RoadFormer: Road-Anchored Adversarial Dynamic Graph Transformer for Unlimited-Range Traffic Incident Impact Prediction

Yanshen Sun
Virginia Tech
yansh93@vt.edu

Kaiqun Fu
South Dakota State University
kaiqun.fu@sdstate.edu

Chang-Tien Lu
Virginia Tech
clu@vt.edu

2023 IEEE International Conference on Big Data (BigData) | 979-8-3503-2445-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/BigData59044.2023.10386290

***Abstract*—The prompt estimation of traffic incident impacts (TIIs) plays a crucial role in guiding commuters' trip planning and enhancing the decision-making resilience of transportation agencies. Despite the strong capability of spatiotemporal modeling, the gap between the TII prediction and the dynamic data mining approaches has not been seamlessly filled. (1) The TII evaluation metrics have never been well-defined, although many criteria for TII exist in research works. (2) Previous attempts heavily rely on predefined road network structures and underscore vital features, leading to inaccurate TII predictions. (3) Predicting the spatiotemporal TII using dynamic road networks is more challenging as it requires extracting both abnormal sub-graph and long-range dependencies due to the large variation of incident clearance time. This research proposes RoadFormer, a novel *Road-Anchored Adversarial Dynamic Graph Transformer*, for predicting unlimited-range spatiotemporal TIIs. (1) We introduce novel criteria for assessing spatiotemporal TIIs and construct two new benchmark datasets to validate the performance of our methods. (2) RoadFormer leverages a road-anchored spatial transformer and an importance-score temporal transformer to form an encoder-decoder framework. The road-anchored spatial transformer prunes unnecessary edges between nodes with a road-anchored cascade attention mechanism, accurately pinpointing the affected sub-graphs. (3) The importance-score temporal transformer highlights abnormal changes in node features with a score-based adversarial training mechanism, enabling predictions to rely on informative feature changes after the accident occurrence. Extensive experiments on real-world datasets demonstrate that RoadFormer outperforms the state-of-the-art methods, especially in capturing spatiotemporal dependency patterns and predicting unlimited-range spatiotemporal TIIs.**

***Index Terms*—intelligent transportation System, traffic incident impact, spatiotemporal, dynamic graph transformer, prediction**

I. Introduction

Accurately predicting the spatiotemporal impact of traffic accidents is crucial for efficient Intelligent Transportation Systems (ITS) due to the inevitable traffic congestion caused by such sudden events [1]. However, effectively pinpointing traffic incidents remains challenging due to their uncertain causes, random occurrence times and locations, the lack of incident labels, and the demand for ultra-low inference times. Many efforts have been made to alleviate the above problems [2]–[4]. Generally, these methods either emphasize predicting TIIs solely by counting the numbers of incidents within a region [5], [6] or in the temporal domain [7]–[9]. However, our perspective is that the influence of traffic incidents can extend over significant durations (ranging from several minutes to hours) and cover extensive distances (ranging from hundreds of meters to tens of miles), which should not be limited to the prediction of within specific regions or time ranges. Therefore, this paper proposes and tackles a more challenging yet practical scenario: *predicting spatiotemporal TIIs by extracting the anomaly subgraph from dynamic graphs*. This task aims to develop a generalizable and robust predictive model to foresee the unlimited spatiotemporal range of TII through more powerful dynamic graph learning.

Fig. 1: **Temporal and spatial dimensions of TII.** The TIIs can be identified by congestion. The closer a location is to the incident, the longer it is affected. As indicated in the map, the incident on I-210 affected four upstream sensors' speed observations from 15:25 to 16:00. To quantify the impact of the incident, we used impact length (3.8 miles) and duration (35 minutes) as the indicators of the accident's impact. On the accompanying plot (right), the vertical axis illustrates the differences between the historical average speeds and speeds during the incident.

The impact of traffic incidents should be quantified in two dimensions: time and space [10]. As illustrated in Fig. 1, the impact of a traffic accident typically manifests as a decrease in speeds, starting at the accident's time and location and

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on August 07,2024 at 19:38:14 UTC from IEEE Xplore. Restrictions apply.

then propagating upstream before reaching its peak. Following the clearing of the obstruction, the queue of congested vehicles begins to advance, leading to a gradual reduction in the impacted distance until the final congested vehicle has passed the incident site. Although some studies have proposed criteria for assessing spatiotemporal TIIs, they often define impact duration (temporal TII) as the time between incident validation and restoration, while the impact length (spatial TII) is quantified by the number of cars blocked by the incident. However, our observations suggest that vehicles tend to slow down even when they are not in close proximity to the accident site. Given this insight, we propose a new definition for spatial impact length on the road: *the maximum continuous congestion distance immediately upstream of the incident*. In this context, "congestion" is identified by the difference between the average historical speed and the current speed at the location. I.e., the location at the time is considered as "congested" if

$$v_{avg-w} - v > \tau \tag{1}$$

is satisfied. Where v_{avg-w} is the average historical speed of the location at the same time of the week, v is the current speed, and τ is the threshold of the speed differences. The criteria used to determine τ is built on the local, daily, and historical speeds. See Section V-A for details.

Early graph-based deep learning approaches have numerous limitations in TII prediction: **(1) Rarely considered spatiotemporal TII quantification and limited publicly available data support.** Research has primarily focused on the temporal impact of accidents or impacts within limited spatiotemporal ranges. The absence of a well-established spatial TII assessment standard significantly impacts the outcomes of TII evaluation. It is essential to define comprehensive "traffic incident impact" criteria and develop open-source datasets in dynamic graph data mining. **(2) Unleashing adaptive attention mechanisms for dynamic road networks.** Previous works have developed different ways to merge the static road network structure with the dynamic similarities of traffic sensor measurements. Yet, the road network graph was constructed with strong human assumptions, while sensor measurements falter in providing ample data for precise similarity assessment. Therefore, it's crucial to create a new method that effectively uses attention mechanisms and sufficiently considers the characteristics of road networks in predicting TII. **(3) Unexplored task-focused sub-graph and sub-time-series extraction in dynamic graph learning.** In the context of abnormal events, the propagation of congested traffic flow shockwaves has undergone changes [11]. Minor traffic incidents can be easily cleaned up and affect only a small region of the traffic network. While for severe incidents, given the long time required to clear incident blockages, the variation and interaction among different road sections can be very complex. Many of the existing traffic forecasting models have been primarily developed for sensor-level forecasting, and their potential for sensor-network level (i.e., sub-graph level) forecasting has yet to be thoroughly unexplored.

To address these challenges, we present RoadFormer, an innovative method for predicting the unlimited-range spatiotemporal impact of traffic incidents. It leverages novel criteria for spatiotemporal TIIs and validates its performance using new benchmark datasets. RoadFormer's key innovation lies in its ability to effectively extract spatiotemporal relations within the traffic network, identifying variations in patterns through a combination of a road-anchored spatial transformer (RAS-Transformer) and an importance-score temporal transformer (IST-Transformer). These components create a powerful encoder-decoder framework. The RAS-Transformer precisely identifies affected sub-graphs by efficiently pruning unnecessary edges with a road-anchored cascade attention mechanism. Meanwhile, using an importance-score-based adversarial training approach, the IST-Transformer effectively detects abnormal node feature changes. In summary, the main contributions of this paper include:

(1) Quantifying the concept of "traffic incident impact" and sharing two open-source datasets. We redefine TII by quantifying spatiotemporal relationships between accident records and sensor measurements and offer two open-source datasets. These datasets encompass not only data for our proposed task but vital auxiliary information like sensor networks, accident records, metadata, and road structures, for future research of this field.

(2) Presenting the RAS-Transformer for accurate affected sub-graph identification. The RAS-Transformer module employs a novel road-anchored cascade attention approach. This approach eliminates the need for human assumptions in constructing road network graphs. Moreover, it optimally utilizes the correlations existing among sensors and effectively integrates various types of relationships among traffic sensors.

(3) Designing the IST-Transformer with importance-score-based adversarial training to highlight sensors affected by accidents. Inspired by TranAD [12], we employ anomaly detection techniques to facilitate precise localization during prediction. The IST-Transformer effectively accentuates abnormal changes in node features by incorporating an importance-score-based adversarial training.

(4) Evaluating RoadFormer's performance with extensive experiments. Experiments on real-world datasets demonstrate that RoadFormer outperforms the state-of-the-art methods, especially in capturing spatiotemporal dependency patterns and predicting unlimited-range spatiotemporal TIIs. Each module in RoadFormer is proven effective. The effectiveness of the importance score is validated through our case study.

II. Related Work

Traffic incident impact prediction has been a significant focus in traffic management for decades. Traditional studies often treated it as a 1-D propagation task, and designed deterministic queuing diagrams [13] and shockwave theory [14]. [15] summarizes statistical and machine learning methods, and argues that quantile regression (QR), finite mixture (FM), and random parameters hazard-based duration (RPHD) perform the best.

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on August 07,2024 at 19:38:14 UTC from IEEE Xplore. Restrictions apply.

Grid-based TII Prediction. As machine learning methods have evolved, TII prediction models have become more sophisticated, considering a more comprehensive range of factors. Lin et al. [16] perform incident duration predictions with a decision tree. Zou et al. [17] utilize Bayesian Model Averaging (BMA) to merge the prediction results of multiple machine learning models. [18] provides three traffic incident datasets. [19] analyzes how road attributes affect congestion durations. [20] discusses the relations between incident report time and congestion occurring time. Zhu et al. [21] propose a model that updates incident duration prediction every five minutes. However, these models fail to consider the complexity of road networks, instead considering only one road.

Deep Learning based TII Prediction. Some researchers treat traffic forecasting as a downstream task of deep learning. RadNet [22] considers TII prediction as an anomaly detection problem and employs a combination of transformer and GCN (Graph Convolutional Network). PrePCT [23] utilizes a CNN and LSTM for TII prediction. DIGC-Net [24] predicts traffic flow speeds by considering incident records and similarities among flow segments in the same time window. SLCNN [25] utilizes static and dynamic graphs, incorporating top-K attention for a C3D-like [26] temporal convolution. Yoo et al. [27] propose a covariance loss considering the basis function space and the targeted variable space. Huang et al. [28] utilize a generative adversarial network (GAN) to predict TIIs by directly learning speed heatmaps.

Attentive Dynamic Graph Representation Learning. The attention mechanism is a popular method for capturing relationships between graph nodes by considering their feature similarities. ADN [29] merges the spatial and temporal dimensions with attention across all elements. STAWnet [30] employs gated TCN (Temporal Convolutional Network) to extract temporal dependencies. Other studies have explored transformers for spatiotemporal data mining. Transformers can harness graph structure and positional features like position encoding [31]. [32] consider a GNN as an auxiliary module for transformers. GraphGPS [33] offers three different ways to encode graph structures and node positions. Some dynamic graph representation learning models like Graph WaveNet [34], AGCRN [2], DMSTGCN [35], DL-Traff [36], MegaCRN [37], and STAEformer [38] have shown efficiency in traffic flow forecasting tasks. However, these methods mainly focus on node-level or whole-graph tasks, leaving the extraction of significant sub-graphs unexplored.

Task-Specific Dynamic Graph Representation Learning. Several studies have tackled the "sub-graph extraction" problem using aggregation or denoising techniques. Titan [7] performs the TII prediction with a shared-parameter multitask model [8]. Meng et al. [9] show that proper graph aggregation techniques improve the performance of dynamic graph learning models for TII prediction. Nevertheless, these models focus solely on the temporal impact of incidents and overlook the significance of spatial impact prediction.

As the survey above reveals, there is a notable gap in the formal definition of spatiotemporal traffic incident impact prediction within dynamic graph representation learning. There is a pressing need for a fresh problem definition, benchmark datasets, and avenues to explore sub-graph impacts.

III. Problem Definition

A. Traffic Graph

A typical traffic performance measurement system utilizes static traffic loop sensors on arterial roads to collect traffic data. Previous research often linked groups of sensors based on their proximity in the geographic or feature space. However, we argue that *the common assumption in previous research is indeed valid.* For instance, if two sensors are recording traffic in opposite directions on the same freeway, they might be physically close to each other due to the road's width. However, it is inappropriate to link these two sensors since making a "U" turn is not allowed on freeways. Moreover, two sensors on different roads could be geographically close at an intersection. However, the vehicle's path to switch to another road might involve a long ramp, resulting in a significantly longer distance than the Euclidean distance would suggest.

Considering the challenge of quantifying relations between sensors on different roads, we propose a road-anchored graph to represent the traffic network. This graph includes three mappings: "sensor-to-road", "road-to-road", and "road-to-sensor". The "sensor-to-road" and "road-to-sensor" mappings connect sensors to the roads they are located on. The "road-to-road" mapping links two roads if they intersect. It is important to note that our graph representation treats two directions on one freeway as separate roads.

Definition 1. Road-anchored traffic network graph.

Consider a road-anchored traffic network graph as G, where $G = (S, R, E^{sr}, E^{rr}, E^{rs}, A^{sr}, A^{rr}, A^{rs})$. S is the sensor node set of size $|S|$. R is the road node set of size $|R|$. E^{sr}, E^{rr}, E^{rs} are the edges linking sensor-road, road-road, and road-sensor, respectively. $A^{sr} \in \mathbb{R}^{|S|\times|R|}, A^{rr} \in \mathbb{R}^{|R|\times|R|}, A^{rs} \in \mathbb{R}^{|R|\times|S|}$ are the adjacent matrix form of E^{sr}, E^{rr}, E^{rs}. Note that the edges here may continuously change during the model training process, as these adjacency matrices are learnable. The road-anchored traffic graph representation at time t is denoted as G_t. t is the timestamp around the accident validation time.

We assume that T timestamps around the incident validation time are utilized for prediction as they are typically the most relevant to the incident impact. We observed that traffic behaviors before and after an incident can be quite different. Hence, we split the T timestamps into "before-validation" (comprising T_{bv} timestamps) and "after-validation" (comprising T_{av} timestamps).

Definition 2. Dynamic traffic network graph.

A dynamic traffic graph can be denoted as $\mathcal{G} = \{\mathcal{G}_{bv}, \mathcal{G}_{av}\} = \{G_0, ..., G_{T_{bv}-1}, G_{T_{bv}}, ..., G_{T_{bv}+T_{av}-1}\}$. $G_0, ..., G_{T_{bv}+T_{av}-1}$ indicates road-anchored traffic network graphs at timestamp $0, ..., T_{bv} + T_{av} - 1$. $\mathcal{G}_{bv}$ and $\mathcal{G}_{av}$ are dynamic graphs before and after the validation time, respectively.

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on August 07,2024 at 19:38:14 UTC from IEEE Xplore. Restrictions apply.

B. Incident Impact Prediction

As stated in Section I, the TII can be characterized in two dimensions: spatial and temporal. Here, we use $\mathbf{Y}_{Dur}$ to represent the temporal TII, which is the difference between the accident validation time and restoration time. This definition aligns with the formal accident duration provided by the Department of Transportation (DOT). Regarding the spatial dimension, our definition differs from the traditional approach used in transportation research, which involves counting the number of blocked cars upstream of the incident. It is essential to consider that cars' speed can be influenced by congestion even if they are not completely blocked. To address this aspect, we define the impact length $\mathbf{Y}_{Len}$ as the maximum continuous congested road distance in the immediate upstream of the accident.

Definition 3. Traffic incident impact precision.

Given a dynamic traffic network graph for an accident $\mathcal{G}$ and corresponding sensor feature tensor $\mathbf{X} \in \mathbb{R}^{|S| \times T \times C_{in}}$, the aim is to find a model $\mathcal{F}$ so that

$$\mathcal{F} : (\mathcal{G}, \mathbf{X}) \rightarrow (\mathbf{Y}_{Dur}, \mathbf{Y}_{Len}) \tag{2}$$

where C_{in} is the number of input channels recorded by the sensors, Y_{Dur} is the impact duration, and Y_{Len} is the impact length. Given that the number of channels can vary across layers within RoadFormer, we will use the variable C to denote an arbitrary number of channels in the subsequent sections.

IV. Methodology

As shown in Fig. 2, the proposed RoadFormer contains three main parts. The Road-Anchored S-Transformer effectively integrates spatial information among sensors at each timestamp. The Importance-Score T-Transformer aggregates temporal information for each sensor, assigning higher "importance-scores" to sensors that undergo sudden feature changes after the validation time. A Pooling module projects the learned dynamic graph representation to the expected output.

A. Road-Anchored S-Transformer

The design of the RAS-Transformer draws inspiration from the concepts of anchored graphs and hypergraphs. Anchors have been effectively utilized in various studies to reduce attention complexity [39]. In our approach, roads serve as anchor nodes for the sensors located on them. This not only addresses the issues discussed in Section III-A but also reduces the rank of the adjacency matrix from $|S|$ to $|R|$, making multi-hop message-passing relevant only among roads. The model maintains a global latent feature tensor $\mathbf{H}^r \in \mathbb{R}^{|R| \times T \times C}$ for each road at each timestamp, which does not vary by case. For each accident case, the sensor features are initially used to update the case-irrelevant road features, obtaining case-relevant road features $\mathbf{H}^{sr}$. Then, message-passing among roads further updates the road features to $\mathbf{H}^{rr}$. At this stage,the output road features can be considered both an intermediate pooling of the sensor features and a spatial feature fusion of the sensors. Finally, the road features are propagated back to the sensors for the next step.

Denote sensors as S and roads as R. A vanilla self-attention exploring spatial relations between sensors can be summarized as (3):

$$a_{ij}^{ss} = \sigma\left(\frac{(\mathbf{Q}\mathbf{h}_{s_j})^T(\mathbf{K}\mathbf{h}_{s_i})}{\sqrt{d}}\right) \tag{3}$$

where s_i and s_j are two different sensor nodes, $\mathbf{Q}$ and $\mathbf{K}$ are query and key projection parameters, $\mathbf{h}_{s_i}$ and $\mathbf{h}_{s_j}$ are embeddings of s_i and s_j, $d = C \div n\_head$ is the dimension of each attention head, and σ is the row-wise softmax activation function.

In contrast to the vanilla approach, in RAS-Transformer, the attentive message-passing is accomplished with three transformer layers. The first transformer layer contains a "sensor-to-road" attention layer and a linear layer. The "sensor-to-road" attention computes the correlation between sensors and roads and masks out the edge between s_i and r_j if s_i is not on r_j, as (4):

$$\begin{aligned} a_{ij}^{sr} &= \sigma(mask_{sr}(\frac{(\mathbf{Q}^{sr}\mathbf{h}_{r_j})(\mathbf{K}^{sr}\mathbf{h}_{s_i})^T}{\sqrt{d}}, m_{ij}^{sr})), \\ mask(\mathbf{x}, \lambda) &= \begin{cases} \mathbf{x} & \lambda = 1 \\ -\infty & \lambda = 0 \end{cases} \end{aligned} \tag{4}$$

where s_i and r_j represent an arbitrary sensor and a road. $\mathbf{Q}^{sr}$ and $\mathbf{K}^{sr}$ are query and key projection parameters for r_j and s_i. $\mathbf{h}_{s_i}$ and $\mathbf{h}_{r_j}$ are embeddings of s_i and r_j. $\mathbf{h}_{r_j} \in \mathbf{R}^{T \times C}$ is a learnable parameter matrix. $\mathbf{M}^{sr} \in \{0,1\}^{|S| \times |R|}$ is the adjacent matrix between sensors and roads. $m_{ij}^{sr} = 1$ if sensor s_i is on road r_j, otherwise $m_{ij}^{sr} = 0$. $\mathbf{M}^{sr}$ performs as the mask of all the attention heads. With a_{ij}^{sr}, the road embedding can be updated as $\mathbf{h}_{r_j}^{sr} = \mathbf{W}^{sr} \sum_{i=1}^{|S|} a_{ij}^{sr}(\mathbf{V}^{sr}\mathbf{h}_{r_j}) + \mathbf{b}^{sr}$.

The second layer of the Road-Anchored S-Transformer is a "road-to-road" self-attention layer designed to extract road intersection information. The attention can be expressed as (5):

$$a_{ij}^{rr} = \sigma(mask_{rr}(\frac{(\mathbf{Q}^{rr}\mathbf{h}_{r_j}^{sr})(\mathbf{K}^{rr}\mathbf{h}_{r_i}^{sr})^T}{\sqrt{d}}, \mathbf{M}_{ij}^{rr})) \tag{5}$$

where $\mathbf{M}^{rr} \in \{0,1\}^{|R| \times |R|}$ represents four levels of adjacency between roads: $m_{ij}^{rr} = 1$ if (1) $i = j$, (2) r_i intersects r_j, (3) r_i intersects r_k and r_j intersects r_k, and (4) fully connected. The four different masks can be applied to different attention heads. In our design, each of the four masks was applied to one attention head. A spatial-relation awared road feature tensor is then computed as $\mathbf{h}_{r_j}^{rr} = \mathbf{W}^{rr} \sum_{i=1}^{|R|} a_{ij}^{rr}(\mathbf{V}^{rr}\mathbf{h}_{r_j}^{sr}) + \mathbf{b}^{rr}$.

The last transformer layer contains a "road-to-sensor" attention, meaning that sensor vectors are queries, and road vectors are keys and values. The output of this step propagates the aggregated road features to the sensors. The equation is essentially the reversed version of (4):

$$\mathbf{H}^{rs} = \sigma(\frac{(\mathbf{Q}^{rs}\mathbf{H}^{rr})(\mathbf{K}^{rs}\mathbf{H}^{rr})^T}{\sqrt{d}})(\mathbf{V}^{rs}\mathbf{H}^{rr}) \tag{6}$$

Note that there is no mask in this step, which preserves the attention mechanism's flexibility. Previous works have usually

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on August 07,2024 at 19:38:14 UTC from IEEE Xplore. Restrictions apply.

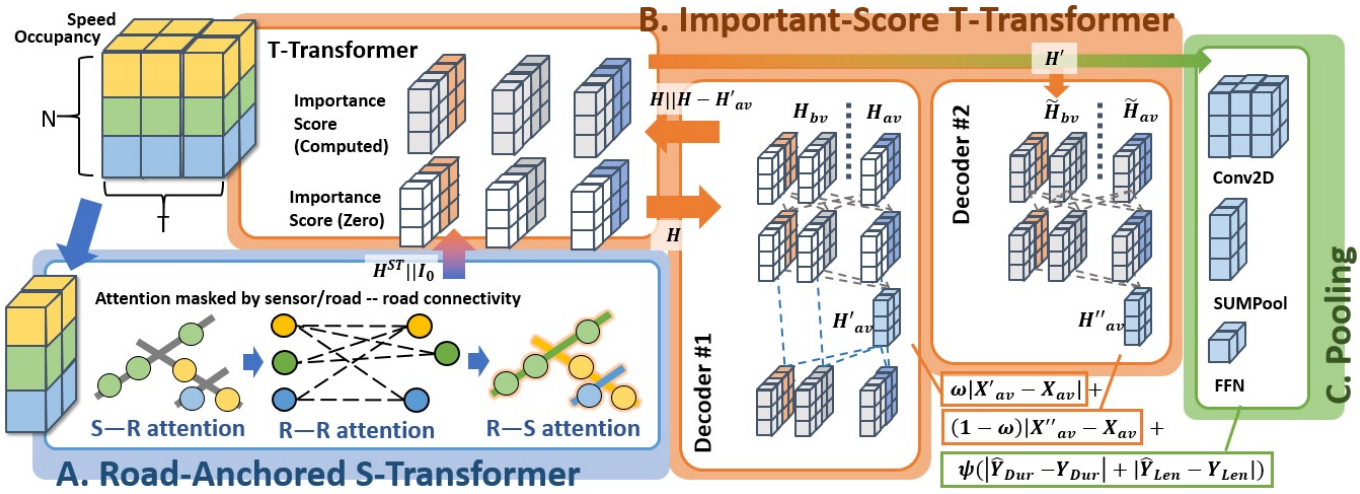


Fig. 2: The architecture of RoadFormer. The blue rectangle labeled **A.** Road-Anchored S-Transformer first encodes the input tensor by performing "sensor-to-road", "road-to-road",and "road-to-sensor" attentions. The orange rectangles labeled **B.** Importance-Score T-Transformer further process the output of A $\mathbf{H}^{ST}$ with three modules. $\mathbf{H}^{ST}$ is split into $\mathbf{H}_{bv}$ and $\mathbf{H}_{av}$ by the validation time of the accident, then fed to the T-Transformer with importance-scores initialized as 0. The outputs $\mathbf{H}$ and $\mathbf{H}_{av}$ are sent to Decoder #1 to reconstruct $\mathbf{X}_{av}$ and compute the importance-score as $\mathbf{H}-\mathbf{H}'_{av}$. With the updated importance-score, $\mathbf{H}^{ST}$ is fed to T-Transformer again. The output is further processed by Decoder #2 to become $\mathbf{H}''_{av}$. The green rectangle labeled **C.** Pooling shows how the latent features are projected to the desired outputs. Finally, the loss is computed as the weighted combination of 1) the reconstruction from $\mathbf{H}'_{av}$ to $\mathbf{X}_{av}$, 2) the reconstruction from $\mathbf{H}''_{av}$ to $\mathbf{X}_{av}$, and 3) the prediction loss of impact duration and length.

linked the top-k closest sensors in the geographic or feature space, weighting the links with the distances. However, we observe that attention is good at learning weights but weak at learning graph structures. In this case, we simply need to find nodes that are definitely linked to each other and leave the weight learning task to attention. Therefore, we chose to partially control the graph structure with unweighted adjacent matrices. Finally, we applied skip-connection, layer-normalization, and dropout to $\mathbf{H}^{rs}$.

Essentially, RAS-Transformer stresses the effects of sensors on high-degree-centrality roads. The strengths of the RAS-Transformer can be summarized as follows. (1) It avoids the human error introduced by manually choosing "road-to-road" for "top-k", (2) it allows long-range message-passing as all sensors on intersected roads are linked, (3) it preserves the flexibility of attention with a relatively small number of edges ($|S|+|R|^2+|S||R|, |R| \ll |S|$ at most) (4) it is more time efficient ($\mathcal{O}(|S||R|^2+|R|^3+|R|^2|S|), |R| \ll |S|$) than traditional attention mechanisms ($\mathcal{O}(|S|^3)$) and requires fewer layers as the spatial message-passing is performed sufficiently by the "road-to-road" self-attention module.

B. Importance-Score T-Transformer

Accidents typically impact a limited section of the traffic network. Treating all sensors equally for prediction can add unnecessary noise. Yet, manually selecting sensors near the accident might overlook broader, intricate impact patterns and essential features due to the traffic network's prompt or delayed response. In this case, a method that dynamically locates the region and time window affected by the accident is important. Based on the assumption that *the traffic measurements of sensors affected by incidents show more obvious changes than other sensors*, we locate the accidents with anomaly detection techniques (i.e., assigning sensors with larger variance a higher "importance-score").

Inspired by [12], our IST-Transformer contains three modules: a temporal transformer (T-Transformer) and two decoders. The T-Transformer module encodes the output of the RAS-Transformer with and without the importance-score along the time dimension. The first decoder computes the importance-score and reconstructs $\mathbf{X}_{av}$, while the second reconstructs $\mathbf{X}_{av}$ from the combination of the importance-score and the graph embedding. Denote the "after-validation" section of $\mathbf{H}^{ST}$ as $\mathbf{H}_{av}$ and the "before-validation" section of $\mathbf{H}^{ST}$ as $\mathbf{H}_{bv}$. The combination of the T-Transformer and any one of the decoders is equivalent to a classic transformer network when considering $\mathbf{H}^{ST}$ as the input sequence and $\mathbf{H}_{av}$ as the output sequence.

Assume the importance-score is $\mathbf{I} \in \mathbb{R}^{|S|\times T\times C}$ and the output of RAS-Transformer as $\mathbf{H}^{ST} \in \mathbb{R}^{|S|\times T\times C}$. The output of the T-Transformer can be written as (7):

$$\mathbf{H} = \mathit{TTrans}([\mathbf{H}^{ST}||\mathbf{I}_0]) \tag{7}$$

where $\mathit{TTrans}()$ indicates T-Transformer, which is a block sequentially performing temporal self-attention and skip-connection. $\mathbf{I}_0$ is the initialized importance-score (which is an all-zero tensor).

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on August 07,2024 at 19:38:14 UTC from IEEE Xplore. Restrictions apply.

The task for both Decoder #1 and Decoder #2 is to reconstruct $\mathbf{X}_{av}$. Each decoder has a self-attention layer and a mutual-attention layer. Decoder #1 attempts to achieve its goal using $\mathbf{H}$ and $\mathbf{H}_{av}$, as (8):

$$\mathbf{H}'_{av} = \textit{mu-attn}_1(\mathbf{H}, \mathbf{H}, \textit{self-attn}_1(\mathbf{H}_{av})) \tag{8}$$

The three parameters in *mu-attn*$_1$ are placeholders for value, key, and query.

Then, the importance-score is updated as $\mathbf{I} = \mathbf{H}^{ST} - \mathbf{H}'_{av}$. Note that $\mathbf{H}^{ST}$ has T timestamps while $\mathbf{H}'_{av}$ has T_{av} timestamps. We examined various methods, such as repeating timestamps in $\mathbf{H}'_{av}$ and getting mean/min/max of $\mathbf{H}'_{av}$ along the time axis. All the methods resulted in similar performance. Accordingly, we adopt the general form to represent the difference between $\mathbf{H}^{ST}$ and $\mathbf{H}'_{av}$. Replacing $\mathbf{I}_0$ with $\mathbf{I}$, we apply the T-Transformer and Decoder #2 the same way as the previous steps, with the concatenated $\mathbf{H}^{ST}$ and $\mathbf{I}$ as the input,as (9).

$$\begin{aligned} \mathbf{H}' &= \textit{TTrans}([\mathbf{H}^{ST}||\mathbf{I}]) \\ \mathbf{H}''_{av} &= \textit{mu-attn}_2(\mathbf{H}', \mathbf{H}', \text{self-attn}_2(\mathbf{H}_{av})) \end{aligned} \tag{9}$$

Finally, $\mathbf{H}'_{av}$ and $\mathbf{H}''_{av}$ are used to reconstruct $\mathbf{X}_{av}$ separately with the same two-layer feed-forward network (FFN):

$$\begin{aligned} \mathbf{X}'_{av} &= \mathbf{W}_{1,2}(\phi(\mathbf{W}_{1,1}\mathbf{H}'_{av} + \mathbf{b}_{1,1})) + \mathbf{b}_{1,2} \\ \mathbf{X}''_{av} &= \mathbf{W}_{1,2}(\phi(\mathbf{W}_{1,1}\mathbf{H}''_{av} + \mathbf{b}_{1,1})) + \mathbf{b}_{1,2} \end{aligned} \tag{10}$$

where $\mathbf{W}_{1,1}, \mathbf{b}_{1,1}, \mathbf{W}_{1,2}, \mathbf{b}_{1,2}$ are parameters of the two linear layers and ϕ is the activation function, which is the LeakyReLu in the FFN in (10).

C. Pooling and Loss

After the spatial and temporal encoding, the spatiotemporal representation $\mathbf{H}'$ is considered well-learned and ready for prediction. To further refine the features, the importance-score is used to weight the elements in $\mathbf{H}'$ through element-wise multiplication. This is followed by a temporal dimension aggregation through a 2-D convolution layer and a spatial dimension aggregation through SUMPooling. Lastly, three linear layers with LeakyReLU activations are used to project the features into the desired dimension. This process can be summarized as (11):

$$\begin{aligned} &\mathbf{H}'_T = \textit{Conv2d}(\mathbf{H}'), \mathbf{H}'_T \in \mathbb{R}^{|S|\times C} \\ &\mathbf{H}'_S = \textit{SUMPool}(\mathbf{H}'_T), \mathbf{H}'_T \in \mathbb{R}^{C}) \\ &\hat{Y} = \mathbf{W}_{p3}\phi(\mathbf{W}_{p2}\phi(\mathbf{W}_{p1}\mathbf{H}'_S + \mathbf{b}_{p1}) + \mathbf{b}_{p2})) + \mathbf{b}_{p3} \end{aligned} \tag{11}$$

where $\hat{Y} \in \mathbf{R}^{C_{out}}, C_{out} = 2$. The first value represents the impact duration, and the second is the impact length. SUMPooling is chosen because it is the most expressive pooling method [40], and we want the pooled representation to be capable of representing all possible accident scenarios. L1 loss is chosen as the prediction loss function so that the model focuses on the overall trends of $\mathbf{Y}$ instead of outliers. The prediction loss can then be written as $Loss_1 = |\hat{\mathbf{Y}}_{Dur} - \mathbf{Y}_{Dur}| + |\hat{\mathbf{Y}}_{Len} - \mathbf{Y}_{Len}|$.

The second part of the loss is the reconstruction loss. This loss is for regularization purposes and is self-supervised. The only objective of the model is to predict $\mathbf{Y}_{Dur}$ and $\mathbf{Y}_{Len}$. L1 loss is employed for both $\mathbf{X}'_{av}$ and $\mathbf{X}''_{av}$, i.e., the loss functions are $Loss_2 = |\mathbf{X}'_{av} - \mathbf{X}_{av}|$ and $Loss_3 = |\mathbf{X}''_{av} - \mathbf{X}_{av}|$. To generate the importance-scores correctly, the IST-Transformer is trained in an adversary way. Here, we explain our design in terms of a GAN in order to make it understandable. Consider $\mathbf{H}_{av}$ as the true "image", $\mathbf{H}_{bv}$ as the fake "image", Decoder #1 as the discriminator, and Decoder #2 as the generator. In the first several epochs, the weight of $Loss_2$ is far larger than the weight of $Loss_3$. As a result, Decoder #1 discriminates $\mathbf{H}_{bv}$ and $\mathbf{H}_{av}$ better by assigning $\mathbf{H}_{av}$ larger attentive weights in *TTrans*(). As the weight of $Loss_3$ increases, Decoder #2 is trained to "generate" a new $\mathbf{H}_{bv}$ by concatenating it with $\mathbf{I}_{bv} \approx \mathbf{H}_{bv} - \mathbf{H}_{av}$. The new $\mathbf{H}_{bv}$ – $\mathbf{H}'_{bv}$ receives more attention in *TTrans*() with the importance-score and thus enlarge $Loss_2$. This way, as the weight of $Loss_3$ grows, the attentive weights in *TTrans*() finally stabilize at some point that slightly inclines to $\mathbf{H}_{bv}$, which makes Decoder #1 produce a larger importance-score for $\mathbf{H}_{av}$ and a smaller score for $\mathbf{H}_{bv}$.

The loss during training can be written as (12):

$$Loss = \psi Loss_1 + \omega Loss_2 + (1-\omega)Loss_3 \tag{12}$$

where ψ is the weight of the prediction loss. $\omega \in (0, 1]$ is the weight of the reconstruction loss of Decoder #1, while $(1-\omega)$ is the weight of the reconstruction loss of Decoder #2. ω is initialized as 1 and decreases as the number of epochs increases.

V. Experiment

A. Dataset and Data Preprocessing

The traffic loop sensor data used for this research are collected from the Caltrans Performance Measurement System (PeMS)[1]; The incident record data are from RITIS [2]; The road networks are downloaded from Tiger Priscroads [3]. Based on the location of sensors and incidents, we selected several freeway segments in Los Angeles and San Bernardino regions, which have well-constructed and complex road networks.

In addition to the data used in this research (i.e., sensor measurements, adjacency information, and incident impact records), we also provided auxiliary data to broaden the potential problem domain. Specifically, our datasets contain three basic elements: roads, sensors, and accidents. For roads, we provide a matrix indicating whether two roads intersect and the location of the intersections (in the form of the distance from the start point of the road). We collect their positions on the roads and their five-minute speed and occupancy measurements for sensors. As the rate of missing records is less than 0.005%, we filled those values with daily

[1]https://pems.dot.ca.gov/
[2]https://www.ritis.org/
[3]https://www2.census.gov/geo/tiger/

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on August 07,2024 at 19:38:14 UTC from IEEE Xplore. Restrictions apply.

average speed/occupancy. Finally, for incidents, we provide the position on the road, the DateTime (number of five minutes from 2019/09/01 00:00:00), the incident category, the impact duration, and the impact length.

Since we lacked the ground truth of impact length, we acquired the label with three steps. We first extracted the regular weekly traffic pattern by averaging the speeds from 2019/06/01 to 2019/08/31. In this case, the value of "average historical speed" for each sensor at each timestamp of the week is acquired as the average of approximately 4 weeks $\times$ 3 months $=$ 12 speed measurements. The differences between the current speed and the corresponding average historical speed are then computed to distinguish incident-caused congestion from regular traffic patterns.

The second step is to determine the threshold of speed differences, so that speeds significantly lower than the regular pattern are considered an indicator of non-recurrent congestion. For each incident, we performed a binary 1D k-means classification of the upstream speed differences between the incident occurrence and clear time. The low-speed cluster is then marked as affected by the incident. To reduce the impact of special events and adjacent congestion events, we removed markers of the speed differences smaller than $0.5\times$ standard deviation of the daily differences and sensors further than $7\times$ duration values. The impact length is zero if the selected sensor is the nearest sensor to the incident. Otherwise, the impacted length is defined by the distance from the accident to the upstream sensor before the picked sensor.

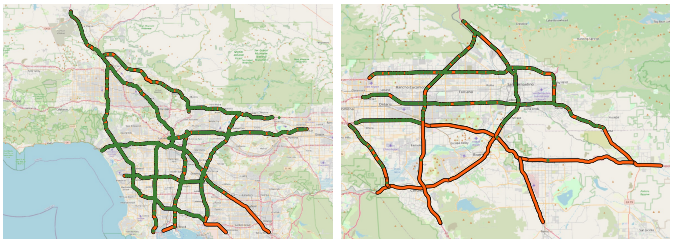
To construct appropriately scaled datasets, we selected several inter-state freeways as illustrated in Fig. 3a and 3b. The time was set to one month (2019/09/01 – 2019/09/30). Sensors and accidents not on the chosen freeways were filtered out. We also removed accidents with a duration of fewer than 30 minutes due to their limited temporal impact. Fig. 3c shows the distribution of impact duration and length. The orange dashed lines indicate the fluctuations of log event counts across different durations and impact lengths. The shades of the bars show the magnitude of event counts. All the label values follow power-law distributions except that the impact length in the San Bernardino region is relatively noisy compared with the others. Finally, we present two datasets for TII prediction: Incident-LA and Incident-SB. The detailed attributes of the two dynamic networks are illustrated in Table I.

TABLE I: Dataset Properties

Dataset	# event	# node (R/S)	# edge (R-R/S-S)	# 0 length
Incident-LA	5,668	32/1,663	142/869,640	2,062
Incident-SB	1,452	28/1,150	140/390,822	454

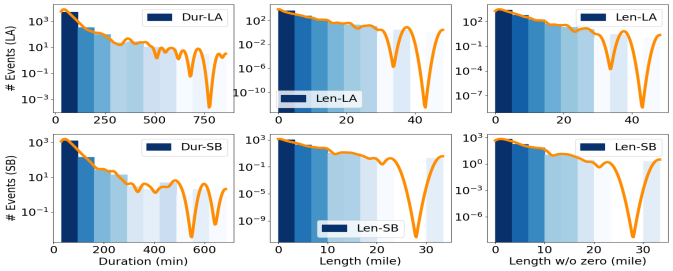
B. Baselines

To evaluate the efficiency of RoadFormer, we chose nine representative models to perform the same task. Considering that only a few models target similar jobs, we choose the



(a) Included freeways with incident locations (Incident-LA)

(b) Included freeways with incident locations (Incident-SB)



(c) Distribution of event duration/impact length/ non-zero impact length

Fig. 3: Basic Dataset Information

two most recently published incident impact prediction models, two conventional models, and five state-of-the-art traffic forecasting models. Two of the later five models leverage attentive graph representation learning techniques, while the other three use graph convolution. For the traffic forecasting models, we employ a pooling module identical to the one used in RoadFormer to obtain the desired output.

- **L-1 regularized linear regression (LASSO) [41].** As LASSO only accepts one-dimensional inputs, we explored various feature aggregation approaches: 1) averaging across spatiotemporal dimensions, 2) selecting the nearest upstream/downstream sensor and averaging temporally, and 3) choosing the closest upstream/downstream sensor along with the five minutes after validation. For parameter selection, we examined λ values of 0.1, 1.0, 10.0, and 100.0.
- **Support vector regression (SVR) [41].** Similar to LASSO, we examined three different feature aggregation methods. We used the default parameters ($C = 1, \epsilon = 0.1$) in the sklearn [42] package of Python.
- **HastGCN [8].** It is a spatiotemporal attention model for incident duration prediction. We reproduce the model with the help of the author. All the settings are the same as the original model.
- **AGWN [9].** It preprocesses the adjacency matrix with a wavelet filter before the graph convolution operation. All the hyperparameters are the same as in the original paper.
- **STTN [31].** STTN leverages transformers for spatial and temporal message-passing, focusing on predicting node-level speeds. The code is derived from the official STTN

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on August 07,2024 at 19:38:14 UTC from IEEE Xplore. Restrictions apply.

GitHub repository [4]. All the hyperparameters remain the same as is mentioned in the paper.

- **STAWnet [30].** It employs attentive graph message-passing and gated TCN. We used the original code [5] and kept the hyperparameters unchanged.
- **DMSTGCN [35].** It decomposes the adjacency matrix into four trainable embeddings for graph convolution. All the hyperparameters are the same as in the original paper. [6].
- **Graph WaveNet [34].** It contains a self-adaptive GCN module and a dilated TCN module. All the hyperparameters are the same as in the original paper [7].
- **AGCRN [2].** AGCRN performs node-adaptive graph convolution and GRU-like temporal message-passing. All the hyperparameters are the same as in the original paper [8].

C. Hyperparameter and Metrics

In the problem settings of this paper, we assume that the objective is to predict the incident's impact within a short time after the event. To do so, nine timestamps (six for "before-validation" and three for "after-validation") were adopted. As a result, the model could not see the full traffic pattern during the incident.

For the training process of RoadFormer, we adopted a batch size of 8 due to the GPU memory limitation. The learning rate was 0.0005 with a 0.001 weight decay. The number of attention heads was 4. The LeakyReLU factor was set to 0.2, and the dropout rate was 0.1. In the loss function (Equation 12), the prediction loss weight ψ was 1.0. To adjust $Loss_2$ and $Loss_3$ as described in Section IV-C, ω was set to $\frac{1}{i+1}$, where i is the index of the current epoch.

Following our previous works [7]–[9], we adopted root mean squared error (RMSE) and mean absolute error (MAE) as metrics. However, the impact length introduces labels of zero value making mean absolute percentage error (MAPE) invalid. Therefore, we replaced MAPE with symmetric mean absolute percentage error (sMAPE). Based on the definition ($RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$, $MAE = \frac{1}{N} sum_{i=1}^{N} |y_i - \hat{y}_i|$, $sMAPE = \frac{1}{N} sum_{i=1}^{N} |\frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}|$), RMSE penalizes large gaps more harshly than MAE, while sMAPE focuses more on the magnitude of the differences from the true values.

D. Prediction Result Analysis

The input to the models includes the adjacent matrices and sensor measurements one hour before and half an hour after the validation time. The output is the impact duration and length. We evaluated the duration and length separately as they are of different units. Table II illustrates the performance of RoadFormer against the baselines.

[4]https://github.com/wubin5/STTN
[5]https://github.com/CYBruce/STAWnet
[6]https://github.com/liangzhehan/DMSTGCN
[7]https://github.com/nnzhan/Graph-WaveNet
[8]https://github.com/LeiBAI/AGCRN

Conventional baselines. For LASSO and SVR, the best performance is achieved with the closest upstream sensor and the first timestamp after the validation time as inputs. The results appeared to be insensitive to hyperparameters. As shown in Table II, even though they produced competitive results in RMSE and MAE impact length prediction, LASSO and SVR performed poorly in predicting impact duration prediction. This result matches our hypothesis that the "closest" sensors and timestamps are not optimal for prediction.

Spatiotemporal neural network baselines. RoadFormer surpasses the other baselines (i.e., spatiotemporal neural networks) in most of the metrics and achieves competitive performance on the other metrics if it is not the best. Specifically, RoadFormer outperforms another spatiotemporal transformer, STTN, by about 10% in duration prediction and 5% in length prediction. RoadFormer also beats previous incident impact prediction models, HastGCN and AGWN, by approximately 10% in performance.

We find that RoadFormer cannot beat the other models in all metrics. To prove the efficiency of our model, we rank all ten models by each criterion and list the average rank in Table III. The table shows that our model performs the best on average.

E. Ablation Study

We conducted four ablation experiments to evaluate the contributions of each component of the RoadFormer model. The result is shown in Table IV. We removed the corresponding modules for the "No-STrans" and "No-TTrans" experiments. For "No-Score", we skipped the concatenation steps in Equation 7 and 9 and removed the reconstruction losses during training. For "No-Road", we replaced the RAS-Transformer with the transformer used in STTN. Initially, we assumed that "No-Road" would outperform RoadFormer as our graph had too many edges removed. However, the results in Table IV show that our model cannot achieve its performance without any of its components. We observed that the performance drops less in "No-Score" and "No-TTrans" than in "No-Road" and "No-STrans". This may be because the number of timestamps in the input is too small for a transformer to work.

F. Case Study

We explored several incident cases to examine whether our importance-score helps identify incidents. One example is incident #18693 (Fig. 4). The map plots the five-minute average speed immediately after the validation time and the importance-scores within the same time slot. Obviously, sensors that detect lower speeds also have higher importance-scores, while high importance-scores cluster around the incident (red star). However, we also observe that incident-irrelevant speed drops also lead to high importance-scores. The model utilizes both the score and the embedded features for prediction.

Beyond This Task. The other parts of the dataset (i.e., the data not used in this paper) can be used to increase the prediction

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on August 07,2024 at 19:38:14 UTC from IEEE Xplore. Restrictions apply.

TABLE II: RMSE, MAE, and sMAPE for duration and impact length prediction of Incident-LA and Incident-SB. This table lists the performance of nine state-of-the-art baselines and our proposed model. Bolded and blue results indicate the best and the second.

Method	Incident-LA (dur (min))			Incident-LA (len (mile))			Incident-SB (dur (min))			Incident-SB (len (mile))		
	RMSE	MAE	sMAPE	RMSE	MAE	sMAPE	RMSE	MAE	sMAPE	RMSE	MAE	sMAPE
LASSO	59.773	51.776	0.760	**8.270**	6.794	1.211	58.921	51.263	0.757	**10.743**	8.934	1.112
SVR	60.073	50.763	0.743	8.559	6.300	1.299	59.560	50.924	0.761	11.632	**8.812**	1.218
HastGCN	31.719	20.372	0.319	8.421	6.402	1.272	35.936	25.078	0.381	13.053	9.299	1.350
AGWN	31.934	20.720	0.341	10.840	6.594	0.874	32.864	22.391	0.365	11.730	8.736	**0.743**
STTN	31.826	21.098	0.322	9.644	6.317	0.893	31.235	20.631	0.326	12.346	9.025	0.812
STAWnet	31.400	20.315	0.318	8.619	6.311	1.310	**29.280**	20.212	0.320	11.994	8.945	1.260
DMSTGCN	31.555	20.342	0.319	10.638	7.784	0.880	29.810	20.263	**0.312**	12.929	9.846	0.791
Graph WaveNet	31.880	20.415	0.320	10.489	7.672	**0.861**	30.765	20.455	0.324	14.093	10.807	0.914
AGCRN	**31.253**	20.363	0.319	8.662	**6.212**	0.899	30.905	20.652	0.324	12.808	9.093	1.172
RoadFormer (ours)	31.413	**20.310**	**0.318**	9.494	6.226	1.477	29.726	**20.140**	0.319	11.731	8.818	1.235

TABLE III: Average rank of baseline and RoadFormer performance. Abbreviated model names are used due to the limited space. The smaller the number, the higher the average rank. Bolded and blue results indicate the best and the second.

Model	**RoadFormer**	STTN	STAW.	DMST.	AGCRN	AGWN	G.W.	LASSO	Hast.	SVR
Rank	**3.58**	3.75	4.08	5.25	5.25	5.75	6.67	6.83	6.92	6.92

TABLE IV: RMSE, MAE, and sMAPE for Duration and Impact Length Prediction of Incident-LA and Incident-SB. The results of the ablation study include our model without the RAS-Transformer (No-STrans), our model without the T-Transformer (No-TTrans), our model without the road anchors (No-Road), and our model without the importance-score (No-Score).

Method	Incident-LA (dur (min))			Incident-LA (len (mile))			Incident-SB (dur (min))			Incident-SB (len (mile))		
	RMSE	MAE	sMAPE	RMSE	MAE	sMAPE	RMSE	MAE	sMAPE	RMSE	MAE	sMAPE
No-STrans	31.245	20.320	0.319	9.608	6.251	1.503	28.793	21.240	0.338	11.922	8.877	1.255
No-TTrans	31.674	20.371	0.319	9.496	6.215	1.570	32.024	27.265	0.422	12.432	8.903	1.344
No-Road	31.611	20.351	0.319	9.909	6.345	1.616	28.568	20.928	0.333	13.240	9.286	1.496
No-Score	31.846	20.321	0.319	9.496	6.215	1.473	30.000	20.236	0.320	14.000	9.743	1.688
RoadFormer (ours)	**31.413**	**20.310**	**0.318**	**9.494**	**6.226**	**1.477**	**29.726**	**20.140**	**0.319**	**11.731**	**8.818**	**1.235**

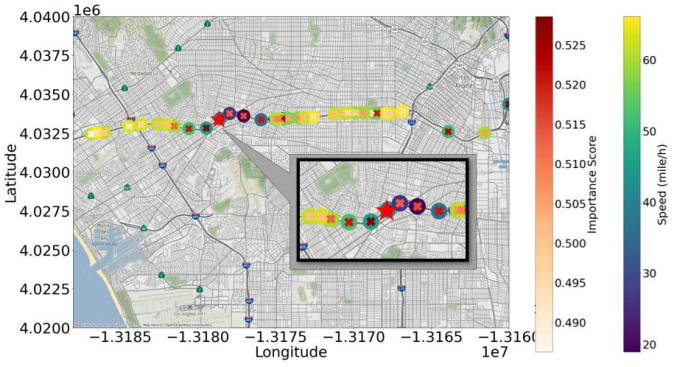


Fig. 4: Case study of incident #18693 on the Freeway I–10. This is the map of an incident and the surrounding traffic loop sensors on I–10 in the Incident-LA dataset. The red star indicates the location of the accident, and the yellow–green–blue dots are sensors colored according to speed measurements. The orange–red "X"s indicate the importance-scores assigned by RoadFormer.

accuracy. We examined simple methods of merging the auxiliary information into the traffic network, such as adding an accident classification task, using road and sensor position for position encoding, and embedding accident metadata. While none of those methods worked, all of these attempts are also uploaded to our GitHub repository [9] for others to investigate. In the future, we will explore more about the relationship between incident impact prediction and related tasks, such as traffic flow/congestion propagation prediction. Additionally, we will explore additional sources of data that could enhance incident impact prediction and will continuously update our datasets accordingly.

VI. Conclusion

In this paper, we present a novel definition of TII prediction tasks within the context of spatiotemporal data mining and introduce two enhanced traffic incident impact datasets. We design a new transformer-based TII prediction model, which contains a novel road-anchored spatial transformer encoder and an importance-score temporal transformer decoder to assist in identifying measurements affected by the accidents. We evaluated the performance of two conventional models and six deep graph traffic forecasting models on our datasets. The experiments show that RoadFormer outperforms the other models. Moreover, the ablation studies show that the road-anchored attention strategy outperforms general attention layers, and the case study shows that the importance-score transformer can identify incident-relevant sensors.

[9]https://github.com/styxsys0927/RoadFormer.git

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on August 07,2024 at 19:38:14 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] M. W. Adler, J. van Ommeren, and P. Rietveld, "Road congestion and incident duration," *Economics of transportation*, vol. 2, no. 4, pp. 109–118, 2013.

[2] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," *Advances in neural information processing systems*, vol. 33, pp. 17 804–17 815, 2020.

[3] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," *arXiv preprint arXiv:1707.01926*, 2017.

[4] F. He, X. Yan, Y. Liu, and L. Ma, "A traffic congestion assessment method for urban road networks based on speed performance index," *Procedia engineering*, vol. 137, pp. 425–433, 2016.

[5] Z. Zhou, Y. Wang, X. Xie, L. Chen, and H. Liu, "Riskoracle: A minute-level citywide traffic accident forecasting framework," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 1258–1265.

[6] B. Wang, Y. Lin, S. Guo, and H. Wan, "Gsnet: learning spatial-temporal correlations from geographical and semantic aspects for traffic accident risk forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 5, 2021, pp. 4402–4409.

[7] K. Fu, T. Ji, L. Zhao, and C.-T. Lu, "Titan: A spatiotemporal feature learning framework for traffic incident duration prediction," in *27th ACM SIGSPATIAL*, 2019, pp. 329–338.

[8] K. Fu, T. Ji, N. Self, Z. Chen, and C.-T. Lu, "A hierarchical attention graph convolutional network for traffic incident impact forecasting," in *2021 IEEE Big Data*. IEEE, 2021, pp. 1619–1624.

[9] G. Meng, Q. Jiang, K. Fu, B. Lin, C.-T. Lu, and Z. Chen, "Early forecasting of the impact of traffic accidents using a single shot observation," in *2022 SIAM SDM*. SIAM, 2022, pp. 100–108.

[10] T. Afrin and N. Yodo, "A survey of road traffic congestion measures towards a sustainable and resilient transportation system," *Sustainability*, vol. 12, no. 11, p. 4660, 2020.

[11] M. Abbasi, A. Shahraki, and A. Taherkordi, "Deep learning for network traffic monitoring and analysis (ntma): A survey," *Computer Communications*, vol. 170, pp. 19–41, 2021.

[12] S. Tuli, G. Casale, and N. R. Jennings, "Tranad: Deep transformer networks for anomaly detection in multivariate time series data," *arXiv preprint arXiv:2201.07284*, 2022.

[13] M. Ben-Akiva, M. Bierlaire, D. Burton, H. Koutsopoulos, and R. Mishalani, "Network state estimation and prediction for real-time traffic management," *Networks and Spatial Economics*, vol. 1, pp. 293–318, 09 2001.

[14] N. Motamedidehkordi, M. Margreiter, and T. Benz, "Shockwave suppression by vehicle-to-vehicle communication," *Transportation research procedia*, vol. 15, pp. 471–482, 2016.

[15] J. Tang, L. Zheng, C. Han, W. Yin, Y. Zhang, Y. Zou, and H. Huang, "Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review," *Analytic Methods in Accident Research*, vol. 27, p. 100123, 2020.

[16] Y. Lin and R. Li, "Real-time traffic accidents post-impact prediction: Based on crowdsourcing data," *Accident Analysis & Prevention*, vol. 145, p. 105696, 2020.

[17] Y. Zou, B. Lin, X. Yang, L. Wu, M. Muneeb Abid, and J. Tang, "Application of the bayesian model averaging in analyzing freeway traffic incident clearance time for emergency management," *Journal of Advanced Transportation*, vol. 2021, pp. 1–9, 2021.

[18] A. Grigorev, A.-S. Mihaita, S. Lee, and F. Chen, "Incident duration prediction using a bi-level machine learning framework with outlier removal and intra–extra joint optimisation," *Transportation Research Part C: Emerging Technologies*, vol. 141, p. 103721, 2022.

[19] H. Zhang, W. Zhang, J. Li, and X. Li, "Analysis of spatiotemporal impact of traffic incidents on road networks," in *International Conference on Intelligent Transportation Engineering*. Springer, 2022, pp. 763–772.

[20] J. Lee, J. Kwak, Y. Oh, and S. Kim, "Quantifying incident impacts and identifying influential features in urban traffic networks," *Transportmetrica B: Transport Dynamics*, pp. 1–22, 2022.

[21] W. Zhu, J. Wu, T. Fu, J. Wang, J. Zhang, and Q. Shangguan, "Dynamic prediction of traffic incident duration on urban expressways: A deep learning approach based on lstm and mlp," *Journal of intelligent and connected vehicles*, vol. 4, no. 2, pp. 80–91, 2021.

[22] S. Tuli, M. R. Wilkinson, and C. Kettell, "Radnet: Incident prediction in spatio-temporal road graph networks using traffic forecasting," *arXiv preprint arXiv:2206.05602*, 2022.

[23] M. Bai, Y. Lin, M. Ma, P. Wang, and L. Duan, "Prepct: Traffic congestion prediction in smart cities with relative position congestion tensor," *Neurocomputing*, vol. 444, pp. 147–157, 2021.

[24] Q. Xie, T. Guo, Y. Chen, Y. Xiao, X. Wang, and B. Y. Zhao, "Deep graph convolutional networks for incident-driven traffic speed prediction," in *29th ACM CIKM*, 2020, pp. 1665–1674.

[25] Q. Zhang, J. Chang, G. Meng, S. Xiang, and C. Pan, "Spatio-temporal graph structure learning for traffic forecasting," in *AAAI*, vol. 34, no. 1, 2020, pp. 1177–1185.

[26] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[27] B. Yoo, J. Lee, J. Ju, S. Chung, S. Kim, and J. Choi, "Conditional temporal neural processes with covariance loss," in *ICML*. PMLR, 2021, pp. 12 051–12 061.

[28] Z. Huang, A. Arian, Y. Yuan, and Y.-C. Chiu, "Using conditional generative adversarial nets and heat maps with simulation-accelerated training to predict the spatiotemporal impacts of highway incidents," *Transportation research record*, vol. 2674, no. 8, pp. 836–849, 2020.

[29] D. Drakulic and J.-M. Andreoli, "Structured time series prediction without structural prior," *arXiv preprint arXiv:2202.03539*, 2022.

[30] C. Tian and W. K. Chan, "Spatial-temporal attention wavenet: A deep learning framework for traffic prediction considering spatial-temporal dependencies," *IET Intelligent Transport Systems*, vol. 15, no. 4, pp. 549–561, 2021.

[31] M. Xu, W. Dai, C. Liu, X. Gao, W. Lin, G.-J. Qi, and H. Xiong, "Spatial-temporal transformer networks for traffic flow forecasting," *arXiv preprint arXiv:2001.02908*, 2020.

[32] Z. Wu, P. Jain, M. Wright, A. Mirhoseini, J. E. Gonzalez, and I. Stoica, "Representing long-range context for graph neural networks with global attention," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 266–13 279, 2021.

[33] L. Rampasek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, and D. Beaini, "Recipe for a general, powerful, scalable graph transformer," *arXiv preprint arXiv:2205.12454*, 2022.

[34] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph wavenet for deep spatial-temporal graph modeling," *arXiv preprint arXiv:1906.00121*, 2019.

[35] L. Han, B. Du, L. Sun, Y. Fu, Y. Lv, and H. Xiong, "Dynamic and multi-faceted spatio-temporal deep learning for traffic speed forecasting," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 547–555.

[36] R. Jiang, D. Yin, Z. Wang, Y. Wang, J. Deng, H. Liu, Z. Cai, J. Deng, X. Song, and R. Shibasaki, "Dl-traff: Survey and benchmark of deep learning models for urban traffic prediction," in *Proceedings of the 30th ACM international conference on information & knowledge management*, 2021, pp. 4515–4525.

[37] R. Jiang, Z. Wang, J. Yong, P. Jeph, Q. Chen, Y. Kobayashi, X. Song, S. Fukushima, and T. Suzumura, "Spatio-temporal meta-graph learning for traffic forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, 2023, pp. 8078–8086.

[38] H. Liu, Z. Dong, R. Jiang, J. Deng, J. Deng, Q. Chen, and X. Song, "Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 4125–4129.

[39] Y. Chen, L. Wu, and M. Zaki, "Iterative deep graph learning for graph neural networks: Better and robust node embeddings," *Advances in neural information processing systems*, vol. 33, pp. 19 314–19 326, 2020.

[40] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *arXiv preprint arXiv:1810.00826*, 2018.

[41] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on August 07,2024 at 19:38:14 UTC from IEEE Xplore. Restrictions apply.