Inferring Intersection Traffic Patterns with Sparse Video Surveillance Information: An ST-GAN method

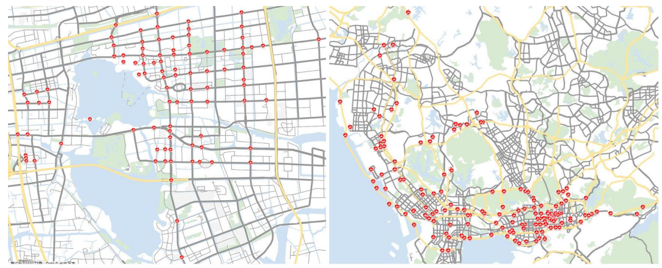
Pengkun Wang, Chaochao Zhu, Xu Wang, Zhengyang Zhou, *Student Member, IEEE,* Guang Wang, and Yang Wang*, *Senior Member, IEEE*

***Abstract*—Traffic patterns of urban road intersections are important in traffic monitoring and accident prediction, thus play crucial roles in urban traffic management. Although real-time traffic information is consistently provided by surveillance cameras equipped at road intersections, the sparsity of surveillance distribution poses great challenges in performing a complete real-time traffic pattern analysis. To tackle that, existing works either assume that the traffic patterns are static, or assume a multi-variant distribution model for intersection traffic volumes. The former assumption neglects the temporal features of traffic patterns, and the latter is limited in capturing fine-grained spatiotemporal dependencies. To tackle the problem, we propose a novel framework, SpatioTemporal-Generative Adversarial Network (ST-GAN), that exploits deep spatiotemporal features of urban networks and offers accurate traffic pattern inferences with incomplete surveillance information. The ST-GAN framework incorporates a modified GCN network wired with the encoder-decoder mechanism and an LSTM network, which are further boosted by an iterative adversarial training process. Comprehensive experiments on real datasets show that ST-GAN achieves better inference accuracies than state-of-the-art solutions.**

***Index Terms*—Inference, intersection, traffic pattern, sparse surveillance, GAN.**

I. Introduction

THE proliferation of road video surveillance systems [1]–[3] gives prominence to intelligent transportation services [4]–[6], including optimization of urban vehicle driving [7]–[10] and analysis of road network traffic flows [1], [2], [11], [12]. Most traffic analysis with surveillance systems assumes a dense coverage of surveillance distribution over road network intersections. However, the sparsity of surveillance distribution can hardly be avoided in real applications, due to the high deployment cost and dynamic characteristics of urban road networks. For instance, Figure 1 shows the distributions of road surveillance cameras of two leading cities in China, Suzhou Industrial Park (SIP) and Shenzhen. In this figure, only 3.0% (103) of the 3,468 road intersections in SIP are surveillance-equipped, while only 0.8% (129) of the 16,264 road intersections in Shenzhen are surveillance-equipped.



(a) SIP (3.0%, 103/3,468) (b) Shenzhen (0.8%, 129/16,264)

Fig. 1. Sparse distributions of road surveillance cameras in SIP and Shenzhen. Red dots indicate intersections where a surveillance camera is deployed. The surveillance camera coverage rates of SIP and Shenzhen are 3.0% (103/3,468) and 0.8% (129/16,264), respectively.

There have been studies [13]–[15] on forecasting traffic statuses with data incompleteness caused by the data sparsity issue or networking failure. However, these seemingly similar techniques cannot be directly used for inferences with the permanent incomplete traffic information caused by the sparse coverage of road surveillance cameras. Recently, there have also been studies [16]–[21] on modeling and inferring citywide traffic statues with sparse surveillance information, which can be clustered into two categories, discrete road segment similarity based methods [16], [18], [19] and holistic road network spatiotemporal correlation based methods [17]. The former makes inferences based on the calculation of similarities between surveillance-equipped and surveillance-free road segments with contextual information, such as velocities, road segment length, and Point of Interest (POI) features. However, these methods simplify the profound natures of spatiotemporal correlations into pair-wise similarity score comparisons, thus fall short in making accurate inference [22]. The latter infers traffic volumes for surveillance-free intersections with the assumption of multi-variant distribution models [17]. Nevertheless, the assumption may yield biased estimation due to the lack of parameters of surveillance-free intersections.

To tackle the challenges mentioned above, we propose a novel framework, SpatioTemporal-Generative Adversarial

This work is partially supported by Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR-005), NSFC (No.62072427), Anhui Science Foundation for Distinguished Young Scholars (No.1908085J24), and Jiangsu Natural Science Foundation (No.BK20191193). *(Corresponding author: Yang Wang.)*

Pengkun Wang, Xu Wang, Zhengyang Zhou, Yang Wang are with University of Science and Technology of China, Heifei, China (e-mail: pengkun@mail.ustc.edu.cn; wx309@mail.ustc.edu.cn; zzy0929@mail.ustc.edu.cn; angyan@ustc.edu.cn).

Chaochao Zhu is with HUAWEI TECHNOLOGIES Inc, Shanghai, China (e-mail: cczhu@mail.ustc.edu.cn).

Guang Wang is with Massachusetts Institute of Technology, Cambridge, MA, USA (e-mail: guangw@mit.edu).

Network (ST-GAN), inspired by recent advances in face completion techniques [23]. Our ST-GAN consists of a modified Encoder-Decoder based Graph Convolution Network (ED-GCN) and a Long Short-Term Memory (LSTM) neural network, for learning latent correlations in graph-structure data like road network [24]–[26] and temporal dependencies of traffic volumes [24], [27], [28], respectively. The iterative adversarial training process of GAN enables our framework to improve the quality of volume inference within surveillance-free intersections.

Our work is a sub-system of a real project, i.e., the integrated urban computing system, in cooperating with the traffic administrative agency of SIP, as shown in Figure 2. However, the information is incomplete in the sense that the distribution of surveillance cameras is sparse, as shown in SIP and Shenzhen in Figure 1. We also collect the third-party GPS data of 4,367 and 8,572 taxicabs with an average sampling rate of 20 seconds for Shenzhen and SIP to generate the training data, we mask a set of randomly selected intersections for the GPS data, in order to imitate the incomplete video surveillance scenarios. We then train the generator of our ST-GAN framework to reconstruct the original data with incomplete training data, which captures the deep spatiotemporal correlations through ED-GCN and LSTM modules. The ability of the generator is further enhanced by an iterative adversarial training process with the discriminator in ST-GAN. At last, the trained generator can be used to infer traffic volumes of surveillance-free intersections, with only real-time and sparse surveillance information collected from surveillance-equipped intersections. Experiments show that our proposal can improve the inference accuracy at least 10.43% and 13.85% on two real-world datasets, respectively.

Our main contributions are summarized as follows.

- To the best of our knowledge, this is the first work that utilizes the GAN-based deep learning framework to tackle the sparse-surveillance based real-time urban traffic pattern inference problem, by modeling the holistic urban traffic patterns of the entire urban road network from a third-party dataset and using the learned holistic patterns to infer traffic volumes of surveillance-free intersections only based on real-time and reliable inferred volumes of sparse surveillance-equipped intersections.
- The proposed generative adversarial network, ST-GAN, takes the well-designed ED-GCN and LSTM integrated module as the generator, to jointly capture spatial correlations and temporal dependencies. Through adversarial training on a dynamically masked third-party dataset, the generator of our ST-GAN is capable of inferring traffic volumes for surveillance-free intersections, and the seamlessly combined generator and discriminator can iteratively improve the performance of our ST-GAN.
- We evaluate the performance of our proposal with real-world large-scale monitoring datasets collected from two cities, i.e., SIP and Shenzhen. Extensive experiments cross-validate that our proposal significantly outperforms other alternative state-of-the-art solutions. Furthermore, we perform a case study to demonstrate that our ST-GAN can effectively capture the dynamic and diverse traffic patterns well and tackle the permanent sparse challenges by visualizing the inferred results of ST-GAN.

The rest of this paper is organized as follows. Section II reports recent related works. Section III introduces preliminaries and formalizes the problem. Section IV investigates the proposed ST-GAN framework. Section V presents empirical studies. Section VI further discusses issues related to our problem and Section VII concludes the paper.

II. Related Work

In recent years, tons of works [13], [14], [16]–[21] have been achieved to address the data sparsity problem in urban traffic analysis. And the data sparsity problem in urban traffic surveillance can be divided into two categories, temporal missing, and spatial sparsity.

Regarding the issue of temporal missing which is mainly caused by the data sparsity issue or network failure, many methods of time series analysis and forecasting [13], [14] have been raised to address the problem. Obviously, these kinds of time series analysis and forecasting technologies, which highly rely on the spatial completeness of data, cannot be used to solve the problem of spatial sparsity in our task by making inferences with the permanent incomplete traffic information.

The problem of spatial sparsity is caused by the sparse coverage of road surveillance cameras, and there are also a small number of recent novel studies [16]–[19] aim at solving this problem. We can also summarize existing efforts on this field into two categories, discrete road segment similarity based methods [16], [18], [19] and holistic road network spatiotemporal correlation based methods [17], [20].

Regarding discrete road segment similarity based methods, [16] calculates and ranks the similarities within road segments to determine whether they should be selected into a candidate set, then infers the traffic volumes of those surveillance-free road segments based on the combination of the candidates by a key-value attention method. [18] proposes a Spatiotemporal Semi-Supervised Learning network (ST-SSL) to solve the problem of citywide traffic volume inference. It first constructs spatial and temporal affinity matrices to represent the correlations within road segments by taxicab trajectories as well as some other static features of road segments, then infers segment traffic volumes based on the assumption that two segments should have similar lane volume patterns if they share similar urban features. [19] first collects traffic speeds and volumes from original GPS data, then solves the problem of speed missing with the method of collaborative matrix factorization and abstracts training traffic features with the bayesian network, and finally infers citywide traffic volumes with the K-Nearest Neighbor (KNN) algorithm. In practice, the road traffic volumes of individual road segments can be significantly influenced by the topology and traffic statuses of the entire road network, so this kind of discrete road segment similarity based methods should have very poor performances on inferring traffic statuses of complex urban road networks. Besides, these discrete road segment similarity based methods mostly focus on the traffic volume completion issue of individual road segments, while the traffic statuses of

Fig. 2. Urban Computing System of SIP. The size of points represents the relative value of the traffic volume at the corresponding intersection, and the point color of red or purple demonstrates the traffic volume of an intersection is monitored by the pre-deployed surveillance cameras or inferred by our method, respectively.

urban intersections are more important for urban traffic administrative departments since it has been proved that most urban hazards and traffic problems concentrate on intersections [22].

For holistic road network spatiotemporal correlation based methods, [17] first models the traffic volume of the entire road network with transferred transition probabilities from a third-party GPS dataset, uses a multi-variate normal distribution model that takes transition probabilities as inputs to make the incomplete surveillance space approximately complemented, and finally infers real-time traffic volumes in road networks with only partial intersections equipped with surveillances. However, the hypothesis that the traffic volumes of urban road networks follow a multi-variant distribution is too idealistic for real-world data research. Further, this statistical model based method cannot truly address the challenge of surveillance-free intersection traffic volume inference since it still has to fill the parameters of surveillance-free intersections by the parameters of the nearest surveillance-deployed intersections.

In summary, existing works on addressing the problem of spatial sparsity cannot effectively and deeply capture the holistic inter-intersection spatial correlations which are the essential elements in inferring citywide traffic volumes when some parts of the surveillance information are unavailable. To this end, we should tackle the problem of spatial sparsity with a new holistic and deep learning perspective.

III. Problem Definition

In this section, we formally define basic concepts as well as the problem studied in the work.

Definition 1 (Road Network): Given an urban road network, it can be formalized as a directed graph $G(\mathcal{V},\mathcal{E})$ where vertex $v_i \in \mathcal{V}$ denotes urban intersection v_i and edge $e_{ij} \in \mathcal{E}$ indicates the directed road segment from intersection v_i to v_j.

In practice, as demonstrated in Figure 1, traffic surveillance cameras are pre-deployed on the road intersections to obtain intersection traffic volumes by analyzing and comprehending captured images and videos. Based on the fact that whether surveillance devices have been deployed, urban intersections can be divided into two classes, monitored intersections $\mathcal{V}_m$ and unmonitored intersections $\overline{\mathcal{V}_m}$ where $\mathcal{V}_m \cup \overline{\mathcal{V}_m} = \mathcal{V}$ and $\mathcal{V}_m \cap \overline{\mathcal{V}_m} = \emptyset$.

Definition 2 (Taxicab Traffic Volume): Given an intersection v_i and a time interval Δt, we can compute the traffic volume of this intersection v_i within the given interval Δt and denote it as $f_i^{\Delta t}$. Therefore, the traffic volumes of the entire road network can be formulated by:

$$\mathcal{F}^{\Delta t} = \left\{ \begin{array}{cccc} f_1^{\Delta t} & f_2^{\Delta t} & \cdots & f_{|\mathcal{V}|}^{\Delta t} \end{array} \right\} \tag{1}$$

Definition 3 (Surveillance Traffic Volume): Given road network $G(\mathcal{V},\mathcal{E})$ and the pre-deployed road surveillance system, the surveillance volume of intersection v_i during time interval Δt can be written as $s_i^{\Delta t}$. The surveillance traffic volumes of the entire road network can be defined by:

$$\mathcal{S}^{\Delta t} = \left\{ \begin{array}{cccc} s_1^{\Delta t} & s_2^{\Delta t} & \cdots & s_{|\mathcal{V}|}^{\Delta t} \end{array} \right\} \tag{2}$$

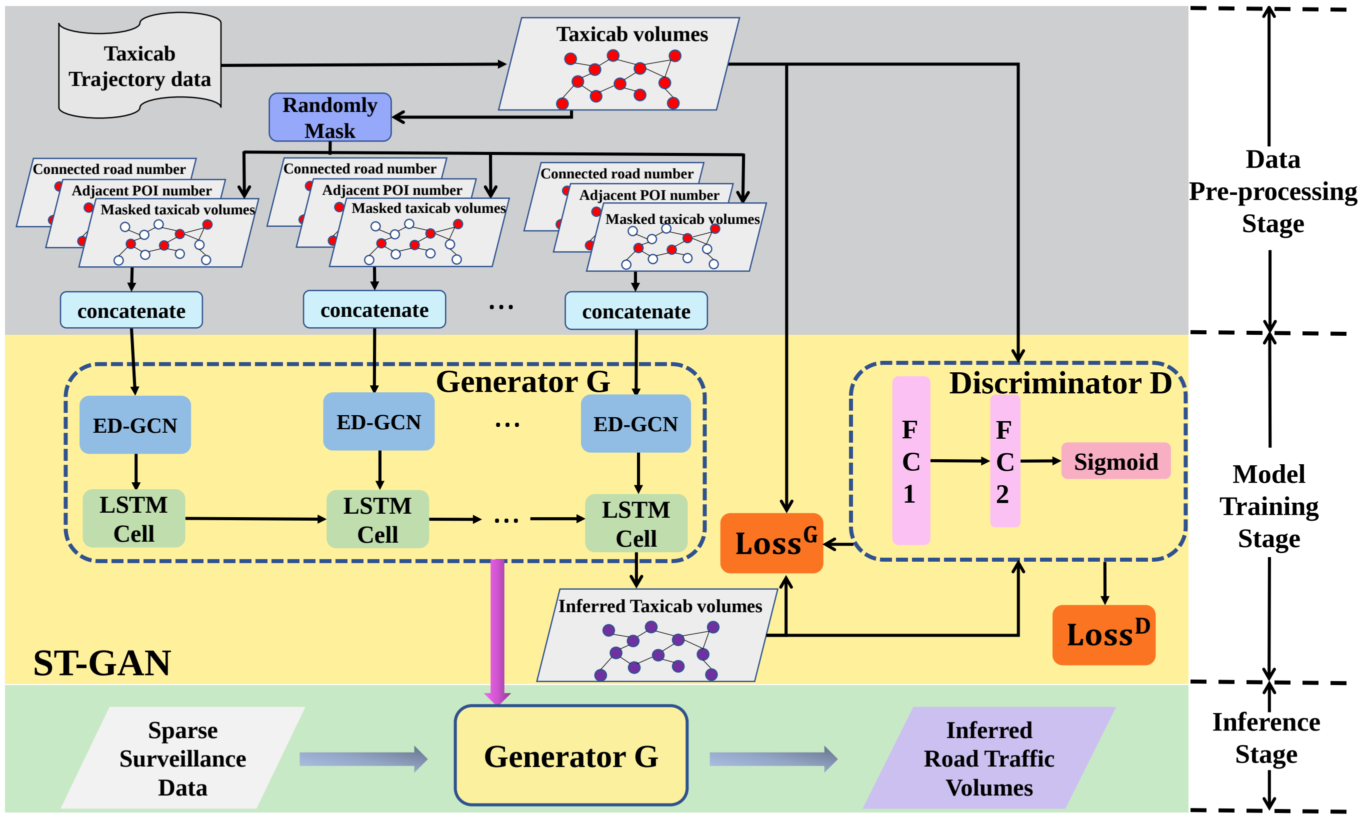


Fig. 3. Solution overview.

Here, the surveillance traffic volume of a surveillance-free intersection i is null ($s_i = null$ iff $s_i \in \overline{\mathcal{V}_m}$) regardless the setting of time interval Δt.

Worth noting that the traffic volume of an unmonitored intersection is $null$, while the traffic volume of a monitored intersection which has no vehicle cross by during a given time interval should be 0. Notice that it is commonly accepted that urban traffic flows have obvious time-varying patterns, and the setting of the time interval can significantly influence the understanding of urban traffic patterns [29]–[32]. With this preliminary, we define taxicab volumes and surveillance volumes with the time-varying traffic features[1].

Definition 4 (Inference with Sparse Surveillance): In the road network $G(\mathcal{V}, \mathcal{E})$, given sparse surveillance information from monitored intersection set $\mathcal{V}_m$ and a time interval Δt, our purpose is to design an algorithm to estimate the traffic volume of intersection $v_i \in \overline{\mathcal{V}_m}$ during the same time interval Δt.

Assuming $\vartheta_i^{\Delta t}$ and $\widehat{\vartheta_i^{\Delta t}}$ are the actual and estimated traffic volumes of intersection v_i during Δt respectively, if $v_i \in \mathcal{V}_m$, we have $\vartheta_i^{\Delta t} = s_i^{\Delta t}$. The accuracy of traffic volume inference can be estimated by equations 3.

$$Inference\ Accuracy\ (IA) = \frac{\vartheta_i^{\Delta t}}{\vartheta_i^{\Delta t} + \left|\widehat{\vartheta_i^{\Delta t}} - \vartheta_i^{\Delta t}\right|} \quad (3)$$

According to this equation, the accuracy of a monitored intersection is 100%, and for an unmonitored intersection, the accuracy is determined by the ratio of the real value to the summation of the real value and the estimation error, and notice that such a setting of the denominator is to normalize the accuracy to 1.

IV. ST-GAN for Traffic Volume Inference

A. Solution Overview

The overview of our proposed solution is illustrated in Figure 3. The main approach includes three stages, the data pre-processing stage, the ST-GAN training stage, and the inference stage. Details about each stage are illustrated as follows.

B. Data Pre-processing

Since the surveillance traffic data are inherently incomplete, we use a third-party taxicab dataset for learning traffic patterns of the entire road network. Figure 4 demonstrates the analysis of similarities of traffic volumes between taxicab and surveillance data in SIP. Figure 4 (a) illustrates the Pearson coefficient analysis with different volumes, where positive correlations can be observed between taxicab and surveillance data for

[1] Δt should be set with considering the equilibrium within the inference accuracies and temporal granularity. We here divide the temporal data into 30-minute slots according to common knowledge [17]. The setting of Δt has obvious correlations with the results of accuracy, and meanwhile, restricts the pervasiveness of our model.

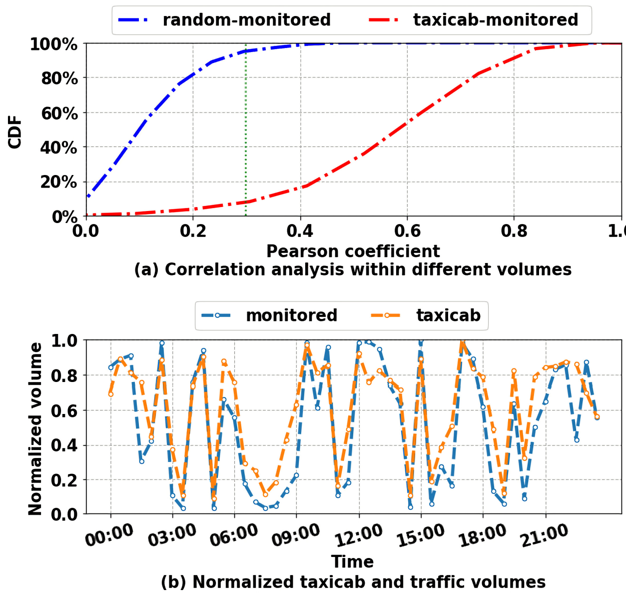


Fig. 4. Analysis of the similarities of taxicab and monitored traffic volumes in Suzhou

intersection traffic volumes. Figure 4 (b) shows the variation tendencies of normalized average traffic volumes for both taxicab and surveillance data, which also shows significant correlations between the two. Based on such observations, we use taxicab traffic data as training data for the learning of traffic patterns of urban vehicles, and for further inferring the traffic volumes of surveillance-free intersections. The benefits gained from adopting taxicab traffic data are for its full coverage of all urban intersections. For making it adaptive to the incomplete surveillance scenario, we randomly mask a set of intersections for making the training data sparse. Thus, the masked taxicab data for traffic volumes are as follows.

$$\mathcal{F}\_mask^{\Delta t} = \left\{ \begin{array}{cccc} f\_mask_1^{\Delta t} & f\_mask_2^{\Delta t} & \cdots & f\_mask_{|\mathcal{V}|}^{\Delta t} \end{array} \right\} \quad (4)$$

where $f\_mask_i^{\Delta t}$ is the after-masking taxicab volume of intersection v_i during Δt, satisfying:

$$f\_mask_i^{\Delta t} = \begin{cases} f_i^{\Delta t} & v_i \textit{ is unselected} \\ null & v_i \textit{ is selected to be masked} \end{cases} \quad (5)$$

With the random masking method, we can enhance the robustness and generalization of our trained model, supporting to capture the dynamic patterns of urban surveillance systems. After being masked, the taxicab data is concatenated with other static features of intersections, such as the numbers of connected road segments and the surrounding POIs, to generate an incomplete graph snapshot $x^{\Delta t}$ for time interval Δt. By doing so, we can use a series of incomplete graph snapshots $\mathcal{X} = \{x^{\Delta t-(m-1)}, x^{\Delta t-(m-2)}, \cdots x^{\Delta t}\}$ as inputs of the ST-GAN network to infer the complete citywide volumes in Δt, where m is the number of input time intervals[2].

[2]According to the settings in [33], we set the value of m as 3.

C. ST-GAN for Traffic Volume Inference

Our ST-GAN includes two modules following the conventional GAN framework, a generator G and a discriminator D. Generator G consists of two submodules, an ED-GCN for spatial correlation learning and an LSTM for temporal correlation learning. The encoder of ED-GCN first extracts and maps the spatial correlations of the inputted incomplete graph snapshots into high dimensional graphs. The decoder of ED-GCN then decodes the mapped high dimensional graphs to complete graph snapshots. Finally, the outputted complete graph snapshots are fed into the LSTM to learn and exploit the temporal correlations of intersection volumes. Regarding the discriminator D, it contains two Fully Connected (FC) layers and a Sigmoid activation layer. We then feed the generated complete graph snapshots and the real graph snapshots into the discriminator D to distinguish whether it is fake or real. With this minimax two-player game, this adversarial process can eventually force G to generate plausible and high-quality recovery of surveillance-free intersection volumes.

1) Generator G: As above mentioned, G contains two parts, ED-GCN and LSTM, for extracting the spatial and temporal correlations of intersection volumes respectively. We hereby introduce detailed implementations of this generator.

ED-GCN for spatial correlation learning: The detailed architecture of ED-GCN is illustrated in Figure 5. Here, we use a multi-layer modified GCN to exploit the spatial correlations within urban intersections in an encoder-decoder manner. The convolution can only affect 1-hop neighbors of an intersection vertex, while the distribution of monitored intersections is sparse. Thus we modify multi-layer convolutions to extract the correlations within multi-hop neighbors[3]. Specifically, the encoder and decoder are two three-layer symmetric GCNs. Two additional ReLU activation functions are employed in the second and fifth layer to make sure the results are non-linearized. For calculating this multi-layer GCN network, instead of calculating the adjacent matrix of urban intersections, we compute the weighted adjacent matrix $\mathcal{M}_a$ for all urban intersections by the following equation.

$$\mathcal{M}_a = \left\{ \begin{array}{ccc} \alpha_{11} & \cdots & \alpha_{1|\mathcal{V}|} \\ \vdots & \ddots & \vdots \\ \alpha_{|\mathcal{V}|1} & \cdots & \alpha_{|\mathcal{V}||\mathcal{V}|} \end{array} \right\} \quad (6)$$

$$where\ \alpha_{ij} = \begin{cases} Lane\ number\ of\ e_{ij} & e_{ij} \in \mathcal{E} \\ 0 & otherwise \end{cases}$$

Here, the element α_{ij} in the matrix $\mathcal{M}_a$ indicates the potential traffic intensity from intersection v_i to v_j. Notice that the fact $\alpha_{ij} = \alpha_{ji}$ may not hold, so that matrix $\mathcal{M}_a$ maybe not symmetric. We thus generate a new matrix $\mathcal{A}$ by setting $\mathcal{A} = \mathcal{M}_a + I_{|\mathcal{V}|}$. Here, $I_{|\mathcal{V}|}$ is the identity matrix of $|\mathcal{V}| \times |\mathcal{V}|$. Next, we generate the degree diagonal matrix $\mathcal{D}$ of all inter-

[3]Considering the scale of the urban road network and the sparsity of surveillance devices, we here set the number of layers to 6.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2022.3176653, IEEE Transactions on Vehicular Technology

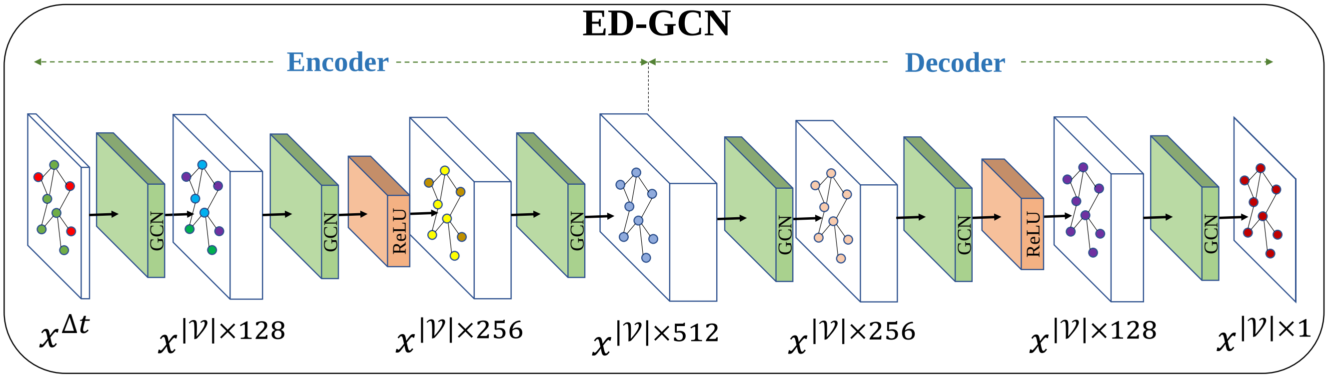


Fig. 5. Architecture of ED-GCN

sections by the following equation.

$$\mathcal{D} = \begin{Bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{|\mathcal{V}||\mathcal{V}|} \end{Bmatrix} \; where \; d_{ii} = \sum_{j=1}^{|\mathcal{V}|} \alpha_{ij} \tag{7}$$

Here α_{ij} is the i_th row and j_th column element of matrix $\mathcal{A}$, d_{ii} is the degree of intersection v_i in the road network graph $G(\mathcal{V}, \mathcal{E})$. With these preliminaries, we then calculate the weight laplacian matrix $\mathcal{M}$ of connections within intersections by:

$$\mathcal{M} = \mathcal{D}^{-\frac{1}{2}} \mathcal{A} \mathcal{D}^{-\frac{1}{2}} \tag{8}$$

For given time interval Δt, we can compute the ED-GCN by:

$$\mathcal{H}_l^{\Delta t} = \begin{cases} \text{ReLU}(\mathcal{M}\mathcal{H}_{l-1}^{\Delta t}\mathcal{W}_l) & l = 2,5 \\ \mathcal{M}\mathcal{H}_{l-1}^{\Delta t}\mathcal{W}_l & otherwise \end{cases} \tag{9}$$

Here, $\mathcal{H}_{l-1}^{\Delta t}$ and $\mathcal{H}_l^{\Delta t}$ are the input and output of the l_th layer, respectively. And $\mathcal{H}_0^{\Delta t} = x^{\Delta t}$. $\mathcal{W}_l$ represents the parameters of the l_th layer.

The encoder sub-part is to learn the spatial correlations between urban intersections by encoding the input incomplete graph snapshots to high-dimensional feature maps. It diffuses the features of intersections to their adjacent neighbors, in accordance to the adjacent matrix $\mathcal{M}_a$, by increasing the dimensionality of features to 128, 256, and 512 respectively. The output of the encoder is $|\mathcal{V}| \times 512$. By using the output of the encoder as the input, the decoder of a 3-layer GCN is to decrease the dimensionality of features. The output of the decoder, denote as $\mathcal{H}_{GCN}^{\Delta t}$, is $|\mathcal{V}| \times 1$. The outputted low dimensional complete snapshots have involved all the initial high dimensional features of urban road networks.

LSTM for temporal correlation learning: Due to the time-varying features of urban traffics, we adopt the LSTM network which is widely used in time sequence issues [34]. By considering the complete graph snapshots $\mathcal{H}_{GCN}^{\Delta t}$ which enclosed with the spatial correlations among all urban intersections, traffic volumes of surveillance-free intersections can be inferred with the time sequence analysis. Given time interval Δt, by using the outputted complete graph snapshots $\mathcal{H}_{GCN}^{\Delta t}$ and the hidden states $\mathcal{I}^{\Delta(t-1)}$ of LSTM cell of the last time interval as inputs, the LSTM equation is defined as:

$$\mathcal{I}^{\Delta t} = \text{LSTM}\left(\mathcal{H}_{GCN}^{\Delta t}, \mathcal{I}^{\Delta(t-1)}\right) \tag{10}$$

The LSTM cells enable our model to learn to retain or discard historical information according to the training data. The final output of the LSTM cell $\mathcal{I}^{\Delta t}$ can be regarded as the inferred citywide volumes at time interval Δt.

Volume loss of generator G: With the outputted inference of traffic volumes $\mathcal{I}^{\Delta t} = \begin{pmatrix} \tau_1^{\Delta t} & \tau_2^{\Delta t} & \cdots & \tau_{|V|}^{\Delta t} \end{pmatrix}$ of all urban intersections, where $\tau_i^{\Delta t}$ corresponds to the inferred taxicab volume of intersection v_i. We define the traffic volume loss function of generator G as:

$$Loss_{vol}^{G} = \text{MSE}\left(\mathcal{F}^{\Delta t}, \mathcal{I}^{\Delta t}\right) = \frac{1}{|\mathcal{V}|} \sum_{i=1}^{|\mathcal{V}|} \left(f_i^{\Delta t} - \tau_i^{\Delta t}\right)^2 \tag{11}$$

2) Discriminator D: The discriminator contains two FC layers and one Sigmoid activation layer. Assuming the input of discriminator D is $\Theta^{\Delta t}$ for time interval Δt, where

$$\Theta^{\Delta t} = \begin{cases} \mathcal{F}^{\Delta t} & The\ input\ is\ real\ taxicab\ volumes \\ \mathcal{I}^{\Delta t} & The\ input\ is\ inferred\ taxicab\ volumes \end{cases} \tag{12}$$

These two FC layers can reduce the input of real or inferred taxicab volumes to a number $y^{\Delta t}$ for evaluating the reliability of the inputs, where $y^{\Delta t} = \text{FC}\left[\text{FC}\left(\Theta^{\Delta t}\right)\right]$. The Sigmoid function of discriminator D in the activation layer can be written as:

$$D(\Theta^{\Delta t}) = \text{Sigmoid}_D(y^{\Delta t}) = \frac{1}{1 + e^{-y^{\Delta t}}} \tag{13}$$

The result of discriminator D is in the range $[0, 1]$. With the discriminator, we calculate the discriminator losses of real and inferred traffic volumes by the following equations.

$$Loss_{real}^{D} = \log(1 - D(\mathcal{F}^{\Delta t})) \tag{14}$$

and

$$Loss_{inferred}^{D} = \log(D(\mathcal{I}^{\Delta t})) \tag{15}$$

Notice that for the two equations, we expect the discriminated results of real traffic volumes can be close to 1, as much as possible. Also, we expect the discriminated result of inferred volumes can be close to 0.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2022.3176653, IEEE Transactions on Vehicular Technology

3) Losses of ST-GAN: The target of the discriminator is to improve generator G on the accuracies of traffic volume inference, until the inferred data is able to deceive the discriminator. Therefore, we expect discriminator D can well distinguish real and inferred data, so the overall loss for training D is as follows.

$$Loss^{D} = Loss^{D}_{real} + Loss^{D}_{inferred} \tag{16}$$

To help the generator G to deceive the discriminator, we have to make sure that the discriminated result of inferred volumes is close to 1. So, the loss function for training G is as follows.

$$Loss^{G}_{dis} = 1 - Loss^{D}_{inferred} \tag{17}$$

Based on that, the overall loss for training the proposed generator G can be formulated as follows.

$$Loss^{G} = Loss^{G}_{dis} + Loss^{G}_{vol} \tag{18}$$

The parameters of ST-GAN are trained iteratively. We fix all parameters of the discriminator during the training of generator G with $Loss^{G}$. We also fix all parameters of the generator while training the discriminator D, similarly.

D. Traffic Volume Inference of Unmonitored Intersections

As illustrated in Figure 3, after the training of ST-GAN, generator G is capable of inferring taxicab volumes for urban intersections with masked taxicab volume dataset. Then, generator G can be used for inferring urban traffic volumes with sparse surveillance information in a transfer learning manner, with the input of $\mathcal{S}^{\Delta t}$. Accordingly, the traffic volumes of surveillance-free intersections can be inferred.

E. Pseudocode of the Training Algorithm of ST-GAN

Algorithm 1 demonstrates the pseudocode of the training pipeline of our ST-GAN model. Algorithm 1 takes the adjacency matrix $\mathcal{M}$, timestep parameter m in the LSTM model and a series of incomplete graph snapshots as inputs. The outputs of Algorithm 1 are parameters in the ST-GAN model, where θ_1 and θ_2 are parameters of the ED-GCN module and the LSTM module in generator G respectively, and θ_3 is the parameter of discriminator D. In Algorithm 1, we first initialize the parameters with the standard normal distribution. In the training phase, we input m incomplete graph snapshots $\{x^{\Delta t}, x^{\Delta t+1}, \cdots x^{\Delta t+(m-1)}\}$ into the ED-GCN at one time, and we can obtain m complete graph snapshots $\{H^{\Delta t}, H^{\Delta t+1}, \cdots H^{\Delta t+(m-1)}\}$ from the output of the ED-GCN. Then, we input these m complete graph snapshots into the LSTM module and get the final complete graph snapshots at $\Delta t + (m-1)$ time slot. According to the ground truth at this time slot, we calculate the loss of generator G via Equation 18 and the loss of discriminator D via Equation 16, respectively. To be specific, when θ_1, θ_2 in G are fixed, we adjust the parameter θ_3 through the loss of D. In the same way, we fix θ_3 in D when we adjust the parameters θ_1, θ_2 in generator G. After the model trained with the training data, the parameters θ_1, θ_2 and θ_3 are obtained finally. To achieve stable training of ST-GAN, we use adaptive momentum estimation

Algorithm 1 Training Algorithm of ST-GAN.

Input: Timestep m;
Adjacency matrix $\mathcal{M}$;
Road Network $G(\mathcal{V}, \mathcal{E})$;
Incomplete graph snapshots $\{x^{\Delta t}, x^{\Delta t+1}, \cdots x^{\Delta t+(m-1)}\}$.
Output: Learned ST-GAN model, all parameters (θ_1, θ_2, θ_3) in this framework .
Initialize θ_1, θ_2, θ_3
for $t \leftarrow 1 \cdots T$ **do**
    $\{H^{\Delta t}, H^{\Delta t+1}, \cdots H^{\Delta t+(m-1)}\} \leftarrow$ ED-GCN($\{x^{\Delta t}, x^{\Delta t+1}, \cdots x^{\Delta t+(m-1)}\}$, $\mathcal{M}$, θ_1);
    $I^{\Delta t+(m-1)} \leftarrow$ LSTM($\{H^{\Delta t}, H^{\Delta t+1}, \cdots H^{\Delta t+(m-1)}\}$, m, θ_2);
    $Loss^{G} \leftarrow 1 - Loss^{D}_{inferred}(\theta_1, \theta_2, \theta_3) + Loss^{G}_{vol}(\theta_1, \theta_2)$;
    $Loss^{D} \leftarrow Loss^{D}_{real}(\theta_3) + Loss^{D}_{inferred}(\theta_1, \theta_2, \theta_3)$;
    Let θ_1, θ_2 fixed, do
        $\theta_3 \leftarrow$ Adamopt($Loss^{D}$, [θ_3]);
    Let θ_3 fixed, do
        (θ_1, θ_2) $\leftarrow$ Adamopt($Loss^{G}$, [θ_1, θ_2]);
end for
return θ_1, θ_2, θ_3

(Adma) optimizer [35] with learning rate of 0.001, β_1 = 0.5, and β_2 = 0.999. For ED-GCN, we set the node number to 16264 and the window size to 3. All our results are generated on 8 NVIDIA Tesla V100 GPUs with a batch size of 4.

V. Experiments

In this section, we conduct extensive empirical studies to evaluate our incomplete volume inference framework on two real-world datasets.

A. Data Description

We use datasets from two different modern cities, i.e., SIP and Shenzhen. The statistics are shown in Table I. Each dataset contains two sub-datasets: GPS data and surveillance data at road intersections as follows.

- **GPS data**: There are 4, 367 and 8, 572 taxicabs that upload their accurate GPS information every 20 seconds via their equipped 4G devices running independently in SIP and Shenzhen, respectively. We collect the GPS data in SIP and Shenzhen from Jan 1, 2017 to Mar 31, 2017, and subsequently generate the corresponding training data.
- **Surveillance data**: For the same period from Jan 1, 2017 to Mar 31, 2017, we use all sparse surveillance information collected from monitoring in SIP and Shenzhen, and match this dataset with the GPS dataset.

TABLE I
DATASETS STATISTICS.

GPS data	SIP	Shenzhen
Time span	1/2017-3/2017	1/2017-3/2017
Number of taxicabs	4,367	8,572
Average sampling rate	20 seconds per record	20 seconds per record
Surveillance data	**SIP**	**Shenzhen**
Time span	1/2017-3/2017	1/2017-3/2017
Number of total intersections	3,468	16,264
Number of surveillance-equipped intersections	103	129
Coverage rate	3.0%	0.8%

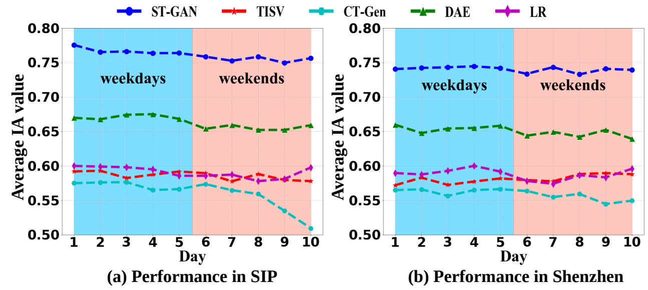


Fig. 6. Performance comparisons on different days.

B. Implementation Details

In the training phase, we first generate citywide taxicab volumes by GPS data. At each time interval, we randomly select to mask part of intersection volumes, leaving the masked intersection volumes as the target data to be inferred. The original citywide volumes are viewed as the ground-truth to train our ST-GAN model, with the Adam optimization in a back-propagation manner.

In the testing phase, we use the traffic volume information obtained by surveillance-equipped intersections. The traffic volume information of surveillance-free intersections can be seen as the masked values in the training phase. Due to the inherent lack of ground-truth data at surveillance-free intersections, we randomly select 20% surveillance-equipped intersections with volumes and assume they are also surveillance-free for numerical comparisons and model evaluations.

C. Evaluation Results and Analysis

1) Baselines: We evaluate the performance of our ST-GAN model by comparing it with the following baseline models.

- Linear Regression (LR) [36]: It is a linear model which learns to infer traffic volumes from previous observations of surveillance-equipped intersections and related road network features.
- Generalization module for citywide volume inference (CT-Gen) [16]: It is a generalized model which infers the volumes by distilling the extrinsic dependencies among existing volume surveillances with neural key-value attention architecture.
- Traffic Volume Inferring with Sparse Video Surveillance Cameras (TISV) [17]: It is a multi-variate distribution based citywide volume inference model by utilizing third-party vehicle GPS data.
- Deep Autoencoder (DAE) [37]: It is an encoder-decoder based method with a deep neural architecture to infer the citywide volumes. In this paper, we use the ED-GCN which is part of our ST-GAN as the deep neural architecture.

2) Performance Comparison: We evaluate the performance of different models on the metric of *Inference Accuracy (IA)* proposed in Equation 3.

Impact of day type: We show the effectiveness of our proposal in Figure 6. It can be observed that the accuracy of our proposed ST-GAN method is steadily above 75% in SIP and 73% in Shenzhen during randomly selected ten days, whether on weekdays or weekends. Compared with the baseline methods (i.e. CT-Gen, TISV, LR and DAE), our solution can increase the accuracy by 35.89%, 29.86%, 28.81%, 10.43% in SIP and 32.41%, 27.42%, 25.90%, 13.85% in Shenzhen. Among four baselines, DAE performs the best with the encoder-decoder mechanism. Since DAE does not consider temporal relationships and lacks the discriminator, the inference accuracy is significantly less than ours. For TISV, the strong assumption of multi-variant normal distribution traps the algorithm into a relatively lower accuracy. LR is a linear model and it fails to capture complex spatial relationships between intersections. As shown, CT-Gen performs the worst due to the lack of spatial correlations in consideration. By contrast, we consider the complex spatiotemporal relationships and solve the sparse problem with the help of third-party data, which takes effect in our spatial sparsity challenge task.

Impact of time slots: We also examine the performance with respect to the effects of time slots in Figure 7. Obviously, our method consistently obtains higher accuracies than others in any time slot even though with little fluctuations. This kind of fluctuation may be related to the complexity and variations in traffic patterns. For example, During the day, especially during the rush hours, since taxis are for-profit and the road conditions are prone to congestion, the travel routes chosen by some drivers may be unconventional, so there is a deviation between the taxi travel pattern and the overall travel pattern. At night, the overall traffic condition is relatively smooth, and the travel choices of drivers are more normal, so the taxi travel pattern is more similar to the overall travel pattern. Further, as shown in Figures 8 and 9, whether on weekdays or weekends, taxicab and monitored traffic volumes are more similar during night times than during rush hours, which more clearly demonstrates the fluctuations in inference accuracy.

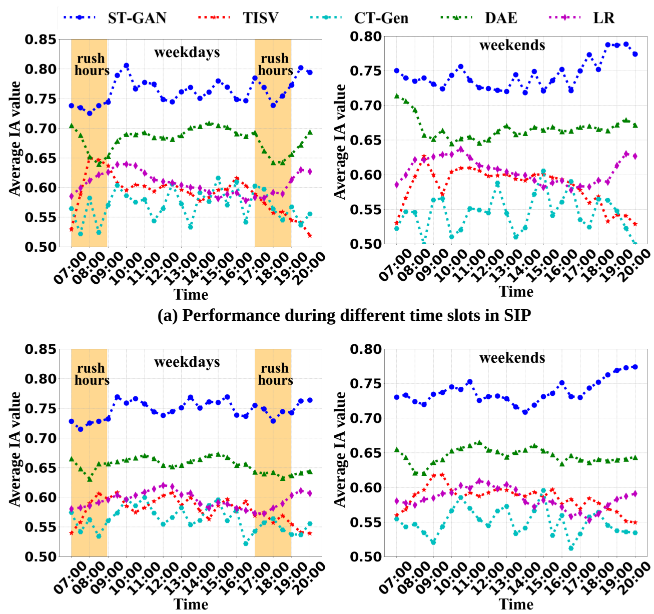
3) Inferring Error Analysis: We also utilize widely used metrics to quantify the inferring errors of different methods, including Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), shown as below.

$$\mathrm{MAE} = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \left| \vartheta_i^{\Delta t} - \widehat{\vartheta_i^{\Delta t}} \right| \tag{19}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \left(\vartheta_i^{\Delta t} - \widehat{\vartheta_i^{\Delta t}} \right)^2} \tag{20}$$

where $\vartheta_i^{\Delta t}$ and $\widehat{\vartheta_i^{\Delta t}}$ are the actual and inferred traffic volumes at intersection v_i during Δt, respectively. $\mathcal{D}$ is the total number of verifying intersections. The experimental results are shown

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2022.3176653, IEEE Transactions on Vehicular Technology



(a) Performance during different time slots in SIP

(b) Performance during different time slots in Shenzhen

Fig. 7. Performance comparisons during different time slots.

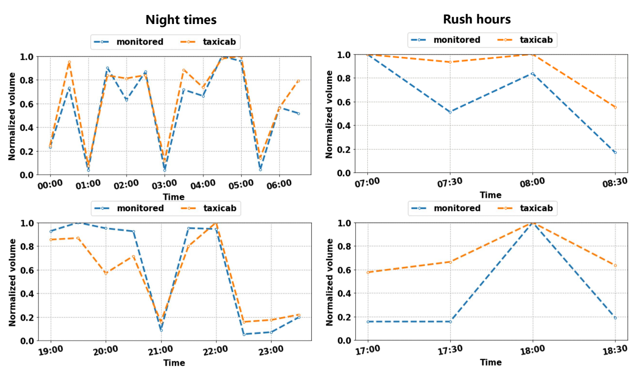


Fig. 8. Similarities between taxicab and monitoring traffic volumes on weekdays.

in Table II. We found our ST-GAN model achieves the best performance on both two real-world datasets.

TABLE II
INFERRING ERROR COMPARISONS.

	SIP / Shenzhen	
Model	MAE	RMSE
LR	206 / 228	211 / 243
TISV	229 /236	250 / 268
CT-Gen	249 / 255	275 / 287
DAE	164 / 187	196 / 214
ST-GAN	**84 / 103**	**105 / 127**

Figure 10 visualizes the inferring errors of all evaluated models in terms of MAE. To achieve a more comprehensive and intuitive understanding of the absolute error values of all methods, we first leverage the Kernel Density Estimation [38] method to calculate the probability density distribution of all intersections' average traffic volumes during all time intervals

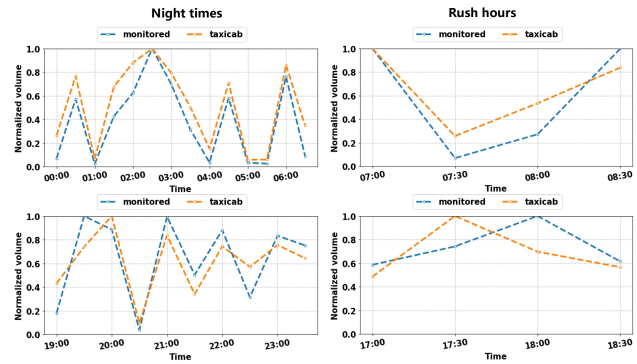
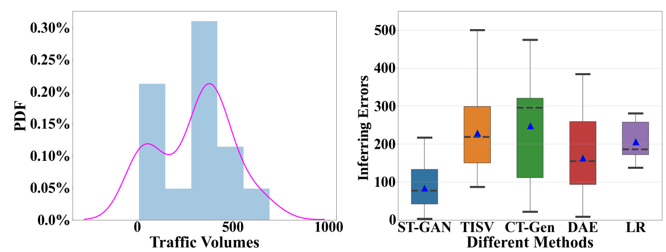


Fig. 9. Similarities between taxicab and monitoring traffic volumes on weekends.



(a) Traffic Volumes Distribution

(b) Inferring Errors Comparison

Fig. 10. Inferring Errors Analysis in SIP.

in 10 different days, and the results are shown in Figure 10 (a). We found that the traffic volumes between 300 and 500 are more than 50% of time slots. Figure 10 (b) is a boxplot that demonstrates the average inferring errors obtained from different methods at time slots in 10 days. We can see that the average inferring error of our ST-GAN model is much smaller than other methods. Moreover, the average inferring errors of other methods (i,e, TISV, CT-Gen, and DAE) are not only large in the average value but also fluctuate greatly. For our ST-GAN, the inferring errors fluctuate in a small range. Although the inferring error range of LR is also small, the value of inferring errors in the range is fairly large.

D. Ablative Studies

In order to evaluate the importance of each component in our ST-GAN, we design the following ablation study. We remove four well-designed components subsequently as follows: (i) LSTM module (Model 1), (ii) Substitute ED-GCN for a traditional GCN layer (Model 2), (iii) Discriminator in GAN (Model 3), (iv) LSTM, and discriminator (Model 4). Except for the changed part(s), all ST-GAN variants have the same structure and parameter settings. We compare the performance of variants both on weekdays and weekends to observe the changes between them. The numerical results are shown in Table III.

Overall, the integrated model consistently outperforms other alternative variants regardless of weekdays or weekends. As illustrated, LSTM and discriminator modules contribute to more than 10.4% improvement in SIP and 13.8% improvement in Shenzhen, respectively. This also verifies the effectiveness

TABLE III
PERFORMANCE ON DIFFERENT VARIANTS OF ST-GAN.

	SIP		Shenzhen	
Variants	Weekdays	Weekends	Weekdays	Weekends
Model 1	0.7033	0.6968	0.6829	0.6653
Model 2	0.7180	0.6940	0.6983	0.6914
Model 3	0.7228	0.7063	0.7039	0.6827
Model 4	0.6709	0.6553	0.6550	0.6453
Integrated	**0.7668**	**0.7551**	**0.7424**	**0.7380**

of our consideration of temporal effects and the generative-adversarial process. Further, as the result of Model 2 shows, incorporating the encoder-decoder mechanism in traditional GCN also makes sense in our integrated model.

E. Case Study

As Figure 2 shows, our work is a sub-research based on a real project in cooperating with the traffic administrative agency of SIP. Figure 11 shows our real application within three time intervals of two typical subregions, i.e., (i) Jinji CBD and (ii) Xietang Residential Community. In the figure, the point color of red or purple demonstrates the traffic volume of an intersection is monitored by the pre-deployed surveillance camera or inferred by our method. In addition, the size of points represents the relative value of the traffic volumes. The visualization results show that the inferred traffic volumes have achieved the expected effect, and we will interpret it from the following three perspectives:

- Spatial similarity: Whether in Jinji CBD or Xietang Residential Community, the distribution of inferred traffic volumes of surveillance-free intersections and volumes of surveillance-equipped intersections are consistent. If the traffic volumes at these intersections are integrated, we find that the overall distribution of traffic volumes across the region is reasonable. Especially in CBD area, the traffic flow shows a distribution that spreads to the surrounding area.
- Temporal dynamics: In Jinji CBD, for surveillance-equipped intersections, the actual traffic volumes during the interval of 7:00 $\sim$ 8:00 a.m. show an upward trend, which indicates that this interval is rush hour. For surveillance-free intersections, the inferred traffic volumes during this interval also show an upward trend, which is consistent with the actual situation. In Xietang Residential Community, the actual traffic volumes show a stable trend, which is also in line with the characteristics of residential areas. In addition, the inferred traffic volumes change smoothly, which is consistent with the actual situation. The above changes indicate that our model can learn this dynamic trend of traffic over time. The above information indicates that our model can learn the trend of dynamic change of traffic volume.
- Mobility tendency: In Jinji CBD, the actual traffic volumes during the interval of 7:00 $\sim$ 7:30 is small. As officers move from various residential areas mostly located in the southern and western in SIP to business blocks during peak hours in the morning, the actual traffic volumes during the time interval of 7:30 $\sim$ 8:00 increase significantly, and traffic volumes tend to move from south to north and from west to east in these time intervals. Obviously, the inferred traffic volumes also conform to this trend.

According to the above analysis, ST-GAN already has the ability to capture spatial similarity, temporal dynamics, and mobility tendency. The visualized results not only corroborate other experimental results but also show that our model can tackle the permanent sparse challenges effectively.

VI. DISCUSSION

In this section, we discuss some practical issues and lessons learned in this paper.

Inferring traffic volumes with sparse surveillance information: In this work, we propose a novel ST-GAN to exploit the spatiotemporal correlations within urban intersections, and then infer traffic volumes with only sparse surveillance information in a transfer learning manner. Experiments show that our approach can effectively infer traffic volumes for unmonitored intersections with the information obtained from fixed sparse urban traffic surveillance cameras, which only cover 3.0% and 0.8% of all intersections in SIP and Shenzhen, respectively. Further, the time complexity of each GCN layer is $O(|\mathcal{E}|CF)$ [39], where $|\mathcal{E}|$ is the number of graph edges, C is the number of input channels, and F is the dimension of feature maps in the output layer. Our modified multi-layer GCN component can finish one inferring in 0.129 seconds on average with 8 NVIDIA Tesla V100 GPUs.

The superiority of the technique for urban computing applications: In most existing intelligent transportation applications, urban traffic information is usually retrieved on the crowdsourcing platforms [40]–[42], or provided by telecommunication suppliers [43]. The results are somehow untrustworthy due to the inherent unreliable nature of the low-deployment-cost crowdsourcing platforms. Figure 12 demonstrates a case of cheating existing monitoring Apps, which originated from a performance art by the German artist Simon Weckert [44]. Specifically, in this case study, 99 used smartphones are transported in a handcart to generate virtual traffic jams in Google Maps. Through this activity, it is possible to turn a green street into red, which has an impact on the physical world. In our work, the information collected by traffic video surveillance systems is obtained in real-time and accurately for the intersections with equipped devices. Combined with advanced communication technology [45], [46], we believe that it makes a better and more reliable basis for advanced urban traffic intelligent systems.

Scalability of ST-GAN network: Our work is cross-validated in two typical cities in China. Further, it can also be a paradigmatic solution in various spatiotemporal applications, ranging from regional epidemics predictions to masked human action detection in vision tasks where sparse surveillance data is collected permanently [47], [48]. Specifically, the encoder and decoder of GCN empower to extract the node-wise correlations in graph-structure data, such as infected populations in cities or detected human skeletons in the graph form. Then the nodes

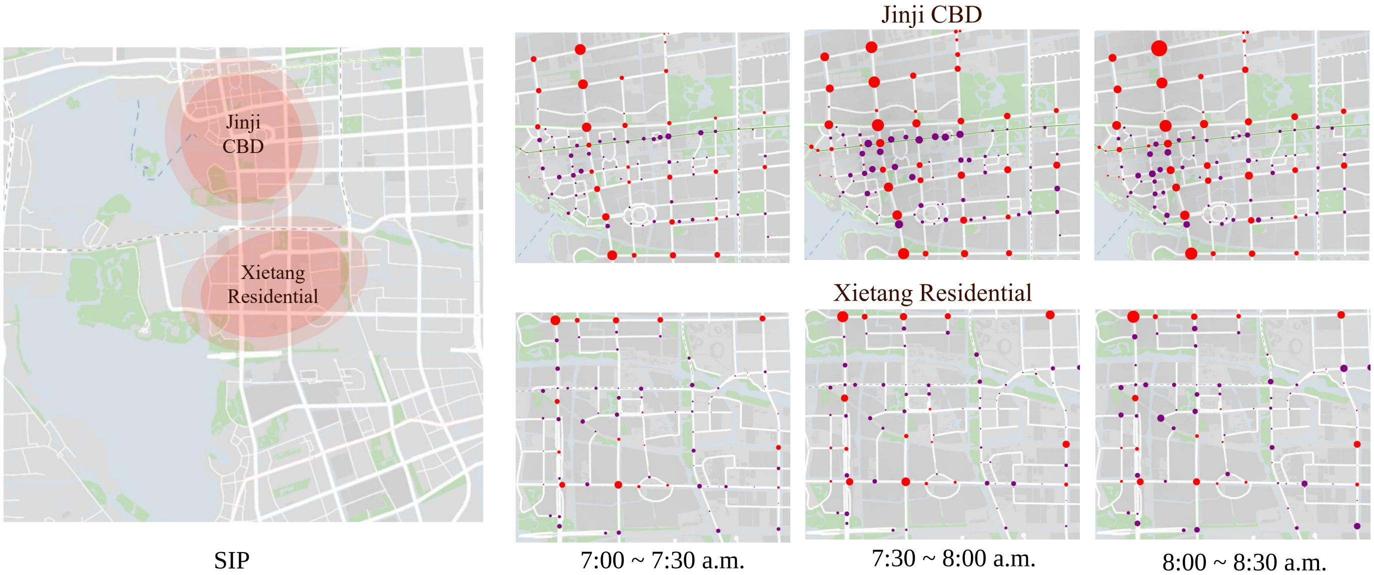


Fig. 11. Traffic volumes visualization of typical regions. The size of points represents the relative value of the traffic volume at the corresponding intersection, and the point color of red or purple demonstrates the traffic volume of an intersection is monitored by the pre-deployed surveillance camera or inferred by our method, respectively.

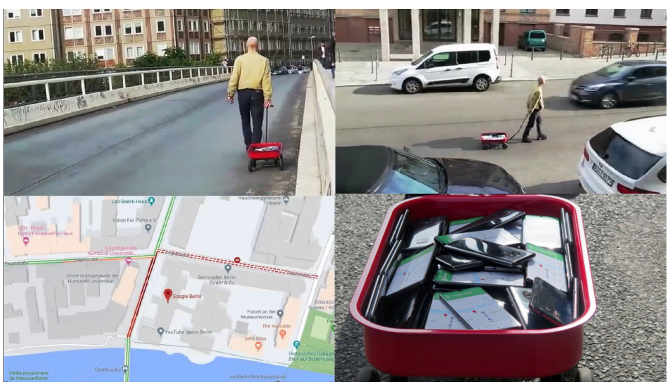


Fig. 12. A case of cheating existing monitoring Apps with a small toy trailer of mobile phones: Google map shows that the street is heavily congested while the traffic of the street is quite smooth [44].

that need to be predicted in the objective graph can be inferred by the GAN architecture with an auxiliary dataset, advancing the deeper applications of the graph-level management like population flow controlling and action prediction.

Possibility to integrate with federated learning: Federated learning has recently been widely used in intelligent transportation [49]–[51] and the Internet of Vehicles [52], [53] due to the ability to break down isolated data islands and protect data privacy. Integrating federated learning with ST-GAN is a potential means to improve model accuracy and generalization in the future. Inspired by federated learning, we can leverage distributed organizations to cooperatively train local traffic datasets in different regions to obtain a globally shared traffic pattern inference model without exchanging raw data, which can maximize the available resources of the model and ensure the privacy and security of users.

Further issues of the inferring model: Even though our proposed model ST-GAN can alleviate the overfitting on local neighborhood structures for graphs with very wide node degree distributions, the possible influence of the percentage that intersections with stationary surveillance cameras account for has not been discussed since the case of intensive traffic surveillance devices in urban areas has not been found. We will further investigate what will happen if the coverage of monitored intersection decreases, and where is the lowest boundary of the coverage ratio if we want to push the proposed algorithm to become practical.

VII. Conclusion

In this paper, we propose a novel integrated network ST-GAN to infer the traffic volumes for surveillance-free intersections with only sparse surveillance information. Based on highly positive correlations between taxicab and surveillance traffic patterns, we generate the training data with masked taxicab traffic volumes obtained from third-party trajectory datasets of reliable floating vehicles. With the well-designed ED-GCN and LSTM incorporated, our ST-GAN has the ability to capture the spatiotemporal traffic patterns between intersections. We further enhance the deep representations by taking advantage of the iterative improved adversarial mechanism. And finally, we infer the traffic volumes of surveillance-free intersections with only sparse surveillance by using the generator of the trained ST-GAN independently in a transfer learning manner. Performance evaluations on real-world datasets demonstrate the effectiveness of our proposal. Therefore, our work provides a brand-new solution to tackle the permanent spatial sparsity challenge from a deep-learning perspective.

In the future, our possible improvement directions include task-specific and task-independent. Task-specific promotion is to leverage multi-source data rather than just taxicab trajectories to further establish the knowledge graph with various auxiliary information for spatiotemporal fusion. Thus, the sparsity challenge of monitored traffic data can be alleviated subsequently, and the inference accuracy of our model can also be improved. Task-independent modification is to further investigate and understand the uncertainty caused by the sparsity of spatiotemporal data, and to support more general predictions like mobility-based pandemic controlling problem and the cold-start problem in recommender systems.

REFERENCES

[1] E. Bas, A. M. Tekalp, and F. S. Salman, "Automatic vehicle counting from video for traffic flow analysis," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2007, pp. 392–397.

[2] H. Liu, S. Chen, and N. Kubota, "Intelligent video systems and analytics: A survey," *IEEE Trans. Ind. Informat.*, vol. 9, no. 3, pp. 1222–1233, Aug. 2013.

[3] P. W. Patil, A. Dudhane, and S. Murala, "End-to-end recurrent generative adversarial network for traffic and surveillance applications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14 550–14 562, Dec. 2020.

[4] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec 2011.

[5] Y. Zhang, B. Wang, Z. Shan, Z. Zhou, and Y. Wang, "Cmt-net: A mutual transition aware framework for taxicab pick-ups and drop-offs co-prediction," in *Proc. 15th ACM WSDM*, 2022, pp. 1406–1414.

[6] X. Ding, J. Wang, C. Dong, and Y. Huang, "Vehicle type recognition from surveillance data based on deep active learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 2477–2486, Mar. 2020.

[7] H. Gonzalez, J. Han, X. Li, M. Myslinska, and J. P. Sondag, "Adaptive fastest path computation on a road network: a traffic mining approach," in *Proc. VLDB*, 2007, pp. 794–805.

[8] E. Schmitt and H. Jula, "Vehicle route guidance systems: classification and comparison," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2006, pp. 242–247.

[9] G. Leduc, "Road traffic data: Collection methods and applications," *Working Papers on Energy, Transport and Climate Change*, vol. 1, no. 55, pp. 1–55, 2008.

[10] J. Liu, H. Guo, J. Xiong, N. Kato, J. Zhang, and Y. Zhang, "Smart and resilient ev charging in sdn-enhanced vehicular edge computing networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 1, pp. 217–228, Jan. 2020.

[11] K. Suzuki and H. Nakamura, "Trafficanalyzer-the integrated video image processing system for traffic flow analysis," in *Proc. 13th World Congr. Intell. Transp. Syst.*, 2006, pp. 1–8.

[12] X. Song, Y. Guo, N. Li, and L. Zhang, "Online traffic flow prediction for edge computing-enhanced autonomous and connected vehicles," *IEEE Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2101–2111, Mar. 2021.

[13] R. Yasdi, "Prediction of road traffic using a neural network approach," *Neural Comput. Appl.*, vol. 8, no. 2, pp. 135–142, May. 1999.

[14] C. De Fabritiis, R. Ragona, and G. Valenti, "Traffic estimation and prediction based on real time floating car data," in *Proc. 11th IEEE Intell. Transp. Syst. Conf.*, 2008, pp. 197–203.

[15] A. Koesdwiady, R. Soua, and F. Karray, "Improving traffic flow prediction with weather information in connected cars: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9508–9517, Dec. 2016.

[16] X. Yi, Z. Duan, T. Li, T. Li, J. Zhang, and Y. Zheng, "Citytraffic: Modeling citywide traffic via neural memorization and generalization approach," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 2665–2671.

[17] Y. Wang, Y. Xiao, X. Xie, R. Chen, and H. Liu, "Real-time traffic pattern analysis and inference with sparse video surveillance information," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 3571–3577.

[18] C. Meng, X. Yi, L. Su, J. Gao, and Y. Zheng, "City-wide traffic volume inference with loop detector data and taxi trajectories," in *Proc. 25th ACM SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, 2017, pp. 1–10.

[19] X. Zhan, Y. Zheng, X. Yi, and S. V. Ukkusuri, "Citywide traffic volume estimation using trajectory data," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 2, pp. 272–285, Feb. 2016.

[20] M. A. M. Izhar, A. J. Aljohani, S. X. Ng, and L. Hanzo, "Joint decoding and estimation of spatio-temporally correlated binary sources," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6690–6694, Jul. 2018.

[21] C. Zhao, X. Duan, L. Cai, and P. Cheng, "Vehicle platooning with non-ideal communication networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 18–32, Jan. 2020.

[22] S. Anowar, M. D. Alam, and M. A. Raihan, "Analysis of accident patterns at selected intersections of an urban arterial," in *Proc. 21st ICTCT Workshop*, 2008.

[23] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3911–3919.

[24] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 3634–3640.

[25] X. Wang, C. Chen, Y. Min, J. He, B. Yang, and Y. Zhang, "Efficient metropolitan traffic prediction based on graph recurrent neural network," *arXiv preprint arXiv:1811.00740*, 2018.

[26] 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 *Proc. 33st AAAI Conf. Artif. Intell.*, vol. 33, no. 01, 2019, pp. 3656–3663.

[27] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. 31st AAAI Conf. Artif. Intell.*, vol. 31, no. 1, Feb. 2017.

[28] D. Jo, B. Yu, H. Jeon, and K. Sohn, "Image-to-image learning to predict traffic speeds by considering area-wide spatio-temporal dependencies," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1188–1197, Feb. 2018.

[29] Y. Wang, L. Huang, T. Gu, H. Wei, K. Xing, and J. Zhang, "Data-driven traffic flow analysis for vehicular communications," in *Proc. IEEE Conf. Comput. Commun.* IEEE, 2014, pp. 1977–1985.

[30] Z. Xu, Z. Li, Q. Guan, D. Zhang, Q. Li, J. Nan, C. Liu, W. Bian, and J. Ye, "Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 905–913.

[31] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "T-drive: Enhancing driving directions with taxi drivers' intelligence," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 220–232, Jan. 2011.

[32] N. J. Yuan, Y. Zheng, and X. Xie, "Segmentation of urban areas using road networks," *Tech. Rep.*, 2012.

[33] X. Dai, R. Fu, Y. Lin, L. Li, and F.-Y. Wang, "Deeptrend: A deep hierarchical neural network for traffic flow prediction," *arXiv preprint arXiv:1707.03213*, 2017.

[34] S. Hochreiter and J. Schmidhuber, "Lstm can solve hard long time lag problems," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 1997, pp. 473–479.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence, "Improving traffic prediction with tweet semantics," in *Proc. 23th Int. Joint Conf. Artif. Intell.*, 2013.

[37] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[38] H. Schaeben, "Exploratory orientation data analysis: Kernel density estimation and clustering," in *Mater. Sci. Forum*, vol. 157, 1994, pp. 431–438.

[39] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[40] Z. Liu, L. Chen, and Y. Tong, "Realtime traffic speed estimation with sparse crowdsourced data," in *Proc. IEEE 34th ICDE*, 2018, pp. 329–340.

[41] Y. Jiao, P. Wang, D. Niyato, B. Lin, and D. I. Kim, "Mechanism design for wireless powered spatial crowdsourcing networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 920–934, 2019.

[42] Y. Lin, Z. Cai, X. Wang, F. Hao, L. Wang, and A. M. V. V. Sai, "Multi-round incentive mechanism for cold start-enabled mobile crowdsensing," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 993–1007, Jan. 2021.

[43] S. Tao, V. Manolopoulos, S. Rodriguez Duenas, and A. Rusu, "Real-time urban traffic state estimation with a-gps mobile phones as probes," *J. Transp. Technol.*, vol. 2, no. 1, pp. 22–31, Nov 2012.

[44] S. Weckert, "Google maps hacks," http://simonweckert.com/googlemapshacks.html.

[45] H. Guo, X. Zhou, J. Liu, and Y. Zhang, "Vehicular intelligence in 6g: Networking, communications, and computing," *Veh. Commun.*, p. 100399, 2021.

[46] S. Zhang, J. Liu, T. K. Rodrigues, and N. Kato, "Deep learning techniques for advancing 6g communications in the physical layer," *IEEE Wirel. Commun.*, 2021.

[47] J. Liu, Z. Shi, S. Zhang, and N. Kato, “Distributed q-learning aided uplink grant-free noma for massive machine-type communications,” *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2029–2041, Jul. 2021.

[48] H. Yu, Z. Li, G. Zhang, P. Liu, and J. Wang, “Extracting and predicting taxi hotspots in spatiotemporal dimensions using conditional generative adversarial neural networks,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 3680–3692, Apr. 2020.

[49] T.-C. Chiu, W.-C. Lin, A.-C. Pang, and L.-C. Cheng, “Dual-masking framework against two-sided model attacks in federated learning,” in *Proc. IEEE GLOBECOM*. IEEE, 2021, pp. 1–6.

[50] Y. Liu, J. James, J. Kang, D. Niyato, and S. Zhang, “Privacy-preserving traffic flow prediction: A federated learning approach,” *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7751–7763, 2020.

[51] L.-Y. Chen, T.-C. Chiu, A.-C. Pang, and L.-C. Cheng, “Fedequal: Defending model poisoning attacks in heterogeneous federated learning,” in *Proc. IEEE GLOBECOM*. IEEE, 2021, pp. 1–6.

[52] X. Li, L. Cheng, C. Sun, K.-Y. Lam, X. Wang, and F. Li, “Federated-learning-empowered collaborative data sharing for vehicular edge networks,” *IEEE Network*, vol. 35, no. 3, pp. 116–124, 2021.

[53] A.-C. Pang, E. Au, B. Ai, and W. Zhuang, “Guest editorial introduction to the special section on fog/edge computing for autonomous and connected cars,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3059–3060, 2019.

Pengkun Wang is now a doctoral student in the School of Data Science, University of Science and Technology of China. He received his bachelor degree from Jilin University in 2017. His research interests mainly include data mining, multi-modal fusion and computer vision.

Chaochao Zhu now works at HUAWEI TECHNOLOGIES Inc. He received his master degree from University of Science and Technology of China. His research interests include machine learning, data mining in traffic applications.

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

Zhengyang Zhou is now a doctoral student in the School of Computer Science and Technology, University of Science and Technology of China. His research interests include machine learning, spatiotemporal data mining as well as artificial intelligence in traffic applications. He is a student member of AAAI and IEEE.

Guang Wang is now a postdoctoral research associate at MIT Media Lab, working with Prof. Alex ‘Sandy’ Pentland and his group for the Connection Science Program. He interests in data-driven research, e.g., finding meaningful patterns from data and then designing applications to enhance systems the data generated.

Yang Wang is now an associate professor at University of Science and Technology of China (USTC). He received his Ph.D. degree from USTC in 2007, under supervision of Professor Liusheng Huang. His research interests mainly include wireless networks, distributed systems, data mining, and machine learning.