Los Angeles, California, to Jacksonville, Florida. With a total length of approximately 2,460 miles (3,960 km), it is the fourth-longest highway in the United States, following I-90, I-80, and I-40. It is also one of the longest and busiest east-west highways in the country. I-10E traverses key economic, cultural, and logistical hubs, playing a crucial role in freight transportation and interstate travel. It connects major regions in the southwestern and southeastern United States, serving as an essential corridor for east-west transportation.

### I80-E

The I80-E<sup>[33]</sup> highway is part of the US Interstate Highway System, spanning over 4,600 km and connecting California and New Jersey, which is commonly used in identifying normal traffic flow. This road traverses 11 states, with I80-E being its eastbound segment. It features a wide roadway, typically comprising three to five lanes, with the specific number of lanes varying by state and region. As a

vital transportation corridor, I80-E serves a large volume of commuter and freight vehicles, exhibiting significant variations in traffic characteristics across its segments. From urban core areas to suburban and rural regions, this highway provides a wealth of sample data for studying dynamic behaviors under different traffic environments. Moreover, due to its complex traffic conditions, frequent peak flows, and diverse traffic patterns, I80-E is an ideal choice for traffic simulation studies.

Overall, I80-E and I10-E differ significantly in both traffic scale and functional context. I80-E, an urban freeway segment, carries approximately 110,000 vehicles per day, while I10-E, located in a more suburban to rural area, sees around 80,000. These structural and environmental differences introduce a substantial geographic domain shift, making the spatial transfer task more challenging and meaningful.

## Data preprocessing techniques

To better demonstrate the generalizability of the model, basic information that can typically be extracted by conventional highway detectors is selected as the model input. These inputs usually include fundamental statistical measures such as the average and standard deviation of traffic flow. Specifically, based on previous experience<sup>[34–36]</sup>, data collected by two detectors is used upstream and downstream of the accident point to extract traffic flow parameters within 0 to 15 min after the accident for analysis, as shown in Fig. 1. When an accident occurs, it often causes congestion and sudden deceleration upstream, generating shock waves that propagate backward in space. Monitoring the upstream area allows us to capture early indicators of abnormal traffic conditions, such as a sudden drop in speed or an increase in occupancy. In contrast, the downstream detectors serve as a control reference, where traffic may be unaffected or return to normal faster.

For analysis, the time is divided into three intervals, each with a five-minute gap, labeled as  $T_1$ ,  $T_2$ , and  $T_3$ . Specifically, the time from 0 to 5 min before the accident corresponds to  $T_1$ , 5 to 10 min corresponds to  $T_2$ , and 10 to 15 min corresponds to  $T_3$ . Additionally, the detectors are numbered sequentially from upstream to downstream in relation to the accident point, designated as  $S_1$ ,  $S_2$ ,  $S_3$ , and  $S_4$ . The selected traffic flow parameters include flow rate, speed, and occupancy, represented by the symbols  $Q$ ,  $V$ , and  $O$ , respectively. To provide a more comprehensive description of the data characteristics, the average values and standard deviations of these parameters are further calculated, denoted as  $Aver$  and  $Std$ , and serve as the initial input variables. In total, 72 variables were constructed to support the model's input and analysis.

## Methodology

To enhance the performance of traffic accident detection systems, the SUMO was utilized for data augmentation and advanced deep learning techniques for traffic accident detection. The following section first introduces SUMO, followed by an introduction to the deep learning models used.

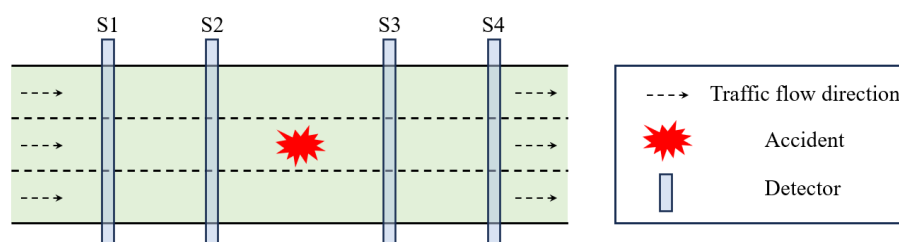


Fig. 1 Accident location and detector selection.

## Introduction to SUMO

SUMO is a microscopic traffic simulation platform designed for modeling multi-modal transportation systems, including vehicles, public transit, pedestrians, and cyclists, developed by the Dresden University of Technology in Germany<sup>[37]</sup>. Since its inception in 2001, the SUMO software suite has been widely applied in numerous domestic and international research projects. Its primary applications include traffic signal evaluation<sup>[38]</sup>, route selection and re-planning<sup>[39]</sup>, traffic monitoring method assessment<sup>[40]</sup>, vehicle communication simulation<sup>[41]</sup>, and traffic forecasting<sup>[42]</sup>. As an open-source software, SUMO allows users to customize and extend it according to their needs. It supports multiple traffic modes, such as cars, pedestrians, and public transport, and can be integrated with tools like Python and MATLAB, facilitating data analysis and model training. Its high efficiency enables SUMO to quickly process large volumes of traffic data and simulate complex traffic scenarios, making it particularly suitable for tasks such as traffic anomaly detection and model optimization. Additionally, SUMO possesses powerful data generation and enhancement capabilities, enabling the creation of simulation data with complex traffic patterns and diverse scenarios, which are especially well-suited for training deep learning models. Finally, SUMO simulates based on real-world traffic network data and employs precise vehicle behavior modeling, ensuring that the simulation results closely align with real-world traffic conditions, demonstrating applicability in real-world scenarios. Due to the various advantages of SUMO mentioned above, SUMO was ultimately chosen as the simulation environment in this paper to generate traffic accident augmented data.

## Simulation calibration

The Intelligent Driver Model (IDM), a microscopic traffic simulation model designed to replicate vehicle dynamics and driver behavior, was calibrated using an exhaustive grid search method. IDM employs precise mathematical formulations to simulate how vehicles adjust their speed and spacing in response to leading vehicles under various traffic conditions, considering factors such as desired speed, current speed, gap distance, and traffic density to determine optimal acceleration or deceleration strategies.

The IDM model assumes that the driver will dynamically adjust the acceleration according to the behavior of the vehicle in front and the current traffic conditions. The acceleration of the vehicle is determined by the following factors: (1) Expected free travel speed: When there is no interference from the vehicle in front, the vehicle tends to travel at a desired speed (such as the speed limit); (2) Safety distance: The vehicle will adjust its speed according to the distance and relative speed to the vehicle in front to avoid a rear-end collision; and (3) Acceleration and deceleration capabilities: The physical properties of the vehicle also constrain driving behavior.

The acceleration formula of the IDM model is as follows:

$$\dot{v} = a \left[ 1 - \left( \frac{v}{v_0} \right)^\delta - \left( \frac{s^*(v, \Delta v)}{s} \right)^2 \right] \quad (1)$$

where,  $a$  represents the maximum acceleration,  $v$  represents the current vehicle speed,  $v_0$  represents the expected speed,  $\delta$  represents the acceleration index. Specifically,  $\delta$  controls how aggressively or smoothly a vehicle approaches its desired speed. A larger value of  $\delta$  leads to smoother, more gradual acceleration, promoting comfort and stability, while a smaller value results in faster, more abrupt acceleration responses.  $s$  represents the distance between the current vehicle and the vehicle in front, and  $s^*(v, \Delta v)$  represents the expected distance, which can be expressed as:

$$s^*(v, \Delta v) = s_0 + \max \left( 0, vT + \frac{v\Delta v}{2\sqrt{ab}} \right) \quad (2)$$

where,  $\Delta v$  represents the relative speed, that is, the speed difference between the current vehicle and the vehicle in front;  $s_0$  represents the minimum safe distance,  $b$  represents the comfortable deceleration. If the current vehicle suddenly decelerates, the vehicle will adjust the current speed with the deceleration of  $b$  to avoid a rear-end collision;  $T$  represents time headway, which refers to the driver's reaction time.

Parameter calibration is used to determine the IDM parameters. The objective of calibration was to minimize the Root Mean Square Error (RMSE) between simulated traffic flow and real-world detector readings. The calibration process was conducted over a defined search space constrained by physically reasonable ranges for each IDM parameter, including maximum acceleration and comfortable deceleration. The upstream section of the three-lane road section 180-E, which is monitored by four detectors (402241, 402243, 402245, 402246) was selected. The vehicle data from the upstream detectors is used as input, and the difference between the simulated and actual readings from the downstream detectors is used to fine-tune the IDM parameters to ensure that the simulation accurately reflects the observed traffic behavior.

## Vehicle inject

After adopting the calibrated parameters, let the probability of a vehicle appearing per second be  $p$ . Then, the number of vehicles departing on this lane within 5 min,  $X$ , follows a binomial distribution:

$$X \sim B(300, p) \quad (3)$$

The mean of this distribution can be expressed as  $E(X) = 300p$ , which equals the actual number of vehicles  $N$  detected by the detector during the 5 mins. Therefore, the input probability  $p = N/300$ .

## Accident detection methods

MLP is a classic feedforward neural network architecture composed of multiple layers, typically including an input layer, one or more hidden layers, and an output layer. Each neuron in one layer is fully connected to the neurons in the previous layer, processing data through weighted sums, biases, and activation functions to transmit information through the network. The goal of MLP is to accurately model the complex nonlinear relationships between input data and outputs, enabling precise classification or regression tasks.

CNNs are a class of deep learning models primarily used for processing and analyzing grid-like data structures, such as images or time-series data. CNNs excel at automatically learning spatial hierarchies of features, making them particularly effective in tasks like image classification, object detection, and anomaly detection.

LSTM networks are a specialized type of Recurrent Neural Network (RNN) designed to handle and predict long-term dependencies in sequential data. Traditional RNNs often struggle with long sequences due to the vanishing or exploding gradient problem, making it difficult to effectively retain information from distant points in the sequence. LSTM addresses this issue by introducing 'memory cells' and 'gating mechanisms,' which enable the model to manage information flow more effectively.

This paper implements the prediction of normal traffic flow using LSTM to validate the feasibility of using SUMO simulation data, and accident detection of abnormal traffic flow using MLP, CNN, and LSTM, using multiple neural networks with the aim of eliminating chance. These models are chosen for their simplicity, interpretability, and compatibility with small-sample datasets. While more advanced models such as spatiotemporal graph convolutional networks (STGCNs) or graph neural networks (GNNs) can more effectively capture complex spatiotemporal dependencies, they usually require larger-scale graph-structured data and more extensive training samples. Given the small-sample and data-scarce scenarios emphasized in this work, simpler models provide more



robust and interpretable benchmarks to evaluate the effectiveness of simulation-based data augmentation. Moreover, the consistent improvements demonstrated in these commonly used models reinforce the practical value of our proposed approach.

## Experimental workflows

The experimental workflows comprises two key components:

(1) Normal traffic flow prediction: By comparing the differences in prediction results generated using real-world data alone versus hybrid data (real-world data + SUMO simulation data), the feasibility of SUMO simulation data in model training was validated;

(2) Abnormal traffic flow detection: Three neural network models were first brought to convergence, followed by joint fine-tuning with SUMO simulation data and real-world data, demonstrating the effectiveness of simulation data in enhancing detection performance.

### Generation and prediction of normal traffic flow

To demonstrate that SUMO simulation can be effectively used for traffic flow simulation, traffic flow simulation was first performed using SUMO, and its effectiveness was validated by comparing the simulation results with real-world data.

Traffic flow data was used as the input for the LSTM model to predict future traffic flow. The input sequence is represented as  $[q_t, q_{t+1}, q_{t+2}, q_{t+3}, q_{t+4}]$ , where  $q_t$  denotes the traffic flow at time  $t$ . A 25 min traffic flow sequence is used as the input, and the model predicts the traffic flow for the next 5 min. This approach enables effective modeling of short-term traffic fluctuations and dynamic changes.

To validate the effectiveness of SUMO simulation data, three training datasets were used, as shown in Fig. 2, where the dataset A included 28 d of real traffic flow data (from December 1, 2016, to December 29, 2016); the dataset B comprised 1 d of real traffic flow data (from December 28, 2016, to December 29, 2016); and the dataset C combined simulated traffic flow data (from December 1, 2016, to December 28, 2016) with an additional day of real data (from December 28, 2016 to December 29, 2016). The test set consists of 7 d of real traffic flow data (from December 29, 2016, to January 5, 2017).

## Accident simulation and transfer experiment

### Accident simulation

As shown in Fig. 3, the accident simulation is conducted by closing lanes in the simulation. When an accident occurs, the lanes in

the fixed areas upstream and downstream of the accident point are closed. Two types of accident simulation are defined: existing accident enhancement and virtual accident enhancement.

Existing accident enhancement involves altering the impact range and affected lanes based on the existing accidents, used for time transfer experiments. Virtual accident generation involves randomly specifying a non-accident time and creating virtual accidents at random locations and times, used for spatial enhancement in simulation experiments. The simulation process is shown in Fig. 4, and the details will be explained later.

To simulate rare accident scenarios and enrich the training dataset, SUMO was used to generate synthetic traffic data based on real-world detector inputs. By calculating the error under different numbers of lanes and impact ranges, the accident scenario was simulated by blocking a single lane. The simulation incorporated five core physical parameters—minimum vehicle gap, maximum acceleration, comfortable deceleration, acceleration exponent, and simulation time step—which govern vehicle behavior and interaction. The resulting traffic patterns exhibit realistic congestion effects, including upstream queuing, abrupt speed drops, and downstream flow disruptions, which are key indicators of anomaly conditions. Subsequently, these simulated data will be spliced with real data for use in subsequent research. The core of this research is to combine real data with simulation data for data augmentation, thereby improving the model's transfer ability in different temporal and spatial contexts, particularly in scenarios with small samples and data scarcity. This approach primarily focuses on two types of transfer: temporal transfer and spatial transfer.

### Temporal transfer

Temporal transfer refers to the model's ability to transfer knowledge across different time periods (e.g., 2017 vs 2022). To validate the model's performance in temporal transfer scenarios, the study uses real data from the I80-E road in 2017 as the training set and trains a baseline model without data augmentation for the accident detection task. Subsequently, the SUMO simulation software was used to generate augmented simulation data by combining the accident data and normal traffic flow data from the 2022 I80-E road. These simulation data are then used to fine-tune the model to enhance its transfer capability to a new time period.

During the temporal transfer model training process, the Adam optimizer was used with the initial learning rate set to  $1e-5$ . Due to the limited size of the input data, the batch size was set to eight.

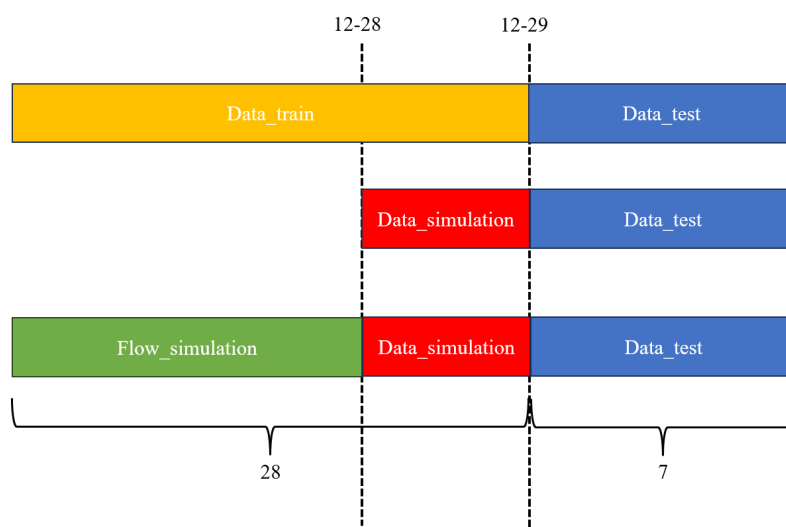
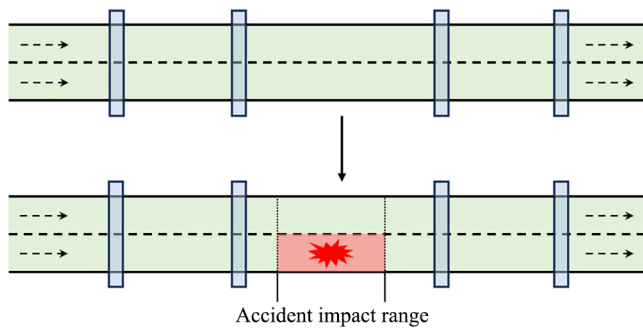


Fig. 2 Mixed dataset strategy for model training.

## Traffic anomaly detection using data augmentation



**Fig. 3** SUMO accident simulation diagram.

When the model was trained to a converged state, it was saved. In the subsequent fine-tuning phase, the learning rate was further reduced to  $1e-6$ , and the batch size was adjusted to 16, 32, or 64 depending on the size of the new training set. During the fine-tuning process, the previously saved model was loaded, and training was continued based on it.

In the experiment, 100 real data points were randomly selected from the I80-E road in 2017 (including 20 accident data and 80 non-accident data, with a 1:4 positive-to-negative sample ratio) as the training set, and the model was trained using neural networks to obtain the baseline model. Then, 25 real data points from the I80-E road in 2022 were randomly selected as the first training set, and samples were extracted from the simulated data at ratios of 1:1, 1:9, and 1:19, resulting in three training sets with simulated data volumes of 25, 225, and 475, respectively. These four training sets were then used for model fine-tuning. Finally, the real data from the I80-E road in 2022 was uniformly used as the test set to evaluate the performance of the baseline model and the fine-tuned models.

### Spatial transfer

Spatial transfer refers to the model's ability to transfer knowledge across different geographical environments or traffic flow characteristics. To validate the model's spatial transfer ability, a test involving transfer between different roads (I80-E and I10-E) was conducted. Initially, accident data from the I80-E road in 2017 and normal traffic flow data from the I10-E road were used to train a baseline model.

Subsequently, the SUMO simulation software, combining the accident data from the 2017 I80-E road with the normal traffic flow data from the I10-E road, was used to generate augmented data for the I10-E road, which was then used to fine-tune the model. The specific experimental details will be described in the results section.

For spatial transfer model training, the optimizer used was Adam with a learning rate of  $1e-5$  and batch size set to eight. The model was saved when training converged, and for fine-tuning, the learning rate was set to  $1e-6$ , the batch size was adjusted to 64 based on the size of the new training set, and the previously saved model was read to perform the training.

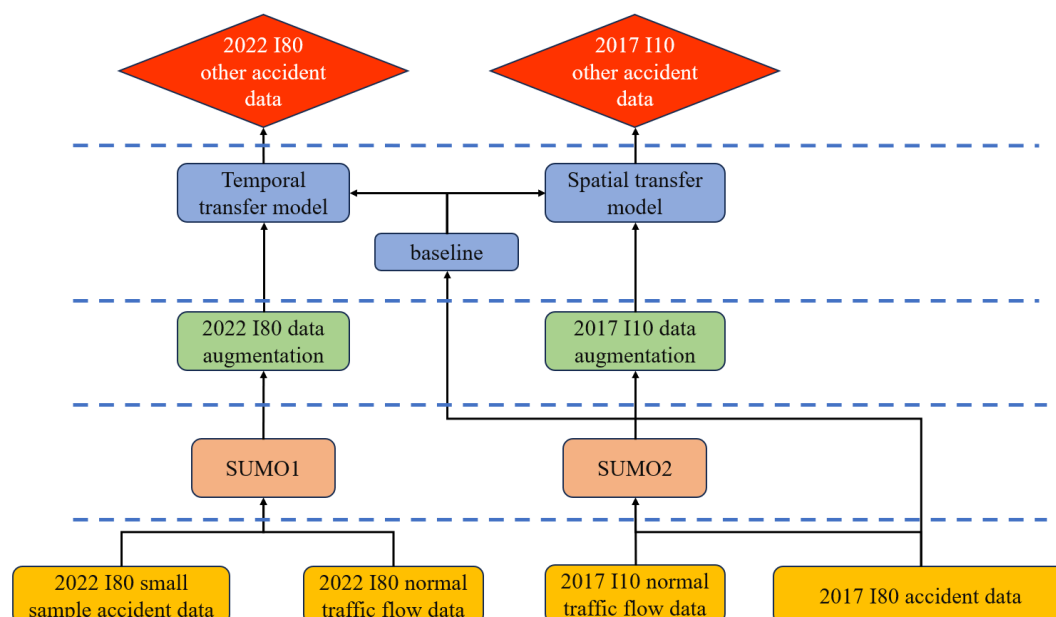
In the spatial transfer testing experiment, it was assumed that the I10-E road in 2017 had just begun operation, resulting in a scarcity of normal traffic flow data and a complete lack of accident data. To address this, a baseline model was first constructed by randomly selecting 100 real data samples from the 2017 I80-E road—comprising 20 accident samples and 80 non-accident samples (a case matching ratio of 1:4, which was consistently used thereafter)—and trained neural networks on this set. To generate training data for fine-tuning under spatial transfer conditions, the accident data from the 2017 I80-E road was combined with the limited normal traffic flow data from the 2017 I10-E road, and the SUMO simulation software was used to simulate accidents, applying simulation-based data augmentation to generate enhanced I10-E road samples. From this augmented dataset, 100 samples (20 accident and 80 non-accident) were then extracted to fine-tune the pre-trained baseline model. Finally, real data from the 2017 I10-E road was used as the test set to evaluate the performance of both the baseline and fine-tuned models, thereby assessing the effectiveness of simulation-based augmentation in enhancing spatial transferability.

## Results

### Generation and prediction of normal traffic flow

#### Traffic flow generation

An exhaustive search with a step size of 0.1 was performed across the parameter space constrained by five physical constraints, resulting in the optimal parameter combination that minimizes the error



**Fig. 4** Temporal and spatial transfer simulation diagram.

as follows:  $a = 2.6 \text{ m/s}^2$ ,  $b = 1 \text{ m/s}^2$ ,  $s_0 = 3 \text{ m}$ ,  $\delta = 4$ ,  $T = 0.5 \text{ s}$ . Afterwards, real traffic data was used to generate the number of vehicles; using binomial distribution to control whether the vehicle is injected at a certain second to form natural fluctuations; and IDM physical parameters to control the micro-behavior of each vehicle to ensure the physical rationality of the behavior. Using those as input, a seven-day simulation is illustrated in Fig. 5 and conducted to obtain the traffic flow data extracted by the detector during the seven-day simulation.

### Traffic flow prediction

Using the three training datasets mentioned in Experimental Workflows for training and predicting the traffic flow data in the next 7 d, the results are shown in Fig. 6. Where the 'Real data (28 d)' curve shows the prediction results for the next 7 d after training the model with dataset A; the 'Real data (1 d)' curve shows the prediction results for the next seven days after training the model with dataset B; and the 'SUMO and real data' curve shows the prediction results for the next 7 d after training the model with dataset C. The 'Ground truth' curve shows the actual observed changes in traffic flow over the 7 d, which is used as a baseline. It was found that analyzing SUMO simulation data in combination with a small amount of real-world data yields predictions that differ very little from predictions based on a large amount of real-world data, compared to using only one day of real data. Additionally, Fig. 7 shows that during non-peak hours, the combined dataset demonstrates particularly significant improvements, with errors approaching the level achieved by training on the complete dataset. This further validates the reliability and utility of the simulated data in relevant analyses.

### Abnormal traffic flow detection

#### Temporal transfer testing experiment

Normal traffic flow data from January 1, 2022, was first selected. The IDM model was calibrated based on the error between the simulation detector data and the real data, yielding the following model parameters:  $a = 2.3 \text{ m/s}^2$ ,  $b = 2.6 \text{ m/s}^2$ ,  $s_0 = 5 \text{ m}$ ,  $\delta = 4$ ,  $T = 1 \text{ s}$ . Next, the SUMO simulation was used to simulate traffic flow after accidents, and by calculating errors under different lane numbers and impact ranges, the simulation settings that best replicated real-world conditions were selected: the impact range was 20 m, and the affected lane number was 1. The accident time and location were randomly selected from five real accidents that occurred on the I80-E road in 2022. Based on the times and locations of these real accidents, combined with the aforementioned settings, simulations were performed using SUMO, resulting in five simulated datasets based on real accident data. The five randomly selected accidents

represent approximately 1.85% of the available accident sample for 2017. This significant reduction highlights the challenge of data scarcity and underscores the need to use simulations to supplement data under such constraints.

The experimental results are shown in Fig. 8. The results show that the samples generated through data augmentation not only significantly improved the model's performance but also demonstrated good temporal transferability. Although real data from the I80-E road in 2017 were used during training, the testing was still based on real data from the I80-E road in 2022. The fine-tuned models showed significant improvements in key performance indicators such as accuracy and AUC. Specifically, the MLP, CNN, and LSTM neural network models all exhibited varying degrees of performance improvement after fine-tuning.

For the MLP model, when using the 2022 small sample with a 1:1 sample ratio for training, the model's accuracy was relatively poor. However, when the sample ratio was adjusted to 1:9, the accuracy increased from 69.39% to 73.23%, and further increased to 73.29% when the sample ratio was 1:19. In terms of AUC, all fine-tuning methods showed improvement, with the most significant increase seen at a 1:19 sample ratio, where the AUC rose from 52.02% to 52.96%.

For the CNN model, all fine-tuning strategies outperformed the baseline model, further confirming the performance-boosting effect of data augmentation. Specifically, when using the 2022 small sample, the accuracy improved from 76.55% to 76.81%. As the sample ratio increased, the accuracy continued to improve, reaching 77.19%, 78.36%, and 78.4%. In terms of AUC, the AUC for the

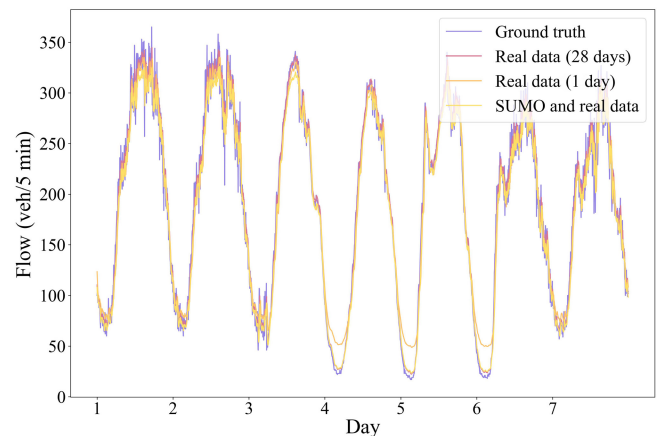


Fig. 6 Comparison of prediction results.

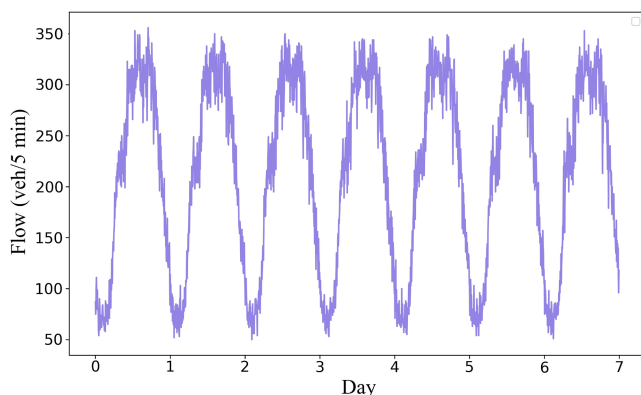


Fig. 5 Seven-day simulation traffic flow variation curve.

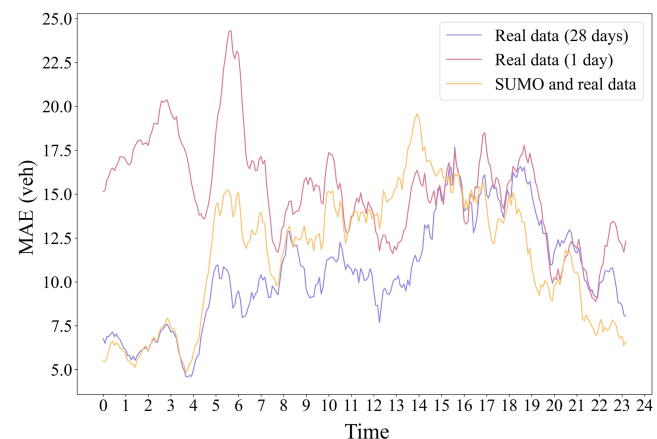
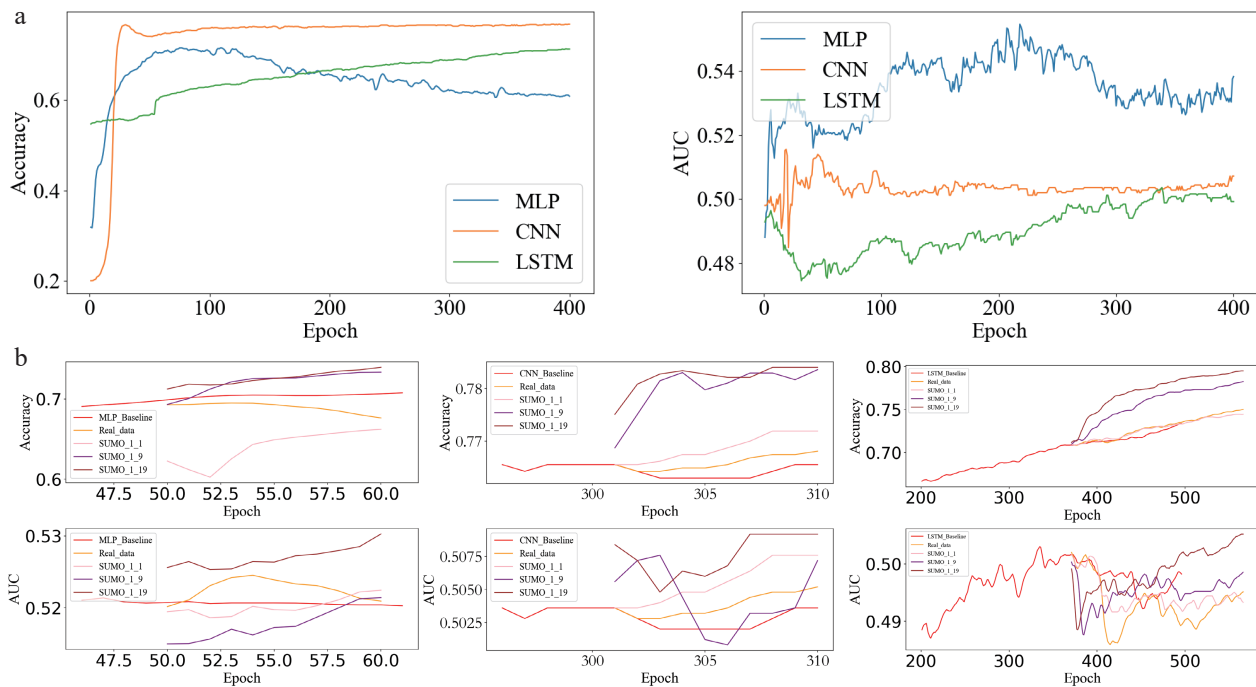


Fig. 7 Twenty four hour traffic flow MAE variation.



**Fig. 8** Temporal transfer experiment simulation results. (a) The convergence curves of accuracy and AUC under MLP, CNN, and LSTM neural networks. (b) The comparison of accuracy and AUC curves before and after fine-tuning the model using the 2022 small sample dataset with a 1:1, 1:9, and 1:19 sample ratio as the fine-tuning training set.

2022 small sample increased from 50.36% to 50.52%, and the AUC for different sample ratios also showed gradual improvements: from 50.36% to 50.76% for a 1:1 ratio, from 50.36% to 50.72% for a 1:9 ratio, and from 50.36% to 50.92% for a 1:19 ratio, showing stable and progressive enhancement.

For the LSTM model, there was also a significant improvement in accuracy. When using the 2022 small sample, accuracy increased from 70.86% to 75.02%. As the sample ratio changed, accuracy reached 74.5% for a 1:1 ratio, 78.27% for a 1:9 ratio, and 79.49% for a 1:19 ratio. AUC also increased from 50.16% to 50.52%. These results further indicate that data augmentation significantly improved the LSTM model's performance in temporal transfer scenarios.

Overall, the experimental results validate the effectiveness of the simulation-based data augmentation training method in temporal transfer scenarios. Especially in small sample situations, data augmentation not only enhances the model's generalization ability but also effectively improves transfer performance across time periods. These findings provide strong empirical support for model optimization using data augmentation techniques, further proving the potential and advantages of simulation data as augmented samples in data-scarce situations. In addition, as the sample proportion increases, the diversity of the data gradually improves, and the performance of the model shows an overall upward trend. This result suggests that in the absence of real data, simulated data helps to improve the generalization ability and performance of the model. Therefore, combining real data with simulation data for data augmentation is an effective strategy to improve model performance and address data scarcity, offering significant guidance for future research and applications.

### Spatial transfer testing experiment

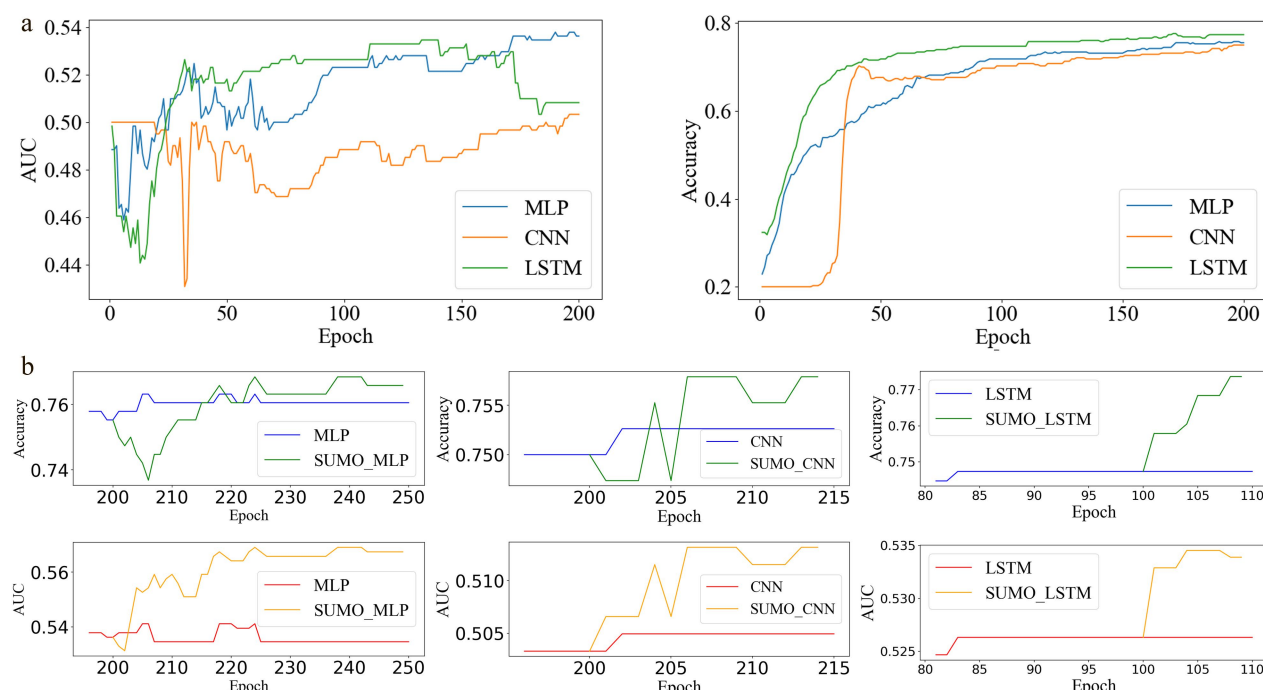
Data was selected from January 1, 2017, and the IDM model was calibrated based on the error between simulated detector data and real data, obtaining parameters:  $a = 3 \text{ m/s}^2$ ,  $b = 2.6 \text{ m/s}^2$ ,  $s_0 = 5 \text{ m}$ ,  $\delta = 4$ ,  $T = 0.25 \text{ s}$ . Next, SUMO was used to simulate traffic flow after

an accident and selected the most realistic simulation settings by calculating the error under different lane counts and influence ranges, obtaining an influence range of 20 and an affected lane number of 0. The time and location of accidents were randomly generated to meet the requirements of the spatial transferability test. The time and location of the accident are determined using a completely randomized approach. Combining the above setup, a simulation dataset is obtained. This simulation setup is consistent with the definition of zero-shot scenarios introduced in the abstract. Since no real accident data exists for the I10-E road, the experiment tests whether models fine-tuned with simulated accident data—originating from different roads—can generalize to such unseen environments.

The experimental results are shown in Fig. 9. From Fig. 9, it is seen that the samples generated were validated through data augmentation not only improving the model's performance but also demonstrating good spatial transferability. Although accident data from the I80-E road and normal traffic flow data from the I10-E road were used for augmentation during training, the testing was still based on real data from the I10-E road. The results show that the fine-tuned model showed significant improvement in key metrics such as accuracy (acc) and AUC value. Specifically, the three neural network models, MLP, CNN, and LSTM, all showed performance improvements after fine-tuning. The accuracy of MLP increased from 75.53% to 76.58%, and AUC increased from 53.62% to 56.74%; CNN's accuracy improved from 75% to 75.79%, and AUC increased from 50.33% to 51.32%; LSTM's accuracy rose from 74.74% to 77.37%, and AUC increased from 52.63% to 53.39%.

Combining the results of the temporal transfer experiments with the spatial transfer experiments, it can be found that in the temporal transfer experiment, LSTM and CNN perform better because they are able to capture spatial and temporal features, respectively. In the spatial transfer experiment, there is not much difference among the three models because the performance improvement of MLP mainly relies on the increase of data diversity, and more training data than





**Fig. 9** Spatial transfer experiment simulation results. (a) The convergence curves of accuracy and AUC under MLP, CNN, and LSTM neural networks. (b) Comparison of accuracy and AUC convergence curves between the fine-tuned model using SUMO simulation data and the baseline model.

in the temporal transfer experiment leads to the improvement of the performance of the MLP model.

Overall, the experimental results have demonstrated the effectiveness of the data augmentation training method based on simulation in spatial transfer scenarios. Even in the absence of accident samples, the performance after fine-tuning has also been improved, indicating that data augmentation, not only improved the training data for the model but also enhanced the model's transfer learning ability, demonstrating the advantage of generated data over the original data in terms of spatial transferability and performance improvement. Moreover, despite the training and testing data coming from different geographical environments and traffic flow characteristics, the data augmentation method was still able to help the model effectively transfer and improve its performance in the new environment, further validating the method's effectiveness and generalization ability.

To further illustrate the effectiveness of the method, the model performance was compared under different data configurations under the cross-model (MLP, CNN, LSTM) setting. Specifically, the experimental results of the spatial migration experiment and the experimental results of the time migration experiment were selected when the ratio of real data to simulated data was 1:19, a total of six groups of paired data, and paired *t*-tests were performed on the accuracy and AUC indicators, respectively. The test results show that the *p*-value of 'accuracy' is 0.0046 and the *p*-value of AUC is 0.0121, which are significantly lower than the commonly used significance level ( $\alpha = 0.05$ ), indicating that the proposed method based on simulation data enhancement is statistically significant in improving model performance.

## Discussion and conclusions

The main contribution of this paper is to demonstrate the effectiveness of simulation data augmentation in improving the performance of traffic anomaly detection models. The experimental results provide strong empirical evidence across multiple dimensions:

**Time transfer performance:** using simulated data generated from 2017 I80-E road data improved model performance when tested on 2022 I80-E road data. After fine-tuning with simulation data, models showed improvements in key metrics such as accuracy and AUC, effectively addressing time-related traffic pattern changes.

**Spatial transfer performance:** the spatial transfer experiment verified that combining simulated data generated from accident data on the I80-E road with normal traffic flow data on the I10-E road enhanced model performance on the I10-E road, despite differences in geographic environment and traffic flow characteristics between the two roads.

**Statistical significance:** To further illustrate the effectiveness of the method, model performance was compared under different data configurations across three architectures (MLP, CNN, LSTM). Using experimental results from spatial and time transfer experiments with a 1:19 ratio of real to simulated data, the test results show that the *p*-value of 'accuracy' is 0.0046 and the *p*-value of AUC is 0.0121, which are significantly lower than the commonly used significance level ( $\alpha = 0.05$ ), confirming the statistical significance of the simulation data augmentation approach.

**Cross-architecture benefits:** all three evaluated models—MLP, CNN, and LSTM—showed varying degrees of improvement after using augmented data, demonstrating that different deep learning architectures benefit from data augmentation, though to different extents.

## Limitation

This study acknowledges that other critical factors—such as adverse weather conditions, road surface quality, lighting, and variability in driver behavior—can significantly influence accident occurrence and traffic dynamics. These were not included in the current simulation due to limitations in SUMO's built-in capabilities and our study's focus on evaluating the core feasibility of physically constrained data augmentation.

## Future directions

Future research should focus on incorporating the aforementioned factors by extending SUMO through external modules or coupling it with more sophisticated behavior models to construct more comprehensive simulation environments that better reflect real-world driving conditions, diversity, and uncertainty.

Additional research directions include exploring the integration of other simulation tools, such as the CityFlow platform based on the SUMO framework with built-in self-driving decision models, and the AIMSUN system, which natively supports hybrid micro-meso-macro modeling with real-time data interfaces. The simulation setup for accident scenarios can be further optimized by introducing more physical constraints, using neural networks to construct more accurate following models, and employing reinforcement learning for parameter selection. Mixed traffic flow environments containing self-driving vehicles could be constructed, incorporating more complex neural network architectures to improve model performance. Expanding the dataset to cover a wider range of accident scenarios and traffic conditions could further strengthen this method's applicability in various traffic management tasks.

## Conclusions

This paper provides strong empirical support for simulation data augmentation as an optimization method for traffic anomaly detection models, particularly when real-world data is scarce or difficult to obtain. Using SUMO simulation software to generate accident scenarios and augment real data has proven statistically significant and effective in improving model robustness and transferability, especially in small sample scenarios.

## Author contributions

The authors confirm contributions to the paper as follows: study conception and design: Yang H, Liao C; data collection: Liao C; analysis and interpretation of results: Yang H, Chen C; draft manuscript preparation: Yang H, Dong Z, Chen C. All authors reviewed the results and approved the final version of the manuscript.

## Data availability

The data that support the findings of this study are available in the Caltrans Performance Measurement System repository. These data were derived from the following resources available in the public domain: <https://pems.dot.ca.gov>.

## Acknowledgments

This research is financially supported by the Zhejiang Provincial Natural Science Foundation of China (Grant Nos LDT23F01011F01, LZ23E080002), 'Pioneer' and 'Leading Goose' R&D Program of Zhejiang (Grant Nos 2023C01038, 2023C03155), and National Natural Science Foundation of China (Grant No. 72171210).

## Conflict of interest

The authors declare that they have no conflict of interest.

## Dates

Received 2 January 2025; Revised 29 July 2025; Accepted 22 September 2025; Published online 31 December 2025

## References

1. World Health Organization. 2023. *Road traffic injuries*. [www.who.int/zh/news-room/fact-sheets/detail/road-traffic-injuries](http://www.who.int/zh/news-room/fact-sheets/detail/road-traffic-injuries) (Retrieved December 21, 2024)
2. Ghosh B, Smith DP. 2014. Customization of automatic incident detection algorithms for signalized urban arterials. *Journal of Intelligent Transportation Systems* 18(4):426–41
3. Li L, Qin L, Qu X, Zhang J, Wang Y, et al. 2019. Day-ahead traffic flow forecasting based on a deep belief network optimized by the multi-objective particle swarm algorithm. *Knowledge-Based Systems* 172:1–14
4. Al Kuwaiti A, Nazer K, Al-Reedy A, Al-Shehri S, Al-Muhanna A, et al. 2023. A review of the role of artificial intelligence in healthcare. *Journal of Personalized Medicine* 13(6):951
5. Wang H, Fu T, Du Y, Gao W, Huang K, et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature* 620:47–60
6. Abdel-Aty M, Uddin N, Pande A, Abdalla MF, Hsia L. 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record* 1897:88–95
7. Abdel-Aty M, Uddin N, Pande A. 2005. Split models for predicting multi-vehicle crashes during high-speed and low-speed operating conditions on freeways. *Transportation Research Record* 1908(1):51–58
8. Oh C, Oh JS, Ritchie SG. 2005. Real-time hazardous traffic condition warning system: Framework and evaluation. *IEEE Transactions on Intelligent Transportation Systems* 6(3):265–72
9. Lee C, Abdel-Aty M. 2006. Temporal variations in traffic flow and ramp-related crash risk. In *Applications of Advanced Technology in Transportation*. Chicago: American Society of Civil Engineers. pp. 244–49 doi: 10.1061/40799(213)40
10. Yu R, Abdel-Aty M. 2013. Multi-level Bayesian analyses for single-and multi-vehicle freeway crashes. *Accident Analysis & Prevention* 58:97–105
11. Yang K, Wang X, Yu R. 2018. A Bayesian dynamic updating approach for urban expressway real-time crash risk evaluation. *Transportation Research Part C: Emerging Technologies* 96:192–207
12. Huang T, Wang S, Sharma A. 2020. Highway crash detection and risk estimation using deep learning. *Accident Analysis & Prevention* 135:105392
13. Pourroostaei Ardakani S, Liang X, Mengistu KT, So RS, Wei X, et al. 2023. Road car accident prediction using a machine-learning-enabled data analysis. *Sustainability* 15(7):5939
14. Yu L, Du B, Hu X, Sun L, Han L, et al. 2021. Deep spatio-temporal graph convolutional network for traffic accident prediction. *Neurocomputing* 423:135–47
15. Guo M, Zhao X, Yao Y, Yan P, Su Y, et al. 2021. A study of freeway crash risk prediction and interpretation based on risky driving behavior and traffic flow data. *Accident Analysis & Prevention* 160:106328
16. Zhang S, Abdel-Aty M. 2022. Real-time crash potential prediction on freeways using connected vehicle data. *Analytic Methods in Accident Research* 36:100239
17. Yuan C, Li Y, Huang H, Wang S, Sun Z, et al. 2022. Using traffic flow characteristics to predict real-time conflict risk: a novel method for trajectory data analysis. *Analytic Methods in Accident Research* 35:100217
18. Payne HJ, Tignor SC. 1978. Freeway incident-detection algorithms based on decision trees with states. *Transportation Research Record* 1978(682):30–37
19. Parkany E, Xie C. 2005. A complete review of incident detection algorithms & their deployment: what works and what doesn't. *Technical Report NETCR37*. New England Transportation Consortium, USA. <https://onlinepubs.trb.org/onlinepubs/trispdfs/00988875.pdf>
20. Jin X, Srinivasan D, Cheu RL. 2001. Classification of freeway traffic patterns for incident detection using constructive probabilistic neural networks. *IEEE Transactions on Neural Networks* 12(5):1173–87
21. Dogru N, Subasi A. 2018. Traffic accident detection using random forest classifier. *2018 15th Learning and Technology Conference (L&T), Jeddah, Saudi Arabia, 25–26 February 2018*. USA: IEEE. pp. 40–45 doi: 10.1109/LT.2018.8368509

22. White J, Thompson C, Turner H, Dougherty B, Schmidt DC. 2011. Wreck-Watch: Automatic traffic accident detection and notification with smart-phones. *Mobile Networks and Applications* 16:285–303
23. Ozbayoglu M, Kucukayan G, Dogdu E. 2016. A real-time autonomous highway accident detection model based on big data processing and computational intelligence. 2016 *IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA, 5–8 December 2016. USA: IEEE. pp. 1807–13 doi: [10.1109/BigData.2016.7840798](https://doi.org/10.1109/BigData.2016.7840798)
24. Gu Y, Qian Z, Chen F. 2016. From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies* 67:321–42
25. Mehrannia P, Bagi SSG, Moshiri B, Al - Basir OA. 2023. Deep representation of imbalanced spatio-temporal traffic flow data for traffic accident detection. *IET Intelligent Transport Systems* 17(3):606–19
26. Li L, Lin Y, Du B, Yang F, Ran B. 2022. Real-time traffic incident detection based on a hybrid deep learning model. *Transportmetrica A: Transport Science* 18(1):78–98
27. Xie T, Shang Q, Yu Y. 2022. Automated traffic incident detection: Coping with imbalanced and small datasets. *IEEE Access* 10:35521–40
28. Jilani U, Asif M, Rashid M, Siddique AA, Talha SMU, et al. 2022. Traffic congestion classification using GAN-based synthetic data augmentation and a novel 5-layer convolutional neural network model. *Electronics* 11(15):2290
29. Huang Y, Wei W, Yang H, Wu Q, Xu K. 2023. Intelligent algorithms for incident detection and management in smart transportation systems. *Computers and Electrical Engineering* 110:108839
30. Dabboussi AH, Jammal M. 2024. Traffic data augmentation using GANs for ITS. 2024 *20<sup>th</sup> International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, Abu Dhabi, United Arab Emirates, 29 April 2024–01 May 2024. USA: IEEE. pp. 66–73 doi: [10.1109/DCOSS-IoT61029.2024.00020](https://doi.org/10.1109/DCOSS-IoT61029.2024.00020)
31. Benabdallah Benarmas R, Beghdad Bey K. 2024. Improving road traffic speed prediction using data augmentation: a deep generative models-based approach. *Annals of Data Science* 11(6):2199–216
32. Liao C, Chen XM. 2024. A meta-learning approach to improving transferability for freeway traffic crash risk prediction. *Digital Transportation and Safety* 4(1):21–30
33. Avila A M, Mezić I. 2020. Data-driven analysis and forecasting of highway traffic dynamics. *Nature Communications* 11(1):2090
34. Qu Q, Shen Y, Yang M, Zhang R. 2024. Towards efficient traffic crash detection based on macro and micro data fusion on expressways: a digital twin framework. *IET Intelligent Transport Systems* 18(12):2725–43
35. Yang K, Quddus M, Antoniou C. 2022. Developing a new real-time traffic safety management framework for urban expressways utilizing reinforcement learning tree. *Accident Analysis & Prevention* 178:106848
36. Zaitouny A, Fragkou AD, Stemler T, Walker DM, Sun Y, et al. 2022. Multiple sensors data integration for traffic incident detection using the quadrant scan. *Sensors* 22(8):2933
37. German Aerospace Center. n.d. SUMO at a glance. *SUMO Documentation*. [https://sumo.dlr.de/docs/SUMO\\_at\\_a\\_Glance.html](https://sumo.dlr.de/docs/SUMO_at_a_Glance.html) (Retrieved December 21, 2024)
38. Krajzewicz D, Brockfeld E, Mikat J, Ringel J, Rössel C, et al. 2005. Simulation of modern traffic lights control systems using the open source traffic simulation SUMO. *Proceedings of the 3<sup>rd</sup> Industrial Simulation Conference 2005*. Berlin, Germany: EUROSIS-ETI. pp. 299–302 <https://elib.dlr.de/21012>
39. Koh SS, Zhou B, Yang P, Yang Z, Fang H, et al. 2018. Reinforcement learning for vehicle route optimization in SUMO. 2018 *IEEE 20<sup>th</sup> International Conference on High Performance Computing and Communications; IEEE 16<sup>th</sup> International Conference on Smart City; IEEE 4<sup>th</sup> International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, Exeter, UK, 28–30 June 2018. USA: IEEE. pp. 1468–73 doi: [10.1109/HPCC/SmartCity/DSS.2018.00242](https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00242)
40. Kastner KH, Pau P. 2015. Experiences with SUMO in a real-life traffic monitoring system. *SUMO 2015–Intermodal Simulation for Intermodal Transport 28*, Berlin. pp. 1–10 [www.researchgate.net/publication/291339917](http://www.researchgate.net/publication/291339917)
41. Fernandes P, Nunes U. 2010. Platooning of autonomous vehicles with intervehicle communications in SUMO traffic simulator. 13<sup>th</sup> *International IEEE Conference on Intelligent Transportation Systems*, Funchal, Portugal, 19–22 September 2010. USA: IEEE. pp. 1313–18 doi: [10.1109/ITSC.2010.5625277](https://doi.org/10.1109/ITSC.2010.5625277)
42. Shamsashtiany R, Ameri M. 2018. Road accidents prediction with multi-layer perceptron (MLP) modelling case study: roads of Qazvin, Zanjan and Hamadan. *Journal of Civil Engineering and Materials Application* 2(4):181–92



Copyright: © 2025 by the author(s). Published by Maximum Academic Press, Fayetteville, GA. This article is an open access article distributed under Creative Commons Attribution License (CC BY 4.0), visit <https://creativecommons.org/licenses/by/4.0/>.