CreST: A Credible Spatiotemporal Learning Framework for Uncertainty-aware Traffic Forecasting

Zhengyang Zhou
University of Science and Technology of China (USTC)
Hefei, China
Suzhou Institute for Advanced Research, USTC
Suzhou, China
zzy0929@ustc.edu.cn

Jiahao Shi
University of Science and Technology of China
Hefei, China
shijiahao@mail.ustc.edu.cn

Hongbo Zhang
University of Science and Technology of China
Hefei, China
zhanghongbo@mail.ustc.edu.cn

Qiongyu Chen
University of Science and Technology of China
Hefei, China
chenqy96@mail.ustc.edu.cn

Xu Wang*
University of Science and Technology of China
Hefei, China
wx309@mail.ustc.edu.cn

Hongyang Chen
Zhejiang Lab
Hangzhou, China
dr.h.chen@ieee.org

Yang Wang
University of Science and Technology of China, Hefei, China
angyan@ustc.edu.cn

ABSTRACT

Spatiotemporal traffic forecasting plays a critical role in intelligent transportation systems, which empowers diverse urban services. Existing traffic forecasting frameworks usually devise various learning strategies to capture spatiotemporal correlations from the perspective of volume itself. However, we argue that previous traffic predictions are still unreliable due to two aspects. First, the influences of context factor-wise interactions on dynamic region-wise correlations are under exploitation. Second, the dynamics induce the credibility issue of forecasting that has not been well-explored. In this paper, we exploit the informative traffic-related context factors to jointly tackle the dynamic regional heterogeneity and explain the stochasticity, towards a credible uncertainty-aware traffic forecasting. Specifically, to internalize the dynamic contextual influences into learning process, we design a context-cross relational embedding to capture interactions between each context, and generate virtual graph topology to dynamically relate pairwise regions with context embedding. To quantify the prediction credibility, we attribute data-side aleatoric uncertainty to contexts and re-utilize them for aleatoric uncertainty quantification. Then we couple a dual-pipeline learning with the same objective to produce the discrepancy of model outputs and quantify model-side epistemic uncertainty. These two uncertainties are fed through a spatiotemporal network for extracting uncertainty evolution patterns. Finally, comprehensive experiments and model deployments have corroborated the credibility of our framework.

CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**; **Data mining**; • **Computing methodologies** → *Knowledge representation and reasoning*; *Artificial intelligence.*

KEYWORDS

Traffic prediction, urban computing, uncertainty quantification, conditional prediction.

ACM Reference Format:
Zhengyang Zhou, Jiahao Shi, Hongbo Zhang, Qiongyu Chen, Xu Wang, Hongyang Chen, and Yang Wang. 2024. CreST: A Credible Spatiotemporal Learning Framework for Uncertainty-aware Traffic Forecasting. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24), March 4–8, 2024, Merida, Mexico.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3616855.3635759

1 INTRODUCTION

Rapid urbanization has introduced large-scale increases of transportation demands, posing great challenges to sustainability and urban management in modern cities [1, 2]. Therefore, Intelligent Transportation System (ITS), especially traffic forecasting, has attracted huge attention from both academia [3–7] and industry [8, 9]. As a pivotal part of ITS, traffic forecasting is powerful to promote the efficiency of urban travelling [3], increase profits of ride-sharing

*Dr. Xu Wang is the corresponding author.

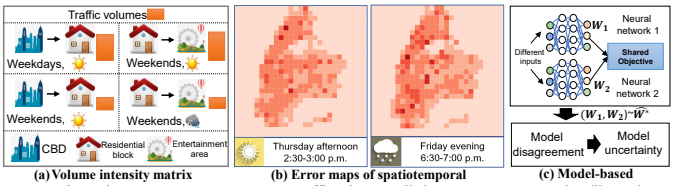
WSDM '24, March 4–8, 2024, Merida, Mexico

ACM ISBN 979-8-4007-0371-3/24/03…$15.00
https://doi.org/10.1145/3616855.3635759

platforms [10], and facilitate the management of urban safety [11], thereby enabling intelligent and efficient urban life.

In the literature, existing traffic forecasting frameworks learn to establish the mapping functions from historical observations to targets, based on various spatial aggregations [12, 13], and sequence learning schemes [3, 4, 14]. Specifically, to extract multi-level and high-level correlations, literature [15] devises a multi-range attentive bicomponent graph convolutional network while [16] proposes a 3-dimensional adjacent matrix, to respectively endow multi-scale correlation learning and spatial-temporal fully-connected correlation construction. Even though, all these methods construct the correlations from the volume itself, ignoring two critical issues, i.e., the dynamic region-wise correlations induced by interactive urban circumstances, and the credibility of these predicted values. Hence, lacking these considerations contribute to unreliable and irresponsible learning frameworks.

In real scenarios, traffic volumes are the reflection of human daily routines thus influenced by various urban circumstances. We formalize all external urban circumstances including functionality, weather, timestamps and day types as contexts, and illustrate a series of region-wise volume patterns under context factor-wise interactions in Figure 1(a). As observed, traffic volumes between CBD and residential blocks are sensitive to day of week while volumes between residential blocks and entertainment areas are sensitive to weather on weekends. Such observations demonstrate that different regional functionalities can interact with the same weather context to induce various traffic patterns, leading to heterogeneous region-wise correlations. A more interpretable reason is that weather can influence the non-necessary activities but have fewer effects on indispensable travelling. To this end, even some pioneering works involve contexts to correct the final prediction intensity [3, 6, 17, 18] still fail to consider the root causes of such dynamic correlations, i.e., interactions of various contexts, yielding suboptimal performances. Furthermore, the complex and dynamic correlations can directly induce another critical issue, i.e., the prediction stochasticity and credibility [19]. Seriously, an inaccurate prediction will provide misleading numerical decision-making basis for police force assignment and route planning of autonomous driving [20, 21], resulting in irreversible crisis on human safety. Therefore, we formally raise the credibility issue in traffic forecasting.



(a) Volume intensity matrix under various contexts (b) Error maps of spatiotemporal traffic volume prediction (c) Model-based uncertainty illustration

Figure 1: Motivation of our solutions. $W_1, W_2, \widetilde{W^*}$ are model parameters of neural network 1, neural network 2 and the potential distributions of model parameters.

The solution to quantifying the prediction credibility is to measure the potential differences between groundtruth and predicted values, which can be named as uncertainty [21–23]. Prediction uncertainty can be generally classified into two categories as follows. The aleatoric one, induced by unobservable factors and data noise, characterizes the difficulty of learning tasks, while the epistemic one captures potential distributions of model parameters, which can be explained away with increasing training samples [21, 24]. Since two categories of uncertainty play different roles in learning systems, explicitly identifying them can better understand the algorithm, hence increasing its robustness and reliability. Unfortunately, existing works of uncertainty-aware spatiotemporal learning, even the latest two studies [22, 25] cannot explicitly distinguish the two sources of uncertainty, resulting in under exploration of credible learning systems. To perform a detailed analysis of uncertainty, considering the significant role of context factors, we further draw maps of prediction bias on a well-known traffic forecasting model (STG2Seq) [3], a.k.a. error maps, associated with various contexts in Figure 1(b). We observe that the prediction accuracy is varied with context factors, and the reason lies in the fact that different contexts can induce heterogeneous occurrence possibility of accidental but unobservable events on the road network. This observation provides insights that context factors can partially explain the uncertainty, but it is still unclear on how to leverage the contexts and techniques to disentangle two types of uncertainty.

As discussed above, a credible traffic forecasting system requires highly efficient data exploitation and uncertainty type-aware prediction. By resorting to informative contexts, two challenges still remain unresolved for credible forecasting, 1) capturing context interactions to model dynamic region-wise correlations, 2) disentangling and quantifying two types of uncertainty.

In this paper, we develop a Credible SpatioTemporal learning framework (CreST), exploring the traffic-related context factors to cooperatively overcome above challenges. Specifically, to internalize the dynamic contextual influences into region-wise correlation learning, we propose a Context-Condition SpatioTemporal network (C2ST), which captures context-wise relational interactions to achieve context embedding and generates region-wise proximity based on above embedding. To enable high-quality context embedding and the awareness of context-target regularity, we introduce a context-target highway for representation enhancement and task regularization. From the perspectives of human behavior patterns and model uncertainty on ensembling disagreement, we propose a Context-Discrepancy Uncertainty Quantification (CDUQ) to respectively quantify model-side epistemic uncertainty and data-side aleatoric uncertainty. First, we take contexts to interpret data-side aleatoric uncertainty, as contexts can reflect the occurrences of unobservable events. We then re-utilize the context embedding to construct data-side uncertainty. Second, inspired by dropout [21, 26] and ensemble-based uncertainty learning [24], we conclude that different models with the same objective can imitate the parameter distributions and thus the disagreement in ensemble models describes the prediction confidence and model uncertainty [27, 28]. Therefore, we couple two learning pipelines and devise an ensemble scheme to capture model-side discrepancy, which is illustrated in Figure 1(c). Finally, since residual measures how much the prediction result deviates from groundtruth, we thereby introduce residuals as an indicator proxy to allow uncertainty predictable. Coherently, we exploit the semantic context embedding and dual learning pipeline to support our CDUQ for type-aware uncertainty disentanglement. We make the following contributions,

- We propose a credible spatiotemporal learning model for ITS, which simultaneously enjoys high-quality data exploitation and fine-grained type-aware uncertainty quantification.
- CreST exploits context factors by coupling two designed components, C2ST and CDUQ,. Specifically, C2ST captures cross interactions between context factors and generates virtual topology while CDUQ leverages the ensembling to imitate the model output discrepancy, and harnesses contexts to achieve the data-aspect uncertainty.
- Experiments on three traffic datasets reveal that our solution not only outperforms other uncertainty quantification by 10%, but also demonstrates the uncertainty regression can boost learning performances from 18.20% to 28.50%.

2 RELATED WORK

2.1 Traffic forecasting

Traffic forecasting is a classic spatiotemporal learning task, where its solutions can be categorized into traditional machine learning and deep learning. Regarding traditional learning-based methods, context-aware matrix factorization [29] and network kernel density estimation [30] are proposed to capture the spatial dependencies, while various autoregression methods including ARIMA [31], Moving Average [32] are devised for sequence pattern extractions. However, these methods capture the regularity only from one single view, failing to consider the joint spatial-temporal correlations. Deep learning-based solutions are capable of making up the fitting capacity issue, and these works can be classified as context-agnostic and context-aware. For context-agnostic ones, [33, 34] and [5] respectively design an ST-attention and an adaptive graph learning to respectively enable the model to focus on the most beneficial spatial-temporal features. Besides, diverse temporal learning solutions including GRU [5, 12], LSTM [11] and TCN [3] are proposed to enhance spatiotemporal learning. Unfortunately, context-agnostic methods neglect the significance of contexts in forecasting, yielding suboptimal performance. Latest works employ a fully-connected neural network to encode the contextual information and then aggregate them with main stream observations with element-wise addition [3, 6, 17, 18] or concatenation [35]. Although these methods incorporate the context to correct intensity of predictions, they still fail to enable contexts to guide dynamic aggregations. Given heterogeneous interactions between contexts, the aggregation patterns of spatiotemporal elements should be reformed.

2.2 Uncertainty quantification for deep learning

Early work [36] categorizes uncertainty into epistemic and aleatoric, which has been widely recognized by subsequent studies. By assuming the learnable variables follow Gaussian distribution, off-the-shelf literature captures the learned model parameter distributions for epistemic uncertainty quantification. They devise various techniques, including Dropout [23, 37, 38], Ensemble [24] and imitated Brownian motions [39] to derive the variances of multiple predictions as their epistemic uncertainty. Regarding aleatoric one, existing methods often construct mappings from input data to such uncertainty and maintain the consistency between errors and learnable aleatoric uncertainty with loss functions. However, since these methods are limited in image-like static data, it is naturally difficult to adapt them to spatiotemporal learning.

Encouragingly, pioneering works of numerical weather [37, 38] and meteorology forecasting [23] advance spatiotemporal framework towards uncertainty quantification. In detail, [37] employs a sampling-computing strategy to identify different categories of uncertainties while [23, 38] speculate the potential uncertainty of results by adjusting the confidence quantiles. However, [23, 38] cannot adaptively identify uncertainty categories and [37] fails to perform real-time speculation considering multiple contexts. More recently, [22] provides comprehensive benchmarks on spatiotemporal uncertainty learning, and [25] devises an ST-variance-based uncertainty indicator. Given context-induced heterogeneous uncertainty, these above-mentioned works either fail to internalize the context into spatiotemporal uncertainty [22, 23, 38], or ignore the model fitting capacity [25], leading to under exploration of a real-time and type-aware uncertainty learning system.

3 NOTATIONS AND PROBLEM DEFINITION

Definition 1 (**Urban graph and urban regions**). *The whole city is equally discretized into N regions by longitudes and latitudes, and constructed as a directed urban graph $G(\mathcal{V}, \mathcal{E})$, where each region in the city constitutes of the vertex set $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$. $\mathcal{E} = \{e_{ij} | 1 \leqslant i, j \leqslant N\}$ is a set of edges describing the dynamic region-wise proximity.*

Definition 2 (**Observations of regional traffic elements**). *Time domain can be equally discretized into an interval set $\mathcal{T} = \{1, 2, 3..., T, T+1, ...\}$. We formulate the main observations of spatiotemporal traffics as $\{\mathbb{X} = \boldsymbol{X}^t | t \in \mathcal{T}\}$, where the element $\boldsymbol{X}^t$ denotes the citywide traffic observation, and $x_i^t \in \boldsymbol{X}^t (1 \leqslant i \leqslant N)$ is the traffic observation at region v_i during interval t.*

Definition 3 (**Context factors**). *We define the traffic covariates, which are related with traffics but not for predictions, as context factors. Given M types of context factors $\mathbb{C} = \{C_1, C_2, \cdots, C_M\}$, the descriptor of m-th context type at region v_i during interval t can be denoted as the $c_{(m,i)}^t \in \mathbb{R}^{1 \times d_m}$, where d_m is the vector dimension of m-th type context factor. The concrete context types can include temporal context, spatial context, as well as weather where each type can involve multiple observations.*

Problem 1 (**Credible spatiotemporal traffic forecasting task**). *Our credible learning is interpreted as the predicting main observations and potential uncertainty by boosting utilization of context factors. Therefore, given historical spatiotemporal observations, $\boldsymbol{X}^1, ..., \boldsymbol{X}^T$ and their counterpart time-varying context factors $\boldsymbol{c}^1, \boldsymbol{c}^2, ..., \boldsymbol{c}^T$, we perform credible spatiotemporal forecasting in the following l time steps, $(\widehat{\boldsymbol{Y}}^{T+t}, \widehat{\boldsymbol{\sigma}}^{T+t})(t = 1, 2, ..., l)$ by simultaneously capturing dynamic region-wise correlations induced by context interactions and tackling type-aware uncertainty learning challenge.*

4 METHODOLOGY

4.1 Framework overview

Illustrated in Figure 2, the Credible SpatioTemporal learning framework (CreST) consists of two well-designed neural networks, Context-conditioned SpatioTemporal network (C2ST) for context conditioned aggregations based on dynamically relating citywide regions,

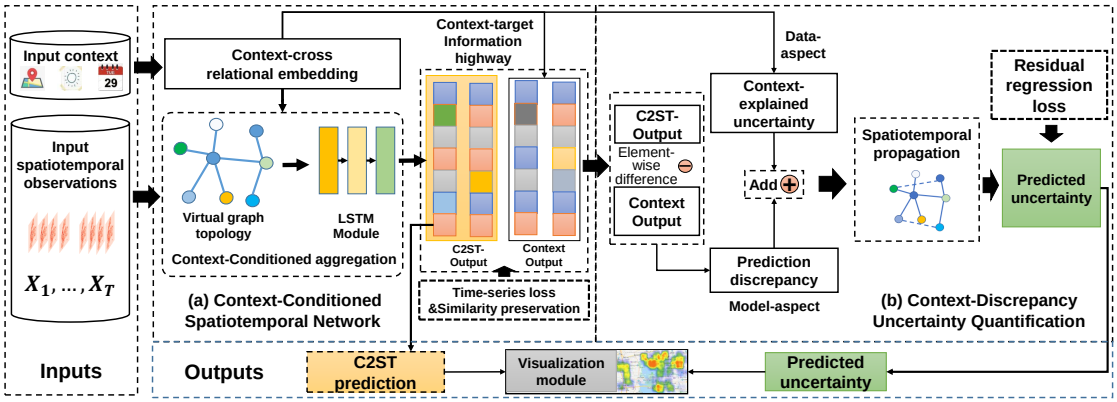


Figure 2: Overview of Credible SpatioTemporal learning framework (CreST)

and Context-Discrepancy Uncertainty Quantification (CDUQ) for disentangling epistemic-aleatoric uncertainty.

4.2 Context-conditioned spatiotemporal network

Our Context-conditioned SpatioTemporal network (C2ST) is proposed to learn the dynamic region-wise proximity conditioned on contexts, and perform the context-conditioned aggregations. It is composed of three components below.

4.2.1 Context-cross relational embedding. The embedding of context factors plays a vital role in context-aware forecasting, but existing embedding strategies usually neglect the influences of context-wise interactions, leading to suboptimal and unreliable predictions. In our work, we devise a novel context-cross relational embedding to quantify context-wise mutual interactions, by drawing the inspiration from relational GNN [40]. Our context-cross relation learning takes each context factor as an individual entity and feeds them into a relational neural network. By denoting $\widetilde{c}_m \in \mathbb{R}^{N\times d_m}$ as the citywide context tensor for type m, we then learn the citywide context-cross latent embedding $\mathbf{Z_c} \in \mathbb{R}^{N\times K}$ by,

$$\mathbf{Z_c} = \prod_{m=1}^{M} (\widetilde{c}_m + \mathbf{b_m})\mathbf{w_{(m,m+1)}} \tag{1}$$

where each region is with a K-dimension embedding and the learnable weight $\mathbf{b_m} \in \mathbb{R}^{1\times d_m}$ accounts for obtaining vectors of continuous values. Then the relations and interactions between each context factor can be well captured by $\mathbf{w_{(m,m+1)}} \in \mathbb{R}^{d_m\times d_{m+1}} (m < M)$. The last weight $\mathbf{w_{(M,M+1)}} \in \mathbb{R}^{d_M\times K}$ is a linear transformation that converts the context dimension from d_M to K, achieving citywide context embedding $\mathbf{Z_c}$. We denote $\mathbf{Z_c}(i) \in \mathbf{Z_c}$ as the i-th row of $\mathbf{Z_c}$ and consider it as the integrated context embedding of i-th region. As a result, the context-wise interactions and target-related semantic correlations can be progressively aggregated into our context-cross relational representation.

4.2.2 Context-target information highway. To encode the awareness of context-target regularity, we design another learning pipeline, context-target information highway by establishing mappings from contexts to main prediction targets. This context-target highway treats regional context factor combinations and targeted time series as sample pairs. Actually, we realize the highway in a simple but effective way, i.e., directly reusing learned $\mathbf{Z_c}$ to predict the targeted time series. For each region v_i, let the output of context-target highway be $\widehat{\mathbf{Y}_\mathbf{c}}(i)$, the predicted sequence can be obtained by the context-to-sequence linear transformation $\boldsymbol{h_{cs}}$,

$$\widehat{\mathbf{Y}_c}(i) = \boldsymbol{h_{cs}}(\mathbf{Z_c}(i)) = \mathrm{ReLU}(\mathbf{Z_c}(i) * \mathbf{w_{cs}} + \mathbf{b_{cs}}) \tag{2}$$

where $\mathbf{w_{cs}} \in \mathbb{R}^{K\times l}$ and $\mathbf{b_{cs}} \in \mathbb{R}^{N\times l}$ are learnable parameters. The training objective of this information highway is minimizing the element-wise MAPE of the predicted sequence, which is shared with the main task in the next subsection. This learning process will backpropagate the gradient to weights on trainable context embedding and force them to gain semantics conforming to targets.

4.2.3 Context-conditioned spatiotemporal aggregation. Since we have obtained the interaction-involved context embedding, we can exploit such semantic embedding to perform spatiotemporal aggregations. As adjacent matrix describes the element-wise proximity and guides spatial aggregations, we thus formulate the context-conditioned adjacent matrix as the function of context-cross relational embedding to dynamically relate pairwise regions. To this end, virtual graph topology is generated with $\mathbf{Z_c}$ by,

$$\widetilde{\mathbf{A}} = \boldsymbol{Trans}(\mathbf{Z_c}) = \mathrm{Softmax}(\mathbf{Z_c}\mathbf{Q} + \mathbf{q}) \tag{3}$$

In particular, ***Trans*** is a linear transformation parameterized by $\mathbf{Q} \in \mathbb{R}^{K\times N}, \mathbf{q} \in \mathbb{R}^{N\times N}$, transferring the context embedding into the formation of adjacent matrix and assigning the node-wise proximity in a data-driven manner. Softmax serves to normalize the adjacent proximity regarding each node and constrains the summation of all neighboring proximity to 1.

After that, we stack several GNN blocks to perform message passing through the graph topology. Here we take one of the GCN blocks to demonstrate the graph convolution by denoting the k-th hidden layer of GNN as $\mathbf{H}^k$,

$$\mathbf{H}^k = \mathrm{ReLU}(\tilde{\mathbf{D}}_A^{t\,-1/2}\widetilde{\mathbf{A}}^t\tilde{\mathbf{D}}_A^{t\,-1/2}\mathbf{H}^{k-1}\mathbf{W_{gc}}^{k-1}) \tag{4}$$

where $\widetilde{\mathbf{A}}^t$ is the context-conditioned adjacency of $\widetilde{\mathbf{A}}$ at interval t and $\widetilde{\mathbf{A}}^t = \mathbf{A}^t + \mathbf{I}_N$. Matrices $\mathbf{I}_N$ and $\tilde{\mathbf{D}}_A^t$ are an N-order identity matrix and a degree matrix for $\widetilde{\mathbf{A}}^t$. We respectively instantiate

$\mathbf{H}_G^0$ as $\mathbf{X}^{\mathcal{P}_c}$, $\mathbf{X}^{\mathcal{P}_d}$ and $\mathbf{X}^{\mathcal{P}_l}$ in each parallel GNN block, consisting of the consecutive observations on the levels of closeness, daily periodicity, and long-term trends, following common settings in traffic forecasting [17, 18]. $\mathbf{W}_{gc}^k$ are a series of learnable parameters.

Regarding temporal learning, we leverage an LSTM parameterized by $\mathbf{W}_{\text{lstm}}$ to capture temporal evolutions, and output the spatiotemporal learning-based prediction results $\widehat{\mathbf{Y}}_f^{T+t}$,

$$\widehat{\mathbf{Y}}_f^{T+t} = \text{LSTM}((\mathbf{H}^{\mathcal{P}_c}, \mathbf{H}^{\mathcal{P}_d}, \mathbf{H}^{\mathcal{P}_l}), \mathbf{W}_{\text{lstm}}) \tag{5}$$

We name $\widehat{\mathbf{Y}}_f^{T+t}$ as C2ST-Output and consider it equivalent to the outputs of our main spatiotemporal forecasting task $\widehat{\mathbf{Y}}^{T+t}$.

4.2.4 Shared objectives of daul-pipeline learning. Since the context-target information highway aims to plug the awareness of context-target regularity into our network, we take the training objectives of both information highway and spatiotemporal prediction into the same one, i.e., regressing the main target observations. Besides, to enable the learned information to share across these two pipelines, we introduce a similarity regularization to preserve the similarity between these two predicted sequences. Then the shared objectives of these two pipelines can be three-fold, two regression losses, and one similarity constraint. We formalize it as follows,

$$\begin{aligned} Loss_{ST} &= \text{MAPE}_{\text{f}} + \beta\text{MAPE}_{\text{c}} + \gamma sim(\widehat{Y_c}(i), \widehat{Y_f}(i)) \\ &= \frac{1}{Nl}\sum_{t=1}^{l}\sum_{i=1}^{N}\{(\frac{\widehat{y_f^t}(i) - y_i^t}{y_i^t})^2 + \beta(\frac{\widehat{y_c^t}(i) - y_i^t}{y_i^t})^2\} \\ &- \frac{\gamma}{N}\sum_{i=1}^{N}\frac{\widehat{Y_c}(i)\cdot\widehat{Y_f}(i)}{||\widehat{Y_c}(i)||||\widehat{Y_c}(i)||} \end{aligned} \tag{6}$$

where sim is the cosine similarity measurement, β, γ are two hyper-parameters for balancing the losses among MAPE of the main task, context-target MAPE and cosine similarity. So far, we have realized a daul-pipeline learning where they share the same learning objectives, and we can respectively consider context-target outputs $\widehat{Y_c}$ and C2ST-Outputs $\widehat{Y_f}$ as coarse-grained predictions from contexts and fine-grained predictions from all historical observations.

4.3 Context-discrepancy uncertainty quantification

4.3.1 Motivations. Firstly, residual measures the differences between model outputs and groundtruth, implicitly representing the model fitting capacity and prediction bias of the trained model. We take the residual derived from the training process as an uncertainty indicator, and let it as an additional regression objective to enable online uncertainty inference. We disentangle the uncertainty into data-side and model-side. Specifically, we attribute the data-side aleatoric uncertainty to context factors due to their potential of reflecting occurrences of unobservable events. For epistemic uncertainty, inspired by ensembling [24] and dropout [21, 26] uncertainty quantification, we argue that different models with the same objective can imitate the parameter distributions, thus the disagreement across diverse sub-models can capture the model uncertainty. Therefore, we formulate our CDUQ with context-explained uncertainty learning, model discrepancy-based uncertainty learning and spatiotemporal uncertainty propagation.

4.3.2 Context-explained uncertainty. First, we capture aleatoric uncertainty $\mathbf{u_c}$ by a context-explained uncertainty network,

$$\mathbf{u_c} = \text{ReLU}(\mathbf{Z_c}\mathbf{R}) \tag{7}$$

where $\mathbf{R} \in \mathbb{R}^{K\times l}$ is the neural weight converting the dimension of context embedding into the same as the time step, and we can obtain the context-explained uncertainty $\mathbf{u_c} \in \mathbb{R}^{N\times l}$.

4.3.3 Model discrepancy-based uncertainty. We resort to model ensembling to capture the potential distribution of model parameters. Actually, the differences between coarse-grained context-target learning and fine-grained spatiotemporal predictions are two different instantiations with the same objectives, thus they can imitate the parameters distributions of models. We then introduce the third role of context-target highway, i.e., constructing an ensembled model by involving C2ST outputs, and extracting model discrepancy to derive model-side uncertainty. We then calculate the differences between C2ST outputs and context-target highway, and impose a learnable weight $\mathbf{V} \in \mathbb{R}^{l\times l}$ to allow achieving model-side uncertainty $\mathbf{u_p} \in \mathbb{R}^{N\times l}$,

$$\mathbf{u_p} = \text{ReLU}((\widehat{\mathbf{Y_c}} - \widehat{\mathbf{Y_f}})\mathbf{V}) \tag{8}$$

where ReLU preserves the positive definiteness. This uncertainty is the function of predicted results, which can be further deemed as prediction-correlated uncertainty. After that we aggregate them with an adjustable parameter α to achieve the overall uncertainty $\mathbf{u_o} \in \mathbb{R}^{N\times l}$ conditioned on three factors, contexts, model property, as well as prediction results,

$$\mathbf{u_o} = \alpha\mathbf{u_c} + (1-\alpha)\mathbf{u_p} \tag{9}$$

4.3.4 Spatiotemporal uncertainty propagation. The evolution of spatiotemporal uncertainty is composed of two components, spatial propagation and temporal evolution. Firstly, to imitate the spatial propagation of uncertainty, we design one GNN-based uncertainty propagation layer by leveraging the topology information. To achieve the topology matrix, we first derive a distance-based adjacent matrix $\mathbf{A_{dist}} \in \mathbb{R}^{N\times N}$ by considering region-wise Euclidean distance as the proximity,

$$A_{dist}(i,j) = e^{-\frac{dist(v_i,v_j)}{\tau}} \tag{10}$$

where $dist(v_i, v_j)$ represents the Euclidean distance between regions v_i and v_j, and τ is the scalar controlling the bandwidth of distance metric. Assuming $\mathbf{D}$ is the degree matrix for $\mathbf{A_{dist}}$, the topology matrix $\mathbf{\Gamma}$ can be calculated based on $\mathbf{A_{dist}}$,

$$\mathbf{\Gamma} = \mathbf{I} + \mathbf{S},\ \mathbf{S} = \mathbf{D}^{-1/2}\mathbf{A_{dist}}\mathbf{D}^{-1/2} \tag{11}$$

Thus, we have the spatially correlated overall uncertainty $\mathbf{u_{os}}$ with one-layer message passing,

$$\mathbf{u_{os}} = \text{GNN}(\mathbf{u_o}, \mathbf{\Gamma}) = \text{ReLU}(\mathbf{\Gamma}\mathbf{u_o}\mathbf{w_G}) \tag{12}$$

where $\mathbf{w_G}$ are learnable weights for spatial filters. Secondly, to simulate the temporal propagation, we perform the step-aware uncertainty transformation to learn the gated scalar by incorporating the embedding of timestamp $\boldsymbol{ts}_i$ for step i. We can obtain the final predicted overall uncertainty $\mathbf{u_o^*}$ as follows,

$$\mathbf{u_o^*} = \mathbf{u_{os}} \odot \text{TempGate}(\boldsymbol{ts}_i) = \mathbf{u_{os}} \odot \tanh(\mathbf{w_e}\cdot\boldsymbol{ts}_i) \tag{13}$$

Table 1: Dataset statistics (m: million, k: thousand)

Dataset	Category of datasets	# of records	Time Span	# of regions
SIP	Surveillance	2.7 m	01/01/2017-03/31/2017	108
	Weather	4.3k		
NYC	Taxi trips	7.5 m	01/01/2017-05/31/2017	354
	Weather	7.4k		
Metr-LA	Loop detectors	4.9 m	03/01/2012-06/30/2012	207
	Weather	5.7k		

where $\odot$ denotes element-wise product, $\mathbf{w_e}$ is the learnable parameter for time step embedding, and TempGate is a tanh function accounting for temporal evolution learning.

Finally, in our CDUQ, taking the residual res_i^j as the learning objective regarding v_i and j, we have loss function,

$$Loss_{unc} = \sum_{j=1}^{N} \sum_{i=1}^{l} (u_{o,i}^{j*} - res_i^j)^2 \tag{14}$$

Noted that $u_{o,i}^{j*}$ is equivalent to the predicted uncertainty, $\widehat{\sigma_i^j}$.

4.4 Optimization

The integrated loss of our credible spatiotemporal learning are threefold, context-conditioned spatiotemporal learning, context-target learning, and uncertainty quantification. Then it is formulated as,

$$\begin{aligned} Loss =& \text{MAPE}_f(\mathbf{Y}, \widehat{\mathbf{Y_f}}) + \beta \text{MAPE}_c(\mathbf{Y}, \widehat{\mathbf{Y_c}}) + \gamma \cos(\widehat{\mathbf{Y_c}}, \widehat{\mathbf{Y_f}}) \\ &+ \lambda Loss_{unc} \end{aligned} \tag{15}$$

In particular, we consider ST forecasting as the main task, and (β, γ, λ) are parameters balancing importances among context-target highway, cosine similarity and uncertainty learning.

4.5 Discussions of CreST

Summary and distinctions. CreST is a hybrid network to cooperatively perform both context and uncertainty-aware learning. Our CreST is capable of dynamically relating region-wise proximity based on the context embedding, and further derives the aleatoric and epistemic uncertainty with contexts and model discrepancy. **Model efficiency.** Our CreST introduces two additional costs compared with traditional GNN. (1) To realize the context-wise interaction extractions, we devise a context-wise learning weight. Let $\widetilde{d_m}$ be the expected dimension of all M-type context embedding where $\widetilde{d_m}, M \ll N$, the additional costs can be $O(M\widetilde{d_m^2})$ that is ignorable to overall architecture. (2) To achieve message passing of uncertainty, we perform one-layer GNN and the additional costs are linear to number of edges of $\mathbf{A}_{dist}$, which will not bring much burden to overall learning. **Uncertainty learning increases performance.** Since less residual indicates higher accuracy, minimizing the regressed residual can in fact bring in performance gains, and alleviate model variations, thus improving prediction credibility.

5 EXPERIMENTS

5.1 Dataset description

We collect diverse datasets of traffics including Suzhou Industry Park Surveillance, NYC taxi trips [1] and highway loop detectors of Metr-LA [2]. The statistical descriptions of datasets are figured in Table 1. In particular, we collect three types of context factors, i.e., spatial context of randomly initialized location descriptions, temporal contexts of day of week and timestamps, weather contexts including weather categories, precipitation and wind speeds [3].

5.2 Implementation details

5.2.1 Experimental setting. For each dataset, we divide well-organized samples into 60%, 30% and 10% for training, testing and validation. All methods are implemented in Tensorflow 1.14.0 or Pytorch 1.10.0, on one Tesla v100. The categorical context is encoded with one-hot embedding and transferred into fixed-length vectors. In the implementation, we feed each type of contexts into our network to capture context-wise interactions. We adjust hyperparameters of all baselines to adapt to these datasets, and feed the raw context embedding into them if they have corresponding placeholders. For our hyperparameters, we stack 3 GCN layers on SIP and Metr-LA, and 2 GCN layers on NYC, and set 1 LSTM layer across all datasets. We instantiate the hyperparameter of task-wise weight as $\beta = 0.3, \gamma = 0.3, \lambda = 1.0$ on SIP and Metr-LA, $\beta = 0.2, \gamma = 0.2, \lambda = 1.0$ on NYC. The uncertainty aggregation weight is set as $\alpha = 0.5$ across all datasets for simplification. We apply our model to predict the 6 future points given past 12 points.

5.2.2 Evaluation metrics. Given region v_i and interval t, we denote the predicted point estimation $\widehat{Y_i^t}$, predicted uncertainty $\widehat{\sigma_i^t}$, and groundtruth Y_i^t. For ST task, we employ RMSE and MAPE as metrics. For evaluation of UQ task, we borrow the prediction interval coverage probability (PICP) [38], and introduce the metric of uncertainty percentage (UP), to evaluate whether the uncertainty can accurately capture the groundtruth and whether the predicted uncertainty is rational to represent uncertainty, where $UP = (\sum_i \sum_t \widehat{\sigma_i^t} / Y_i^t)/(N \times l)$.

5.3 Competitors

ST learning. (1) Traffic transformer: A variant of Transformer and captures temporal continuity, periodicity as well as the spatial dependency [41]. **(2) STG2Seq:** A hierarchical graph convolution to capture spatial and temporal dependencies for passenger demand forecasting [3]. **(3) MDL :** A collective human mobility forecasting method which simultaneously models nodes and edges in a multi-task scheme [18]. **(4) Graph-WaveNet:** An improved version of DCRNN [12] by constructing learnable dynamic region-wise proximity and replacing the GRU with dilated convolutions [42]. **Uncertainty quantification competitors.** We evaluate the uncertainty learning by plugging our credible spatiotemporal learning framework with existing uncertainty quantification methods. **(1) Dropout-based BNN:** Realize this BNN method with dropout [43].

[1] https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
[2] https://github.com/liyaguang/DCRNN
[3] Collected from API: https://api.weather.com

Table 2: Performance comparisons on three datasets

	Methods	SIP			NYC			Metr-LA		
		MAPE	RMSE	UP	MAPE	RMSE	UP	MAPE	RMSE	UP
Baseline for spatiotemporal forecasting	Traffic transformer	23.29%	163.55	-	54.67%	76.64	-	9.71%	1.333	-
	STG2Seq	31.66%	184.23	-	17.61%	22.16	-	22.20%	3.873	-
	MDL	34.45%	192.96	-	18.62%	24.98	-	36.67%	4.552	-
	Graph-WaveNet	45.91%	218.11	-	27.39%	36.66	-	12.06%	2.399	-
	CreST	19.05%	144.32	-	12.59%	22.07	-	10.63%	2.125	-
	Methods	MAPE	PICP	UP	MAPE	PICP	UP	MAPE	PICP	UP
Baseline for uncertainty learning	SDE	26.14%	60.96%	0.587	11.87%	67.97%	0.558	12.15%	79.13%	0.895
	DeepEnsembles	25.80%	69.72%	0.381	23.84%	74.25%	2.710	24.22%	77.60%	2.937
	Dropout BNN	35.83%	61.86%	0.419	26.51%	74.41%	0.450	28.15%	75.44%	0.749
CreST(Ours)		**19.05%**	**84.74%**	**0.294**	**12.59%**	**85.37%**	**0.419**	**10.63%**	**86.57%**	**0.379**

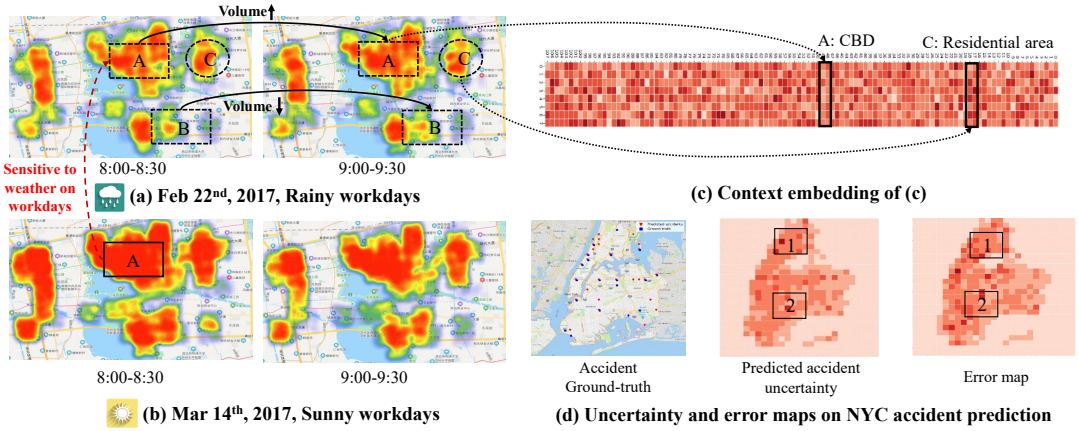


Figure 3: Deployed system and visualized case studies

Table 3: Performances on ablative spatiotemporal learning

Variants	MAPE		
	SIP	NYC	Metr-LA
CreST-FW	28.07%	25.96%	26.39%
CreST-CG	21.95%	18.42%	23.93%
CreST-CT	23.54%	20.20%	16.78%
Integrated CreST	**19.05%**	**12.59%**	**10.63%**

Table 4: Performances with uncertainty samples removed

Quantile	SIP	NYC	Metr-LA
	MAPE/PICP	MAPE/PICP	MAPE/PICP
10%	18.35%/85.07%	13.88%/86.37%	10.20%/87.88%
20%	17.21%/86.59%	13.02%/87.42%	9.31%/88.54%
30%	15.44%87.39%	12.47%/88.75%	8.43%/90.12%

(2) DeepEnsembles: Train a series of neural networks with different initializations [24][4]. **(3) SDE:** An uncertainty learning model with injections of noise and out of distribution samples [39].

[4]The number of ensembled networks is set as 5, according to [24].

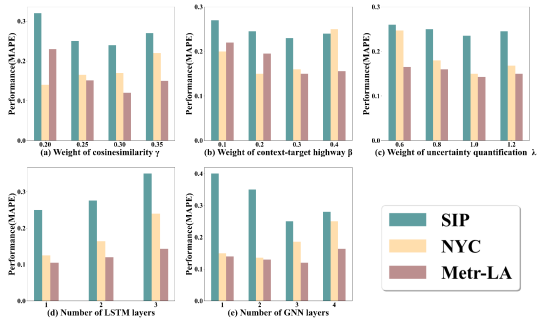


Figure 4: Performance on different hyperparameter settings

5.4 Spatiotemporal forecasting evaluation

5.4.1 Effectiveness of spatiotemporal forecasting. The results are reported in the upper half of Table 2. CreST outperforms the best competitor by 18.20% (transformer), 28.50% (STG2Seq) on SIP and NYC, and achieves comparable performance with traffic transformer on Metr-LA. Specifically, Traffic transformer achieves the best performance among baselines on SIP and Metr-LA, but still fails to model the taxicab trip records on NYC, probably due to the larger fluctuations and some specific patterns of NYC. In contrast, STG2Seq,

Table 5: Performances on different neural dimensions

		LSTM			GNN		
SIP	Dim	80	**96**	108	32	**64**	96
	MAPE	0.30	**0.25**	0.38	0.25	**0.23**	0.32
NYC	Dim	**56**	60	64	**2**	4	8
	MAPE	**0.13**	0.26	0.26	**0.13**	0.16	0.24
Metr -LA	Dim	160	**180**	192	16	32	**64**
	MAPE	0.12	**0.11**	0.14	0.18	0.16	**0.11**

tailored for taxi demand prediction, obtains the best results on NYC. STG2Seq and traffic transformer incorporate the timestamp and weather contexts into frameworks, but they cannot capture factor-wise interactions and fail to guide directional aggregations.

5.4.2 Ablative study for spatiotemporal forecasting. The ablative variants are as below: **(1) CreST-FW:** Learn context embedding with concatenations of context vectors rather than context-wise interactions. **(2) CreST-CG:** Replace the context conditioned dynamic topology with distance-based adjacency. **(3) CreST-CT:** Remove context-target highway, remain GNN learning. Tables 3 lists the performances of ablative variants. As shown, context-cross embedding is the most effective module in our framework (gain performance of 32.13%, 43.79%, 52.52% on three datasets). The context-target highway leads to a little promotion as it can be viewed as the regularization overcoming the overfitting issue.

5.5 Uncertainty quantification evaluation

5.5.1 Comparison of prediction intervals. The results are listed in the bottom half of Table 2. Our uncertainty-based prediction intervals capture the most groundtruth and achieve the least UP for uncertainty indicator on all datasets. For SDE, it is a relatively robust uncertainty learning scheme and surpasses other baselines on forecasting metrics. This is because that it can be viewed as a denoise AutoEncoder where pure and noise-incorporated observations can be trained alternately. Even though, all baseline methods are inferior to ours on two aspects. First, they provide collective and statistical value based sampling techniques, which are context-agnostic and cannot be learned individually. Thus they tend to overestimate the variations on some results and fail to provide context-specific uncertainty. e.g., DeepEnsembles have 2.710 and 2.937 UPs on NYC and Metr-LA. Second, these methods cannot internalize both data and model dependency into uncertainty indicators, and they are not tailored for spatiotemporal learning.

5.5.2 Quality of uncertainty learning. To investigate the quality of learned uncertainty, we remove the prediction results with high uncertainty and re-compute the performance accuracy without re-training the model. We remove top-10%, 20%, 30% predicted uncertainty samples on each dataset and the results are shown in Table 4. As expected, ruling out highly uncertain predictions could improve the overall accuracy, verifying that our uncertainty learning is of high quality and benefits selecting crucial samples.

5.6 Case study

In this section, we retrieve a series of intermediate and predicted results to answer two questions. (Q1) Can CreST exactly capture the context interactions and guide the spatiotemporal aggregation? (Q2) How our uncertainty quantification benefit urban management in this credible transportation system?

For Q1, we illustrate the predicted traffic volumes and context embedding of SIP, in Figure 3(a)-(c). Overall, traffic volumes on sunny days reveal more regular and active transitions than on rainy days. And volumes of CBD experience a drop at 9:00 a.m. on sunny days while they show an increase at the same period on rainy days, this is because employees tend to put off their working time during rainy days. These observations confirm the intuitions of context interaction-induced dynamic correlations. Regarding context embeddings, embeddings around CBDs are with smaller values while embeddings at residential areas are with larger values, delivering that volumes at residential areas are more sensitive to context at that step. This is consistent with the fact that on rainy days, officers must commute for working as usual while freelance workers can individually plan their urban travelling.

For Q2, as accidents are highly uncertain and sporadic, we especially incorporate an accident dataset in NYC [5] to perform the uncertainty-aware accident forecasting. The selected prediction uncertainty and error maps are illustrated in Figure 3(d). We observe that: i) The maps of uncertainty and errors share spatial similarities in both Region 1 and 2. ii) For Region 1, Bronx District, the combined context tripe (nightclubs, nights, rainy) contributes to high uncertainty of accidents due to ambiguous unseen risk factors and few existing context-accident correlations, corresponding to data-side aleatoric and model-side epistemic uncertainty. In fact, the inferred uncertainty, indicating prediction quality, can provide the basis for transferring risk alerts to human inspection in emergency.

5.7 Hyperparameter study

The hyperparameters are five-fold here, i.e., the number and hidden dimension of GCN layers, the number and hidden dimension of LSTM layers, and the β, γ, λ to balance losses across context-target highway, cosine similarity and UQ learning. The dynamic evolution of performances is illustrated in Table 5 and Figure 4. Finally, we can achieve the optimized hyperparameter settings in Sec. 5.2.1. For efficiency, it takes an average of 1.0 seconds to do one round of forecasting, sufficiently satisfying real-time credible forecasting.

6 CONCLUSION

In this paper, we shed light on coupling context factors with main learning streams to realize credible traffic predictions. To internalize context-wise interactions into inter-region correlations, we devise a context-cross relational embedding and a context-target information highway to achieve semantic context representation. We then generate virtual region-wise proximity based on semantic contexts for spatiotemporal aggregation. To decouple two uncertainties, we attribute data-side uncertainty to context factors and take model discrepancy as model-side uncertainty. Further, the residual is regressed and enables real-time uncertainty inference, alleviating model variation. Experiments demonstrate our CreST can boost prediction performance and provide credibility.

[5] https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes

ACKNOWLEDGMENTS

This paper is partially supported by the National Natural Science Foundation of China (No.62072427, No.12227901), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR-005), Academic Leaders Cultivation Program, USTC, the National Key Research and Development Program of China (No.2022YFB4500300), and the Key Research Project of Zhejiang Lab (No.2022PI0AC01).

REFERENCES

[1] T. Li, J. Zhang, K. Bao, Y. Liang, Y. Li, and Y. Zheng, "Autost: Efficient neural architecture search for spatio-temporal prediction," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 794–802.

[2] J. Liu, L. Deng, H. Miao, Y. Zhao, and K. Zheng, "Task assignment with federated preference learning in spatial crowdsourcing," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 1279–1288.

[3] L. Bai, L. Yao, S. S. Kanhere, X. Wang, and Q. Z. Sheng, "Stg2seq: spatial-temporal graph to sequence model for multi-step passenger demand forecasting," in *28th International Joint Conference on Artificial Intelligence, IJCAI 2019*. International Joint Conferences on Artificial Intelligence, 2019, pp. 1981–1987.

[4] 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.

[5] 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, 2020.

[6] M. Li and Z. Zhu, "Spatial-temporal fusion graph neural networks for traffic flow forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, 2021, pp. 4189–4196.

[7] H. Miao, J. Shen, J. Cao, J. Xia, and S. Wang, "Mba-stnet: Bayes-enhanced discriminative multi-task learning for flow prediction," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[8] R. Barnes, S. Buthpitiya, J. Cook, A. Fabrikant, A. Tomkins, and F. Xu, "Bustr: Predicting bus travel times from real-time traffic," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3243–3251.

[9] S. Wang, J. Cao, and S. Y. Philip, "Deep learning for spatio-temporal data mining: A survey," *IEEE transactions on knowledge and data engineering*, vol. 34, no. 8, pp. 3681–3700, 2020.

[10] D. Zhang, T. He, S. Lin, S. Munir, and J. A. Stankovic, "Online cruising mile reduction in large-scale taxicab networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 11, pp. 3122–3135, 2014.

[11] Z. Yuan, X. Zhou, and T. Yang, "Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 984–992.

[12] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *International Conference on Learning Representations*, 2018.

[13] Y. Zhao, J. Qi, Q. Liu, and R. Zhang, "Wgcn: Graph convolutional networks with weighted structural features," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 624–633.

[14] S. Wang, M. Zhang, H. Miao, Z. Peng, and P. S. Yu, "Multivariate correlation-aware spatio-temporal graph convolutional networks for multi-scale traffic prediction," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 3, pp. 1–22, 2022.

[15] W. Chen, L. Chen, Y. Xie, W. Cao, Y. Gao, and X. Feng, "Multi-range attentive bicomponent graph convolutional network for traffic forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3529–3536.

[16] C. Song, Y. Lin, S. Guo, and H. Wan, "Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 914–921.

[17] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[18] J. Zhang, Y. Zheng, J. Sun, and D. Qi, "Flow prediction in spatio-temporal networks based on multitask deep learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 3, pp. 468–478, 2019.

[19] Q. Zhaowei, L. Haitao, L. Zhihui, and Z. Tao, "Short-term traffic flow forecasting method with mb-lstm hybrid network," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[20] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 33–47, 2020.

[21] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in neural information processing systems*, 2017, pp. 5574–5584.

[22] D. Wu, L. Gao, X. Xiong, and M. Chinazzi, "Quantifying uncertainty in deep spatiotemporal forecasting," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2021.

[23] Y. Liu, H. Qin, Z. Zhang, S. Pei, Z. Jiang, Z. Feng, and J. Zhou, "Probabilistic spatiotemporal wind speed forecasting based on a variational bayesian deep learning model," *Applied Energy*, vol. 260, p. 114259, 2020.

[24] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in neural information processing systems*, 2017, pp. 6402–6413.

[25] Z. Zhou, Y. Wang, X. Xie, L. Qiao, and Y. Li, "Stuanet: Understanding uncertainty in spatiotemporal collective human mobility," in *Proceedings of the Web Conference 2021*, 2021, pp. 1868–1879.

[26] P. L. McDermott and C. K. Wikle, "Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data," *Entropy*, vol. 21, no. 2, p. 184, 2019.

[27] J. Caldeira and B. Nord, "Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms," *Machine Learning: Science and Technology*, vol. 2, no. 1, p. 015002, 2020.

[28] W. A. Bomberger, "Disagreement as a measure of uncertainty," *Journal of Money, Credit and Banking*, vol. 28, no. 3, pp. 381–392, 1996.

[29] Q. Chen, X. Song, Z. Fan, T. Xia, H. Yamada, and R. Shibasaki, "A context-aware nonnegative matrix factorization framework for traffic accident risk estimation via heterogeneous data," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2018, pp. 346–351.

[30] Anderson and K. Tessa, "Kernel density estimation and k-means clustering to profile road accident hotspots," *Accident Analysis & Prevention*, vol. 41, no. 3, pp. 359–364, 2009.

[31] G. P. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[32] C. Chiarella, X.-Z. He, and C. Hommes, "A dynamic analysis of moving average rules," *Journal of Economic Dynamics and Control*, vol. 30, no. 9-10, pp. 1729–1753, 2006.

[33] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 922–929.

[34] S. Guo, Y. Lin, H. Wan, X. Li, and G. Cong, "Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[35] J. Ye, L. Sun, B. Du, Y. Fu, X. Tong, and H. Xiong, "Co-prediction of multiple transportation demands based on deep spatio-temporal neural network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 305–313.

[36] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?" *Structural safety*, vol. 31, no. 2, pp. 105–112, 2009.

[37] T. Vandal, E. Kodra, J. Dy, S. Ganguly, R. Nemani, and A. R. Ganguly, "Quantifying uncertainty in discrete-continuous and skewed data with bayesian deep learning," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2377–2386.

[38] B. Wang, J. Lu, Z. Yan, H. Luo, T. Li, Y. Zheng, and G. Zhang, "Deep uncertainty quantification: A machine learning approach for weather forecasting," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2087–2095.

[39] L. Kong, J. Sun, and C. Zhang, "Sde-net: Equipping deep neural networks with uncertainty estimates," *arXiv preprint arXiv:2008.10546*, 2020.

[40] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European semantic web conference*. Springer, 2018, pp. 593–607.

[41] L. Cai, K. Janowicz, G. Mai, B. Yan, and R. Zhu, "Traffic transformer: Capturing the continuity and periodicity of time series for traffic forecasting," *Transactions in GIS*, vol. 24, no. 3, pp. 736–755, 2020.

[42] S. Shleifer, C. McCreery, and V. Chitters, "Incrementally improving graph wavenet performance on traffic prediction," *arXiv preprint arXiv:1912.07390*, 2019.

[43] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

7 ETHICAL ISSUE

Given that we only extract the mobility intensity for research, there is none identity information regarding human or specific persons. Therefore, we have no ethical issues to declare in this work.