Condition-Guided Urban Traffic Co-Prediction with Multiple Sparse Surveillance Data

Binwu Wang, Pengkun Wang*, Yudong Zhang, Xu Wang, Zhengyang Zhou and Yang Wang*, *Senior Member, IEEE*

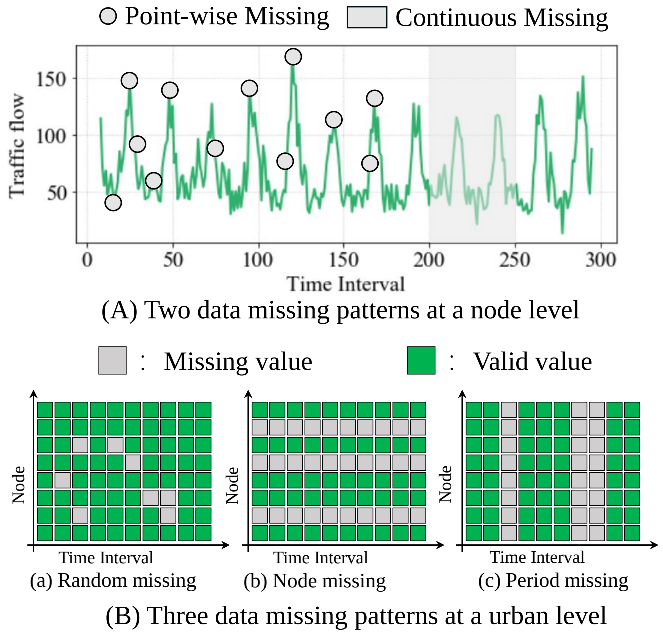
***Abstract*—Traffic prediction is one of the important research directions in Intelligent Transportation Systems, with positive implications for vehicle dispatching and vehicle management. In reality, due to the unreliability of data transmission and the volatility of storage devices, data sparsity limits the stability and prediction performance of existing methods that rely on high-quality observed data. To achieve robust sparse traffic prediction, we first investigate two findings as our motivation: the influence of external geographical features on shaping traffic distribution and the coupling dependence of multi-source urban data. Specifically, we develop a condition-guided collaborative learning network for traffic prediction with sparse data. The core idea is to exploit both the informative external geographic features and multiple urban data as auxiliary information to cooperatively learn mobility patterns from sparse data. First, we design an attention-based bilateral filter, which explicitly models the influence of external geographic features on spatial-temporal targets, and exploits such patterns as conditions to further estimate the missing elements. Secondly, a collaborative-learning framework, including a graph fusion module and a memory-preserved mechanism is devised to adaptively extract and aggregate fragments of similar spatial-temporal sequences from multiple urban data, helping the model to learn comprehensive mobility patterns from sparse data. We verify the excellent effectiveness of our model on multi-source traffic datasets collected in modern urban transportation systems.**

***Index Terms*—Vehicle data, public transportation, sparse surveillance, traffic prediction.**

I. INTRODUCTION

In recent years, emerging technologies (e.g 5G technology [1], [26], [15], [27], [48] and global position system (GPS)) have provided new empowerment for the transportation systems [6], [16], [19], [38], [7], [8], in particular, the prosperity of the vehicle networking has enabled the interconnection of vehicles, which can provide real-time ITS data covering the entire road network. This has promoted the development of the field of traffic prediction, which is intended to estimate the future traffic conditions (e.g., traffic flow or taxi demand) of an entire city, which can benefit many downstream vehicle applications, such as vehicle scheduling, vehicle management, and autonomous driving [23].

Recently, researchers have committed to the development of traffic prediction models based on deep learning [4], [29]. Existing models focus on analyzing human mobility patterns from historical traffic data, and then the learned patterns are used to predict future traffic conditions. While significantly achieving impressive success, they all rely on large amounts of high-quality and intensive historical traffic data to achieve a comprehensive understanding of traffic patterns and dynamics.



(A) Two data missing patterns at a node level

(B) Three data missing patterns at a urban level

Fig. 1. Different missing pattern scenarios at different levels.

However, in practical applications, the sparsity of traffic data is indeed an inevitable challenge. Existing studies typically classify the missing data into two main categories: random missing and continuous missing (as shown in Figure.1.)Random missing occurs when point-wise data points are lost randomly, which may result from unstable network connections and transmission packet losses. Continuous missing happens when continuous or correlated data are missing for a certain period, which can be further categorized into two distinct types at the urban level: period missing and node missing [47]. Period missing refers to a continuous absence of data for an entire city over a long period, and this results from storage volatility, data corruption, and power failure. Node missing (as shown in Figure.1 (B)) refers to the gaps in historical data of some regions or sections of urban, and this could be due to equipment failure or insufficient coverage of monitoring devices. In fact, the high deployment overheads

Dr. Pengkun Wang and Prof. Yang Wang are the corresponding authors.

Binwu Wang, Yudong Zhang, Xu Wang, Pengkun Wang, Zhengyang Zhou, and Yang Wang are with University of Science and Technology of China, Hefei 230022, China (e-mail: wbw1995@mail.ustc.edu.cn; zyd2020@mail.ustc.edu.cn; wx309@mail.ustc.edu.cn; pengkun@mail.ustc.edu.cn; zzy0929@mail.ustc.edu.cn; angyan@ustc.edu.cn).

of urban surveillance equipment hinder the dense coverage of monitoring sensors. For example, in highly-developed cities like Suzhou Industrial Park (SIP) and Shenzhen, only a small percentage (e.g., 3.0% and 3.2% respectively) of intersections are equipped with stationary surveillance cameras. Moreover, the rising requirements for privacy protection impose stricter limitations on the utilization of personal privacy-related data. As a result, the availability of certain types of data that could contribute to urban data analysis may be further limited in the future. This increased emphasis on privacy protection may exacerbate the severity of the sparse data issues in urban traffic analysis.

When existing traffic state prediction models deal with sparse data, they usually use local statistical features (such as mean or zero values) for linearly imputing missing values, which is a naive approach. Because as the sparsity of the data increases, these local statistics become more and more unreliable and even noisy. As the sparsity of the data increases, the reliability of these local statistics diminishes, leading to inaccurate and noisy imputations. Especially for continuous missing scenarios, these preprocessed data would provide spurious mobility patterns. In addition, the dilemma of zero-inflation [50] or breakdown [39] can significantly impact the ability of models to comprehensively capture mobility patterns from sparse data, resulting in instability and decreased prediction accuracy. Thus, how to achieve accurate traffic prediction with sparse data is an open question.

There are some overlooked observations and enlightenment that are beneficial to solving the dilemma of sparse data. First, external geographical features have a crucial influence in shaping traffic data, and regions with similar external geographical features (such as POI, road network structure, etc.) show similar traffic patterns. For instance, as illustrated in Figure 2 A), i.e., heatmap of taxi flow at different POIs from 7:00 a.m. to 5:00 p.m., taxi flow is relatively stronger in residential areas since people need to leave residential areas to regional areas during the morning, and while the situation during the afternoon is completely contrary to what we see during the morning. Based on this finding, it becomes possible to estimate the traffic state of a missing region by perceiving regions that share similar external features. This approach goes beyond relying solely on local statistics and instead utilizes the knowledge gained from regions with comparable features. By incorporating this knowledge, the model's robustness is enhanced, leading to more reliable estimations of the traffic state, even in situations where specific local data is unavailable or incomplete.

Secondly, there is coupling dependency among multi-source urban data due to the interaction of multimodal transportation. In fact, these data fundamentally reflect people's mobility patterns [42], [50]. For example, as can be observed in Figure.2 A), both taxi flow and sharing-bike flow exhibit similar layout-driven patterns in the residential areas during morning and afternoon traffic rush hours. Moreover, Figure.2 B) and C) illustrate the striking similarity in volume distributions across various event categories, including taxi flow, bike flow, and traffic accident numbers, across different days of the week and months of the year. Additionally, Figure 2 D) provides evidence of long-term temporal similarities. By considering the correlations among multiple urban data sources, we can integrate them to capture comprehensive patterns and enhance the learning effect of sparse data. Leveraging the complementary information from different data sources allows for an improved understanding of traffic dynamics by the model. This integration enables more robust and accurate predictions of traffic flow, even in scenarios where data sparsity presents a challenge.

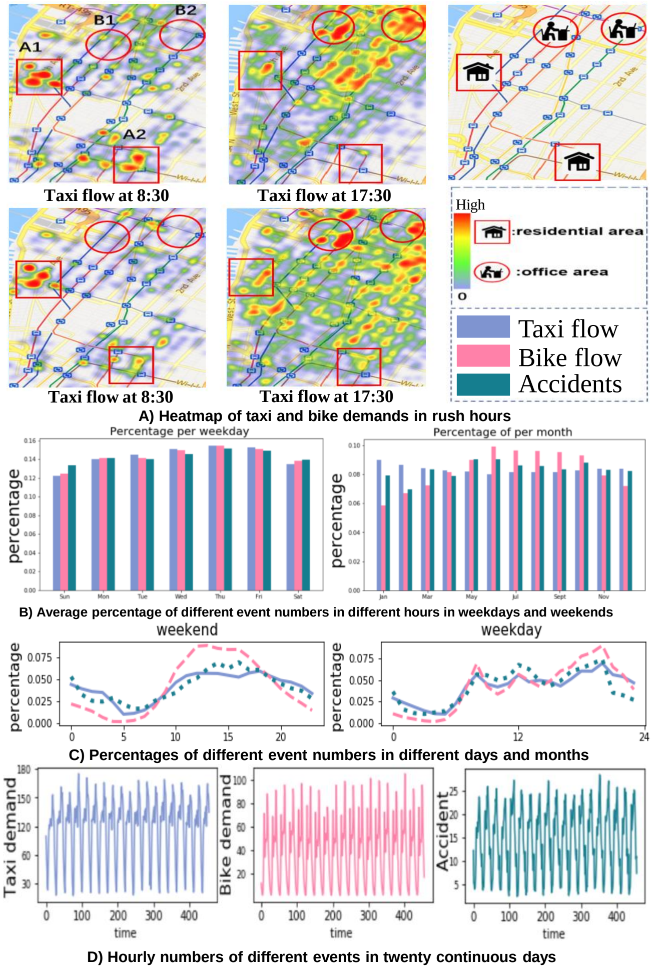

Taxi flow at 8:30
Taxi flow at 17:30
Taxi flow at 8:30
Taxi flow at 17:30
A) Heatmap of taxi and bike demands in rush hours
B) Average percentage of different event numbers in different hours in weekdays and weekends
C) Percentages of different event numbers in different days and months
D) Hourly numbers of different events in twenty continuous days

Fig. 2. Demonstration of spatial-temporal correlations among multiple categories of urban data: Subfigure A) demonstrates the spatial distribution similarities among the flow of taxi and sharing-bike during rush hours, B) illustrates the average daily or monthly percentages that different categories of events account for in a week or year, C) shows the average hourly percentages that different categories of events account for in a day, and D) the hourly numbers of different categories of events in about 20 continuous days.

Based on these two findings, in this paper, we attempt to exploit both the informative external geographic features and correlations among multiple urban data to cooperatively complement traffic patterns from sparse data. Specifically, we propose a Condition-guided Spatial-Temporal Network (CSTN) for traffic prediction with only sparse data, which explicitly models the influence of external geographic features on traffic distribution to estimate missing traffic patterns by examining regions with similar external geographic features instead of local statistical features. Moreover, we transfer the

learned knowledge from multi-source urban data to complement the comprehensive patterns. This framework consists of two phases:
i) Spatial-temporal learning on the single sparse data: We formalize urban traffic as a graph structure, and then design a condition-guided spatial-temporal network to learn spatial-temporal correlations from sparse data, which combines the advantages of the graph convolutional neural network and Transformer. Specifically, we first design an attention-based bilateral filtering to explicitly extract how external geographic features affect urban flow and take these results as conditional guidance to generate the graph edge weights of the graph attention network (GAT). Based on such imaginative use of external geographic features, those inaccessible or missing data points can be dynamically estimated based on the similarities of external geographic features among data points rather than empirically assumed. It is worth mentioning that even if the corresponding data is not missing, such external geographic features can also be used as supplements to further enhance spatial-temporal representations, thus effectively benefiting the extraction of mobility patterns from single sparse data from both sides. Simultaneously, we especially equip Transformer with causal convolutions to enable multiple receptive fields and eventually capture multi-scale temporal trends from historical data.
ii) Collaborative learning on multiple-source sparse data: Considering that multiple-source urban data is closely correlated in both spatial and temporal perspectives, we design a collaborative learning framework to complement missing patterns by leveraging the power of different attention mechanisms. In particular, our collaborative learning internalizes the attention mechanism respectively into a graph fusion scheme and memory-preserved mechanism, to achieve information synergy in both spatial and temporal perspectives. For the graph fusion scheme, we implement adaptive spatial aggregations on node levels with regard to multiple-source urban data, ie., cross-domain elements. Meanwhile, from the temporal perspective, given that different traffic elements share similar long-term traffic patterns, we design an attention-based memory-preserved mechanism to extract similar long-term patterns to enhance the spatial-temporal representations of sparse data.

The main contributions of this paper are as follows:

- To the best of our knowledge, this paper is the first one targeting the issue of traffic learning with only sparse data, and this is also an effective attempt that deep learning based data analysis is liberated from the dependence on a long-term accumulation and intensive collection of data.
- We propose a condition-guided spatial-temporal model (CSTN) for traffic prediction with sparse data, and this model exploits both informative external geographic features and correlations among multiple-source urban data to cooperatively complement traffic patterns among sparsely observed traffic datasets.
- CSTN integrates an attention-based bilateral filter, which is designed to learn the influence patterns of external geographic features on traffic targets and is further exploited to estimate the missing elements. Second, a collaborative learning method seamlessly incorporates a node-level attention-based graph fusion and a memory-preserved mechanism to respectively allow adaptive node-level element aggregations and long-term fragment combinations from multiple-source data, enabling the information fusion in spatial-temporal perspectives and finally achieving our co-predictions.
- Extensive experiments on real-world datasets demonstrate that our CTSN outperforms alternative baselines in the traffic prediction task.

The remainder of this paper is organized as follows. We introduce the existing studies on traffic learning and review methodology limitations in Section II. Then Section III identifies some key definitions and the task we focus on. Next, we describe the details of our proposed model condition-guided collaborative prediction network (CSTN) in Section IV. Section V uses real-world traffic datasets to evaluate the proposed model, which mainly includes two parts: the accuracy of traffic prediction and the contribution measure of each component. Next, we provide the limitations of the paper and the potential avenues for future research in Section VI. Finally, we make a conclusion for this paper in Section VII.

II. Related Work

Traffic prediction is an important task for ITS and benefits advanced applications, such as Internet of Things (IoT) applications [45], [13], [24], [36], [31], traffic management [32] and autonomous driving [38], [37]. Traffic prediction aims to analyze traffic patterns from observed traffic data and predict future traffic conditions based on the learned patterns [41], [43].

A. Traffic prediction

The early algorithms mainly include mathematical or statistical analytical methods, such as time series models, the autoregressive integrated moving average model (ARIMA), and the Kalman filtering model. However, these algorithms are inefficient because they fail to model complex nonlinear relationships.

Recently, inspired by deep learning techniques in computer vision and natural language understanding, researchers have moved to study models based on neural networks, and these models are generally composed of spatial components and temporal components to model spatial and temporal correlations, respectively. For example, ST-ResNet [46] and DMVST [41] use Convolutional Neural Networks (CNN) to capture spatial correlations among regions or nodes, and they also integrate Long Short-Term Memory (LSTM) networks to learn dynamic temporal trends. H-CNN [14] develops a hexagon-based convolutional neural network (HCNN) as a spatial component. However, CNNs fail to process non-European data (i.e., graph-structured data), and traffic data can be naturally formed as graph-structured data. Thus, the current most popular models are based on Graph Convolutional Networks (GCNs). For example, ASTGCCN [44] integrates Graph Attention Convolutional Network to capture dynamic spatial correlations among urban regions or nodes. STSGCN [25] designs synchronization-GCN to capture temporal and spatial correlations simultaneously. To capture complex temporal correlations from historical data, the transformer has been widely used as a temporal component due to its powerful ability to model long-term dependencies [16], [35].

After achieving promising results, it is evident that these models heavily depend on the availability of abundant high-quality data to accurately capture mobility patterns. However, data sparsity is an inevitable challenge arising from communication failures, fluctuations in storage facilities, damage to monitoring equipment, and various other factors. As a consequence, the sparse nature of the collected traffic data presents significant hurdles for these models in learning comprehensive mobility patterns. The presence of zero inflation and sparse learning dilemmas further exacerbates the issue, leading to a noticeable decline in their performance.

B. Traffic prediction with sparse data

Recent studies have emerged that specifically address traffic prediction using sparse data. These works [12], [3] aim to leverage the knowledge acquired from data-rich cities, where extensive and comprehensive data collection is in place, to enhance the learning capabilities of models for data-sparse target cities. For example, MetaST [39] designs a meta-learning traffic framework to transfer the traffic pattern knowledge of other similar cities which are intensively and integrally monitored to predict the future traffic state of the target city. AreaTransfer [33] selects the appropriate source city from multi-source candidate cities and establishes area-matching relationships between the target city and source cities. This facilitates the transfer of relevant knowledge for traffic prediction. ST-GFSL [18] presents a method that enables multi-level knowledge transfer by matching parameters associated with similar traffic meta-knowledge. This approach allows for enhanced prediction performance by leveraging the shared characteristics and patterns between different cities.

However, these methods still rely on at least one instance of intensive data collection, the ideal scenario of having comprehensive data in every target city is often challenging to achieve. Moreover, the collection of appropriate auxiliary data from other cities introduces additional complexities related to data privacy and access, particularly when dealing with multiple municipalities. These factors may limit the widespread deployment of these methods.

III. PPRELIMINARIES

In this section, we explicitly explain the explanatory variables and formally define the traffic prediction problem using multi-source sparse data studied in this paper.

Definition 1 (Spatial Region): We partition an area of interest (e.g., a city) evenly into $N = H \times W$ disjoint geographical grids, in which each grid is considered as a spatial region $r_n\,(1 \leq n \leq N)$. And $\mathcal{V} = \{r_1, \cdot, r_N\}$ is used to denote the spatial region set in a city.

Definition 2 (Urban traffic graph): In this paper, we define the traffic of a urban as an undirect graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V}$ means the set of nodes and $\mathcal{E}$ indicate the set of edges with $(|\mathcal{V}| = H \times W)$ nodes, and $v_i \in \mathcal{V}(1 \leq i \leq |\mathcal{V}|)$ corresponds to the i-th node in $\mathcal{V}$, and $e_{ij} \in \mathcal{E}$ indicates that there is a direct edge within node v_i and v_j. The corresponding adjacency matrix $\mathcal{A} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ can be denoted as where

$$a_{ij} = \begin{cases} 1 \quad \text{iff} \quad e_{ij} \in \mathcal{E} \\ \\ 0 \quad \text{otherwise} \end{cases} \tag{1}$$

Definition 3 (Sparse traffic data): Traffic data is the traffic status (e.g., traffic flow) recorded by devices deployed on roads (e.g., vehicle detectors). We use $\mathcal{X} \in \mathbb{R}^{T \times |\mathcal{V}| \times D} = \{X_1 \cdots X_T\}$ to denote observed traffic data over previous T time steps, and $X_t = \{x_{t,1}, \cdots, x_{t,|V|}\}$ means the data collected from the t-th time-step of all nodes, where $x_{t,v} \in \mathbb{R}^D$ is the data of v-th node at t-th time-step with D traffic features (e.g., traffic outflow or inflow). Due to the reasons mentioned above, collected traffic data is sparse, thus, we use a mask tensor $\mathcal{M} \in \mathbb{R}^{T \times |\mathcal{V}| \times D} = \{M_1 \cdots M_{|\mathcal{V}|}\}$ to indicate the presence of the data point. If a data point $x_{i,t}$ is missing, the corresponding $m_{i,t}$ is equal to 0, else $m_{i,t}$ is equal to 1.

Definition 4 (Multi-source sparse urban data): The urban multiple sparse data can be defined as a tensor $\mathbb{X} = \{\mathcal{X}^1 \cdots \mathcal{X}^C\}$ where $\mathcal{X}^i (1 \leq i \leq C)$ indicates the i-th category of collected sparse urban data. Accordingly, the mask tensor $\mathbb{M} = \{\mathcal{M}^1, \cdots, \mathcal{M}^C\}$ is used to denote the the presence of $\mathbb{X}$. In this paper, we introude the sharing-bike data $\mathcal{X}^B$ and the traffic accident data $\mathcal{X}^A$ as auxiliary, i.e., $\mathbb{X} = \{\mathcal{X}^\mathcal{T}, \mathcal{X}^\mathcal{B}, \mathcal{X}^\mathcal{A}\}$, where $\mathcal{X}^T$ means the traffic data.

Definition 5 (Traffic prediction with multi-source sparse data): Given the urban traffic network $\mathcal{G}$ and the corresponding urban multiple sparse data $\mathbb{X}$ with the masking matrix $\mathbb{M}$ during the T historical time spots, we aims to train a prediction model $\mathcal{F}$ which can effectually process sparse traffic data and accurately predict the traffic state at the next T_p time steps.

$$\mathcal{F} : \left(\mathcal{X}^\mathcal{T} | \mathbb{X}, \mathcal{G}\right) \rightarrow \left[X^\mathcal{T}_{T+1}, \cdots, X^\mathcal{T}_{T+T_P}\right] \tag{2}$$

IV. METHOD

In this section, we will elaborate on the details of the model CSTN. The model contains two phases: i) Learning on single category of sparse data, ii) Collaborative-interactive learning on multiple sparse datasets. For more clarity, Figure 3 shows details of each block.

A. Learning on single category of sparse data

Modeling spatial-temporal correlations from historical data is critical for traffic flow forecasting. The traffic conditions of different locations influence each other and the mutual influence is highly dynamic. Hence, Graph Attention Network (GAT), which uses an attention mechanism to adaptively capture dynamic correlations between nodes in the spatial dimension, has attracted more attention[10]. However, when the data is sparse, the attention mechanism becomes inefficient due to the inaccessibility of some nodes' historical data. To reverse the inefficiency, the attention-based bilateral filtering mechanism is proposed to calculate the similarity in terms

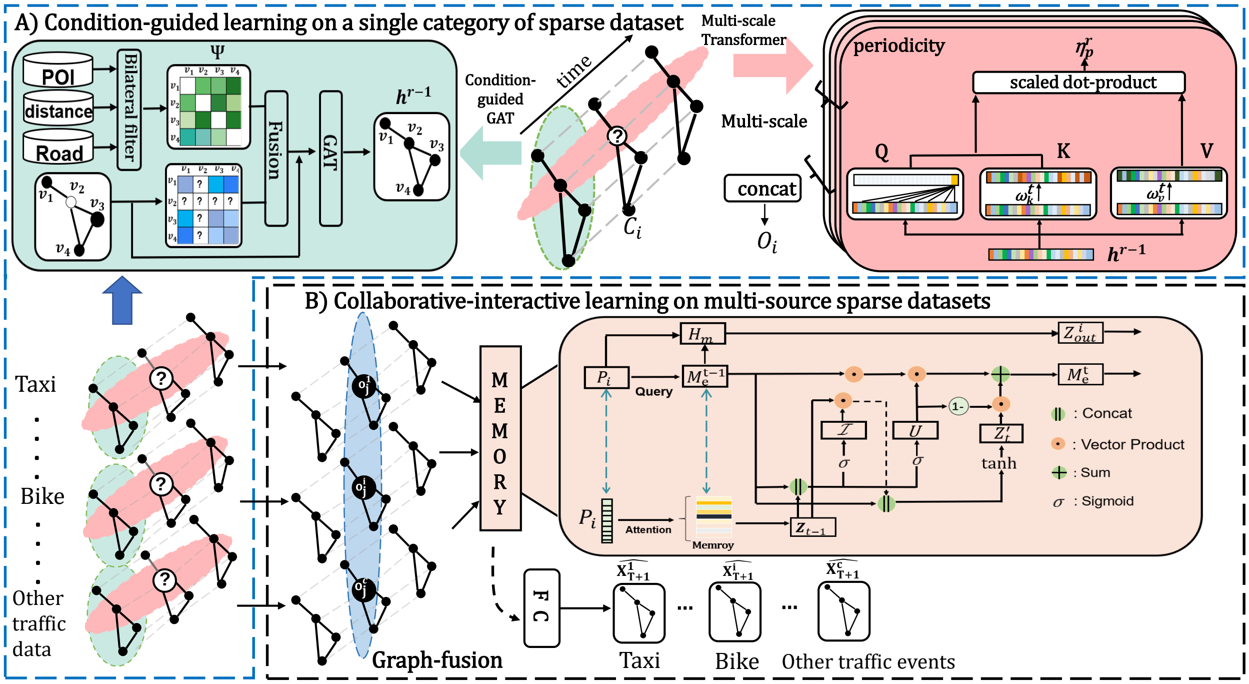


Fig. 3. **Solution overview:** Subfigure A) and B) respectively correspond to two different phases of CSTN, i.e., condition-guided learning on single category of sparse data (within the top-left red dashed wireframe) and collaborative-interactive learning on multiple sparse datasets (with the bottom-right black wireframe).

of geo-context features. The result will be as a condition to generate edge weights in GAT to guide the model to learn spatial correlations. Meanwhile, we integrate causal convolution and transformer and design a multi-scale transformer to mine temporal correlation from different temporal granularity.

Note that in this phase, to simplify the expression, we select a single category of sparse data $\mathcal{X}^c$ as input. It is simply expressed as $\mathcal{X} = \{X_1 \cdots X_{|\mathcal{V}|}\} \in \mathbb{R}^{|\mathcal{V}| \times T \times D}$. The detailed design of this phase is shown in Subfigure A) of Figure 3.

External factor learning. The conventional practice of learning on the sparse dataset is to use local statistics (e.g. zero values or the last observation) to directly complement missing values. Honestly, such methods have certain effects to some degree. However, in case the sparsity of the dataset increases to a threshold or there exist continuous missing values, these methods become unreliable, and it is almost impossible for the model to extract the complete spatial dependencies from sparse datasets. Based on our analysis, i.e., there exist similar traffic patterns among nodes with similar external geographic features, regarding a missing value of a specific node, an intuitive approach involves simulating this value by referring to nodes with comparable external geographic features. The bilateral filter is particularly adept at addressing this task. Widely employed in the realm of image processing, the fundamental concept underlying the bilateral filter revolves around employing a weighted average of neighboring pixel values to substitute the value of a given pixel [11]. We introduce it to learn the node with missing value by accommodating the similarities of geographic-context features among different nodes. To adaptively quantify the influencing weights of different nodes with similar geographic-context features on this specific node, it uses an attention-based bilateral filtering mechanism where a score function is designed to accommodate the inter-node similarities in terms of geographic-context features, i.e.,

$$e_{(i,j)} = \exp\left(-\frac{1}{2}\left(\frac{\phi\left(i,j\right)}{\sigma_r}\right)^2\right)\exp\left(-\frac{1}{2}\left(\frac{d\left(i,j\right)}{\sigma_d}\right)^2\right) \tag{3}$$

where σ_r and σ_d are the trainable parameters of bilateral filtering. $\phi\left(i,j\right)$ is the Pearson correlation coefficient of geographic features(e.g. POI and road network structure) of node i and j, $d\left(i,j\right)$ means the Euclidean distance between these two nodes. Worth noting that the bilateral filtering based parameters are also trainable parameters. So far, the static similarity between nodes i and j in terms of their geographic-context features can be obtained by:

$$\psi_{i,j} = \frac{e_{(i,j)}}{\sum_{j\in\mathcal{V}} e_{(i,j)}} \tag{4}$$

and $\Psi = \{\psi_{ij}\} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ indicates the entire similarity matrix.

Condition-guided GAT. GAT has been proved its superiority in processing traffic data. Attention scores is calculated by the spatial-temporal features of neighboring nodes in road network:

$$\begin{cases} \beta_{i,j} = \mathrm{FC}(\left[\omega_1^s h_i^{r-1} \,||\, \omega_2^s h_j^{r-1}\right]) \\ s_{i,j} = \frac{\exp(\mathrm{LeakyRelu}(\beta_{i,j}))}{\sum_{k\in\mathrm{Nei}(i)}\exp(\mathrm{LeakyRelu}(\beta_{i,k}))} \end{cases} \tag{5}$$

where ω_1^s and ω_2^s are learnable parameters. h_i^{r-1} means the hidden state of node v_i in the (r-1)-th layer and h_i^0=X_i. The value of an element $s_{i,j}$ in attention matrix $\mathbf{S} = \{s_{ij}\} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ semantically represents the similarity of traffic information between node i and node j. However, in case that data is sparse, the calculated attention scores are correspondingly sparse. It's the obstacle of message delivery. As discussed earlier, similar external features lead to similar traffic patterns. Therefore, we consider the external factor relation matrix Ψ as conditions to guide aggregation. First, two matries are adaptively fused to obtain enhanced similarity matrix:

$$\phi_{i,j} = \omega_3^s s_{i,j} + \omega_4^s \psi_{i,j} \tag{6}$$

where ω_3^s and ω_4^s are the learnable parameters. $\Phi = \{\phi_{ij}\} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is reinforced because it is calculated from the spatial-temporal features and external features instead of just relying on the spatial-temporal features. Then we aggregate node features from their neighbors:

$$h_i^r = \mathrm{ELU}\left(\sum_{j\in \mathrm{Nei}(i)} \phi_{i,j}(\omega_2^s h_j^{r-1})\right) \tag{7}$$

where ELU$(\cdot)$ known as Exponential Linear Unit [5] is activation function.

Multi-scale Transformer. Recently, Transformer has proven a strong ability to model time trends, which consists of multi-head attention layers, shared feed-forward neural layers, and batch normalization layers between them. To deal with the dilemmas caused by data sparsity, we redesign the multi-head attention layer, and the other layers remain the same as the regular transformer [28]. Specifically, for node v_i, we take the output h_i^r from the condition-guided GAT layer as input to show the learning process.

Traffic flow is believed to be multi-periodic. Researchers [46], [41] tend to categorize them into closeness patterns, period patterns, and trend patterns according to temporal frequencies. Thus, the fusion of multiple views with different time resolutions is beneficial to learning robust representations of variation patterns from sparse data. Regular transformers can't capture multi-resolution temporal dependencies, because it takes the same strategy (i.e. linear projection) to obtain query vectors (Q), key vectors (K), and value vectors (V). We introduce one-dimensional dilated convolution (1D-CNN) to replace the linear projection layer. In this way, the multi-head attention layer can configure different receptive fields. Specifically, 1D-CNN denoted as DC$(\cdot)$ is used to obtain query vectors Q:

$$Q = \mathrm{DC}(h_i^r, \rho) \tag{8}$$

$$\text{where} \qquad \mathrm{DC}(\eta, \rho) = \sum_{i=0}^{L-1} \mathrm{Conv}\,(i) \bullet \eta_{T-di} \tag{9}$$

where ρ means the dilation factor. L indicates the kernel size of convolutions, opertion $\bullet$ corresponds to the dot product operation. η_{T-di} indicates the some layer's hidden state of the entire graph in time point $T - di$. We then linearly project the h_i^r into key vectors K and value vectors V:

$$K = \omega^k h_i^r, V = \omega^v h_i^r \tag{10}$$

where ω^k and ω^v are learnable parameters. $K \in \mathbb{R}^{T\times d_k}$ and $V \in \mathbb{R}^{T\times d_v}$ are key vector and value vector respectively. Then we calculate the attention scores for the queries of all positions by dot product operation of K vector and Q vector, and a self-attention layer can be denoted as:

$$\text{Attention}\,(h_i^r, \rho) = \mathrm{softmax}\left(\frac{Q(K^{\mathrm{T}})}{\sqrt{d_k}}\right) V \tag{11}$$

The multi-head attention mechanism is preferred to enhance the presentation capabilities of the model. However, the problem of information redundancy between multiple heads [20] could weaken this advantage. As mentioned above, the collaboration of multiple views with different time resolutions is benefit, thus we suggest that different heads could be equipped with different receptive fields (i.e., different ρ) to learn different temporal trends. The result of the multi-head attention is the concatenation of the output of each attention function. The learning process can be formulated as:

$$\begin{aligned} \mathrm{O}_i &= Multi - head\,(h_i^r)\,; \\ &= \text{ Concat } (\text{ head }_1, \ldots \text{ head}_{N_S})\,\omega^P; \\ \text{where head }_s &= \text{ Attention } (h_i^r, \rho_s) \end{aligned} \tag{12}$$

where ω^P is learnable parameters of linear projection to adaptively coordinate the attention heads. N_S is the number of heads and different heads. We expect the dilation factor ρ is not equal in different heads. Note that in the first phase: Learning on single category of sparse data, we employ consolidated but independent models on different categories of sparse datasets separately. This means that they do not share learning parameters of models. So we obtain the outputs sequence of all categories of datasets, it is expressed as $\mathbb{O} = \left\{\mathbf{O}^1, \cdots, \mathbf{O}^C\right\} \in \mathbb{R}^{C\times|\mathcal{V}|\times T\times d_o}$ where $\mathbf{O}^c = \left\{O_1, \cdots, O_{|\mathcal{V}|}\right\}$.

B. Collaborative-learning on multi-source sparse data

As analyzed above, there are correlations between multiple traffic data. We design a graph fusion mechanism and memory-preserved mechanism for collaborative learning of multiple datasets to further estimate missing patterns and enhance spatial-temporal representations.

Attention-based graph fusion.

For a node (region), there are close relationships between multiple traffic events. Inspired by the heterogeneous graph representation learning task, we can regard different types of spatial-temporal features as the independent attribute channels of each node. Thus, we can achieve preliminary collaborative spatial learning from the node perspective by interacting with these channels of each node.

Thus, the graph fusion mechanism based on the attention mechanism is proposed. Specifically, considering the output of the first phase $\mathbb{O}$, it can be reshape as $\overline{\mathbb{O}}$=$\{o_1, \cdots, o_{|\mathcal{V}|}\}$ where $o_i \in \mathbb{R}^{C\times(T\times d_o)}$. For node v_i, we first use a nonlinear transformation to embed its corresponding output after phase one, i.e.,

$$\mu_i = \mathrm{Tanh}\,(\omega_i^\mu o_i + b_i^\mu) \tag{13}$$

where ω_i^μ and b_i^μ are learnable parameters. And then we then normalize learned semantic correlations by

$$\Omega_i^{c,n} = \frac{\exp((\mu_i^c)^T \mu_i^n)}{\sum_{k=1}^{C}\exp((\mu_i^c)^T \mu_i^k)} \tag{14}$$

where $\Omega_i^{c,n}$ represents the correlations between node v_i in the c-th category of data and node v_i in the n-th category of data. μ_i^c represents the c-th row of μ_i. Then we achieve aggregation:

$$\sigma_i^c = \sum_{n=1}^{C} \Omega_i^{c,n} o_i^n \tag{15}$$

where $\sigma_i^c \in \mathbb{R}^{T \times d_p}$, d_p is the dimension of feature. Thus, for node v_i, we obtain the output sequence $P_i = \left\{\sigma_i^1, \cdots, \sigma_i^C\right\} \in \mathbb{R}^{C \times T \times d_p}$.

Memory-preserved mechanism. Long-term traffic patterns (e.g., periodic patterns) are critical for traffic forecast models. As discussed above, these patterns of multi-traffic events may be similar (as shown in Figure.2 D), which can be regarded as a global property. To find these shared long-term pattern fragments, we design the memory-preserved mechanism to explicitly learn and store them, and then we achieve further collaborative interactive learning between multiple sparse datasets by dynamically interacting the long-term shared spatial-temporal information in the memory to improve the effect of collaborative learning.

Specifically, the memory is a parameterized matrix and denoted as $M_e \in \mathbb{R}^{N_m \times d_m}$, where N_m is a hyperparameter and means the number of stored patterns in memory. We use M_e^{t-1} to denote the state of the memory unit during the $(t-1)$ time step. For the hidden features P_i of node v_i from the graph fusion part, we map it to a high-dimensional space to get the query vector Q_m, and then Q_m is used to access memory to obtain hidden feature vector H_m.

$$H_m^i = \operatorname{softmax}\left(\frac{Q_m (M_e^{t-1})^T}{\sqrt{d_m}}\right) M_e^{t-1} \tag{16}$$

then we use the residual term to integrate H_m^i into P_i for enhancing spatial-temporal representation. It can be formulated as Z_{out}^i=$H_m + P_i$, for all nodes, the final output is denoted as $\mathcal{Z}_{out}$=$\{Z_{out}^1, ..., Z_{out}^{|V|}\}$.

Next we use the gate mechanism inspired by Gate Recurrent Unit (GRU) to filter the noise information and update patterns in the memory. Specifically, we employ the self-attention mechanism to obtain hidden feature vector Z_{t-1}, i.e.,

$$\begin{aligned} Z_{t-1} &= \operatorname{softmax}\left(\frac{M_e^{t-1}(Q_m)^T}{\sqrt{d_m}}\right) V_m \\ \text{where} \quad Q_m &= \omega_k^\pi P_i, \quad V_m = \omega_v^\pi P_i \end{aligned} \tag{17}$$

where ω_q^π, ω_k^π, and ω_q^π are learnable parameters. Then we use Z_{t-1} to renew the memory unit M_e^{t-1} as follow:

$$\begin{cases} \mathcal{I} = \operatorname{Sigmoid}\left(\omega_z^{\mathcal{I}} Z_{t-1} + \omega_\pi^{\mathcal{I}} M_e^{t-1} + b^{\mathcal{I}}\right) \\ U = \operatorname{Sigmoid}\left(\omega_z^U Z_{t-1} + \omega_\pi^U M_e^{t-1} + b^U\right) \\ Z_t' = \operatorname{Tanh}(\omega_{z'}\left[\mathcal{I} \odot Z_{t-1} \,||\, M_e^{t-1}\right] + b_{z'}) \\ M_e^t = U \odot M_e^{t-1} + (1-U) \odot Z_t' \end{cases} \tag{18}$$

where $\mathcal{I}$ is the input gate, U indicates the update gate which controls the information update of the memory component. $\omega_z^{\mathcal{I}}$, $\omega_\pi^{\mathcal{I}}$, ω_z^U, ω_π^U, $\omega_{z'}$, $b^{\mathcal{I}}$, b^U, and $b_{z'}$ are learnable parameters. So far the state of the memory is updated to t time step.

C. Multi-task learning

In order to improve the effect of collaborative learning of sparse datasets, we make predictions for each traffic event, and use the loss sum of multi tasks to train the model. Specifically, we first feed $\mathcal{Z}_{out}$ into FC layers, and obtain the final output of the entire two-phase attention-based network, i.e.,

$$\widehat{\mathbb{Y}} = \left\{\widehat{\mathcal{Y}^1}, \cdots, \widehat{\mathcal{Y}^C}\right\} = \operatorname{FCs}\left(\mathcal{Z}_{out}\right) \tag{19}$$

Where $\widehat{\mathcal{Y}^c}$ $(1 \le c \le C)$represents the predicted value of the c-th urban data. Then we calculate the prediction loss of each task. Considering the various sparsity rates of multiple datasets, we design loss function to dynamically adjust the loss with the sparsities of multiple datasets:

$$\text{Loss} = \sum_{c=1}^{C} \frac{\lambda_c}{(1-\vartheta^c)^2} \left\| \left(\mathcal{Y}^c - \widehat{\mathcal{Y}^c}\right) \odot \mathcal{M}^c \right\|_2 \tag{20}$$

where $\mathcal{Y}^c$ means the ground-truth value. Notice here we use the mask tensor $\mathbb{M}$ to make sure that the calculated loss is to optimize the network only based on the observed data points in multiple sparse datasets. ϑ^c represents the sparsity which can be calculated by the ratio of the number of data points value equal to zero to the total number of all data points in the c-th category of urban data.

V. Experiments

In this section, we evaluate the validity of our proposed model through a series of experiments and address the following concerns:

Q.1: What is the performance for the traffic prediction with sparse data of the proposed model? Please refer to Subsection V-D.

Q.2: Can the model effectively handle various sparse data scenarios? Please refer to Subsection V-E.

Q.3: Can the results confirm the positive effects of these two findings in this paper? Please refer to Subsection V-F.

Q.4: Does each component of the proposed model contribute to the performance? Please refer to Subsection V-G.

Q.5: What is the sensitivity of hyperparameter K? Please refer to Subsection V-H.

A. Experiment setting

Datasets. As an important branch of traffic data, taxi flow has been widely studied by researchers. In the experiment, following the previous work, we select taxi flow as a monitoring indicator of traffic status and evaluate our method on public real-world datasets of NYC [1]. The taxi flow data is recorded as taxi GPS information from April 1st to October 1st, 2018 (183 days in total). Moreover, we use sharing-bike data and traffic accident data as auxiliary. Our goal is to predict future taxi flow by collaborative learning on three datasets and external information. The descriptions of the three datasets are as follows:

(1) NYC taxi. The dataset of NYC taxi consists of more than 3.5 million taxicab trip records including recording time, longitude and latitude.

(2) NYC Bike. The dataset of NYC Bike includes about 3 million transaction records where each record contains the information of sharing-bike order time and location.

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

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2024.3397716

TABLE I
THE PREDICTION PERFORMANCE OF THE MODELS.

Dataset	Metric	ARIMA	SVR	LSTM	STGCN	DMVST-Net	STDN	GMAN	ST-MGCN	Ours
RM-20	MAE	6.81	6.72	6.06	5.04	5.67	5.21	4.94	4.82	**4.60**
	RMSE	11.32	11.40	10.21	8.35	9.10	8.65	8.33	7.89	**7.62**
	MAPE(%)	16.73	14.59	13.80	11.28	12.23	11.67	11.06	10.62	**9.90**
RM-40	MAE	10.23	8.11	7.79	5.92	6.46	6.21	5.57	5.39	**4.98**
	RMSE	17.07	12.98	11.79	9.81	10.34	10.70	9.56	9.23	**8.37**
	MAPE(%)	19.08	17.37	16.78	12.67	15.22	13.89	12.93	12.39	**11.01**
RM-60	MAE	16.31	11.44	11.93	7.67	8.96	8.43	6.32	6.12	**5.66**
	RMSE	26.58	18.05	19.53	11.24	13.25	13.81	10.65	10.49	**9.86**
	MAPE(%)	25.32	20.21	21.12	15.81	18.12	17.24	14.65	14.26	**12.68**
NM-40	MAE	10.68	8.23	8.96	5.87	6.54	6.42	5.73	5.51	**5.15**
	RMSE	17.72	13.91	13.55	10.12	10.75	9.99	9.43	9.65	**8.79**
	MAPE(%)	21.61	19.35	18.81	16.45	15.89	14.21	12.84	12.89	**11.58**
PM-40	MAE	10.81	8.44	8.13	5.99	6.81	6.31	5.73	5.47	**5.03**
	RMSE	14.12	14.50	13.12	11.51	11.24	9.87	9.36	9.12	**8.04**
	MAPE(%)	23.64	22.33	20.32	17.20	16.67	14.01	12.77	13.01	**10.34**

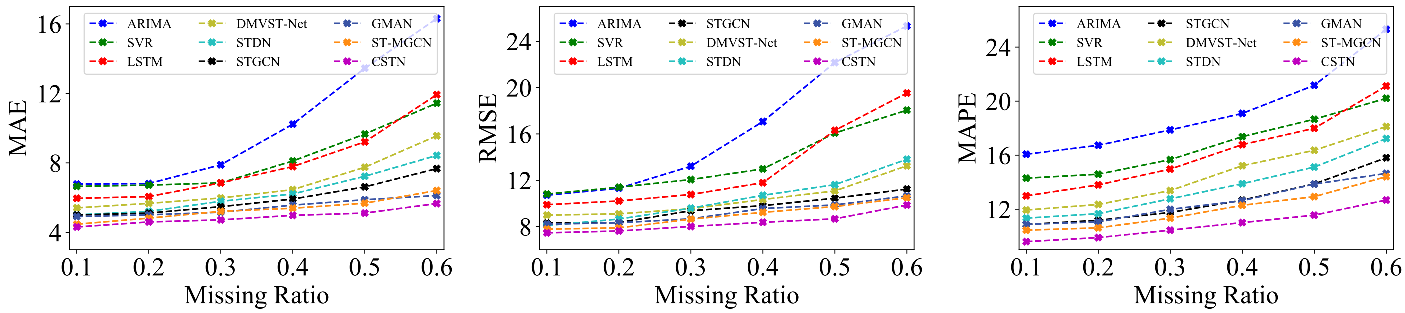


Fig. 4. The prediction performance of the models with different missing ratio of data.

(3) NYC Accident. The dataset of NYC Accident consists of more than 90k accident records which include the information of accident location, time, and etc.

(4) External factors. The two types of external geographic features in our experiments are explained in detail below, POI and road segments. Regarding POI, it includes seven possible options, residence, school, culture facility, recreation, social service, transportation, or commercial. The road segment data includes road length, width and type.

Experiments settings. We partition all data in three pieces in temporal perspective with the ratio of 7:2:1 respectively for training, testing, and validation. Meanwhile, the whole New York City is divided into 5*15 grids. We regard all grids as graph nodes by employing the method proposed in [30], [22]. The length of time interval is set to 1 hour [2]. We used one-hot encoding to transform discrete features (e.g., POI and the type of the road). The training phase is performed using the Adam optimizer with learning rate 10^{-4} and batch size is 32. The early-stop strategy is used. Regarding the predictions on both sharing-bike and taxicab datasets, we focus on the demands of bicycles and taxi, therefore the feature dimensionality D is set to 1. In the multi-head attention module, we achieve the fusion of four heads (i.e. N_S=4). Our model is implemented with PyTorch 1.9 in Python, and executed with Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz and Nvidia Tesla V100 16GB. Metrics. Classic metrics including Mean Absolute Error (MAE), Rooted Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) are used to evaluate the performances of the models for the demand prediction of taxi.

B. Sparse data

As motivated by our paper, we focus on traffic prediction without intensively monitored traffic data. Following previous general settings [17], [21], [47], we mask partially collected data in taxi/Bike datasets to generate sparse data in both spatial and temporal perspectives. For the NYC accident dataset, it is naturally sparse [50], [30].

We use the parameter ξ to measure the ratio of masked data points [3]. For instance, $\xi = 0.3$ means that 30% of all data points in all these datasets are masked. Considering real application scenarios, we formulated three masking rules to construct sparse dataset (as shown in Figure 5):

1. Random Mask strategy (RM). As shown in Figure 3(a), the missing data points are randomly scattered and completely

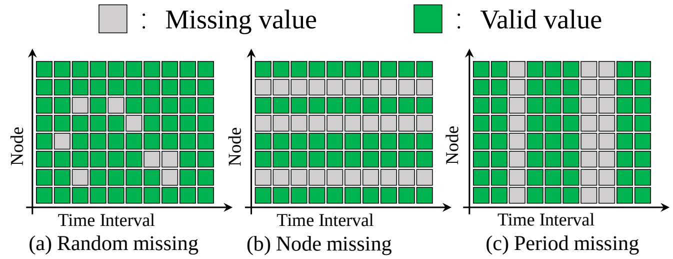
[2] The length of time interval should be set with considering the equilibrium within the prediction accuracy and temporal granularity of different datasets.

[3] Notice here the sparsity of all C categories of data $\left\{\xi^1 \cdots \xi^C\right\}$ is set to the same value, therefore we use a unified ξ to represent the sparsity of all categories of data.

independent. In experiments, we set ξ equal to 0.2, 0.4, and 0.6 respectively, and three datasets are denoted as RM-20, RM-40, and RM-60.

2. Node Mask strategy (NM). As shown in Figure 3(b), we first randomly select some nodes, and current input data of these needs will be masked. For this missing pattern strategy, it may be caused by the privacy policy of a certain area of the city or monitoring equipment that has not been deployed. In this setting, we select 40% of the nodes to mask, which is denoted as NM-40.

3. Period Mask strategy (PM). We mask data in consecutive time horizons (as shown in Figure 3(c)). In this setting, About 40% of flow data is masked, and this dataset is denoted as PM-40.



(a) Random missing (b) Node missing (c) Period missing

Fig. 5. Illustration of RM, NM, and PM scenarios. (a) Random elements of data are masked. (b) Current input data of the node is masked. (c) Consecutive data segments are masked over some time slots.

C. Baselines

We compare our model with the following alternative baseline models. For baselines, we impute missing values with 0 or the mean value, and the better prediction performance of the two methods are shown.

(1). Autoregressive Integrated Moving Average Model (ARIMA). It is an attempt to predict future traffic through autocorrelation and difference of data.

(2). Support Vector Regression (SVR) [2]. It is also a traditional time series data learning model by learning feature mapping functions.

(3). Long short-term memory (LSTM). An LSTM network has input, output, and forget gates, and can capture long-term temporal correlations more effectively than traditional RNNs.

(4). ST-MetaNet [46]. It utilize GNN and RNN to learn spatial-temporal patterns, which employs meta learning strategy to generate the parameters of GAT and RNN.

(5). DMVST-Net [41]. It is a multi-View spatial-temporal network for predicting traffic flows. And semantic information is used to model the spatial-temporal patterns of similar areas.

(6). STDN [40]. It designs flow gating mechanism and periodic shift attention mechanism to learn dynamic spatial-temporal dependencies.

(7). GMAN [49]. It is a graph attention model to capture spatial-temporal patterns for traffic prediction.

(8). ST-MGCN [9]. It develops a multi-graph convolution to explicitly model complex spatial correlation.

D. Experimental result analysis

Table.I and Figure.4 show the performance of different methods on four series of intensive experiments on different datasets, which includes three datasets randomly masked (RM) 20, 40, and 60 percent of data points. The NM-40 dataset and CM-40 dataset are constructed using the node masking (NM) strategy and the node masking (NM) strategy with the sparsity rate 0.4, respectively. Five datasets are respectively expressed as RM-20, RM-40, RM-60, NM-40 and PM-40.

The first three datasets are obtained according to the random masking strategy. ARIMA and SVR can only learn linear mappings from historical data and fail to capture complex spatial-temporal correlations, so they have higher errors than the models based on deep learning. LSTM has poor prediction performance for sparse data because it only captures temporal correlations and is sensitive to missing values. DMVST-Net, STGCN, and STDN have better prediction performance than LSTM because they can efficiently learn more spatial correlations. However, since DMVST-Net and STDN integrate CNN which is also sensitive to missing values when the data missing rate is large, they cannot capture complete traffic patterns from sparse data, and the prediction performance of the two models is not promising. STGCN achieves smaller errors than DMVST-Net and STDN because they use external geographic features (e.g. POI and road network information) as auxiliary information to help models learn complex spatial-temporal correlations from sparse data and enhance spatial-temporal representations. This is beneficial for the model to learn more patterns from sparse data. ST-MGCN is composed of multi-graph convolutional networks based on multiple graphs which are constructed according to external geographic features (i.e. POI information and geographic attributes). This avoids the learning dilemma of models relying only on sparsely observed data. When the missing rate of data becomes larger, the prediction errors of other baselines increase significantly. In contrast, CSTN achieves excellent prediction performance with sparse data. Because CSTN mainly utilizes external geographic features as auxiliary and collaborative learning of multi-related datasets, these two strategies effectively help the model complement missing patterns and improve the ability of the model to model data distributions.

In conclusion, CSTN outperforms state-of-art traffic prediction models in various data missing scenarios. This illustrates the feasibility of using multimodal transport interaction to improve learning performance.

E. Robustness Analysis of CSTN

Real-world applications may encounter various traffic data missing conditions. To understand the robustness of CSTN in dealing with complex data collection challenges, we evaluate the effectiveness of our model on the NM-40 dataset and PM-40 dataset. The NM-40 dataset is constructed by using a node mask strategy to mask the current input of 40% nodes. We mask continuous data segments over some time slots to construct PM-40 dataset. Compared with the element mask strategy, there are continuous missing values in the temporal and spatial dimensions, respectively. The results of experiments on two datasets are shown in Table.I.

For the NM-40 dataset, we observe that the models based on CNN (e.g DMVST-Net and STDN) generally perform worse than the models based on GNN (e.g GMAN and ST-MGCN), and this phenomenon is more obvious in the fourth group of experiments (on NM-40 dataset). On the one hand, GNN is more effective in processing traffic road graph data. On the other hand, GNNs have been proven to have strong inductive learning ability that can generalize messages to unknown nodes by the message passing mechanism [34]. The traffic states of masked nodes are directly related to neighbor nodes and evaluated by them. However, their prediction performance is still not as good as CSTN, because the external factor learning module of CSTN can guide the GAT to globally discover nodes that are more relevant to traffic patterns and capture comprehensive spatial dependencies. Simultaneously, CSTN integrates a multi-graph fusion mechanism, and the information of the missing node can be supplemented by other spatial-temporal event features of this node.

The prediction errors of LSTM increase significantly on the PM-40 dataset, because LSTM is sensitive to continuous missing values and only explores temporal correlations of traffic data. When the traffic data is missing continuously, the temporal information is relatively less preserved, and the spatial information for traffic forecasting is extremely important. The multi-scale Transformer of CSTN can help the model learn more temporal information about traffic data, and the memory mechanism which learns the sharing patterns among multi-source data provides long-term spatial-temporal features.

In conclusion, CSTN has good robustness and can effectively deal with various sparse data in complex scenarios.

F. Experimental analysis with the findings

The influence of external geographical features. We argue that nodes with similar external geographic factors may exhibit similar traffic distributions. In order to investigate the positive role of external geographic factor learning, we develop a variant, CSTN-EF, which means CSTN without External geographic Factors. The prediction performance of the two models is shown in the Table.II. We can observe that CSTN-EF can achieve lower errors than CSTN, especially on the NM-40 dataset, where the data of 40% nodes is missing. CSTN can fill in traffic patterns by detecting other nodes with similar geographic features, which is more robust than relying entirely on local statistics, such as averages or zero values. This also proves that our finding is beneficial for sparse data learning.

TABLE II
THE PREDICTION PERFORMANCE OF CSTN AND CSTN-EF.

Model	RM-20			NM-40		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
CSTN-EF	4.74	7.96	10.14	5.43	10.03	12.95
CSTN	**4.60**	**7.62**	**9.90**	**5.15**	**8.79**	**11.58**

The coupling dependency of multi-source urban data. We develop a variant CSTN-MUD which means CSTN without multi-source urban data (i.e., sharing bike dataset and traffic accident dataset). We show the prediction performance of two models at 8.am and 9.am, the results are shown in Table III. We find that CSTN achieved better predictive performance for the morning peak period. This proves that the correlation of multi-source urban datasets can improve the learning effect of sparse data in the model, especially in the peak period. For example, in residential areas, because people come to work areas with diverse transportation, the flow of taxis and shared bikes both increases sharply at 8 a.m, the model can make accurate predictions for traffic flow by analyzing the demand trend of sharing bikes.

TABLE III
THE PREDICTION PERFORMANCE OF CSTN AND CSTN-MUD.

Model	RM-20			RM-40		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
CSTN-MUD	9.74	13.98	22.15	12.31	17.59	25.08
CSTN	**9.46**	**10.34**	**20.04**	**10.17**	**15.46**	**22.15**

G. Ablation studies

In this subsection, We design some variants to conduct ablation experiments on the datasets RM-40, which randomly mask 40% of data observations, to illustrate the effectiveness of different components. Variants are described as follow:
(1). CSTN-EF: CSTN without external factors as auxiliary information, and we only rely on the observation data to generate edge weights in GAT.
(2). CSTN-CG: CSTN with regular GAT instead of condition-guided GAT, and we use a fully connected layer to embed external features instead of attention mechanism inspired by bilateral filtering. And we only rely on the observation data to generate edge weights in GAT in this variant.
(3). CSTN-MS: we use regular transformer to learn temporal correlations instead of multi-scale transformer.
(4). CSTN-GF: CSTN without graph fusion mechanism in the collaborative-interactive learning phase. The result of the first phase will be as the input of the memory to learn long-term patterns.
(5). CSTN-Me: CSTN without memory to store the long-term patterns.
(6). CSTN-GM: CSTN without both attention-based graph fusion mechanism and memory mechanism. We feed the outputs of the first stage directly to the fully connected layers to make predictions.
(7). CSTN: Our proposed framework uses multi-source urban datasets and external geographical features to support sparse traffic learning.

Figure.6 shows the performances of CSTN and all its variants. As illustrated, the performances of all variants are more or less not as competent as the performances of CSTN. This demonstrates the effectiveness of the individual components in CSTN. Worth noting that, the worse prediction performance

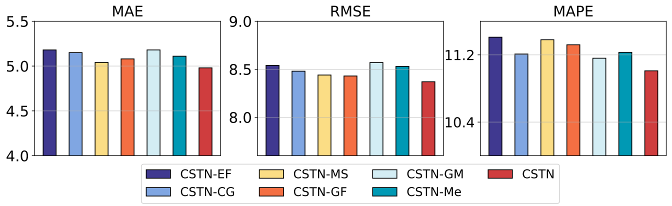


Fig. 6. Ablation studies of CSTN on RM-40.

of CSTN-EF indicates that external geographic features, as auxiliary factors, can assist the model in mining abundant patterns from sparse data, and improve the performance of the model. CSTN-CG which only uses FC layer to model external geographic features into the model achieves higher errors than CSTN, this suggests the necessity of explicitly modeling the impact of external geographic features on traffic flow, i.e. the effective attention mechanism inspired by the bilateral filtering mechanism. CSTN-GM which does not include multiple datasets collaborative learning phase is not as excellent as CSTN, this demonstrates that exploiting correlations across multiple related datasets can help models infer complete traffic patterns from sparse data. The prediction of CSTN is better than CSTN-Me without the memory to store similar pattern fragments, this demonstrates that explicitly preserving temporal patterns is beneficial for estimating missing traffic patterns and can enhance spatial-temporal representations.

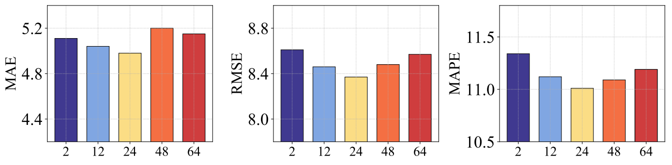


Fig. 7. The prediction performance of CSTN with different N_m on RM-40.

H. Hyperparameter Experiment (Q.5)

The number of patterns in the memory. We evaluate the effect of the hyperparameter N_m value for the prediction performance of the model, and the N_m value represents the number of stores and the number of patterns in the memory. We conducted experiments on the dataset RM-0.4. And the results are shown in Figure.7.

For RM-0.2 and dataset RM-0.4, we find that the performance increases in the beginning but decreases later, and when the N_m is equal to 24, the model has the lowest prediction errors. If the N_m is smaller than 24, the memory fails to provide enough information for the model. When N_m is larger than 24, it means that the memory needs to store too many patterns and can not focus on capturing long-term patterns. Simultaneously, we observe that with the increase of the data sparsity rate, a larger N_m is beneficial to improve the prediction performance of the models.

VI. Discussion and future work

In this section, we discuss some interesting issues, which can be our future research.

First, just as we expected, prediction performances are essentially enhanced by making full use of and transferring relevant knowledge among multi-related datasets. In fact, such a co-prediction framework can be widely extended to address additional spatial-temporal forecasting tasks in other fields and domains. For example, the task of air pollutant forecasting can benefit from the knowledge learned from meteorological data. Hence, in the future, we will explore more spatial-temporal tasks with our model to further verify its generalization ability and universality.

Second, we evaluate the model on three traffic datasets (NYC taxi, NYC Bike, and NYC accidents). However, these three kinds of data are essentially within the same modality. In fact, there are many urban-related datasets with heterogeneous modalities such as high-speed sensor data (e.g. PeMS dataset) and taxi order datasets. In the future, we will additionally focus on the extension of the model to support the collaborative learning of multiple modal data.

VII. Conclusion

In this paper, for the first time, we are concerned about a novel question in the field of deep learning, i.e., whether is intensive data essential for deep learning based models. We preliminarily discuss this issue by proposing a novel framework for traffic forecasting with only sparse data: Condition-guided Spatial-Temporal graph network (CSTN).

With New York City as a case study, we investigate two unexplored findings: (1). Geographic external geographic features, such as points of interest (POI) and road network structure, indeed play a significant role in shaping traffic behaviors. Nodes, such as road sections or regions, that share similar external geographical features tend to exhibit similar traffic patterns. For example, areas with popular tourist attractions, shopping centers, or business districts often experience higher traffic volumes and congestion during peak times. (2). Modern multimodal transportation leads to the coupling correlation of multi-source urban data. Different traffic events may have a similar distribution in specific areas. In addition, these data also show semblable temporal trends over a long time span.

Inspired by these findings, we integrate external geographic features and multi-source urban data to extract comprehensive patterns and enhance the learning effect of sparse data. Specifically, we first design an attention-based bilateral filter, which explicitly learns the influence patterns of external geographic features on spatial-temporal targets, and exploits such patterns as conditions to further estimate the missing elements. Secondly, to fuse multi-source traffic information, a collaborative learning framework which includes a graph fusion module and a memory-preserved mechanism is devised to adaptively detect and aggregate shared patterns from multiple traffic events, enhancing spatial-temporal representation and achieving co-predictions. With the help of external information and correlations between multiple urban events, comprehensive traffic patterns are eventually learned from sparse data, which can provide accurate insights into the future state of transportation.

We evaluated the validity of our model for traffic prediction with sparse data on multi-source urban datasets, which are collected from New York. The experimental results show that CSTN outperforms state-of-art traffic prediction models in various data missing scenarios and has up to 7.52% improvement in MAE, 9.41% improvement in RMSE, and 11.14% improvement in MAPE. We further performed ablation experiments to evaluate the contribution of each component.

VIII. Acknowledgement

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.

References

[1] Y. Cao, S. Xu, J. Liu, and N. Kato, "Toward smart and secure v2x communication in 5g and beyond: A uav-enabled aerial intelligent reflecting surface solution," *IEEE Vehicular Technology Magazine*, vol. 17, no. 1, pp. 66–73, 2022.

[2] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[3] K. Chen, J. Han, S. Feng, and H. Yang, "Cross-city traffic prediction via semantic-fused hierarchical graph transfer learning," *arXiv preprint arXiv:2302.11774*, 2023.

[4] 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.

[5] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[6] J. Dai, J. Liu, Y. Shi, S. Zhang, and J. Ma, "Analytical modeling of resource allocation in d2d overlaying multihop multichannel uplink cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 8, pp. 6633–6644, 2017.

[7] L. D'Acierno, M. Botte, A. Placido, C. Caropreso, and B. Montella, "Methodology for determining dwell times consistent with passenger flows in the case of metro services," *Urban Rail Transit*, vol. 3, pp. 73–89, 2017.

[8] M. Gallo, G. De Luca, L. D'Acierno, and M. Botte, "Artificial neural networks for forecasting passenger flows on metro lines," *Sensors*, vol. 19, no. 15, p. 3424, 2019.

[9] X. Geng, Y. Li, L. Wang, L. Zhang, Q. Yang, J. Ye, and Y. Liu, "Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3656–3663.

[10] 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.

[11] Q. Han, D. Lu, and R. Chen, "Fine-grained air quality inference via multi-channel attention model." in *IJCAI*, 2021, pp. 2512–2518.

[12] Y. Huang, X. Song, Y. Zhu, S. Zhang, and J. James, "Traffic prediction with transfer learning: A mutual information-based approach," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[13] J. Ji, J. Wang, C. Huang, J. Wu, B. Xu, Z. Wu, J. Zhang, and Y. Zheng, "Spatio-temporal self-supervised learning for traffic flow prediction," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 4, 2023, pp. 4356–4364.

[14] J. Ke, H. Yang, H. Zheng, X. Chen, Y. Jia, P. Gong, and J. Ye, "Hexagon-based convolutional neural network for supply-demand forecasting of ride-sourcing services," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4160–4173, 2018.

[15] T. Kitagawa, Y. Kawamoto, and N. Kato, "Communication scheduling with diversity for unmanned aircraft systems using local 5g," *Journal of Communications and Information Networks*, vol. 5, no. 1, pp. 50–61, 2020.

[16] H. Lin, R. Bai, W. Jia, X. Yang, and Y. You, "Preserving dynamic attention for long-term spatial-temporal prediction," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 36–46.

[17] Z. Liu, Y. Yang, W. Huang, Z. Tang, N. Li, and F. Wu, "How do your neighbors disclose your information: Social-aware time series imputation," in *The World Wide Web Conference*, 2019, pp. 1164–1174.

[18] B. Lu, X. Gan, W. Zhang, H. Yao, L. Fu, and X. Wang, "Spatio-temporal graph few-shot learning with cross-city knowledge transfer," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 1162–1172.

[19] B. Mao, F. Tang, Z. M. Fadlullah, and N. Kato, "An intelligent route computation approach based on real-time deep learning strategy for software defined communication systems," *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 3, pp. 1554–1565, 2019.

[20] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?" *arXiv preprint arXiv:1905.10650*, 2019.

[21] U. Mital, D. Dwivedi, J. B. Brown, B. Faybishenko, S. L. Painter, and C. I. Steefel, "Sequential imputation of missing spatio-temporal precipitation data using random forests," *Frontiers in Water*, vol. 2, p. 20, 2020.

[22] Z. Pan, Y. Liang, W. Wang, Y. Yu, Y. Zheng, and J. Zhang, "Urban traffic prediction from spatio-temporal data using deep meta learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1720–1730.

[23] Y. Qiao, Y. Cheng, J. Yang, J. Liu, and N. Kato, "A mobility analytical framework for big mobile data in densely populated area," *IEEE transactions on Vehicular Technology*, vol. 66, no. 2, pp. 1443–1455, 2016.

[24] T. K. Rodrigues, K. Suto, and N. Kato, "Edge cloud server deployment with transmission power control through machine learning for 6g internet of things," *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 4, pp. 2099–2108, 2019.

[25] 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.

[26] Y.-X. Sun, "Large frequency ratio antennas based on dual-function periodic slotted patch and its quasi-complementary structure for vehicular 5g communications," *IEEE Transactions on Vehicular Technology*, 2023.

[27] F. Tang, Y. Zhou, and N. Kato, "Deep reinforcement learning for dynamic uplink/downlink resource allocation in high mobility 5g hetnet," *IEEE Journal on selected areas in communications*, vol. 38, no. 12, pp. 2773–2782, 2020.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[29] B. Wang, X. Luo, F. Zhang, B. Yuan, A. L. Bertozzi, and P. J. Brantingham, "Graph-based deep modeling and real time forecasting of sparse spatio-temporal data," *arXiv preprint arXiv:1804.00684*, 2018.

[30] B. Wang, Y. Lin, S. Guo, and H. Wan, "Gsnet: Learning spatial-temporal correlations from geographical and semantic aspects for traffic accident risk forecasting," 2021.

[31] B. Wang, Y. Zhang, J. Shi, P. Wang, X. Wang, L. Bai, and Y. Wang, "Knowledge expansion and consolidation for continual traffic prediction with expanding graphs," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[32] P. Wang, C. Zhu, X. Wang, Z. Zhou, G. Wang, and Y. Wang, "Inferring intersection traffic patterns with sparse video surveillance information: An st-gan method," *IEEE Transactions on Vehicular Technology*, 2022.

[33] X. Wei, T. Guo, H. Yu, Z. Li, H. Guo, and X. Li, "Areatransfer: A cross-city crowd flow prediction framework based on transfer learning," in *International Conference on Smart Computing and Communication*. Springer, 2021, pp. 238–253.

[34] Y. Wu, D. Zhuang, A. Labbe, and L. Sun, "Inductive graph neural networks for spatiotemporal kriging," *arXiv preprint arXiv:2006.07527*, 2020.

[35] 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.

[36] S. Xu, J. Liu, Y. Cao, J. Li, and Y. Zhang, "Intelligent reflecting surface enabled secure cooperative transmission for satellite-terrestrial integrated networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 2, pp. 2007–2011, 2021.

[37] S. Xu, J. Liu, and J. Zhang, "Resisting undesired signal through irs-based backscatter communication system," *IEEE Communications Letters*, vol. 25, no. 8, pp. 2743–2747, 2021.

[38] Y. Xun, J. Qin, and J. Liu, “Deep learning enhanced driving behavior evaluation based on vehicle-edge-cloud architecture,” *IEEE Transactions on Vehicular Technology*, vol. 70, no. 6, pp. 6172–6177, 2021.

[39] H. Yao, Y. Liu, Y. Wei, X. Tang, and Z. Li, “Learning from multiple cities: A meta-learning approach for spatial-temporal prediction,” in *The World Wide Web Conference*, 2019, pp. 2181–2191.

[40] H. Yao, X. Tang, H. Wei, G. Zheng, and Z. Li, “Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 5668–5675.

[41] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li, “Deep multi-view spatial-temporal network for taxi demand prediction,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[42] 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.

[43] J. Ye, L. Sun, B. Du, Y. Fu, and H. Xiong, “Coupled layer-wise graph convolution for transportation demand prediction,” *arXiv preprint arXiv:2012.08080*, 2020.

[44] C. Zhang, J. James, and Y. Liu, “Spatial-temporal graph attention networks: A deep learning approach for traffic forecasting,” *IEEE Access*, vol. 7, pp. 166 246–166 256, 2019.

[45] H. Zhang, J. Liu, K. Li, H. Tan, and G. Wang, “Gait learning based authentication for intelligent things,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 4450–4459, 2020.

[46] 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.

[47] K. Zhang, F. Zhou, L. Wu, N. Xie, and Z. He, “Semantic understanding and prompt engineering for large-scale traffic data imputation,” *Information Fusion*, p. 102038, 2023.

[48] B. Zhao, X. Dong, G. Ren, and J. Liu, “Optimal user pairing and power allocation in 5g satellite random access networks,” *IEEE Transactions on Wireless Communications*, 2021.

[49] C. Zheng, X. Fan, C. Wang, and J. Qi, “Gman: A graph multi-attention network for traffic prediction,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 1234–1241.

[50] 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.

Binwu Wang is currently working toward the Ph.D. degree in the School of Data Science, University of Science and Technology of China (USTC). He has published several papers on top conferences and journals such as ICLR, AAAI, IJCAI, IEEE TITS, IEEE TMC, KDD, DASFAA, and WSDM. His main research interests include traffic data mining and continuous learning, especially their applications in urban computing.

Pengkun Wang (Graduate Student Member, IEEE) is now a Research Associate Professor at the University of Science and Technology of China (USTC). He got his Ph.D. degree at USTC in 2023, under the supervision of Professor Qi Liu and Yang Wang. His research interest mainly includes generalized machine learning, spatio-temporal data mining, and generalized AI for Science.

Yudong Zhang (Graduate Student Member, IEEE) is now a Ph.D. candidate in the School of Data Science, University of Science and Technology of China (USTC). He received his bachelor’s degree from the University of Electronic Science and Technology of China (UESTC) in 2020. He has published over 10 research papers on top conferences and journals such as IEEE TITS, SIGKDD, WSDM and ICDM. His current research interests include spatial-temporal data mining and intelligent transportation systems.

Xu Wang is now a doctoral student in the School of Data Science, University of Science and Technology of China. He got his bachelor degree of automation at North Eastern University in 2017. His research interest mainly includes data mining, machine learning and computer vision.

Zhengyang Zhou (Graduate Student Member, IEEE) is now an Associate Researcher at the University of Science and Technology of China (USTC). He received his Ph.D. degree at USTC in 2023. He has published over 20 papers on top conferences and journals such as IEEE TKDE, IEEE TMC, IEEE TVT, WWW, AAAI and ICDE. His main research interests include spatial-temporal data mining and urban computing, and he is committed to improving the accuracy, reliability and generalization of deep spatial-temporal learning models to empower the fields of traffic prediction, urban safety and pollution control.

Yang Wang (Senior Member, IEEE) is now an Associate Professor at USTC. He got his Ph.D. degree at the University of Science and Technology of China (USTC) in 2007. He has published over 100 high-level conference and journal papers on IEEE TKDE, IJCAI, AAAI, MOBICOM, ICDE, et, al. His research interest mainly includes wireless sensor networks, spatial-temporal data mining, and data-driven interdisciplinary research. He is a senior member of both ACM and IEEE.