ST-TAR: An Efficient Spatio-Temporal Learning Framework for Traffic Accident Risk Forecasting

Hongyu Wang[1], **Lisi Chen**[1*], **Shuo Shang**[1*], **Peng Han**[1] and **Christian S. Jensen**[2]
[1]University of Electronic Science and Technology of China
[2]Aalborg University
wanghongyu907@gmail.com, lchen012@e.ntu.edu.sg, jedi.shang@gmail.com,
penghan_study@foxmail.com, csj@cs.aau.dk

Abstract

Traffic accidents represent a significant concern due to their devastating consequences. The ability to predict future traffic accident risks is of key importance to accident prevention activities in transportation systems. Although existing studies have made substantial efforts to model spatio-temporal correlations, they fall short when it comes to addressing the zero-inflated data issue and capturing spatio-temporal heterogeneity, which reduces their predictive abilities. In addition, improving efficiency is an urgent requirement for traffic accident forecasting. To overcome these limitations, we propose an efficient Spatio-Temporal learning framework for Traffic Accident Risk forecasting (ST-TAR). Taking long-term and short-term data as separate inputs, the ST-TAR model integrates hierarchical multi-view GCN and long short-term cross-attention mechanism to encode spatial dependencies and temporal patterns. We leverage long-term periodicity and short-term proximity for spatio-temporal contrastive learning to capture spatio-temporal heterogeneity. A tailored adaptive risk-level weighted loss function based on efficient locality-sensitive hashing is introduced to alleviate the zero-inflated issue. Extensive experiments on two real-world datasets offer evidence that ST-TAR is capable of advancing state-of-the-art forecasting accuracy with improved efficiency. This makes ST-TAR suitable for applications that require accurate real-time forecasting.

1 Introduction

Urbanization and the proliferation of vehicles have caused traffic accidents to become one of the world's largest public-health threats. The Global Plan for the Decade of Action for Road Safety 2021–2030 [Organization and others, 2021] reports nearly 1.3 million avoidable deaths and around 50 million injuries worldwide yearly due to road accidents. Moreover, as it stands without intervention measures, there will be an estimated 13 million deaths and 500 million injuries in the next decade. In light of these disconcerting statistics, advanced accident prediction mechanisms are urgently needed to identify and mitigate risks as early as possible, safeguarding public health and safety.

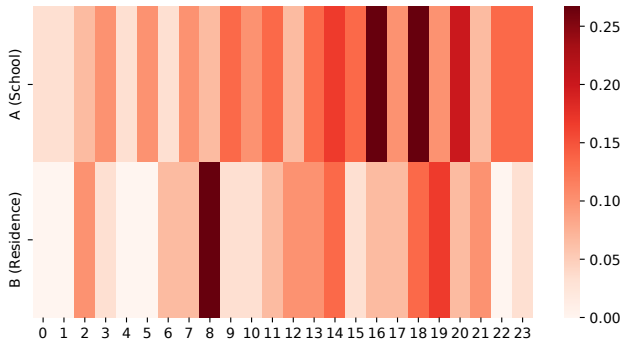


Figure 1: An example of spatio-temporal heterogeneity in traffic accidents. The figure visualizes the frequency of accidents in two areas A and B during each hour of the day.

Traffic accidents are influenced by various spatio-temporal factors [Bergel-Hayat *et al.*, 2013; Trirat *et al.*, 2023]. Traditional statistical learning-based methods [Sharma *et al.*, 2016; Barba *et al.*, 2014] struggle to capture the complex correlations, while deep learning modules such as Graph Convolutional Networks (GCN) [Kipf and Welling, 2017] and Transformers [Vaswani, 2017] have emerged as powerful tools for modeling spatial dependencies and temporal patterns. Deep learning-based methods [Wang *et al.*, 2021; An *et al.*, 2022; Wang *et al.*, 2023; Chen *et al.*, 2024] integrate these modules with customized designs to achieve better performance. Although existing methods have made great advances in traffic accident forecasting, they still have three major limitations.

The first limitation is the problem of zero-inflated data. Traffic accidents are low-probability events in a city, with most regions and time intervals labeled as 0 due to their infrequent occurrence. In the deep learning models, the resulting excessively imbalanced label distribution tends to lead all predictions to be 0, reducing their utility [Bao *et al.*, 2019]. Existing solutions [Wang *et al.*, 2023; Chen *et al.*, 2024] follow a process where they build multi-level graphs and then aggregate fine-grained features to obtain high-level features with a smaller proportion of zero labels. In addition, weighted loss

*Corresponding authors.

function [Wang *et al.*, 2021; Chen *et al.*, 2024] is introduced to assign higher weights to samples with high accident values. However, the weighted loss function only enhances the importance of non-zero data but ignores a large amount of zero-label data that can be further mined.

The second limitation is a lack of modeling of the spatio-temporal heterogeneity. As illustrated in Figure 1, A and B are schools and residential areas in New York City with different urban functions. We collect and calculate the frequency of traffic accidents at each hour of the day during February 2013[1], presenting these data as a heat map. We observe a significant inconsistency in accident frequency between the two regions at the same hour of the day. In area A, the peak hours for traffic accidents are from 4 p.m. to 5 p.m. and from 6 p.m. to 7 p.m., while in area B, the peak hour of traffic accidents occurs from 8 a.m. to 9 a.m. Thus, there are distinct spatial differences between the two areas. Meanwhile, the frequency of accidents in both regions fluctuates significantly over time, indicating obvious temporal dynamics within each region. Existing methods that model spatial differences and temporal dynamics through shared parameters without any explicit handling face difficulty in effectively capturing spatio-temporal heterogeneity.

The third limitation is insufficient efficiency. Traffic accident forecasting is highly time-sensitive. As the spatial and temporal granularity of predictions becomes more refined, existing approaches may struggle to meet the demands for timeliness and resource consumption in the future. Many existing studies [Wang *et al.*, 2021; Wang *et al.*, 2023; Chen *et al.*, 2024] combine long-term and short-term data into single sequential input, and simultaneously feed sequence data into grid-based modules (e.g., CNN) and graph-based modules (e.g., GCN) to enhance model performance. Although these two types of modules each have unique advantages, their simultaneous use not only requires additional hardware resources but also increases the training time, which may be a bottleneck for large-scale data applications.

To overcome these limitations, we propose a Spatio-Temporal Learning framework for Traffic Accident Risk forecasting (ST-TAR). The ST-TAR leverages hierarchical multi-view GCN and long short-term cross-attention mechanism to encode spatio-temporal traffic correlations. More specifically, ST-TAR models long-term and short-term historical data respectively for spatio-temporal contrastive learning to capture spatio-temporal heterogeneity. The zero-inflated data issue is alleviated through hierarchical structure and adaptive risk-level weighted loss function. In particular, we propose an efficient auxiliary label processing algorithm that exploits locality-sensitive hashing to reduce the label distribution imbalance. By exploiting contextual spatio-temporal information, ST-TAR effectively addresses the aforementioned limitations to improve the effectiveness and efficiency of traffic accident risk forecasting.

The main contributions are summarized as follows:

- We propose a novel spatio-temporal learning framework named ST-TAR that effectively utilizes short-term proximity and long-term periodicity with hierarchical multi-view GCN and long short-term cross-attention mechanism for traffic accident risk forecasting.
- We introduce spatio-temporal contrastive learning and adaptive risk-level weighted loss function with an efficient auxiliary label processing algorithm that can capture spatio-temporal heterogeneity and alleviate the zero-inflated data issue.
- We report on an extensive experimental study on two real-world datasets. The results demonstrate that our proposed ST-TAR model outperforms baseline models while utilizing fewer computational resources and requiring less runtime. This underlines the effectiveness and efficiency of ST-TAR in forecasting traffic accident risk from large-scale spatio-temporal data. The source code of our ST-TAR implementation is publicly available at https://github.com/wanghyhy/ST-TAR.

[1]https://opendata.cityofnewyork.us/

2 Related Work

Existing methods can generally be categorized into statistical learning-based methods and deep learning-based methods. Statistical learning-based methods such as K-Nearest Neighbors (KNN) [Lv *et al.*, 2009], decision trees [Lin *et al.*, 2015], Support Vector Machine (SVM) [Sharma *et al.*, 2016], and Autoregressive Integrated Moving Average Model (ARIMA) [Barba *et al.*, 2014] make predictions on small-scale traffic accident data with limited features, and they are unable to capture complex spatio-temporal dependencies in historical data.

To model spatio-temporal correlations in traffic data more effectively, recent methods utilize deep learning modules such as graph convolutional networks [Guo *et al.*, 2021; Luo *et al.*, 2022; Han *et al.*, 2021]. These methods demonstrate advantages by effectively capturing complex temporal and spatial dependencies that are inherent to traffic data. Hetero-ConvLSTM [Yuan *et al.*, 2018] employs an ensemble method with a convolutional long short-term memory network to handle the spatial heterogeneity, but manually-selected regions limit the spatial patterns. Therefore, HintNet [An *et al.*, 2022] devises a multi-level risk-based spatial partitioning with a hierarchical knowledge transfer network to capture irregular spatial heterogeneity patterns. To alleviate the zero-inflated data issue, GSNet [Wang *et al.*, 2021] presents a geographical and semantic spatio-temporal network with a weighted loss function. Further, MVMT-STN [Wang *et al.*, 2023] leverages a multi-task learning framework to predict fine and coarse-grained traffic accidents jointly to contend with data sparsity. More recently, MGH-STN [Chen *et al.*, 2024] introduces multi-level hierarchical structures with multivariate hierarchical loss function, and incorporates remote sensing images to make a comprehensive traffic accident risk prediction. In addition to efficiency, none of these studies explicitly consider the challenges of zero-inflated data issue and spatio-temporal heterogeneity simultaneously for traffic accident forecasting.

3 Preliminaries and Definitions

Definition 1 (Grid and Region). A city is partitioned into a regular grid with $I \times J$ cells based on longitude and latitude.

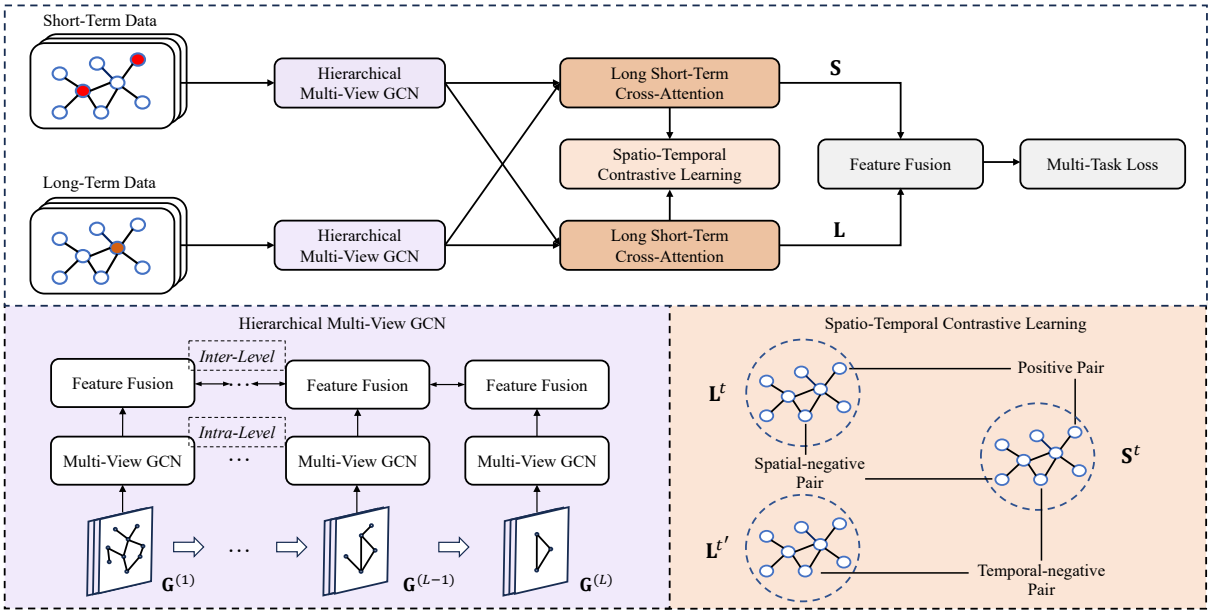


Figure 2: Architecture of the proposed ST-TAR model.

Since cities typically exhibit irregular shapes, only a subset $R = \{r_1, r_2, \ldots, r_N\}$ contains road segments (i.e., $N \leq I \times J$), and each r_i in the subset R is defined as a region.

Definition 2 (**Spatio-temporal Features**). Spatio-temporal features $\mathbf{x}_i^t \in \mathbf{X}$ of region r_i in time interval t are categorized into static and dynamic features. Static features are the POI distribution, while dynamic features include temporal information, weather data, traffic flow, and accident risk values.

Definition 3 (**Traffic Accident Risk**). Three types of traffic accidents are defined based on the number of casualties in traffic accidents, i.e., minor accidents, injured accidents, and fatal accidents, with corresponding risk values set to be 1, 2, and 3, respectively [Wang *et al.*, 2021]. The traffic accident risk $y_i^t \in \mathbf{Y}$ is equal to the sum of the risk values of accidents that occurred in region r_i in time interval t.

Definition 4 (**Hierarchical Multi-view Graph**). To capture spatial dependency at multiple granularities and from various semantic perspectives, a hierarchical multi-view graph $\mathbf{G} = \{\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \ldots, \mathbf{G}^{(L)}\}$ is constructed, where L represents the number of layers. Each layer $\mathbf{G}^{(i)} = (V^{(i)}, E^{(i)}, \mathbf{X}^{(i)})$ is a multi-view graph at a specific granularity level, which consists of three views of graph: road similarity graph $\mathbf{G}_D^{(i)}$, POI similarity graph $\mathbf{G}_P^{(i)}$, and risk similarity graph $\mathbf{G}_K^{(i)}$. $\mathbf{G}^{(1)}$ is the finest-grained graph for prediction (i.e., $\mathbf{X}^{(1)} = \mathbf{X}$ where $\mathbf{X}$ is the original feature).

Definition 5 (**Traffic Accident Risk Forecasting**). Given the historical features $\{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^T\}$, a hierarchical multi-view graph $\mathbf{G}$ and target time information $\mathbf{z}^{T+1}$, our goal is to forecast the traffic accident risk $\hat{\mathbf{Y}}^{T+1}$ in the time interval $T+1$ with the model f:

$$\hat{\mathbf{Y}}^{T+1} \leftarrow f(\mathbf{X}^1, \ldots, \mathbf{X}^T, \mathbf{z}^{T+1}, \mathbf{G}). \quad (1)$$

4 Methodology

In this section, we present our ST-TAR model for traffic accident risk forecasting, as illustrated in Figure 2. The complete historical data is composed of long-term and short-term data. Historical data from the p most recent time intervals is short-term data, and the same time interval in the previous q weeks is long-term data ($T = p + q$). In contrast to most existing methods that concatenate long-term and short-term data into a single sequential input, we treat long-term and short-term data as separate inputs, respectively, to capture spatio-temporal correlations across different time steps.

4.1 Hierarchical Multi-view GCN

High-level features can significantly reduce the proportion of zero-label data by aggregating information from multiple low-level nodes. Moreover, these high-level features enhance fine-grained predictions by providing rich contextual information. The hierarchical multi-view GCN consists of two main stages: intra-level multi-view GCN and inter-level feature fusion.

Intra-level Multi-view GCN. To model neighbor interactions with spatial dependencies, we adopt the GCN [Kipf and Welling, 2017] for each view within a hierarchical level. The adjacency matrix is augmented with a self-connection to obtain matrix $\tilde{\mathbf{A}}$ and degree matrix $\tilde{\mathbf{D}}$. Then we can get the normalized symmetric adjacency matrix $\hat{\mathbf{A}} = \tilde{\mathbf{D}}^{(-\frac{1}{2})}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{(-\frac{1}{2})}$. Then the graph convolution operation can be formulated as follows:

$$\mathbf{H}_c^l = \mathrm{GCN}(\mathbf{A}, \mathbf{H}_c^{l-1}) = \sigma(\hat{\mathbf{A}}\mathbf{H}_c^{l-1}\mathbf{W}_c^l + \mathbf{b}_c^l), \quad (2)$$

where $\mathbf{H}_c^l$ represents the l-th layer of the hidden features matrix in the GCN, $\mathbf{W}_c^l$ and $\mathbf{b}_c^l$ are learnable parameters of the l-th graph convolutional layer.

In general, we employ a two-layer GCN to learn the spatial features of each view and then aggregate them by summation, which is formulated as follows:

$$\mathbf{H}_c = \sum_{i \in \{D,P,K\}} \mathrm{GCN}(\mathbf{A}_i, \mathrm{GCN}(\mathbf{A}_i, \mathbf{X})), \tag{3}$$

where $\mathbf{X}$ denotes the initial embeddings of regions, and $\mathbf{A}_i \in \{\mathbf{A}_D, \mathbf{A}_P, \mathbf{A}_K\}$ denotes the adjacency matrix of each view.

Inter-level Feature Fusion. For two adjacent k-th and k+1-th levels, the nodes at the k+1-th layer aggregate multiple nodes from the k-th layer. The corresponding relationship can be represented by matrix $\mathbf{B}_{k,k+1}$, which is represented as:

$$\mathbf{B}_{k,k+1}(i,j) = \begin{cases} 1 & \text{if } r_i^{(k)} \in r_j^{(k+1)} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

By utilizing matrix $\mathbf{B}_{k,k+1}$ as a bridge to represent the relationships between nodes in adjacent levels, the feature fusion process is executed sequentially from the finest-grained level to the coarsest-grained level, which is consequently represented as follows:

$$\begin{aligned} \mathbf{H}_f^{(k)} &= \mathbf{H}_c^{(k)} + \lambda_f \mathbf{B}_{k,k+1} \mathbf{H}_c^{(k+1)}, \\ \mathbf{H}_f^{(k+1)} &= \mathbf{H}_c^{(k+1)} + (1-\lambda_f) \mathbf{B}_{k,k+1}^T \mathbf{H}_f^{(k)}, \end{aligned} \tag{5}$$

where $\mathbf{H}_f^{(k)}$ and $\mathbf{H}_f^{(k+1)}$ represent the feature fused from the k+1-th layer to the k-th layer and from the k-th layer to the k+1-th layer, respectively. λ_f is hyperparameters that balance the weights between low-level and high-level layers.

4.2 Long Short-term Cross-attention Mechanism

To distinguish the output of long-term and short-term data through hierarchical multi-view GCN module, we use the $\mathbf{H}_{f,l}$ and $\mathbf{H}_{f,s}$ to represent the embeddings learned from long-term and short-term data, respectively (we omit the superscripts indicating the level because the attention mechanism is implemented independently within each level). Given target time information $\mathbf{z}^{T+1}$, we utilize $\mathbf{H}_{f,l}$ and $\mathbf{H}_{f,s}$ as the keys and values in the attention mechanism, while concatenating $\mathbf{z}^{T+1}$ and complementary embeddings as respective queries, which are represented as follows:

$$\begin{aligned} \mathbf{Q}_s = (\mathbf{H}_{f,l} \parallel \mathbf{z}^{T+1})\mathbf{W}_Q,\ \mathbf{Q}_l = (\mathbf{H}_{f,s} \parallel \mathbf{z}^{T+1})\mathbf{W}_Q, \\ \mathbf{K}_s = \mathbf{H}_{f,s}\mathbf{W}_K,\ \mathbf{K}_l = \mathbf{H}_{f,l}\mathbf{W}_K, \\ \mathbf{V}_s = \mathbf{H}_{f,s}\mathbf{W}_V,\ \mathbf{V}_l = \mathbf{H}_{f,l}\mathbf{W}_V, \end{aligned} \tag{6}$$

where $\mathbf{W}_Q$,$\mathbf{W}_K$ and $\mathbf{W}_V$ are learnable parameters, and $\parallel$ refers to the concatenation operation. We process the above features through a multi-head attention method:

$$\begin{aligned} \mathbf{S} = \mathrm{MultiHead}(\mathbf{Q}_s,\mathbf{K}_s,\mathbf{V}_s), \mathbf{L} = \mathrm{MultiHead}(\mathbf{Q}_l,\mathbf{K}_l,\mathbf{V}_l), \\ \mathrm{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_h)\mathbf{W}_O, \\ \mathrm{head}_i = \mathrm{Attention}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right), \\ \mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \end{aligned} \tag{7}$$

where $\mathbf{W}_O$ is the learnable linear transformation to merge the information from various attention heads.

4.3 Feature Fusion and Prediction

In general, short-term proximity and long-term periodicity influence the target time to different degrees. Consequently, we employ two trainable weight matrices $\mathbf{W}_s^{(k)}$ and $\mathbf{W}_l^{(k)}$ in each level to dynamically integrate the outputs of long short-term data and employ a fully-connected (FC) layer to make the risk prediction at each granularity:

$$\hat{\mathbf{Y}}^{(k)} = \mathrm{FC}\left(\mathbf{W}_s^{(k)}\mathbf{S}^{(k)} \parallel \mathbf{W}_l^{(k)}\mathbf{L}^{(k)}\right). \tag{8}$$

4.4 Spatio-Temporal Contrastive Learning

With the proposed hierarchical multi-view GCN and long short-term cross-attention mechanism, we obtain two feature representations of short-term proximity and long-term periodicity as $\mathbf{S}$ and $\mathbf{L}$, which contribute to the prediction at the target time from different perspectives. Therefore, we regard the short-term embedding and long-term embedding in the same region for the same target time as positive pairs $(\mathbf{s}_i^t, \mathbf{l}_i^t)$, while the embeddings at different regions (i.e., $r_i \neq r_j$) and for different target time (i.e. $t \neq t'$) as spatial-negative pairs $(\mathbf{s}_i^t, \mathbf{l}_j^t)$ and temporal-negative pair $(\mathbf{s}_i^t, \mathbf{l}_i^{t'})$.

Following this criterion, the correlation between short-term proximity and long-term periodicity is manifested as a complementary relationship in the contrastive learning process, which promotes representations in both views that retain spatio-temporal heterogeneity to match each other. The spatio-temporal contrastive learning task is optimized with BCEWithLogitsLoss, which is defined as follows:

$$\begin{aligned} \mathcal{L}_s = -\Bigg[\sum_{i=1}^{N} \log D\left(\mathbf{s}_i^t, \mathbf{l}_i^t\right) + \\ \sum_{i=1}^{N} \log\left(1 - D\left(\mathbf{s}_i^t, \mathbf{l}_j^t\right)\right) + \sum_{i=1}^{N} \log\left(1 - D\left(\mathbf{s}_i^t, \mathbf{l}_i^{t'}\right)\right) \Bigg]. \end{aligned} \tag{9}$$

Here, $D(\cdot)$ is a discriminator that evaluates the matching scores between two input embeddings. Specifically, we implement it with a fully connected network fed by the concatenation of the input, which is defined as:

$$D\left(\mathbf{s}_i^t, \mathbf{l}_i^t\right) = \sigma\left(\mathbf{W}_d\left(\mathbf{s}_i^t \parallel \mathbf{l}_i^t\right) + \mathbf{b}_d\right), \tag{10}$$

where $\mathbf{W}_d$ and $\mathbf{b}_d$ are learnable parameters of the discriminator.

4.5 Adaptive Risk-level Weighted Loss Function

Zero-inflated data causes common loss functions (e.g., mean squared error) to disproportionately focus on zero-label data, thereby neglecting non-zero data. To alleviate this issue, we classify the data labeled as 0 at the finest-grained level (i.e., $\mathbf{Y}^{(1)} = 0$, we omit the level superscripts for simplicity) into high-risk and low-risk based on the feature similarity. The resulting auxiliary labels serve as soft targets [Hinton *et al.*, 2015] to implicitly optimize the label distribution to enhance the discriminative ability of the model.

An intuitive approach is to set a threshold for distinguishing high-risk and low-risk labels by calculating pairwise feature similarities (e.g., cosine similarity) between the

accident-occurred and non-accident intervals in a given region. However, this is extremely time-consuming when processing large-scale spatio-temporal data.

To improve efficiency, we utilize the hash collisions of Locality Sensitive Hashing (LSH) [Indyk and Motwani, 1998] for approximate similarity search. LSH is an algorithmic technique that employs hash functions to efficiently map similar data points into the same bucket with high probability, reducing the computational cost of similarity search. Due to its simplicity and effectiveness, we adopt random projection hash [Charikar, 2002] as the specific LSH method. Given the spatio-temporal features $\mathbf{x}_i^t$, we obtain a random vector $\mathbf{u}$ from the Gaussian distribution (i.e., each coordinate is drawn from the 1-dimensional Gaussian distribution). Then the hash value $p(\mathbf{x}_i^t)$ is calculated as follows:

$$p(\mathbf{x}_i^t) = \begin{cases} 1 & \text{if } \mathbf{u} \cdot \mathbf{x}_i^t \geq 0 \\ 0 & \text{if } \mathbf{u} \cdot \mathbf{x}_i^t < 0 \end{cases} \tag{11}$$

In our model, we construct a hash table by sampling K independent random projection hash functions, denoted by $P = \{p_1, p_2, \ldots, p_K\}$. As a single random projection hash function returns a one-bit output (0 or 1), the hash table returns a K-bit output. Given that $\mathbf{A}$ and $\mathbf{B}$ are represented as feature vectors, the collision probability in this condition conforms as follows [Goemans and Williamson, 1995]:

$$\begin{aligned} \Pr[P(\mathbf{A}) = P(\mathbf{B})] &= \left(1 - \frac{\theta}{\pi}\right)^K, \text{ where} \\ \theta &= \arccos\left(\frac{\mathbf{A} \cdot \mathbf{B}}{\sqrt{|\mathbf{A}| \cdot |\mathbf{B}|}}\right). \end{aligned} \tag{12}$$

Therefore, according to the power law, the collision probability declines sharply as K increases, ensuring that similar embeddings have a high probability of obtaining identical outputs. In a specific region, we initialize hash buckets and map the accident-occurred hash values into buckets. For accident-occurred intervals, we label them as accident ($c_i^t = 2$). For non-accident intervals, if the K-bit hash values match a non-empty bucket, we label them as high-risk ($c_i^t = 1$); otherwise, we label them as low-risk ($c_i^t = 0$). We obtain three types of auxiliary labels, and the relationship with traffic accident values is as follows:

$$c_i^t = \begin{cases} 0 \text{ or } 1, & \text{if } y_i^t = 0 \\ 2, & \text{if } y_i^t > 0 \end{cases} \tag{13}$$

Time complexity. The time complexity of auxiliary label processing algorithm is $O(N \times M \times K \times d_h + 2 \times N \times M) = O(N \times M \times K \times d_h)$, where N, M, and K denote the number of regions, time intervals, and hash functions, respectively, and d_h denotes the feature dimensionality. Compared with the intuitive solution, which has a complexity of $O(N \times M_a \times M_n \times d_h)$ (M_a and M_n denote the number of time intervals with and without accident occurrences, which satisfy $M = M_a + M_n$), our method has a better complexity. Moreover, our method is more suitable for online risk prediction due to the $O(K \times d_h)$ complexity of an update operation, while the intuitive solution has a complexity of $O(M \times d_h)$.

Finally, the adaptive risk-level weighted loss function based on Mean Squared Error (MSE) can be divided into two parts: zero-label data and non-zero data. For non-zero data, we classify them into three levels according to the accident risk value $I_{nz} = \{\mathbf{Y}{=}1, \mathbf{Y}{=}2, \mathbf{Y}{\geq}3\}$ and assign them different weights, which are expressed as:

$$\mathcal{L}_{nz} = \frac{1}{N_{nz}} \sum_{k \in I_{nz}} \lambda_k^{nz} \cdot (\mathbf{Y}(k) - \hat{\mathbf{Y}}(k))^2, \tag{14}$$

where N_{nz} is the number of non-zero samples, λ_k^{nz} are hyperparameter weights of three risk value levels.

For zero-label data, we classify them into two levels according to the auxiliary labels $I_z = \{\mathbf{C}{=}0, \mathbf{C}{=}1\}$ and assign them different weights, which are expressed as:

$$\mathcal{L}_z = \frac{1}{N_z} \sum_{k \in I_z} \lambda_k^z \cdot (\mathbf{Y}(k) - \hat{\mathbf{Y}}(k))^2, \tag{15}$$

where N_z is the number of zero-label samples, λ_k^z are hyperparameter weights of high-risk and low-risk data.

The overall adaptive risk-level weighted loss function is defined as follows:

$$\mathcal{L}_r = \mathcal{L}_z + \mathcal{L}_{nz}. \tag{16}$$

4.6 Model Optimization

In the training process, adaptive risk-level weighted loss and spatio-temporal contrastive loss are jointly optimized with the hierarchical loss. Therefore, the total loss of multi-task learning is represented as follows:

$$\mathcal{L} = \lambda_r \mathcal{L}_r + \lambda_c \mathcal{L}_c + \lambda_h \mathcal{L}_h, \tag{17}$$

where $\mathcal{L}_h$ represents the hierarchical loss following existing hierarchical studies [Wang *et al.*, 2023; Chen *et al.*, 2024]. λ_r, λ_c and λ_h are the hyperparameters of the loss function to balance the importance of the different tasks.

5 Experimental Study

We conduct extensive experiments on real-world datasets to evaluate the effectiveness and efficiency of the proposed model. Moreover, we conduct the ablation study and create visualizations to assess and illustrate the impact of the different components on the model.

5.1 Experimental Settings

Datasets. Experiments are conducted on two public real-world traffic accident datasets collected from New York City

Dataset	NYC	Chicago
Time span	1/1/2013 - 12/31/2013	2/1/2016 - 9/30/2016
Traffic accidents	147k	44k
Taxi trips	173,179k	1,744k
POIs	15,625	None
Hours of weather	8,760	5,832
Road segments	103k	56k

Table 1: Statistics of datasets.

Models	*NYC*			*Chicago*		
	RMSE/RMSE*	RECALL/RECALL*	MAP/MAP*	RMSE/RMSE*	RECALL/RECALL*	MAP/MAP*
ARIMA	10.4025/9.4632	26.84%/28.56%	0.1094/0.1187	13.7652/10.6935	16.27%/18.45%	0.0579/0.0637
MLP	8.6526/7.8175	27.25%/29.17%	0.1212/0.1269	12.6740/8.5328	17.12%/19.84%	0.0684/0.0735
ConvLSTM	7.6919/7.3402	30.88%/31.33%	0.1557/0.1629	11.2145/8.4636	18.49%/20.14%	0.0759/0.0823
Hetero-ConvLSTM	7.7904/7.3988	29.74%/30.76%	0.1504/0.1567	11.3112/8.5107	18.72%/19.20%	0.0712/0.0746
GSNet	7.5842/6.7335	33.29%/34.15%	0.1820/0.1788	11.1605/8.5616	20.16%/21.45%	0.0869/0.1021
HintNet	8.0867/7.1372	32.56%/33.40%	0.1702/0.1734	11.3792/8.7764	19.21%/20.53%	0.0806/0.0924
MVMT-STN	10.1781/9.4540	33.56%/34.25%	0.1862/0.1807	13.0438/10.2517	20.27%/21.65%	0.0898/0.1062
MGHSTN	6.8083/6.4137	34.13%/34.54%	**0.1932**/0.1853	7.7628/6.0792	20.87%/22.06%	0.0912/0.1128
ST-TAR	**6.6738/6.2040**	**34.37%/35.11%**	0.1921/**0.1884**	**7.5044/5.5936**	**21.12%/22.39%**	**0.0951/0.1187**

Table 2: Performance comparison of different methods on the NYC and Chicago datasets. RMSE*/RECALL*/MAP* represents the performance during rush hours (7:00–9:00 and 16:00–19:00). We denote the best results in **boldface** and the second best results in underline. To ensure a fair comparison, we modify the MGHSTN model by removing the remote sensing data, which is not available in other models.

(NYC)[2] and Chicago[3]. Statistics of the datasets are presented in Table 1. The POI data includes seven categories: residence, school, culture facility, recreation, social service, transportation, and commercial area. Due to the absence of POI data in the Chicago dataset, we only constructed neighborhood, risk similarity, and road similarity graphs for this dataset. The weather data includes temperature and weather conditions, including sunny, rainy, cloudy, snowy, and misty days. We employ the taxi trip data to calculate the inflow and outflow of each region as the human mobility data.

Experimental Setup. Following existing studies of traffic accident forecasting [Wang *et al.*, 2021; Wang *et al.*, 2023], we divide the data on the timeline into training, validation, and test sets in a ratio of 6:2:2. A city is partitioned into grid cells of size about 2km $\times$ 2km. The length of short-term input data p and of long-term input data q are set to 3 and 4, respectively. The time interval length is set to 1 hour.

Baselines. We compare ST-TAR with eight baselines for traffic accident risk forecasting: (1) statistical methods: **ARIMA**, (2) grid-based learning methods: **MLP**, **ConvLSTM** [Shi *et al.*, 2015], **Hetero-ConvLSTM** [Yuan *et al.*, 2018] (3) graph-based learning methods: **GSNet** [Wang *et al.*, 2021], **HintNet** [An *et al.*, 2022], **MVMT-STN** [Wang *et al.*, 2023], **MGHSTN** [Chen *et al.*, 2024].

Evaluation Metrics. We adopt Root Mean Square Error (RMSE), Recall, and Mean Average Precision (MAP) to assess the performance. Due to the high frequency of traffic accidents during rush hours (i.e., 7:00–9:00 and 16:00–19:00), we independently use RMSE*, Recall*, and MAP* to evaluate the performance during these periods to evaluate the performance of our model more comprehensively.

5.2 Effectiveness Comparison

Table 2 presents the performance comparison of all methods on the two datasets. Our model achieves the best performance in most metrics on all datasets, which demonstrates the superiority of ST-TAR. We attribute the performance improvement to spatio-temporal contrastive learning and adaptive risk-level weighted loss function for capturing spatio-temporal heterogeneity and alleviating the zero-inflated data issue simultaneously. More specifically, the ST-TAR model performs well both throughout the entire day and during rush hours, demonstrating its robustness under diverse conditions.

Furthermore, the statistical learning method ARIMA has the worst performance, which is due to its limited capability in capturing complex spatial and temporal correlations. Among the deep learning models, the performance of basic MLP is inferior to ConvLSTM and Hetero-ConvLSTM, which utilize convolutional networks and LSTMs to capture spatial and temporal dependencies. The graph-based models (GSNet, HintNet, MVMT-STN, MGHSTN, and ST-TAR) perform better than grid-based models (ConvLSTM and Hetero-ConvLSTM), indicating the effectiveness of modeling non-Euclidean spatial structure for traffic accident forecasting. However, these baseline methods do not simultaneously address the zero-inflated data issue and capture spatio-temporal heterogeneity. Therefore, ST-TAR achieves superior performance compared to baselines.

5.3 Ablation Study

To validate the effectiveness of the different components in ST-TAR, we conduct an ablation study with three variants: (1) **ST-TAR-LS** combines long-term and short-term data into a single sequential input. (2) **ST-TAR-CL** removes the spatio-temporal contrastive learning. (3) **ST-TAR-WL** replaces the adaptive risk-level weighted loss function with a classic MSE loss function. Figure 3 shows the experimental results between ST-TAR and these variants. We observe that ST-TAR outperforms the three variants across all metrics, indicating that each component contributes to the model. Specifically, the deteriorated RMSE and MAP of both datasets in ST-TAR-CL suggest that separately considering long-term and short-term data benefits long-term periodicity and short-term proximity modeling. Furthermore, ST-TAR-CL outperforms ST-TAR-WL in regression prediction on the NYC dataset while the converse is observed on the Chicago dataset, proving both capturing spatio-temporal heterogeneity and alleviating the zero-inflated data issue can enhance model performance across different scenarios.

[2]https://opendata.cityofnewyork.us/

[3]https://data.cityofchicago.org/

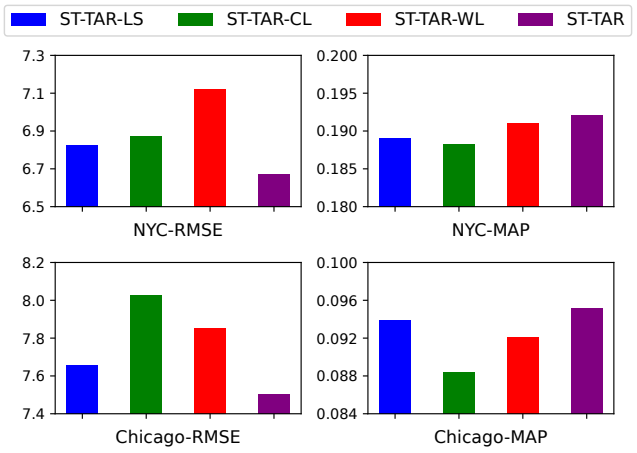


Figure 3: Performance of the ST-TAR model and variants.

Models	*NYC*		*Chicago*	
	Training (s/epoch)	Inference (s)	Training (s/epoch)	Inference (s)
GSNet	14.90	1.28	8.95	0.82
MVMT-STN	23.62	2.15	12.40	0.96
MGHSTN	18.33	1.56	16.13	1.02
ST-TAR	**10.64**	**0.87**	**5.42**	**0.45**
Improvement	28.59%	32.03%	39.44%	45.12%

Table 3: Running time on NYC and Chicago datasets.

5.4 Efficiency Comparison

We compare the training time per epoch and the inference time of our ST-TAR model and the three best prediction performing baselines: GSNet, MVMT-STN, and MGHSTN. We also record the GPU memory usage of these models for training. The results are shown in Table 3 and Figure 4. Moreover, the auxiliary label processing procedure takes 10.8 seconds and 5.2 seconds on NYC and Chicago datasets respectively, roughly equivalent to one training epoch time, indicating that the pre-processing does not impose a significant overhead. Compared to the baselines, ST-TAR saves more than 25% of running time with minimal GPU usage. Different from the baselines that handle long-term and short-term data as a concatenated input and predict their results by fusing the outputs of their respective grid-based and graph-based modules, ST-TAR processes long-term and short-term data separately and performs risk prediction solely based on graph-based feature learning, thus significantly improving the efficiency.

5.5 Visualization

To illustrate the impact of auxiliary labels on the zero-inflated data issue intuitively, we visualize the spatial label distribution before and after processing on the NYC dataset from 10 p.m. to 11 p.m. on February 2, 2013. As presented in Figure 5, accidents only occurred in a small fraction of regions. Our auxiliary label processing algorithm classifies non-accident regions as high-risk and low-risk, resulting in a more balanced label distribution. Notably, high-risk regions are mainly concentrated near the accident-occurred regions, which aligns with Tobler's first law of geography [Tobler, 1970] as geographically close regions tend to have similar characteristics in terms of traffic flow, weather, and other factors, making accidents more likely to occur near the accident-occurred regions.

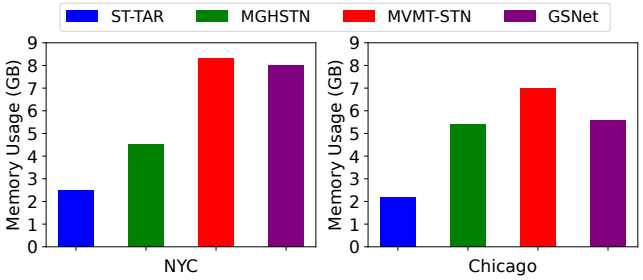
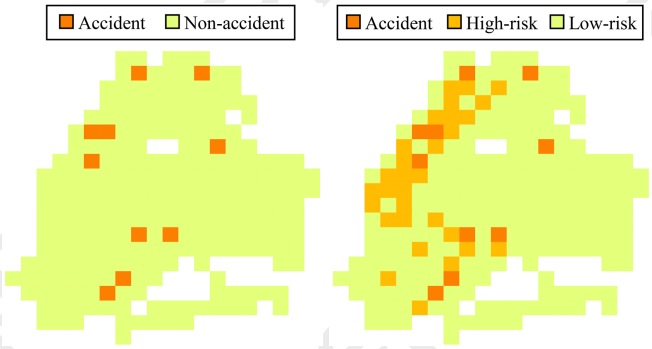


Figure 4: GPU memory usage on NYC and Chicago datasets.



(a) Map before auxiliary labeling (b) Map after auxiliary labeling

Figure 5: Visualization of label distribution of the city before and after the auxiliary label processing procedure on the NYC dataset.

6 Conclusion

We propose an efficient spatio-temporal learning framework, called ST-TAR, for traffic accident risk forecasting. We introduce hierarchical multi-view GCN and long short-term cross-attention mechanism to capture spatio-temporal correlation by considering long-term periodicity and short-term proximity separately. To capture spatio-temporal heterogeneity and contend with the zero-inflated data issue, spatio-temporal contrastive learning and adaptive risk-level weighted loss function are devised to achieve a multi-task framework. Extensive experimental studies on two real-world datasets demonstrate that ST-TAR achieves superior performance with improved efficiency.

Acknowledgments

This work was supported by the National Key R&D Program of China 2024YFE0111800, NSFC U22B2037, and NSFC U21B2046.

References

[An *et al.*, 2022] Bang An, Amin Vahedian, Xun Zhou, W Nick Street, and Yanhua Li. Hintnet: Hierarchical knowledge transfer networks for traffic accident forecasting on heterogeneous spatio-temporal data. In *SIAM International Conference on Data Mining*, pages 334–342, 2022.

[Bao *et al.*, 2019] Jie Bao, Pan Liu, and Satish V Ukkusuri. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. *Accident Analysis & Prevention*, 122:239–254, 2019.

[Barba *et al.*, 2014] Lida Barba, Nibaldo Rodríguez, Cecilia Montt, et al. Smoothing strategies combined with arima and neural networks to improve the forecasting of traffic accidents. *The Scientific World Journal*, 2014, 2014.

[Bergel-Hayat *et al.*, 2013] Ruth Bergel-Hayat, Mohammed Debbarh, Constantinos Antoniou, and George Yannis. Explaining the road accident risk: Weather effects. *Accident Analysis & Prevention*, 60:456–465, 2013.

[Charikar, 2002] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *The annual ACM symposium on Theory of computing*, pages 380–388, 2002.

[Chen *et al.*, 2024] Minxiao Chen, Haitao Yuan, Nan Jiang, Zhifeng Bao, and Shangguang Wang. Urban traffic accident risk prediction revisited: Regionality, proximity, similarity and sparsity. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 281–290, 2024.

[Goemans and Williamson, 1995] Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

[Guo *et al.*, 2021] Kan Guo, Yongli Hu, Yanfeng Sun, Sean Qian, Junbin Gao, and Baocai Yin. Hierarchical graph convolution network for traffic forecasting. In *AAAI conference on artificial intelligence*, volume 35, pages 151–159, 2021.

[Han *et al.*, 2021] Liangzhe Han, Bowen Du, Leilei Sun, Yanjie Fu, Yisheng Lv, and Hui Xiong. Dynamic and multi-faceted spatio-temporal deep learning for traffic speed forecasting. In *SIGKDD conference on knowledge discovery & data mining*, pages 547–555, 2021.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Indyk and Motwani, 1998] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *The annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[Lin *et al.*, 2015] Lei Lin, Qian Wang, and Adel W Sadek. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. *Transportation Research Part C: Emerging Technologies*, 55:444–459, 2015.

[Luo *et al.*, 2022] Guiyang Luo, Hui Zhang, Quan Yuan, Jinglin Li, and Fei-Yue Wang. Estnet: Embedded spatial-temporal network for modeling traffic flow dynamics. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):19201–19212, 2022.

[Lv *et al.*, 2009] Yisheng Lv, Shuming Tang, and Hongxia Zhao. Real-time highway traffic accident prediction based on the k-nearest neighbor method. In *International conference on measuring technology and mechatronics automation*, volume 3, pages 547–550, 2009.

[Organization and others, 2021] World Health Organization et al. Global plan for the decade of action for road safety 2021-2030. Technical report, WHO Regional Office for the Western Pacific, 2021.

[Sharma *et al.*, 2016] Bharti Sharma, Vinod Kumar Katiyar, and Kranti Kumar. Traffic accident prediction model using support vector machines with gaussian kernel. In *International Conference on Soft Computing for Problem Solving, Volume 2*, pages 1–10, 2016.

[Shi *et al.*, 2015] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28:802–810, 2015.

[Tobler, 1970] Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

[Trirat *et al.*, 2023] Patara Trirat, Susik Yoon, and Jae-Gil Lee. MG-TAR: Multi-view graph convolutional networks for traffic accident risk prediction. *IEEE Transactions on Intelligent Transportation Systems*, 24(4):3779–3794, 2023.

[Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[Wang *et al.*, 2021] Beibei Wang, Youfang Lin, Shengnan Guo, and Huaiyu Wan. GSNet: Learning spatial-temporal correlations from geographical and semantic aspects for traffic accident risk forecasting. In *AAAI conference on artificial intelligence*, volume 35, pages 4402–4409, 2021.

[Wang *et al.*, 2023] Senzhang Wang, Jiaqiang Zhang, Jiyue Li, Hao Miao, and Jiannong Cao. Traffic accident risk prediction via multi-view multi-task spatio-temporal networks. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12323–12336, 2023.

[Yuan *et al.*, 2018] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *SIGKDD international conference on knowledge discovery & data mining*, pages 984–992, 2018.