Temporal Multi-view Graph Convolutional Networks for Citywide Traffic Volume Inference

Shaojie Dai[1], Jinshuai Wang[1], Chao Huang[2], Yanwei Yu[1,✉], Junyu Dong[1]
[1]Department of Computer Science and Technology, Ocean University of China, Qingdao, China
[2]Department of Computer Science, The University of Hong Kong, Hong Kong, China
{daishaojie, wangjinshuai}@stu.ouc.edu.cn, {yuyanwei, dongjunyu}@ouc.edu.cn, chaohuang75@gmail.com

***Abstract*—With the development of mobile position techniques, sensing the citywide traffic information has been well recognized as a crucial task for various urban computing applications, such as intelligent transportation system, location-based recommendation, and user mobility modeling. With the consideration of high cost for sensor installment and maintenance, the traffic monitoring spatial coverage is often very limited in practical urban sensing scenarios. The goal of this paper is to perform the traffic inference over road segments which lack of (with very limited) historical traffic observations. Towards this end, we propose a temporal multi-view graph convolutional network for Citywide Traffic Volume Inference (CTVI) which jointly captures the spatial-temporal dependencies across different time intervals and geographical locations. In our CTVI framework, we design our attentive multi-view graph neural architecture based on our generated spatial and feature affinity graphs, to perform the cross-layer message passing with the preservation of road segment-wise topological context. In addition, we develop a temporal self-attention module to encode the evolving traffic patterns over time, which incorporates the time-wise relation contextual signals into the main embedding space. Furthermore, we propose a joint learning objective function that consists of an unsupervised random walk enhancement and a semi-supervised spatio-temporal volume constraint to guide the learning of road segment representations for citywide traffic volume inference. Evaluation results on real-world traffic datasets demonstrate the superiority of our proposed CTVI framework as compared to state-of-the-art baselines.**

I. Introduction

The development of mobile Internet techniques make the real-time traffic monitoring important and valuable for a variety of urban sensing applications, such as intelligent transportation system [1], location-based recommendation [2] and user mobility modeling [3]. For example, accurate citywide traffic volume monitoring could provide efficient and convenient transportation services to the public [4][5]. In addition, the understanding the traffic patterns of different regions, is beneficial for governments to make better decisions for transportation scheduling and traffic jam alleviation [6].

While sensing the citywide traffic volume provides the great benefits for a wide spectrum of data-driven smart city applications, it faces several key challenges that remain to be solved: First, ***Arbitrary Missing Values***: the sensed traffic volume records can be absent at arbitrary time slots and locations, due to the various factors (e.g., sensor errors or inter-networking communication failure) [7]. Such arbitrary missing data will involve the information noise and hinder the traffic pattern modeling, which may lead to the performance degradation of traffic inference. Second, ***Lack of Historical Observations***: despite the growing deployment of various sensors (e.g., surveillance cameras, traffic radars or loop detectors), their geographical coverage is still very limited with the consideration of the whole urban space, due to the high installment and maintenance cost. For instance, only 2% of road segments in Jinan city deploy the surveillance cameras for traffic monitoring [8]. Hence, how to perform automated learning with the exploration of complex spatial-temporal dependencies to make traffic volume inference without (or with very limited) historical data, remains a significant challenge. Third, ***Complex Spatial-Temporal Dependencies***: the complex traffic patterns are exhibited with time-dependent and multi-grained temporal relations. Different granularity-specific variation regularities of traffic data may present various temporal patterns (e.g., hourly, daily, weekly) which are complementary and inter-dependent with each other [9].

To address the aforementioned challenges, in this paper, we propose a novel representation learning framework, named **CTVI**, to collectively learn the spatial and temporal dependencies for citywide traffic volume inference with volume sensing data. Specifically, we first construct multiple spatial and feature affinity graphs based on road network and road contextual features, such as the number of lanes, speed limits, road type, traffic volume, etc. Second, we perform the multi-view graph convolution encoding with the integration of attention mechanisms on both spatial affinity graph and feature affinity graph at each time interval, which extracts the most correlated information from both node features and topological structures substantially. Third, we design a temporal self-attention module to capture the different intra-road segment time-varying dependencies in the embedding space, which jointly includes recently\daily\weekly effects. Furthermore, we propose a joint learning objective function to guide the learning paradigm of road segment representations, which consists of an unsupervised random walk enhancement and a semi-supervised spatial-temporal traffic volume constraint on monitored road segments.

We summarize our main contribution as follows.

- We propose a novel representation learning framework, called CTVI, to infer citywide traffic volume by jointly

2374-8486/21/$31.00 ©2021 IEEE
DOI 10.1109/ICDM51629.2021.00120

modeling complex spatial corrections and temporal dependencies from both intra- and inter-road dimensions.
- We incorporate multi-view graph convolution on spatial and feature affinity graphs with temporal self-attention mechanism to adaptively learn the deep temporal correlations of road segment representation in both topological structures and contextual features.
- We propose a joint learning objective function that consists of an unsupervised random walk enhancement and a semi-supervised spatial-temporal volume constraint to augment the learning of road segment representations for citywide traffic volume inference.
- We conduct extensive experiments on two real-world traffic datasets from two cities to demonstrate the superiority of our CTVI compared with state-of-art baselines for citywide traffic volume inference. We release our source code at: https://github.com/dsj96/CTVI-master.

II. Related Work

Semi-supervised learning (SSL) method has been widely applied for unlabeled data inference, which can be used for inferring missing values in traffic volume. Meng *et al.* [10] propose ST-SSL to predict traffic volume values based on loop detector and taxi trajectories. ST-SSL first learns the speed pattern of each road segment from taxi trajectories, and then constructs a spatiotemporal affinity graph. Then it infers citywide volume by applying SSL method on the spatiotemporal affinity graph. Recently, Yu *et al.* [9] propose CityVolInf for citywide traffic volume inference using surveillance camera data, which combines a SSL-based similarity module with a traffic simulation module (*i.e.*, SUMO [11]) to model spatiotemporal correlations and transitions of traffic volume between adjacent road segments.

Recently, deep learning methods have shown strong ability in modeling complex nonlinear spatio-temporal relationship of traffic volume. Aiming at the problem of traffic volume inference, there are some works based on reinforcement learning, unsupervised or semi-supervised learning, and representation learning. Yi *et al.* [8] propose CT-Gen based on key-value memory neural network for traffic volume inference, which consists of candidate selection module and key-value attention network. The former component selects related road segments with existing volume sensors as candidates and the latter network learns the extrinsic dependencies among volume sensors. Tang *et al.* [12] propose JMDI model, which combines taxi trajectories and surveillance camera data. JMDI first adopts traffic simulator SUMO [11] and deep reinforcement learning to recover complete vehicle movements from incomplete trajectories. Then it constructs a spatio-temporal graph and use multi-view graph embedding to model the correlations between road segments. Finally, JMDI infers the citywide traffic volumes by propagating the traffic volume values of monitored road segments to the unmonitored ones through masked pairwise similarities.

III. Problem Definition

In this section, we first introduce key notations used in this paper and then formally define the studied problem.

Definition 1 (Road Segment)**.** *We utilize intersections to split roads into short road segments. Each road segment connects two adjacent intersections. Notice that road segments are directed. Let* $R = \{r_1, r_2, \ldots, r_n\}$ *denote the set of all road segments in a city.*

For each road segment r_i, we extract its road contexts as road segment features $\mathbf{x}_i = \{x_i^1, x_i^2, \ldots, x_i^f\}$, including road level, road length, number of lanes, speed limitation, stating\ending locations, etc. Let $\mathbf{X}$ denote the feature matrix of all road segments.

Definition 2 (Time Interval)**.** *We split the whole time period into non-overlapping equal-length time intervals and let* $T = \{t_1, t_2, \ldots, t_m\}$ *denote the set of all time intervals.*

Sensing devices (*e.g.*, loop detectors, surveillance cameras) are deployed at the corresponding road segments. We use $\mathcal{M}$ to denote the set of monitored road segments, and $\mathcal{U}$ to denote the set of unmonitored road segments.

Definition 3 (Traffic Volume)**.** *The traffic volume for a road segment is defined the total number of vehicles traversing through it during a specific time interval. We use* y_i^j *to denote the traffic volume values of the road segment* r_i *during the time interval* t_j*.*

Notice that the traffic volume values are only available at monitored road segments $r_i \in \mathcal{M}$. We now state our problem as below:

Problem 1 (Citywide Traffic Volume Inference)**.** *Given a road network, observed traffic volume at the monitored road segments, our goal is to infer citywide traffic volume of any unmonitored road segment,* $r_i \in \mathcal{U}$*, at any time interval.*

IV. Methodology

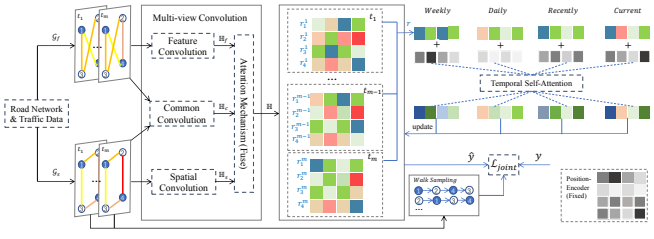


Fig. 1: The overview of the proposed CTVI.

In this section, we present the details of our proposed framework CTVI, as shown in Figure 1.

A. Affinity Graph Construction.

Spatial and temporal correlations on road network play important roles in the traffic volume inference. That is, the traffic volume values of different road segments are correlated

with each other in spatial and temporal perspectives. To model the dynamism of traffic in the road network, we first construct a spatio-temporal affinity graph, as shown in Figure 2.

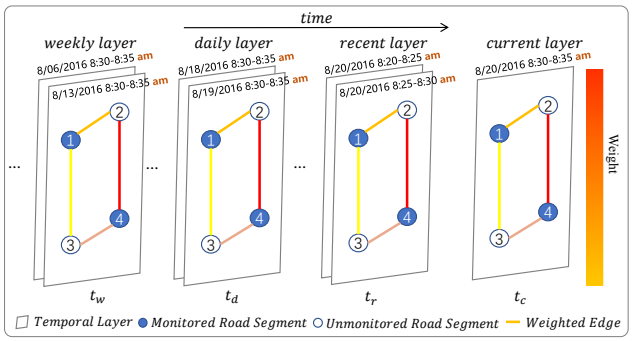


Fig. 2: An illustration of spatio-temporal affinity graph.

The spatio-temporal affinity graph consists of multiple spatial affinity graphs at all time intervals. Each spatial affinity graph $\mathcal{G}_s^i$ is a weighted graph during time interval t_i based on road network, where each road segment is a node in the graph, and edges indicate the connectivity between road segments.

Generally, the difference in the number of lanes at a intersection mainly affects the similarity of the traffic volume. Therefore, we define the weight on e_{ij} as follows:

$$w_{ij} = \sigma(liner(\frac{\min(lane_i, lane_j)}{\max(lane_i, lane_j)})), \tag{1}$$

where $lane_i$ denotes the number of lanes on road segment r_i, $liner$ is a linear function.

In addition, we extract road contexts from road network as road segment features, including road type, the number of lanes, speed limitation, starting/ending locations. Beyond that, we also consider the traffic volume value as a feature of road segments on each time interval, and the unobserved volume is initialized with average value of its spatial k-nearest neighbors (kNN). Next a feature affinity graph $\mathcal{G}_f^i$ is constructed using kNN method based on road feature matrix on each time interval. Based on graph $\mathcal{G}_f^i$, we finally get the adjacency matrix $\mathbf{A}_f^i$.

That is, we construct a spatial affinity graph $\mathcal{G}_s^j$ based road network and a feature affinity graph $\mathcal{G}_f^j$ based on road contextual features on each time interval t_j. For convenience, let $\mathbb{G} = \{\mathcal{G}^1, \mathcal{G}^2, \ldots, \mathcal{G}^m\}$ denote the collection of affinity graphs at all time intervals, where $\mathcal{G}^j = \{\mathcal{G}_s^j, \mathcal{G}_f^j\}$ $(j = 1, 2, \ldots, m)$.

B. Multi-view Graph Convolution Network

Inspired by [13], we first perform a multi-view graph convolution on both constructed spatial and feature affinity graphs on each time interval to learn road segment representations.

1) Spatial Convolution Module: To jointly model road constraint and spatial similarity, we first employ convolution operation on each spatial affinity graph $\mathcal{G}_s$ based on the spectral graph theory [14] to aggregate neighbor road information in Fourier domain. Following [14], multi-layer spatial convolution network performs following layer-wise propagation rule:

$$\mathbf{H}_s^{(l+1)} = ReLU(\tilde{\mathbf{D}}_s^{-\frac{1}{2}}\tilde{\mathbf{A}}_s\tilde{\mathbf{D}}_s^{-\frac{1}{2}}\mathbf{H}_s^{(l)}\mathbf{W}_s^{(l)}), \tag{2}$$

where $\mathbf{W}_s^{(l)}$ is a specific layer trainable weight matrix, $\tilde{\mathbf{A}}_s = \mathbf{A}_s + \mathbf{I}$ and $\tilde{\mathbf{D}}_{s,ii} = \sum_j \tilde{\mathbf{A}}_{s,ij}$. $\mathbf{H}_s^{(0)} = \mathbf{X} \in \mathbb{R}^{n \times f}$, where $\mathbf{X}$ is the feature matrix of all road segments and f is the number of features. $\mathbf{H}_s^{(l)} \in \mathbb{R}^{n \times d}$ is the output of l-th layer and d is the embedding dimension, denoting the hidden representations of all road segments in *spatial space*.

2) Feature Convolution Module: Following [13], we next perform feature convolution with $\mathbf{A}_f$ and $\mathbf{X}$ as input:

$$\mathbf{H}_f^{(l+1)} = ReLU(\tilde{\mathbf{D}}_f^{-\frac{1}{2}}\tilde{\mathbf{A}}_f\tilde{\mathbf{D}}_f^{-\frac{1}{2}}\mathbf{H}_f^{(l)}\mathbf{W}_f^{(l)}), \tag{3}$$

where $\mathbf{W}_f^{(l)}$ is a specific layer trainable weight matrix. In this way, we can obtain road segment feature representation $\mathbf{H}_f^{(l)}$, which captures the specific information in the *feature space*.

3) Common Convolution Module: Next, we adopt common-GCN to perform convolution with parameter sharing strategy. The propagation rules are defined as follows:

$$\mathbf{H}_{cs}^{(l+1)} = ReLU(\tilde{\mathbf{D}}_s^{-\frac{1}{2}}\tilde{\mathbf{A}}_s\tilde{\mathbf{D}}_s^{-\frac{1}{2}}\mathbf{H}_{cs}^{(l)}\mathbf{W}_c^{(l)}), \tag{4}$$

$$\mathbf{H}_{cf}^{(l+1)} = ReLU(\tilde{\mathbf{D}}_f^{-\frac{1}{2}}\tilde{\mathbf{A}}_f\tilde{\mathbf{D}}_f^{-\frac{1}{2}}\mathbf{H}_{cf}^{(l)}\mathbf{W}_c^{(l)}). \tag{5}$$

According to input graphs $\mathcal{G}_s$ and $\mathcal{G}_f$, we could obtain two output embedding $\mathbf{H}_{cs}$ and $\mathbf{H}_{cf}$, and the common embedding $\mathbf{H}_c$ in the *spatial and feature space* is defined as:

$$\mathbf{H}_c^{(l)} = \frac{\mathbf{H}_{cs}^{(l)} + \mathbf{H}_{cf}^{(l)}}{2}. \tag{6}$$

4) Multi-view Fusion: We finally utilize the attention mechanism $att(\mathbf{H}_s, \mathbf{H}_f, \mathbf{H}_c)$ [13] to combine their embedding in a reasonable way as follows:

$$(\mathbf{a}_s, \mathbf{a}_f, \mathbf{a}_c) = att(\mathbf{H}_s, \mathbf{H}_f, \mathbf{H}_c), \tag{7}$$

where $\mathbf{a}_s, \mathbf{a}_f, \mathbf{a}_c \in \mathbb{R}^{n \times 1}$ denotes the attention weight of n road segments *w.r.t.* $\mathbf{H}_s, \mathbf{H}_f$ and $\mathbf{H}_c$, respectively. We then have the learned attention weight $\mathbf{a}_S = diag(\mathbf{a}_s)$, $\mathbf{a}_F = diag(\mathbf{a}_f)$ and $\mathbf{a}_C = diag(\mathbf{a}_c)$. Finally, we fuse the representations as follows:

$$\mathbf{H} = \mathbf{a}_S \cdot \mathbf{H}_s + \mathbf{a}_F \cdot \mathbf{H}_f + \mathbf{a}_C \cdot \mathbf{H}_c. \tag{8}$$

5) Multi-view Graph Convolution on Multiple Time Intervals: Finally, we perform the multi-view convolution operation (Eq. (8)) on each time interval to model the temporal correlations. Specifically, we take the spatial/feature affinity graph as the input of multi-view convolution networks, and the output on $\mathbb{G}$ is the representation $\mathbb{H} \in \mathbb{R}^{m \times n \times d}$ of all road segments at all time intervals in a d-dimensional space.

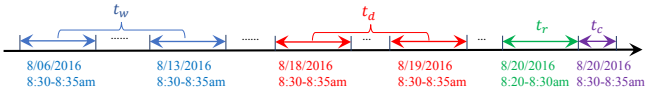


Fig. 3: An illustration of the input of temporal self-attention.

C. Temporal Self-Attention Mechanism

According to previous researches [10][8][9], traffic volume has strong temporal dependencies. Namely, citywide traffic volume usually follows a regular and periodic pattern, such as daily pattern, weekly pattern or even seasonally pattern. To incorporate such periodicities, in this paper, we select some relevant historical records by jointly taking the recent, daily, and weekly patterns into consideration to distinguish the influence of different historical information on the current traffic volume.

As illustrated in Figure 3, we intercept four types of time intervals along the time axis: (i) the current time interval t_c, (ii) the recent time intervals t_r, (iii) the daily time intervals t_d, (iv) the weekly time intervals t_w. Then we extract the hidden representations of the corresponding time intervals as the input of temporal self-attention module. The temporal self-attention score matrix is defined as follows:

$$\mathbf{S}_i = (\mathbf{H}_i + \mathbf{P})\mathbf{W}^Q((\mathbf{H}_i + \mathbf{P})\mathbf{W}^K)^\mathsf{T}, \tag{9}$$

where $\mathbf{H}_i$ denotes the concatenated hidden representation of road segment r_i at all related time intervals. $\mathbf{W}^Q \in \mathbb{R}^{d\times d}$ and $\mathbf{W}^K \in \mathbb{R}^{d\times d}$ denote the linear projection matrices, which transform the node representation $\mathbf{H}_i$ to a different space and can be optimized recursively during the training process. $\mathbf{P}$ denotes the position encoding, which aims to distinguish the sequence position of historical information. Notice that $\mathbf{P}$ is fixed and does not need to train, and is defined as in [15].

Finally, we could obtain the node representations that have considerate the historical information:

$$\mathbf{Z}_i = softmax(\frac{\mathbf{S}_i}{\sqrt{d}})(\mathbf{H}_i + \mathbf{P})\mathbf{W}^V, \tag{10}$$

where $\mathbf{W}^V \in \mathbb{R}^{d\times d}$ is a trainable linear projection matrix.

D. Multi-head Temporal Self-Attention

By stacking temporal self-attention layers, our approach can sufficiently model a single type of temporal relation. Nevertheless, real-world citywide traffic volume typically evolves along multiple latent relations. Therefore, we endow our temporal self-attention mechanism to capture different relations form historical information through multi-head attentions [15].

$$\mathbf{Z}_i = FC(concat(\mathbf{Z}_i^{(1)}, \mathbf{Z}_i^{(2)}, \ldots, \mathbf{Z}_i^{(\#head)})) \tag{11}$$

where $\#head$ denotes the number of head of temporal self-attentions. Afterwards, we adopt a fully connected layer FC to fuse multi-aspect attentions and change the dimension.

E. Joint Learning Optimization

Next, we further design a joint learning objective to enhance the learning of road segment representations, which incorporate both the spatial similarity and the knowledge of spatio-temporal volume patterns in a unit way.

First, we present an unsupervised objective function that captures the dynamic structural and temporal information into the node representations to train our model.

$$\mathcal{L}_{walk} = \sum_{t\in T}\sum_{v_i\in\mathcal{V}}\Big(\sum_{v_j\in\mathcal{N}^t_{walk}(v_i)} -\log(\sigma(s^t_{i,j})) - \sum_{v_k\in Neg^t(v_i)}\log(1-\sigma(s^t_{i,k}))\Big), \tag{12}$$

where $s^t_{i,j}$ is the representation similarity of road segments r_i and r_j, $\mathcal{N}^t_{walk}(v_i)$ is the set of nodes that co-occur with v_i in fixed-length random walks, and $Neg^t(v_i)$ is a negative edge sampling *w.r.t*, node v_i at time interval t.

Afterwards, the second objective aims to make a constraint that traffic volume on each road segment should be similar with the volume inferred by its top-k most similar road segments in the embedding space. The semi-supervised learning loss function is defined as follows:

$$\mathcal{L}_{volume} = \sum_{t\in T}\sum_{r_i\in\mathcal{M}} |y^t_i - \frac{\sum_j^k s^t_{ij}y^t_j}{\sum_j^k s^t_{ij}}|, \tag{13}$$

where y^t_i denotes the ground truth traffic volume of road r_i during the time interval t. Finally, we combine $\mathcal{L}_{walk}$ and $\mathcal{L}_{volume}$ to jointly train our model. Therefore, we minimize the following joint objective $\mathcal{L}_{joint}$:

$$\mathcal{L}_{joint} = \mathcal{L}_{walk} + \mathcal{L}_{volume} + \frac{\lambda}{2}||\Theta||^2, \tag{14}$$

where Θ denotes all parameters need to train, λ is a hyperparameter, and they act as a regular term. By Eq. (14), we can observe that it successfully captures the spatial properties and temporal dependencies of traffic volume patterns.

Taking all the aforementioned factors into consideration, we can infer the traffic volume for unmonitored road segments according to the final learned road segment representations: $\hat{y}^t_i = \frac{\sum_j^k s^t_{ij}y^t_j}{\sum_j^k s^t_{ij}}$.

V. Experiment

A. Datasets

We conduct extensive experiments on two real-world datasets collected from Hangzhou and Jinan cities in China. The traffic volume data in Hangzhou is collected from traffic radar, while the traffic data in Jinan is collected form traffic surveillance cameras. More specifically, there are 46 traffic radars in Yuhang district of Hangzhou and 165 surveillance cameras in the selected region of Jinan city, respectively. The detailed statistics of two datasets are summarized in Table I.

TABLE I: Basic statistics of two datasets.

Dataset	Hangzhou City	Jinan City
Time Spans	2021/01/03-01/03	2016/08/01-08/31
#Road Segments	553	493
#Monitored segments	46	165
#Features	8	7
Time interval (minute)	5	5
Sensor Type	Traffic radar	Surveillance camera

B. Baselines

We compare our CTVI against the following baselines:

- **K-Nearest Neighbors (KNN)** - KNN averages the nearest top k volume values as the inference result.
- **Contextual Average (CA)** - CA adopts the averaged volume values from top k most similar road segments according to features.
- **MLP** - MLP is a multi-layer fully connected neural network. We flattens all the features and then feeds them together into the network.
- **XGBoost** [16] - XGBoost is a boosting-tree-based method which is popular in data mining community. We train each time interval separately by XGBoost.
- **ST-SSL** [10] - ST-SSL is a semi-supervised model based on the fusion of multi-source data, which utilizes affinity graph to capture the temporal and spatial characteristics.
- **CT-Gen** [8] - CT-Gen is based on key-value memory neural network, which follows adjacent roads are likely to have similar volume and road segments with the same road properties share similar volume pattern.
- **JMDI** [12] - JMDI employs a traffic simulator and reinforcement learning to recover vehicle movements from the incomplete trajectories. The masked semi-supervised approach enhanced by multi-view graph embedding is also introduced for citywide traffic volume inference.

Since some baselines cannot handle traffic volume inference on multiple time intervals, we train each time interval separately, and report the average volume values.

C. Evaluation Metric

We use Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) of inferred traffic volume values to evaluate the performance of inference methods, which are defined as follow:

$$RMSE = \sqrt{\frac{1}{n|T|}\sum_{t=1}^{|T|}\sum_{i=1}^{n}(y_i^t - \hat{y}_i^t)^2}, \tag{15}$$

$$MAPE_t = \frac{100\%}{n|T|}\sum_{t=1}^{|T|}\sum_{i=1}^{n}|\frac{y_i^t - \hat{y}_i^t}{y_i^t}|, \tag{16}$$

where n is the number of test samples, $|T|$ is the number of time intervals, and $y_i^t \backslash \hat{y}_i^t$ denotes the ground truth\inferred volume value for road segment r_i during time interval t.

TABLE II: Performance comparison of different baselines.

Dataset / Methods	Hangzhou City $MAPE_t$	Hangzhou City $MAPE_p$	Hangzhou City $RMSE$	Jinan City $MAPE_t$	Jinan City $MAPE_p$	Jinan City $RMSE$
KNN (k=5)	0.6636	0.7139	63.1035	0.6446	0.6306	60.3842
CA (k=5)	0.6879	0.7325	65.4562	0.6568	0.6423	61.2357
MLP	0.6029	0.6561	56.4201	0.8180	0.6808	69.3974
XGBoost	0.4689	0.5243	53.9832	1.5811	0.5917	93.3649
ST-SSL	0.5638	0.5983	44.2793	0.7052	0.6883	59.0377
CT-Gen	0.3602	0.4622	37.9691	0.6727	0.4760	57.4482
JMDI	\	\	\	0.4655	0.5574	42.0020
CTVI	0.3294	0.4037	33.1924	0.4487	0.4389	34.5814

Additionally, we also employ another version of MAPE [8] as follows:

$$MAPE_p = \frac{100\%}{n|T|}\sum_{t=1}^{|T|}\sum_{i=1}^{n}|\frac{y_i^t - \hat{y}_i^t}{\hat{y}_i^t}|. \tag{17}$$

D. Parameter Setting

We randomly split the road segments with traffic volume data into training (80%) and testing (20%), respectively. We further select 20% of the training randomly as validation. For our method, we set learning rate to 0.005, l to 2, d to 128, λ to 5e-3, $\#head$ to 3, k to 5, and the number of negative samples to 5. In each experiment, we repeat 10 runs for all baselines and report average $MAPE_t$, $MAPE_p$ and $RMSE$.

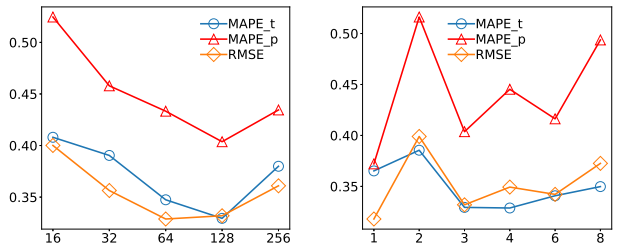
E. Comparison and Result Analysis

The comparison of CTVI with all baselines on two real-world datasets is shown in Table II. Since the transitions of vehicles is not known on the radar data in Hangzhou, JDMI cannot run on Hangzhou dataset.

As we can see, CTVI achieves state-of-the-art performance on both Hangzhou and Jinan datasets. Specifically, CTVI significantly performs better than state-of-the-art baseline CT-Gen, achieving average 20.92%, 10.23% and 26.18% improvements in terms of $MAPE_t$, $MAPE_p$ and $RMSE$ on two datasets, respectively. This is because that CT-Gen only learns dependencies among road segments based on road features, which can not model the complex spatiotemporal correlations well. Our CTVI employs multi-view graph convolution on each time interval and temporal self-attention at all time intervals to learn the complex spatiotemporal correlations of road segments in traffic volume. Moreover, CTVI significantly outperforms the joint embedding baseline JMDI by 3.61%, 21.26% and 17.67% improvements in terms of $MAPE_t$, $MAPE_p$ and $RMSE$ on Jinan, respectively. The main reason behind is that our CTVI learns the road segment representations by jointly considering spatial proximity, feature similarities, temporal dependencies, and spatiotemporal traffic pattern constraint. CTVI also significantly outperforms traditional approaches, such as KNN, CA, MLP and XGboost. This is because that these approaches do not have the ability to capture complex relationships both in spatial and temporal aspects, and thus cannot model the dynamism of traffic volume. Additionally, CTVI is better than all approaches in terms of $MAPE_t$, $MAPE_p$, which indicates CTVI can

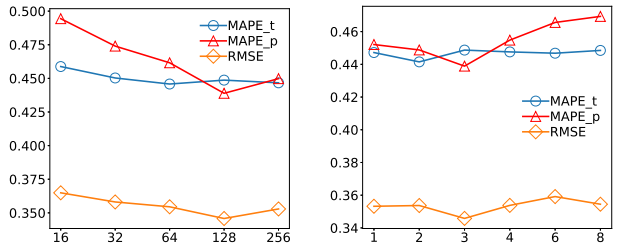
keep a balance between the ground truth and the inferred value, and can make traffic inference more accurately.

F. Parameter Sensitivity



(a) *w.r.t* d (b) *w.r.t* $\#head$

Fig. 4: Parameter sensitivity *w.r.t.* d and $\#head$ on Hangzhou.



(a) *w.r.t* d (b) *w.r.t* $\#head$

Fig. 5: Parameter sensitivity *w.r.t.* d and $\#head$ on Jinan.

We now investigate the sensitivity of our proposed framework CTVI *w.r.t.* the important parameters, including embedding dimension d and the number of head in temporal self-attention module $\#head$. To clearly show the influence of these parameters, we report $MAPE_t$, $MAPE_p$ and $RMSE$ with different parameter settings on both Hangzhou and Jinan datasets. Figures 4 and 5 show the experimental results. Notice that we multiply RMSE values by 0.01 to display all metrics in the same figure in Figures 4 and 5.

As shown in Figures 4(a) and 5(a), CTVI achieves the best performance when dimension $d = 128$ on both Hangzhou and Jinan datasets. Meanwhile, with the increasing of d, the overall performance first increases gradually and then tends to be stable. From the results in Figures 4(b) and 5(b), we can observe that CTVI is more sensitive to $\#head$ on Hangzhou dataset than on Jinan dataset. CTVI achieves the best results when $\#head$=1 on Hangzhou and $\#head$=3 on Jinan dataset. This is expected, because the traffic volume patterns in Hangzhou dataset is simpler than that in Jinan dataset. Specifically, there is no daily and weekly traffic volume patterns in Hangzhou dataset. Our model adopts more attention heads to capture different relations form historical information, and obtains the best performance when $\#head$=3 on Jinan dataset.

VI. Conclusion

In this paper, we propose a novel spatio-temporal representation learning framework CTVI for citywide traffic volume inference. It employs multi-view graph convolution over the spatial and feature affinity graphs to model the spatial and feature similarities, and performs temporal self-attention to differentiate the dependencies of different historical information. Additionally, CTVI tactfully designs a joint learning objective function, which guides the learning of road segment representations for traffic volume inference. Extensive experiments on two real-world datasets demonstrate that CTVI achieves better performance compared to state-of-the-art baselines in inferring citywide traffic volume.

Acknowledgment

This work is partially supported by the National Natural Science Foundation of China under grant Nos. 62176243, 61773331, U1706218 and 41927805, the Fundamental Research Funds for the Central Universities under grant No. 201964022, and the National Key Research and Development Program of China under grant No. 2018AAA0100602. Corresponding author: Yanwei Yu.

References

[1] R. Barnes, S. Buthpitiya, J. Cook, A. Fabrikant, A. Tomkins, and F. Xu, "Bustr: Predicting bus travel times from real-time traffic," in *KDD*, 2020, pp. 3243–3251.

[2] S. Feng, L. V. Tran, G. Cong, L. Chen, J. Li, and F. Li, "Hme: A hyperbolic metric embedding approach for next-poi recommendation," in *SIGIR*, 2020, pp. 1429–1438.

[3] J. Feng, Y. Li, C. Zhang, F. Sun, F. Meng, A. Guo, and D. Jin, "Deepmove: Predicting human mobility with attentional recurrent networks," in *WWW*, 2018, pp. 1459–1468.

[4] C. Chen, B. Liu, S. Wan, P. Qiao, and Q. Pei, "An edge traffic flow detection scheme based on deep learning in an intelligent transportation system," *TITS*, vol. 22, no. 3, pp. 1840–1852, 2021.

[5] C. Chen, H. Wei, N. Xu, G. Zheng, M. Yang, Y. Xiong, K. Xu, and Z. Li, "Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control," in *AAAI*, vol. 34, no. 04, 2020, pp. 3414–3421.

[6] 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 *KDD*, 2019, pp. 1720–1730.

[7] X. Yi, Y. Zheng, J. Zhang, and T. Li, "St-mvl: filling missing values in geo-sensory time series data," in *IJCAI*, 2016, pp. 2704–2710.

[8] X. Yi, Z. Duan, T. Li, T. Li, J. Zhang, and Y. Zheng, "Citytraffic: Modeling citywide traffic via neural memorization and generalization approach," in *CIKM*, 2019, pp. 2665–2671.

[9] Y. Yu, X. Tang, H. Yao, X. Yi, and Z. Li, "Citywide traffic volume inference with surveillance camera records," *IEEE Transactions on Big Data*, 2019.

[10] C. Meng, X. Yi, L. Su, J. Gao, and Y. Zheng, "City-wide traffic volume inference with loop detector data and taxi trajectories," in *SIGSPATIAL*, 2017, pp. 1–10.

[11] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of sumo-simulation of urban mobility," *International journal on advances in systems and measurements*, vol. 5, no. 3&4, 2012.

[12] X. Tang, B. Gong, Y. Yu, H. Yao, Y. Li, H. Xie, and X. Wang, "Joint modeling of dense and incomplete trajectories for citywide traffic volume inference," in *WWW*, 2019, pp. 1806–1817.

[13] X. Wang, M. Zhu, D. Bo, P. Cui, C. Shi, and J. Pei, "Am-gcn: Adaptive multi-channel graph convolutional networks," in *KDD*, 2020, pp. 1243–1253.

[14] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2016.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[16] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *KDD*, 2016, pp. 785–794.