Predicting Urban Region Heat via Learning Arrive-Stay-Leave Behaviors of Private Cars

Zhu Xiao, *Senior Member, IEEE*, Hao Li, Hongbo Jiang, *Senior Member, IEEE*, You Li, Mamoun Alazab, *Senior Member, IEEE*, Yongdong Zhu, and Schahram Dustdar, *Fellow, IEEE*

***Abstract*— Urban region heat refers to the extent of which people congregate in various regions when they travel to and stay in a specified place. Predicting urban region heat facilitates broad applications ranging from location-based services to intelligent transportation management. The region heat is essentially characterized by the 'arrive-stay-leave (ASL)' behaviors, while it is a challenging task to well capture the spatial-temporal evolution of region heat since the following issues remain: i) ASL behaviors of private cars is usually heterogeneous resulting in a hierarchical distribution of region heat. ii) Urban region heat contains complex spatial-temporal correlations hidden in ASL behaviors and how to collaboratively integrate them is challenging. To address these challenges, we propose a Hierarchical Spatial-Temporal Network (HierSTNet) to forecast urban region heat, which contains two representations, namely, grid region from micro perspective and node region from macro perspective. For the grids, three-dimension spatial and temporal convolutional network (3D-STCNN) is proposed to model multi-scale properties in temporal dimension of ASL behaviors. For the nodes, multi-head graph attention networks are utilized to model the periodicity and spatial heterogeneity among macro region. Hierarchical structures are designed for multi-view modeling spatial-temporal distribution of ASL behaviors, by which they capture small-scale features in micro regions and embeds the global representation into graph propagation. Finally, we design an interaction decoder layer to integrate the external factors and aggregate spatial-temporal information across hierarchical structures. Extensive experiments based on real-world private car trajectory dataset demonstrate the superiority and effectiveness of proposed framework.**

***Index Terms*— Urban region heat, private cars, arrive-stay-leave behaviors, trajectory data, hierarchical spatial-temporal network.**

I. Introduction

URBAN region heat refers to the extent of which people congregate in various regions when they travel to and stay in a specified place [1], taking some time to perform their daily activities. With strong spatial and temporal characteristics, the distribution of region heat reveals how people's travels reflect the formation and disappearance of urban hot zones. As such, predicting urban region heat is an intriguing problem from both researchers and policymakers since it benefits broad applications ranging from location-based services to Internet of Vehicles (IoVs) and intelligent transportation management [2], [3], [4], [5].

The region heat is essentially characterized by human mobility, more precisely, the '*arrive-stay-leave (ASL)*' behavior. To fulfill daily travel demands, people always *arrive* at a specified region, *stay* for a certain period participating in their activities, and then *leave* to next destination. Intuitively, a lot of people from many different parts of city move to and stay in several regions; indeed, their ASL behaviors are tightly connected with the spatial-temporal evolution of urban region heat. The ASL behaviors can be retrieved from various types of trajectory data, for instance, the smart card data[1] [6], taxi trajectory data [7], [8], mobile App data [9], [10], and private car trajectory data [11], [12], [13].

Specifically, we capitalize that the ASL behaviors retrieved from private cars are well suited to characterize the

Manuscript received 13 July 2022; revised 29 December 2022, 4 March 2023, and 7 May 2023; accepted 12 May 2023. Date of publication 24 May 2023; date of current version 4 October 2023. This work was supported in part by the NSFC under Grant U20A20181 and Grant 62272152, in part by the National Key Research and Development Program of China under Grant 2022YFE0137700, in part by the Humanities and Social Sciences Foundation of Ministry of Education under Grant 21YJCZH183, in part by the Science and Technology Innovation Program of Hunan Province under Grant 2021RC4023, in part by the Key Research and Development Program of Hunan Province under Grant 2021WK2001 and Grant 2022GK2020, in part by the Hunan Natural Science Foundation of China under Grant 2022JJ30171, in part by the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy [Shenzhen (SZ)] under Grant GML-KF-22-22 and Grant GML-KF-22-23, in part by the Shenzhen Science and Technology Program under Grant JCYJ20220530160408019, in part by the CAAI-Huawei MindSpore Open Fund, and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011915. The Associate Editor for this article was Y. Kamarianakis. *(Corresponding author: Hongbo Jiang.)*

Zhu Xiao, Hao Li, and Hongbo Jiang are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China, and also with the Shenzhen Research Institute, Hunan University, Shenzhen 518055, China (e-mail: zhxiao@hnu.edu.cn; lihao24@hnu.edu.cn; hongbojiang@hnu.edu.cn).

You Li is with the Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen 518060, China (e-mail: liyougis@gmail.com).

Mamoun Alazab is with the College of Engineering, IT and Environment, Charles Darwin University, Darwin, NT 0810, Australia (e-mail: mamoun.alazab@cdu.edu.au).

Yongdong Zhu is with the Zhejiang Lab, Hangzhou 311121, China (e-mail: zhuyd@zhejianglab.com).

Schahram Dustdar is with the Distributed Systems Group, TU Wien, 1040 Vienna, Austria (e-mail: dustdar@dsg.tuwien.ac.at).

Digital Object Identifier 10.1109/TITS.2023.3276704

[1]The smart card data is used to investigate passenger behavior and the demand characteristics of public transport, such as bus and subway.

spatial-temporal distribution of urban region heat. This assertion is explained as follows. For the smart card data, which records the trajectory of public transportation such as bus and subway, the travel routes and staying spots (e.g., the bus stations) are usually fixed, resulting in sparse recordings and not necessarily indicating passengers' staying destination and stay time. For the taxi trajectory data, taxi's routes are of randomness and their 'stay' in ASL behaviors are short for just pick-up or drop-off passengers. Mobile App data mainly collect the check-in information, which has no idea to record stay and leave behavior. Overall, these trajectory retrieved from public transportation, taxi and Mobile App data, are unable to collect complete ASL behaviors, thereby leading to unsatisfied performance on modeling urban region heat.

In comparison, the private car trajectory data provides individual travel preferences [14] and complete information of ASL behaviors [15], [16]. For example, people drive private cars and arrive at the destinations such as working places or leisure centers, staying for a certain amount of time to conduct activities and then leave to the next destination. Apparently, the essential components of ASL behavior, which includes the travel origin/destination and stay time, are explicitly recorded. More than that, the ownership of private cars has surged for years, which contributes the main body participating in urban automobiles. According to [17], nearly 88.6 % of the urban automobiles (223 millions) are private cars in China (and more than 76.6% in the European Union [18]). In this sense, large numbers of private cars travel across urban regions, their ASL behaviors lead to the formation and dissipation of urban region heat. In other words, the ASL behaviors obtained from private car trajectory data best reflect the spatial-temporal distribution of urban region heat [19]. As such, learning the latent temporal and spatial characteristics of ASL behaviors will provide a promising way to understand urban region heat.

In this work, we strive to predict urban region heat via learning ASL behaviors of private cars. Recent advances in deep learning networks can be applied to predict its spatial-temporal evolution. Recurrent Neural Networks (RNN) are usually used to model the temporal dependencies. For instance, Yang et al. [20] utilize long short-term memory (LSTM) networks to model time-series prediction problems. Li et al. [21] employ gated recurrent unit (GRU) networks to capture long-range sequential correlations. Convolution Neural Networks (CNN) are widely developed to build spatial topology of regular gird-based division [22]. After that, researchers model data as spatial-temporal graphs and utilize Graph neural networks (GNN) to deal with non-Euclidean correlations and extract spatial-temporal correlations. Despite the inspiring results, it is not straightforward to apply recent advances in deep learning to foresee urban region heat since following technical issues remain:

- *Hierarchical distribution of region heat*. People's congregation are heterogeneously distributed in terms of ASL behaviors [23]. The evolution pattern of ASL behaviors in urban hot zones obviously differ among areas at different time granularity. In other words, the distribution of urban region heat is hierarchical, where the feature and information in such structure play a vital role in predicting urban region heat. However, the current methods ignore the hierarchical representations and seldom utilize these information.
- *Complex spatial-temporal correlations in region heat*. For modeling the spatial dependencies, region heat is mutually correlated since the sum of heat streaming into a region is usually composed with the outflow nearby. For the temporal correlations, region heat changes dynamically over the time of a day, which are heavily influenced by ASL behavior and other external factors (e.g., meteorological conditions and event information) How to synchronously integrate the spatial correlations with temporal correlations for precise reference is still challenging.

To address these challenges, we propose a Hierarchical Spatial-Temporal Network (**HierSTNet**) for urban region heat prediction. First, aiming to acquire the hierarchical representations, the urban region construction block is designed to model the grid region and node region, which are used to explore the characteristics of region heat in micro and macro view, respectively. Moreover, we propose a three-dimension spatial and temporal convolutional network (3D-STCNN) in grid region. We divide the temporal properties into closeness, period and trend and capture the spatial dependencies at different time granularity. For the node region, we design a spatial-temporal graph network combined with multi-head attention to model the periodicity and spatial heterogeneity in global perspectives. Finally, we add an interaction decoder layer to integrate the external factors and aggregate spatial-temporal information across hierarchical structures. The main contributions in this paper are summarized as follows:

- We propose a Hierarchical Spatial-Temporal Network to predict urban region heat through capturing spatial-temporal correlations of ASL behaviors.
- We design a 3D-STCNN in grid region to capture spatial dependencies at different time granularity. On top of that, we employ a multi-head attention graph network to model the periodicity and spatial heterogeneity synchronously.
- We design an interaction decoder layer to integrate the external factors and aggregate spatial-temporal information across hierarchical structures.
- Extensive experiments are conducted based on real-world private car trajectory dataset. Experimental results demonstrate that our proposed HierSTNet outperforms the baselines with a roughly 7%-10% improvement.

The remainder of this paper is organized as follows. Section II presents the related works. In Section III, we introduce notations and preliminaries used in this paper, and then we formalize the problem of region heat forecasting. The details of the proposed model are presented in Section IV. Section V presents experimental results based on the real-world trajectory dataset. Finally, we conclude the paper in Section VI.

II. Related Works

Along with the rapid development of urbanization, a huge volumes of traffic data from urban areas related to human

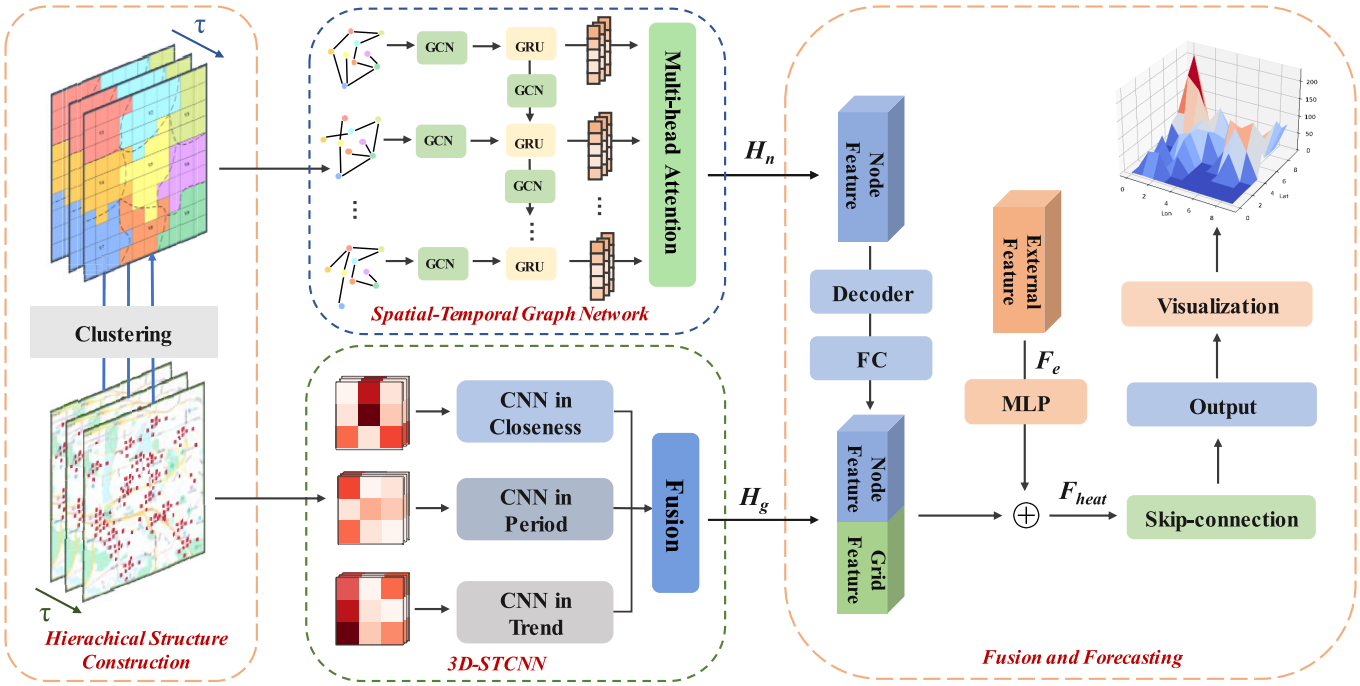


Fig. 1. The proposed HierSTNet for region heat prediction.

mobility have been increasingly collected and mined, which benefits broad application from location-based services to intelligent transportation management, such as traffic flow analysis and (Point of Interest) POI recommendation [2], [3].

Facing various applications and scenarios, continuous researches have paid much attention on mining spatial-temporal characteristic by using various types of trajectory data. Zhao et al. [6] used the bus transaction data to discover the patterns of passenger travel distribution and detect the abnormal passengers based on the empirical knowledge. Yu and He [24] used smart card data and revealed the spatial-temporal patterns of bus travel demand. Wang et al. [7] extracted pick-up and drop-off location points based on taxi data, and uses them as a basis to analyze the popular areas of passengers' trips. Yuan et al. [25] combined the trajectory data of Beijing's taxi car and POI data to analyze the transfer pattern and human mobility in region-scale. Moreover, researchers utilized the taxi data for next destination prediction. For instance, Rossi et al. [26] modeled the taxi drivers' behavior and encoded the semantics of visited locations by using geographical information. Zhang et al. [8] proposed a novel data embedding method for time-related feature pre-processing. Based on check-in data of location-based social apps, Li et al. [10] made recommendations of next POI for check-in users. Yang et al. [9] enriched the check-in data with potential visitors for check-in time prediction. Long et al. [13] investigated the problem of stay location prediction and explore travel regularity and preference for individual vehicles. Xiao et al. [27] derived the attractiveness of different urban areas to citizens by modeling the spatial distribution of the points of stop based on private car trajectory data. Liu et al. [15] attempted to predict private car flows in urban functional regions by exploiting spatiotemporal correlations of arrive-stay-leave (ASL) behaviors of private car users.

In the early stage, lots of efforts have been made in modeling the spatial-temporal characteristics of traffic trajectories. The statistical analysis method based on time series is an technical route, such as Autoregressive Integrated Moving Average model (ARIMA) and its variants [28], [29]. These methods rely on strong assumptions of linearity, which ignore the spatial dependencies among real traffic situation. Later, machine learning of data-driven models have been employed, such as K-nearest Neighbor (KNN) [30] and Support Vector Machine (SVR) [31] to handle nonlinearity in traffic data. However, these hand-crafted features are often shallow in structure with limited performance.

In recent years, deep neural networks provide a promising way to learn the spatial and temporal correlations. Existing deep-learning methods usually construct hybrid networks by combining different architectures to capture the hidden dependencies. Li et al. [22] utilized CNN to capture the spatial characteristics in grid region. Stepdeep [32] incorporate spatial and temporal filters into a 3D-CNN to predict spatial-temporal events. Although CNN excels at traffic data of regular division, it fail to model complex spatial topology structure. After that, graph neural networks [33] step into stage and break the restrictions which are more flexible and mightful to capture spatial correlation in Non-Euclidean data. Researchers model data as spatial-temporal graphs. Feng et al. [34] utilized graph convolutional network (GCN) to extract the spatial dependencies in graph frames. For modeling the temporal dependencies, RNN is powerful in capturing sequential information in time dimensions [35]. For instance, Yang et al. [20] used LSTM

networks to learn the regularity and preference hidden in traffic data. Li et al. [21] employed GRU to capture long-range sequential correlations.

In addition to CNN, GNN and RNN, attention mechanism networks are also introduced into extracting spatial and temporal features. Graph attention network (GAT) [36] is proposed for preserving hidden spatial dependencies and modeling the dynamics in traffic, which combines the self-attention with graph structures. To incorporate more information, Feng et al. [37] proposed RNN with attention mechanism to capture periodicity and preference based on distant hidden states. Guo et al. [38] designed a novel attention-based model, which combines the spatial-temporal attention mechanism with convolution operations.

Most frameworks stack two separate module to capture spatial and temporal information, respectively. Researchers are constantly trying to find ways to handle these correlations, simultaneously. Zhang et al. [39] constructed a CNN-based architecture and stacked it with residual network (ResNet) to jointly capture the spatiotemporal correlations. Yao et al. [40] combined CNN and LSTM to jointly model the nonlinear spatial-temporal correlations relations and capture the dynamics in traffic demand prediction tasks. Wu et al. [41] designed an adaptive graph structure together with gate mechanism for spatial-temporal modeling, which integrate diffusion graph convolution with 1-D dilated convolution. Zhang et al. [42] proposed a multitasak deep learning (MDL) framework to predict the node flow and edge flow, simultaneously. Moreover, rich contextual factors, such as weather, POI information and traffic events are included into modeling the forecasting tasks [43], [44]. Wang et al. [45] integrated GRU with Transformers to capture the local and global temporal dependencies, in which they proposed a position-wise attention to embed the external feature as auxiliary information. Yao et al. [46] proposed a semantics-enriched recurrent model to jointly learn the embeddings of multiple factors in a unified framework. Zhang et al. [47] embedded the semantic information into temporal modeling and propose a multi-graph convolution network for traffic demand forecasting.

Summarizing, existing methods ignore the hierarchical structures in urban traffic system. In addition, these methods for urban region heat prediction overlook the human travel behaviors of private cars, particularly the ASL behaviors. To resolve those issues, we construct a hierarchical structure to capture the spatial-temporal evolution of urban region heat via learning ASL behaviors of private cars.

III. Preliminaries

In this section, we introduce the basic definitions and present on the problem statement.

A. Definitions

Definition 1 (Urban Region): Urban region is divided into *grid region* and *node region*. In the grids, the city map is equally divided into $N = I \times J$ sub-regions according to the longitude and latitude. Each grid is denoted by r_n ($n \in [1, \cdots, N]$). For the nodes, firstly, we retrieve the historical private car trajectory into the grid map and extract the (Points of staying) PoSs information on the basis of grids divided. Secondly, the PoSs are clustered by spatial temporal-clustering of applications with noise (ST-DBSCAN) algorithm according to its density distribution. Each node can be regarded as a global region $r_i^k (k \in [1, \cdots, K])$, which stands for the i-th grid belongs to the k-th node region.

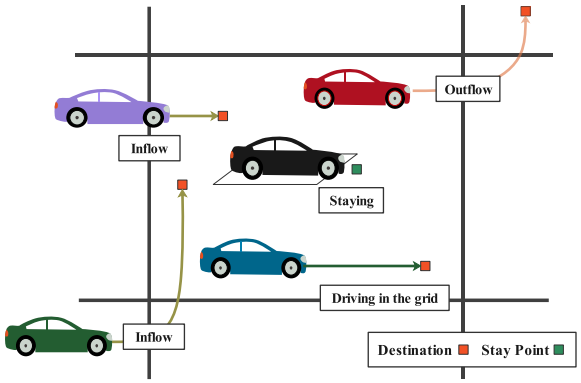


Fig. 2. The representation of ASL behavior in one region, which can be decomposed into following: 1)in/out flows, 2) those staying cars and 3) those cars are still driving in this region during the observed time window are also deemed as staying.

Definition 2 (ASL Behavior): From the private car trajectory $Tr = \{g_1, g_2 \cdots, g_T\}$ of dataset P, the process of ASL behavior can be observed at a sequence of time intervals $\tau = \{t_1, t_2 \cdots, t_T\}$, which contains consecutive spatial-temporal points $g_i = (lon_i, lat_i, t_i)$. Here (lon_i, lat_i) denotes geospatial coordinates. As shown in Fig. 2, the ASL behavior in one region can be decomposed into: *i*) in/out flows, *ii*) staying cars and *iii*) those cars are driving in this region during the observed time window. Let $s_{t,n}^{in}$, $s_{t,n}^{out}$, $s_{t,n}^{drive}$, and $s_{t,n}^{stay}$ denotes the amount of arrive (inflow) and leave (outflow), drive and stay in the region, respectively. These compositions from ASL behaviors are formulated as:

$$s_{t,n}^{in} = \sum_{Tr \in \mathbb{P}} \left| \{k > 1 \mid g_{t-1} \notin (r_n) \wedge g_t \in (r_n)\} \right|, \quad (1)$$

$$s_{t,n}^{out} = \sum_{Tr \in \mathbb{P}} \left| \{k \geq 1 \mid g_{t-1} \in (r_n) \wedge g_t \notin (r_n)\} \right|, \quad (2)$$

$$s_{t,n}^{drive} = \sum_{Tr \in \mathbb{P}} \left| \{k \geq 1 \mid g_{t-1} \in (r_n) \wedge g_t \in (r_n)\} \right|, \quad (3)$$

$$s_{t,n}^{stay} = \sum_{Tr \in \mathbb{P}} \left| \{k \geq 1 \mid g_{t-1} \in (r_n) \wedge g_t = g_{t-1}\} \right|, \quad (4)$$

where t denotes time interval, $g_k \in (r_n)$ denotes the trajectory points located in the region r_n.

Definition 3 (Region Heat): In this work, we strive to quantify the region-level heat by learning ASL behaviors from two perspectives, namely, the grid region and the node region. In the grids, as stated in previous definition, region heat is denoted as a time-ordered sequence of tensors, $H_t = \{h_{t,1}, \cdots, h_{t,N}\}$, $H_t \in \mathbb{R}^{I \times J}$. Let $h_{t,n}$ denote the quantified representation of region heat at timestamp t from ASL behaviors, which can be expressed by:

$$h_{t,n} = s_{t,n}^{in} - s_{t,n}^{out} + s_{t,n}^{drive} + s_{t,n}^{stay}, \quad (5)$$

TABLE I
SAMPLE OF TRIP DATA IN PRIVATE CARS DATASET

UserID	StratTime	StartLat	StartLon	StopTime	StopLon	StopLat
384002	2018/9/8 8:48	113.831488	22.6306	2018/9/8 11:51	113.863917	22.584313
384002	2018/9/8 12:33	113.864038	22.584268	2018/9/8 12:56	113.938868	22.507392
384002	2018/9/8 14:23	114.009668	22.610188	2018/9/8 15:04	114.11147	22.546782

TABLE II
MAIN NOTATIONS AND DEFINITIONS

Notations	Definitions
$s_{t,n}^{in}$	The amount of inflow of grid n
$s_{t,n}^{out}$	The amount of outflow of grid n
$s_{t,n}^{drive}$	The amount of driving of grid n
$s_{t,n}^{stay}$	The amount of staying of grid n
h	The heat value of the grid
h'	The converted total heat in the node region
H	The time-ordered sequence of tensors of grid region
H_c	The time-ordered sequence of tensors of node region
H_d	The hidden states representations in GRU module
H_g	The output representation of grid region
H_n	The output representation of node region
H_e	The external feature
H'	The converted node representation by grid representation H
H'_n	The output transfer representation of node region
H_{heat}	The concatenation representation of region heat

Then, we design a three-order tensor matrix to symbolize it at the past T time periods, where $H = \{H_1, \cdots, H_T\}$, $H \in \mathbb{R}^{T\times I\times J}$. In the node region, we cluster the grid input and define the node representation as $H_c = \{H_{c,1}, \cdots, H_{c,T}\}$, $H_c \in \mathbb{R}^{T\times K}$, which is the total of its grids.

B. Problem Statement

Given a sequence of historical grid input matrices H and node matrices H_c over the past T time slices, our goal is to learn a model Pre, accompanied with external feature H_e to collaboratively predict $\hat{H}_{T+t}$ in the future.

IV. METHODOLOGY

In this section, we introduce the proposed model in detail. The overall framework of HierSTNet is shown in Fig. 1, which mainly contains four parts: *i)* urban region construction; *ii)* 3D-STCNN on grid region; *iii)* spatial-temporal graph network on node region; *iv)* fusion and forecasting.

A. Urban Region Construction

1) Preprocessing: Region heat is largely quantified by ASL behaviors that are extracted from the private car trajectory data. The private car trajectory dataset [11] is collected from real-world urban scenarios via using vehicle positioning technologies [48], [49], [50]. As shown in Table I, the collected trajectory data are stored in individual trips, which contains the user ID, the start and stop time, the start and stop locations, etc. As for analyzing ASL behavior of one trip, the arrive matches the stop point and the leave denotes the start point of next trip, respectively. The stay behavior is reflected by the arrival time, the PoS (e.g., the arrival point), and the duration of stay time. For example, Table I presents the user arrived at a specified place at 11:51 then leave at 12:33, stayed for 42 minutes probably in a restaurant. The next ASL behavior lasts 88 minutes, maybe home for a naps. Each arrive and leave behavior both cause the heat change in corresponding region. In this work, the start and stop points of one ASL behavior are collectively deemed as the PoS to quantify the region heat. Considering the meaningless data, it is required to screen out the real and effective PoSs from private car trajectory since there will be traffic jams and other special circumstances during the driving process. Supplemented with time and speed thresholds, we find the PoSs of each ASL behavior in the trajectory and use the relative distance at the previous moment as the threshold, where the PoSs with a distance of less than 10 meters will be regarded as error value.

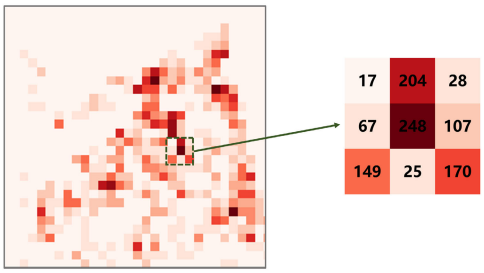


Fig. 3. Grid representation of region heat.

2) Grid Region Modeling: The city-wide area is equally divided into N grids, and each grid could be viewed as a snapshot of time-varying spatial map. Region heat can be detected by ASL behaviors generated by private cars. For a grid, the ASL behavior can be represented within a certain time interval of 1) arrive in the gird, 2) stay in the gird, 3) leave the grid and 4) those have not stayed or left are still driving in the grid. Accordingly, region heat in grids has been quantified by their amount, which is denoted as a time-ordered sequence of tensors. In details, the grid matrix is shown in Fig. 3.

3) Node Region Modeling: As shown in Fig. 4, node region is a wider representation that consists of several grids with same color. The modeling process is reported as follows:

First, we extract the PoSs from the private car trajectory, in which the PoSs set is denoted as $S = \{s_1, \cdots, s_n\}, s_i = (lon_i, lat_i, t_i)$. Then, we employ an improved ST-DBSCAN algorithm to cluster the density based on spatial-temporal distribution of real PoSs. After that, the grids are partitioned according to the density classification. In details, this method finds the grid corresponding to each staying point by its real geographic location and classifies the grid according to the

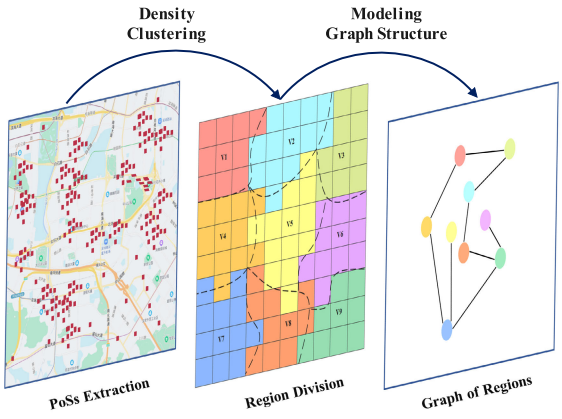


Fig. 4. Node region modeling.

staying point's density distribution. For example, if the grid r_i has the only staying point s_j, which is classified with k, we attribute r_i to be a sub-element of set v_k. Furthermore, if the grid has multiple PoSs, we select the largest proportion of classification as the result.

Moreover, we construct a connectivity-based graph $G = (V, E, A)$, where each clustered region is a vertex in G. The edges indicate the connectivity between node regions. The edge set E indicates the connectivity between two nodes, that is, an edge is created if there exits heat transfer starts from region i/j and to region j/i. The mathematical representation of connectivity graph is denoted by the adjacent matrix A.

B. 3D-STCNN on Grid Region

According to the grid region modeling, region heat at each grid in the certain time interval can be represented as the tensor H. For the short-term dependencies, region heat is easily affected by random event, like big sales in store, causing a sudden increase of the corresponding region's heat. For the long-term, region heat reveals similarity and periodicity. For instance, the morning and evening traffic congestion in rush hours tend to be similar on working days and it will be gradually severer as the seasonal winter comes. As such, Zhang et al. [51] intercepted three time series segments of the close, period and trend component along the time axis, which select different key frames to predict the time interval t. Inspired by this, we divide the temporal properties into closeness, period and trend and capture the spatial dependencies at different time granularity, which is defined as H_{close}, H_{period}, H_{trend}, respectively.

As shown in Fig. 5, we leverage a 3D spatial-temporal convolution-based network to process the historical grid heat tensor. To specify, 1D convolution is used to extract the time dependence, and 2D convolution can extract the spatial dependence. Compared with them, 3D convolution is more appropriate to capture the spatial and temporal correlation of ASL behavior. In such 3D network, we stack multiple 3D convolution units to combine the spatial and temporal information. In particular, we conduct three 3D convolution kernels to extract the spatial-temporal dependencies synchronously,

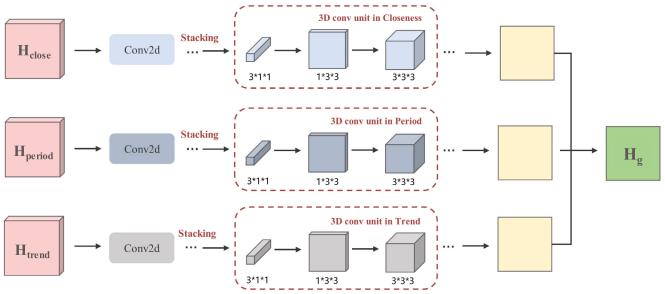


Fig. 5. The architecture of 3D-STCNN.

which includes time sensitive kernels, spatial sensitive kernels, and spatial-temporal kernels. The optimal amount of 3D convolution units is set to 64 (see experiment in Fig. 8(a) for details). In this way, we use 3D convolution kernel to connect multiple 2D tensors from the previous layer of the network and performs feature mapping operation. Formally, we obtain the (x, y, z) element $d_{i,j}^{x,y,z}$ of the convolution neuron matrix as follows, which is on j-th channel and generated by the i-th convolutional filter:

$$d_{i,j}^{x,y,z} = \left(\sum_{m} \sum_{e=0}^{E_i-1} \sum_{f=0}^{F_i-1} \sum_{g=0}^{G_i-1} d_{m(i-1)}^{(x+e),(y+f),(z+g)} \times w_{i,j,m}^{e,f,g}\right) + b_{i,j}, \tag{6}$$

where E_i, F_i, G_i is the size of the input feature in different dimension of i layer, $w_{i,j,m}^{e,f,g}$ is the weight value of the m-th feature channel mapped with (e, f, g) element in $i-1$ layer and $b_{i,j}$ denotes the bias term. After that, we take padding operation to keep the size of the output feature tensor matching the input.

To capture the temporal dependencies of different scales, the historical data of closeness, period and trend are fed into 3D-STCNN, which are denoted as H_{close}, H_{period}, H_{trend}, respectively. We propose using a parametric-matrix-based fusion to merge the spatial-temporal features:

$$H_g = H_{close} \odot W_c + H_{period} \odot W_p + H_{trend} \odot W_t. \tag{7}$$

Here $\odot$ denotes an element-wise multiplication operator, W_c, W_p, W_t are learnable parameters to represent the degree of adjustment affected by the different time dimensions.

C. Spatial-Temporal Graph Network on Node Region

After clustering the grids into node region, we convert the data into graph structures to capture the spatial-temporal dependencies in global perspectives. Accordingly, each node can be deemed a broader but irregular region division. In order to learn the depth representation among macro region, we adopts a spatial-temporal graph network with multi-attention to model the periodicity and spatial heterogeneity.

1) Capture Spatial Dependencies: Graph convolutional network (GCN) is the basic operation of extracting node features given its prior knowledge which has achieved great success in graph representing learning [34]. To propagate the spatial

location information among node regions, we carry out convolution on topological graph G, which achieves the process of filter in the frequency domain as follows,

$$g_\theta * Gx = g_\theta(L)x = g_\theta\left(U\Lambda^T U\right)x = Ug_\theta(\Lambda)U^T x, \quad (8)$$

where $L = U\Lambda^T U$ is the graph Laplacian matrix and $\Lambda = \text{diag}([\lambda_1, \ldots, \lambda_N]) \in \mathbb{R}^{N\times N}$. U is Fourier transform basis of G and g_θ is the convolution kernel. To simplify notations, we summarize the convolutional operation f_g as below:

$$H_c^{l+1} = f_g(A, H_c^l), \quad (9)$$

where H_c^l and H_c^{l+1} are the correlation matrix in layers l and $l+1$, respectively. A is the adjacency matrix, which is crucial in the aggregation of nodes and their neighbors in graph propagation.

2) Capture Temporal Dependencies: Gating mechanisms are excellent in controlling information flow through layers [21], [52]. Motivated by this, we combine GCN with gated recurrent units to capture the spatial-temporal dependencies. Specially, the internal structure of GRU includes a reset gate $r^{(t)}$ that helps to forget dispensable information and an update gate $z^{(t)}$ determines the hidden state passed by the previous node and the input of the current node. At the time step t, Let $H_c{}^{(t)}$ be the input and $H_d^{(t-1)}$ be the hidden states from previous step, we conduct GCN operation for both of them as below,

$$\tilde{H}_c{}^{(t)} = f_g\left(A, H_c{}^{(t)}\right), \quad (10)$$

$$\tilde{H}_d^{(t-1)} = f_g\left(A, H_d^{(t-1)}\right). \quad (11)$$

For each node v_i at time step t, its hidden states will be applied to control the flow of information saved at the next moment [53]. Each node performs the GRU operation independently and its learnable parameters are globally shared for all node regions. Then, the calculating of GRU can be expressed:

$$z^{(t)} = \sigma\left(\tilde{H}_c^{(t)}[i,:], \tilde{H}_d^{(t)}[i,:]\right), \quad (12)$$

$$r^{(t)} = \sigma\left(\tilde{H}_c^{(t)}[i,:], \tilde{H}_d^{(t)}[i,:]\right), \quad (13)$$

$$\tilde{H}_d^{(t)}[i,:] = \sigma\left(\tilde{H}_c^{(t)}[i,:], r^{(t)} \odot \tilde{H}_d^{(t)}[i,:]\right), \quad (14)$$

$$H_d^{(t)}[i,:] = \left(1 - z^{(t)}\right)\tilde{H}_d^{(t)}[i,:] + z^{(t)} \odot \tilde{H}_d^{(t)}[i,:], \quad (15)$$

where $H_d^{(t)}[i,:]$ is the output at time step t. The activation function $\sigma(x, y) = \tanh(Wx + Uy + b)$ and W, U, b are the learnable parameters.

3) Multi-Head Attention: Multi-head attention is widely adopted to integrate the sequence information from different representation dimensions [54]. For node v_i, the output sequence $H_d^{(i)} = \left\{H_d^{(1)}[i.:], \ldots, H_d^{(t)}[i.:]\right\}$ from GRU units is the input for the multi-attention. We first integrate multiple self-attention mechanisms and project the input into them with learnable parameters $W_Q, W_K, W_V \in \mathbb{R}^{I\times J\times d}$. Then, the scaled dot-product attention can be computed as:

$$H_n^{(i)} = \text{Softmax}\left(\frac{\left(H_d^{(i)}W_Q\right)\left(H_d^{(i)}W_K\right)^T}{\sqrt{d_k}}H_d^{(i)}W_V\right), \quad (16)$$

where $H_n^{(i)}$ is the output matrix and scale d_k is used to avoid the saturation of standardized function. We use multi-head attention to jointly learn the dependencies from diverse dimensions. In details, we have:

$$H_n = FC\left(\text{Concat}\left(H_n^{(1)}, H_n^{(2)}, \ldots, H_n^{(i)}\right)\right). \quad (17)$$

We concatenate multiple outputs and perform linear transformation through a fully-connected layer. With the multi-attention mechanism, each input token can be related to the tokens at other time steps. Finally, we generate the output as H_n.

D. Fusion and Forecasting

1) Interaction Decoder Module: Aiming to correspond the grid representations to the nodes, we propose a decoder layer to realize the interaction between grid and node region. First, we construct a transformation matrix $Q \in \mathbb{R}^{N\times K}$ to obtain the mapping between N grid and K node regions. If the grid i belongs to the node j, the corresponding element is set to 1. Formally,

$$Q_{i,j} = \begin{cases} 1, & \text{if } r_i \in v_j \\ 0, & \text{else}, \end{cases} \quad (18)$$

where r_i denotes the i-th grid and v_j is the j-th clustered node. Moreover, given the weight matrix Q_c column-wisely normalized by Q, the grid's feature can be flattened into one dimension:

$$H' = Q_c{}^T \cdot \text{Flatten}(H), \quad (19)$$

where $H \in \mathbb{R}^N$ denotes the grid representation and $H' \in \mathbb{R}^K$ is the converted node representation. Here we multiply it Q_c to obtain corresponding node representation $H' \in \mathbb{R}^K$. Furthermore, we define a novel transfer matrix Q^* as follows,

$$Q^*_{i,j} = \begin{cases} 0, & \text{if } r_i \notin v_j \\ \dfrac{h_i}{h'_j}, & \text{else}, \end{cases} \quad (20)$$

where h_i is the heat value in the i-th grid and h' is the total in the j-th node. Then, for the node feature $H_n \in \mathbb{R}^{K\times T}$, the transfer feature can be formulated as follows,

$$H'_n = \text{reshape}(Q^* \cdot H_n), \quad (21)$$

where the node transfer feature $H'_n \in \mathbb{R}^{N\times T}$. We reshape it to match the grid feature H_g. In this way, the representations of region heat learned by CNN and GCN are both integrated into one network.

2) Fusion Module: External factors such meteorological conditions and event information may have great impact on traffic condition [55], [56] For example, during the Chinese Spring Festival, the traffic flows in the city will have a cliff-like decline. Founded on this analysis, weather conditions, temperature and holiday are considered as external factors for region heat prediction. To reflect the different degrees of influences, weather conditions are categorized into 16 types (i.e., sunny, windy, storm and snowy) and each type is converted into a one-hot vector. Moreover, We empirically divide the temperature into 10 grades and the temperature difference of each grade is 5 °C since tiny changes (e.g., 1°C∼2°C) have little influence on people's activities and are overlooked. The categories of holiday are encoded into a binary vector. Finally, all the external factors are concatenated into a one dimensional tensor H_e.

In this part, we design a fusion module to capture the influences of external factors and fuse them with spatial-temporal dependencies of region heat in hierarchical structures. First, we fuse the grid feature H_g with the node transfer feature H_n'. The fusion is formulated as:

$$H_{\text{heat}} = W_g \odot H_g + W_r \odot H_n', \tag{22}$$

where H_{heat} is the concatenation representation of region heat with spatial-temporal dependencies. W_g, W_r are both learnable weights. Then, we embed the spatial-temporal feature H_{heat} and external factors H_{e}, generating a MLP with $Tanh$ activation function to fuse external factors with spatial-temporal data. Finally, the information for layers of different depth are passed for the output prediction $\hat{H}^{t+1}$:

$$\hat{H}^{t+1} = \text{Tanh}\left(H_{\text{heat}} + W_e \odot H_e\right), \tag{23}$$

where W_e is learnable parameter in MLP layer.

3) Loss Function: In the training process, our goal is to minimize the error between the predicted value $\hat{H}_i$ and the ground truth H_i. The loss function can be represented as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|H_i - \hat{H}_i\right|. \tag{24}$$

V. Experiments and Discussions

A. Dataset

In this section, we conduct the experiments based on the real-world private car trajectory dataset collected in Shenzhen, China. Table I details the sample of trip data, of which the data is collected from July to September 2018, containing 561,534 trips. After removing the incomplete and abnormal data, we retrieve the trajectory data between longitude of (113.48-114.49) and latitude of (22.45-22.84). We make use of location information and trip recordings of private cars to construct the topological graph. To protect user information and prevent privacy leakage, we have desensitized the sensitive information from the raw trajectories. We make the trajectory dataset used in this paper publicly available: https://github.com/HunanUniversityZhuXiao/PrivateCarTrajectoryData.

B. Baselines

We compare the proposed HierSTNet with the following baselines, of which the parameters have been fine-tuned from the original settings.

- **HA**: Historical Average is a classical approach using the average of historical data as the forecast output.
- **SARIMA** [29]: Seasonal ARIMA is a variant of ARIMA based on seasonal periodic improvement.
- **ConvLSTM** [40]: A classic deep learning combinatorial method, which utilize convolutional neural networks and long short-term memory networks to capture the spatial and temporal dependencies of data, respectively.
- **STGNN** [45]: A fine model integrates GCN and GRU, in which GCN is used to capture the spatial correlation and GRU is used to capture the temporal dependence of traffic data.
- **ASTGCN** [38]: A novel attention based model, which combines the spatial-temporal attention mechanism and the spatial-temporal convolution, simultaneously.
- **StepDeep** [32]: StepDeep utilizes the network based on CNN uses 3D convolution kernel to extract the spatiotemporal dependence.
- **MDL** [42]: A multitask deep-learning framework simultaneously predicts the node flow and edge flow throughout a spatial-temporal network.
- **GWNET** [41]: A spatial-temporal graph convolutional network, which integrates diffusion graph convolutions with 1-D dilated convolutions and develops a novel adaptive dependency matrix through node embedding.

C. Setting and Metrics

In the experiments, the setups are presented as follows.

- **Region division**. We first retrieve the location information from private car trajectory dataset and match them into city map. The Shenzhen city is equally divided into 32×32 grids. After that, we use ST-DBSCAN clustering algorithm to obtain corresponding node division. We set the spatial threshold as 800 meters, the time threshold as 30 minutes, and the minimum sample value $MinPts$ as 5. Finally, we obtain 96 node regions after clustering.
- **GRU**. In the graph network, we employ the GRU to extract temporal correlations and stack the GRU units with GCN to capture spatial-temporal dependencies, simultaneously. We set the GRU hidden units to 64.
- **Time segment length**. With the reference to the common demand of time series forecasting, the time segment length $\Delta\tau$ is set to 30min and 60min. It indicates that we divide the day into 24 and 48 time slices, respectively.
- **Training methods**. We employ Adam function to optimize the model and perform all weight updated during the training process. We set the initial learning rate to 0.005 and the batch size to 32. Moreover, the training iteration is set to 200. We adopt Min-Max normalization to process the data. Each dataset is splitted into 80% for training, 10% for testing and 10% for testing with chronological order. We train models in the training-set

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT MODELS

Method	30 minutes			60 minutes		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
HA	8.63	17.62	14.98%	8.77	23.09	18.06%
SARIMA	6.23	13.47	12.23%	6.32	18.37	16.42%
ConvLSTM	4.33	7.83	10.93%	5.28	11.83	13.01%
STGNN	3.23	7.79	10.09%	3.59	10.26	12.98%
StepDeep	2.75	7.21	9.81%	2.94	9.42	13.09%
ASTGCN	2.71	7.17	9.72%	2.90	9.37	12.98%
GWNET	2.57	7.11	9.63%	2.83	9.29	12.91%
MDL	2.68	**7.05**	9.65%	2.84	9.38	12.91%
HierSTNet	**2.42**	7.08	**9.58**%	**2.70**	**9.23**	**12.89**%

and test the model on the test-set according to the optimal parameters of the validation-set.

The experiments are conducted based on the MindSpore framework platform. The performance of each method is measured by Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The MAE is expressed in Eq. (24). Given predicted value $\hat{y}_i$ and corresponding ground truth y_i, the calculations of RMSE and MAPE can be formulated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}, \tag{25}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|y_i - \hat{y}_i\right|}{y_i}. \tag{26}$$

D. Experiment Results

1) Prediction Results: For the overall performance of region heat prediction, we conduct a numerical comparison by using various metrics in terms of MAE, RMSE and MAPE, including the forecasting of next 30 minutes and 60 minutes in Shenzhen dataset. The evaluation is reported in Table III, where the best is marked in bold.

The performance of naive HA and SARIMA methods is worse as they only consider assumptions of linear correlations over time series while ignoring the spatial dependencies among real traffic situation. Compared with traditional methods, deep learning-based models achieve better performance, which demonstrates their superior capacity on learning the nonlinear spatial-temporal correlations. Methods such as ConvLSTM and StepDeep have limited performance since they fail to capture the non-Euclidean correlations among complex regions. STGNN and ASTGCN highly rely on predefined graph which leads to a bad performance. More recent baselines MDL and GWNET obtain competitive results. These methods are inferior to our proposed model due to the following reasons. For the former, it benefits from the strategies for multi-task learning to capture spatial-temporal correlations, in which such training strategy can alleviate the error propagation. The GWNET method benefits from that it designs an adaptive graph to model relationships between nodes and construct different components to model temporal and spatial correlations respectively. However, they only consider the nearest time steps for capturing local correlations, while the long-term correlations and external factors are omitted. The proposed HierSTNet achieves the best results and verifies its superiority based on the following reasons. First, the design of hierarchical modeling helps our model capture the subtler correlations in both micro and macro perspective. Second, it applies a more effective strategy in hierarchical structure to mine the temporal characteristics, in which we model three temporal properties in 3D-STCNN and combine multi-head attention mechanism in graph frames to integrate the sequence information from diverse dimensions. Moreover, we incorporate the external factors into prediction, which boosts the prediction performance.

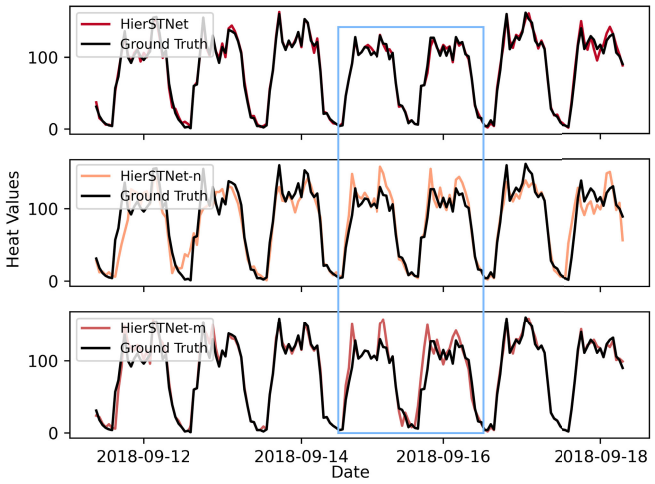


Fig. 6. Ablation study.

2) Ablation Study: To estimate the effect of hierarchical structure, we design two variants: HierSTNet-m and HierSTNet-n:

- **HierSTNet-m**: In this variant, we remove the multi-head attention block to verify the performance of attention for capturing global temporal dependencies.
- **HierSTNet-n**: In this variant, we remove the whole node block and interaction decoder module, i.e., only employing grid region module to demonstrate the importance of hierarchical structure.

We randomly choose a grid region from Shenzhen dataset to conduct the ablation analysis. Fig. 6 presents the predicted results of HierSTNet and its variant against ground truth from 9/12/2018 to 9/18/2018. It can be observed that the components of hierarchical structure are effective as HierSTNet is strongly accurate in tracing the ground truth curves, while HierSTNet-m is slightly worse. There is a large deviation between HierSTNet-n toward the ground truth since the spatial correlations among single network cannot generate a fine-grained prediction for a specified grid. Moreover, neither of these two variants that removed the multi-head attention seem to be sensitive to heat changes in the weekend (see the curve with a blue box), in which they cannot accurately predict the heat change over the weekend.

TABLE IV
PERFORMANCE COMPARISON OF ABLATION VARIANTS

Method	30 minutes			60 minutes		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
HierSTNet-n	2.82	7.35	9.95%	3.12	9.44	13.21%
HierSTNet-m	2.63	7.16	9.70%	2.88	9.37	13.03%
HierSTNet	2.42	7.08	9.58%	2.70	9.23	12.89%

TABLE V
THE COMPARISON OF COMPUTATION TIME

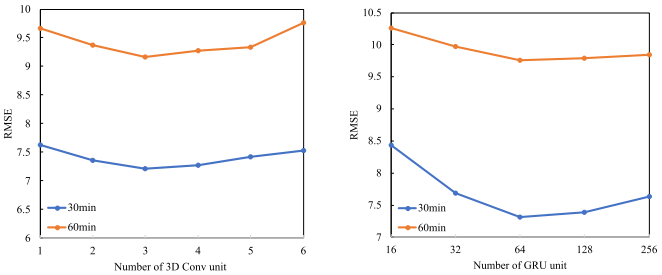
Method	Computation Time	
	Training(s/epoch)	Inference(s)
ConvLSTM	23.04/25.33	2.40/3.09
STGNN	34.23/41.88	4.00/4.76
StepDeep	77.65/99.38	5.61/15.96
ASTGCN	85.70/115.93	6.98/14.23
MDL	110.27/146.80	8.13/14.84
GWNET	175.75/205.05	12.31/35.15
HierSTNet	91.07/114.05	6.13/19.61

In addition, we compare the overall performance including the prediction of 30 minutes and 60 minutes. Compared to HierSTNet-n, HierSTNet-m achieves better prediction results. This indicates the importance and effectiveness of hierarchical structures. The lack of a multi-headed attention component degrades prediction performance, but not nearly as bad as the ablation of hierarchical components.

3) Computation Time: In addition, we compare the computation time of StepDeep, MDL, ASTGCN and GWNET with HierSTNet to analyze the time complexity. The results are shown in Table V. We can find out, although HierSTNet is slower than StepDeep, ASTGCN and slightly slower than HierSTNet-f, it has far better accuracy than those models. Compared to the methods with single structures, HierSTNet models region heat in hierarchical structure expressed as grids and nodes, by which the corresponding training cost is higher but the performance is correspondingly more outstanding. Compared with the latest methods, like GWNET and MDL, HierSTNet is nearly two times faster. The reason behind is the calculation and training of multitask learning in MDL and adaptive learning in GWNET are more complicated, which proves the superiority of hierarchical components with a balanced overhead and the cost of hierarchical components is deserved.

4) Hyper-Parameter Sensitivity: We conduct the sensitivity study to discuss how different choices of parameters affect the performance of the proposed HierSTNet. We report the impact of the number of 3d convolution units and the number of GRU units using RMSE. Moreover, we examine the influence of different prediction time steps from 1 to 12. Each time step denotes 30 minutes. We summarize the results and have the following observations:

- Fig. 8(a) shows that the resulting RMSE decreases with the increase of the number of layers as the number of layers is within 3 layers. When the number of layers exceeds 3, the RMSE increases instead of decreasing caused by the overfitting of networks. This is suggested that the optimal number of layers for 3D spatial-temporal convolution is 3.
- Fig. 8(b) shows that the performance only achieves minor improvement or even degraded gradually when the number of GRU units is larger than 64. It indicates that the optical value of GRU units is around 64.
- We compare the multi-step prediction performance of HierSTNet against baseline models as depicted in Fig. 8. It can be seen that, as wide the prediction horizon achieves, the corresponding difficulty of prediction is becoming greater and the prediction performance is gradually worsen. HierSTNet achieves the start-of-the-art performance whether long-term or short-term prediction, where the reason behind is that our method has great generalization ability to exploit long-interval context.



(a) Number of 3D conv units (b) Number of GRU units

Fig. 7. Parameter sensitivity on Shenzhen dataset.

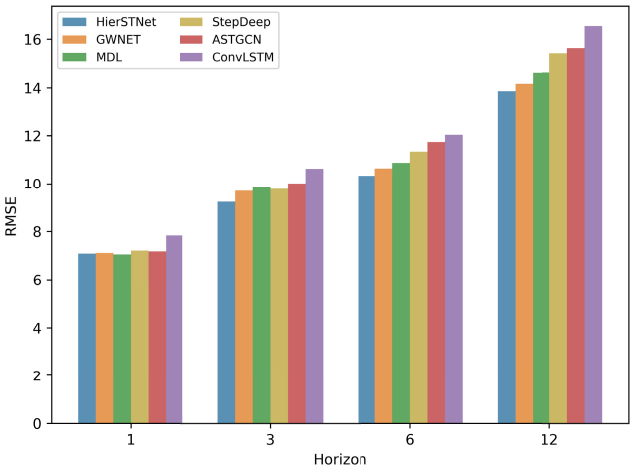
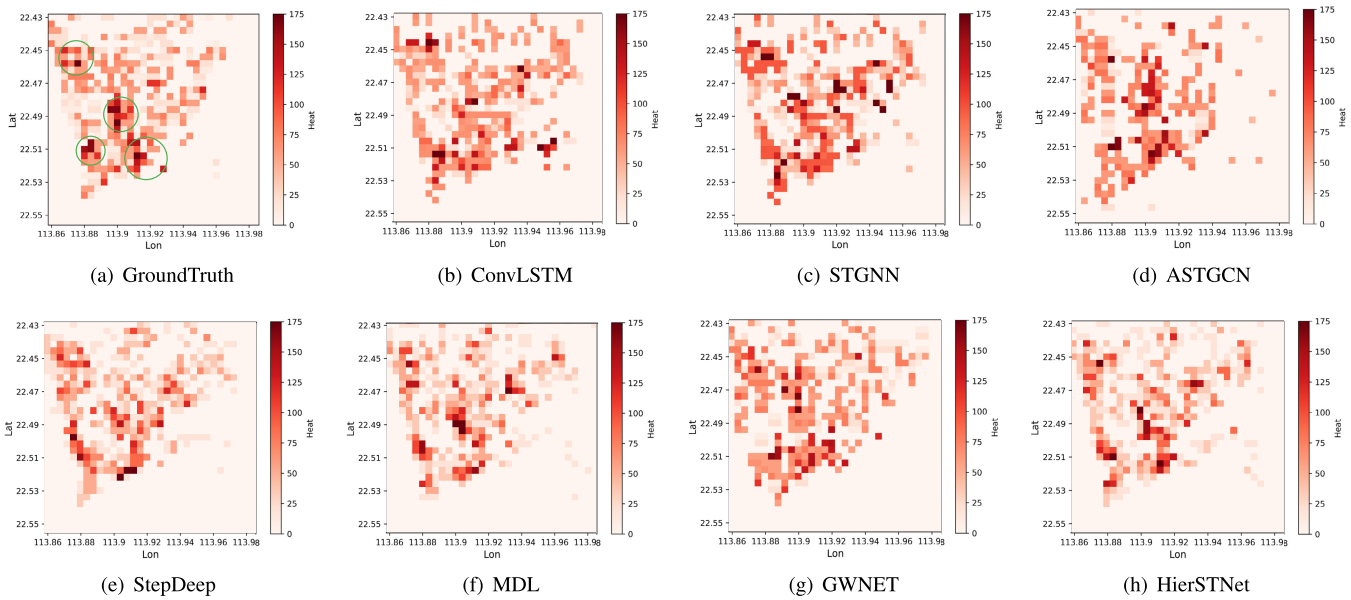


Fig. 8. Performance comparison under different prediction time steps.

5) Case Study: To better understand HierSTNet, we carry out a case study to interpret how does HierSTNet handle the spatial-temporal dependencies in hierarchical structures compared with several competitive baselines. We discuss the predicted results as follows.

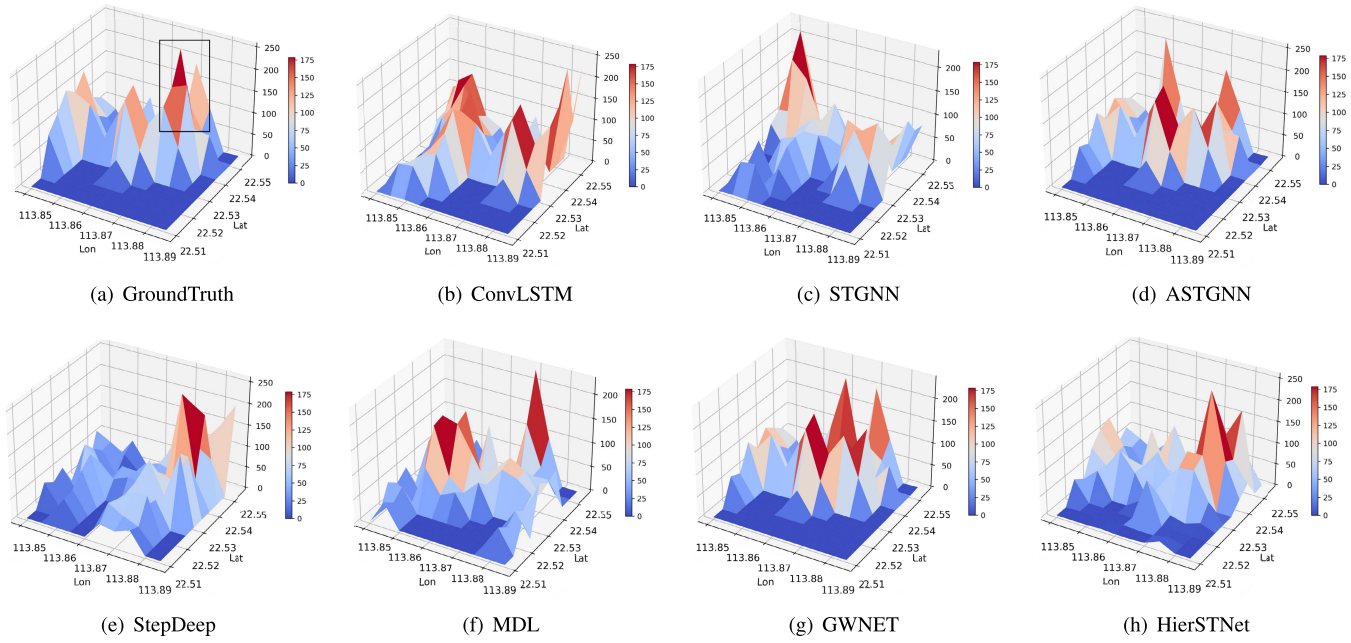
First, we conduct analysis on the region heat comparison. Fig. 9 illustrates the 2D heat map comparison between the predicted value and the real value of the region heat in



(a) GroundTruth (b) ConvLSTM (c) STGNN (d) ASTGCN

(e) StepDeep (f) MDL (g) GWNET (h) HierSTNet

Fig. 9. Region heat at 11:00 on August 23, 2018 (2D view).



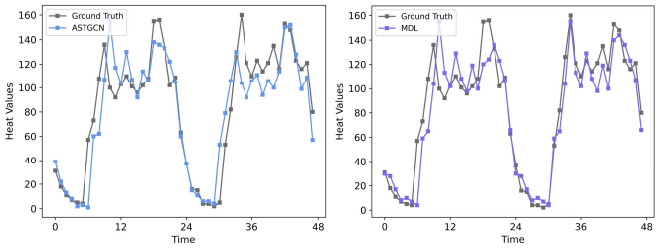
(a) GroundTruth (b) ConvLSTM (c) STGNN (d) ASTGNN

(e) StepDeep (f) MDL (g) GWNET (h) HierSTNet

Fig. 10. Region heat at 19:00 on August 23, 2018 (3D view).

Shenzhen dataset during the morning rush hours of 11:00 on August 23, 2018. In 2D map, we conduct spatial analysis among 32 × 32 grids, in which the horizontal and vertical coordinates are the index number. As depicted in Fig. 10(a), there are four congregated hot zones marked in green circular. The darker the color, the higher the congregation degree and accordingly larger the region heat values in corresponding grid. Fig. 10(b) and Fig. 10(c) present the results of ConvLSTM and STGNN, respectively. It can be seen these results are over-predicting. As depicted in Fig. 10(e) and Fig. 10(d), StepDeep and ASTGCN cannot fully predict the hot zones occur during the selected time period. These models with single structure lack of global consideration, failing to fully capture the spatial distribution of region heat. MDL and GWNET obtain relatively better results. Compared with them, the proposed HierSTNet generates a smoother prediction and its spatial distribution of region heat is most similar to ground truth.

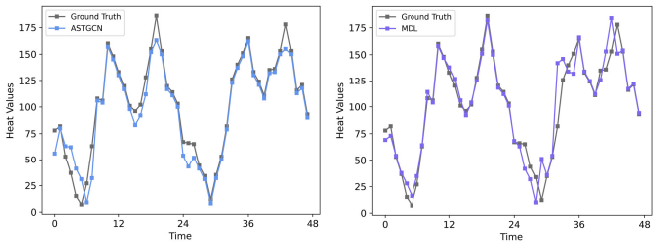
Second, we choose 10 × 10 grids near Nanshan District, Shenzhen to conduct spatial analysis in 3D point of view, where many high-technology enterprises are located in selected region. In 3D map, Fig. 11(a) presents the ground



(a) ASTGCN (b) MDL

(c) GWNET (d) HierSTNet

Fig. 11. Prediction of grid (16, 12) in Nanshan District, Shenzhen dataset.



(a) ASTGCN (b) MDL

(c) GWNET (d) HierSTNet

Fig. 12. Prediction of grid (10, 22) in Furong District, Changsha dataset.

truth at 19:00 on August 23, in which the black rectangular area denotes heat aggregation center. When the off-duty hour is approaching, most office employees are leaving work and the heat is spreading from the office building to the surrounding area. As such, we can observe the peak arising in the aggregation center and potential peaks around. From the 3D visualization of Fig. 11(b), Fig. 11(c) and Fig. 11(d), ConvLSTM, STGNN and ASTGCN completely predicted the wrong location of aggregation region. Compared to these methods, StepDeep behaves slightly better but its predicted peaks regions are still not so accurate as shown in Fig. 11(e). As depicted in Fig. 11(g) and Fig. 11(f), GWNET ignores the potential peaks in upper left and the prediction around peaks in MDL has a deviation against true values. Compared with these methods, the shape of aggregation peaks in HierSTNet best match the real condition since the peak value and those potential peaks are the closest to the ground truth.

Moreover, we choose a specific grid (16, 12) of Nanshan District from Shenzhen dataset. We present the prediction results and true values for a more straightforward loss comparison. Fig. 11 depicts the comparisons starting from August 24 for the next two days. Specifically, we observe there exist several heat changes, i.e., the plunge of heat in the early morning hours and the rise in the commute and off-work rush. MDL and ASTGCN cannot properly capture temporal dependencies, thereby leading to a degraded performance. GWNET performs slightly better than MDL and ASTGCN.

To comprehensively evaluate our model, we conduct experiments based on the trajectory data from different scenario. Specifically, we retrieve the trajectory data from the Changsha dataset, i.e., the trajectory data ranging from longitude of (112.95-113.18) and latitude of (28.01-28.22), which is collected from July to September 2018, Changsha City. All the dividing and training strategies remain unchanged and after clustering, 149 node regions were obtained in the Changsha dataset. Fig. 12 shows the predictions of the chosen grid (10, 22) near Changsha Railway Station from August 24 to August 26. The selected grid in Changsha has similar trend of region heat change, in which the peak values are a little higher than that in Shenzhen dataset. In both scenarios, HierSTNet are strongly accurate in tracing the ground truth curves, profiting from the design of hierarchical networks to extract the temporal feature from different perspectives of grid region and node region.

VI. Conclusion

In this paper, we investigate urban region heat via learning arrive-stay-leave (ASL) behavior of private cars. Specifically, we propose a Hierarchical Spatial-Temporal Network (HierSTNet) to forecast urban region heat, which contains two representations, namely, grid region and node region. For the grid region, the 3DSTCNN is proposed to model multi-scale properties in temporal dimension of ASL behavior. For the other, multi-head graph attention networks are utilized to model the periodicity and spatial heterogeneity among node regions. Benefiting from the design of interaction decoder layer, we integrate the external factors and aggregate spatial-temporal information across hierarchical structures. The proposed HierSTNet is evaluated on a real-world private car trajectory dataset to demonstrate its superiority and effectiveness.

Notably, the property of function areas in the city directly affect people's willingness to travel. In the future, we will consider a series of transportation factors and strive to collect sufficient multisource data such as regions of interests (ROIs), which will provide an important complementary in exploring the aggregation effect. Besides, we will study its impact on the distribution of region heat. Moreover, it is promising to mine the latent spatial relationship between regions under complex and giant traffic systems.

REFERENCES

[1] D. Wang et al., "Stop-and-wait: Discover aggregation effect based on private car trajectory data," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3623–3633, Oct. 2019.

[2] A. L. Alfeo, M. G. C. A. Cimino, S. Egidi, B. Lepri, and G. Vaglini, "A stigmergy-based analysis of city hotspots to discover trends and anomalies in urban transportation usage," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 7, pp. 2258–2267, Jul. 2018.

[3] Y. Gong, Z. Li, J. Zhang, W. Liu, and Y. Zheng, "Online spatio-temporal crowd flow distribution prediction for complex metro system," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 2, pp. 865–880, Feb. 2022.

[4] Z. Xiao, J. Shu, H. Jiang, G. Min, H. Chen, and Z. Han, "Perception task offloading with collaborative computation for autonomous driving," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 2, pp. 457–473, Feb. 2023.

[5] X. Dai et al., "A learning-based approach for vehicle-to-vehicle computation offloading," *IEEE Internet Things J.*, vol. 10, no. 8, pp. 7244–7258, Apr. 2023.

[6] J. Zhao, Q. Qu, F. Zhang, C. Xu, and S. Liu, "Spatio-temporal analysis of passenger travel patterns in massive smart card data," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3135–3146, Nov. 2017.

[7] P. Wang, G. Liu, Y. Fu, Y. Zhou, and J. Li, "Spotting trip purposes from taxi trajectories: A general probabilistic model," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 3, pp. 1–26, Dec. 2017.

[8] X. Zhang, Z. Zhao, Y. Zheng, and J. Li, "Prediction of taxi destinations using a novel data embedding method and ensemble learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 68–78, Jan. 2020.

[9] G. Yang, Y. Cai, and C. K. Reddy, "Spatio-temporal check-in time prediction with recurrent neural network based survival analysis," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 2976–2983.

[10] Y. Li, T. Chen, Y. Luo, H. Yin, and Z. Huang, "Discovering collaborative signals for next POI recommendation with iterative Seq2Graph augmentation," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1491–1497.

[11] Z. Xiao et al., "TrajData: On vehicle trajectory collection with commodity plug-and-play OBU devices," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 9066–9079, Sep. 2020.

[12] C. Liu, Z. Xiao, D. Wang, M. Cheng, H. Chen, and J. Cai, "Foreseeing private car transfer between urban regions with multiple graph-based generative adversarial networks," *World Wide Web*, vol. 25, no. 6, pp. 2515–2534, Mar. 2022.

[13] W. Long et al., "Location prediction for individual vehicles via exploiting travel regularity and preference," *IEEE Trans. Veh. Technol.*, vol. 71, no. 5, pp. 4718–4732, May 2022.

[14] J. Liang, J. Tang, F. Liu, and Y. Wang, "Combining individual travel preferences into destination prediction: A multi-module deep learning network," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 13782–13793, Aug. 2022.

[15] C. Liu et al., "Exploiting spatiotemporal correlations of arrive-stay-leave behaviors for private car flow prediction," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 2, pp. 834–847, Mar. 2022.

[16] Z. Xiao, H. Xiao, H. Jiang, W. Chen, H. Chen, and A. C. Regan, "Exploring human mobility patterns and travel behavior: A focus on private cars," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 5, pp. 129–146, Sep. 2022.

[17] NBS. (2021). *China Statistical Yearbook 2021*. [Online]. Available: http://www.stats.gov.cn/tjsj/ndsj/2021/indexch.htm

[18] *European Energy and Transport Trends to 2030*, Eur. Commission, Apr. 2008. [Online]. Available: http://ec.europa.eu/energy/observatory/trends_2030/doc/trends_to_2030_update_2007.pdf

[19] Z. Xiao et al., "Understanding private car aggregation effect via spatio-temporal analysis of trajectory data," *IEEE Trans. Cybern.*, vol. 53, no. 4, pp. 2346–2357, Apr. 2023.

[20] M. Yang, J. Liu, L. Chen, Z. Zhao, X. Chen, and Y. Shen, "An advanced deep generative framework for temporal link prediction in dynamic networks," *IEEE Trans. Cybern.*, vol. 50, no. 12, pp. 4946–4957, Dec. 2020.

[21] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16.

[22] D. Li, J. Zhang, Q. Zhang, and X. Wei, "Classification of ECG signals based on 1D convolution neural network," in *Proc. IEEE 19th Int. Conf. e-Health Netw., Appl. Services (Healthcom)*, Oct. 2017, pp. 1–6.

[23] B. An, A. Vahedian, X. Zhou, W. N. Street, and Y. Li, "HintNet: Hierarchical knowledge transfer networks for traffic accident forecasting on heterogeneous spatio-temporal data," in *Proc. SIAM Int. Conf. Data Mining*, 2022, pp. 334–342. [Online]. Available: https://epubs.siam.org/doi/abs/10.1137/1.9781611977172.38

[24] C. Yu and Z.-C. He, "Analysing the spatial–temporal characteristics of bus travel demand using the heat map," *J. Transp. Geography*, vol. 58, pp. 247–255, Jan. 2017.

[25] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2012, pp. 186–194.

[26] A. Rossi, G. Barlacchi, M. Bianchini, and B. Lepri, "Modelling taxi drivers' behaviour for the next destination prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 7, pp. 2980–2989, Jul. 2020.

[27] Z. Xiao, H. Fang, H. Jiang, J. Bai, V. Havyarimana, and H. Chen, "Understanding urban area attractiveness based on private car trajectory data using a deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 12343–12352, Aug. 2022.

[28] Q. Ye, W. Y. Szeto, and S. C. Wong, "Short-term traffic speed forecasting based on data recorded at irregular intervals," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1727–1737, Dec. 2012.

[29] S. Shekhar and B. M. Williams, "Adaptive seasonal time series models for forecasting short-term traffic flow," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2024, no. 1, pp. 116–125, Jan. 2007.

[30] H. van Lint and C. P. I. van Hinsbergen, "Short-term traffic and travel time prediction models," *Transp. Res. Circular*, 2012. [Online]. Available: https://onlinepubs.trb.org/onlinepubs/circulars/ec168.pdf

[31] C.-H. Wu, J.-M. Ho, and D. T. Lee, "Travel-time prediction with support vector regression," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 4, pp. 276–281, Dec. 2004.

[32] B. Shen, X. Liang, Y. Ouyang, M. Liu, W. Zheng, and K. M. Carley, "StepDeep: A novel spatial–temporal mobility event prediction framework based on deep neural network," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Jul. 2018, pp. 724–733.

[33] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2017, *arXiv:1609.02907*.

[34] X. Feng, J. Guo, B. Qin, T. Liu, and Y. Liu, "Effective deep memory networks for distant supervised relation extraction," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 4002–4008.

[35] S. Wang, J. Cao, and P. S. Yu, "Deep learning for spatio-temporal data mining: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 8, pp. 3681–3700, Aug. 2022.

[36] B. Sun, D. Zhao, X. Shi, and Y. He, "Modeling global spatial–temporal graph attention network for traffic prediction," *IEEE Access*, vol. 9, pp. 8581–8594, 2021.

[37] J. Feng et al., "DeepMove: Predicting human mobility with attentional recurrent networks," in *Proc. World Wide Web Conf.*, 2018, pp. 1459–1468.

[38] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial–temporal graph convolutional networks for traffic flow forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 922–929.

[39] J. Zhang, Y. Zheng, D. Qi, R. Li, X. Yi, and T. Li, "Predicting citywide crowd flows using deep spatio-temporal residual networks," *Artif. Intell.*, vol. 259, pp. 147–166, Jun. 2018.

[40] H. Yao et al., "Deep multi-view spatial–temporal network for taxi demand prediction," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[41] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph WaveNet for deep spatial–temporal graph modeling," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1907–1913.

[42] J. Zhang, Y. Zheng, J. Sun, and D. Qi, "Flow prediction in spatio-temporal networks based on multitask deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 3, pp. 468–478, Mar. 2020.

[43] Z. Lin, J. Feng, Z. Lu, Y. Li, and D. Jin, "DeepSTN+: Context-aware spatial–temporal neural network for crowd flow prediction in metropolis," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 1020–1027.

[44] B. Xia, C. Wong, Q. Peng, W. Yuan, and X. You, "CSCNet: Contextual semantic consistency network for trajectory prediction in crowded spaces," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108552.

[45] X. Wang et al., "Traffic flow prediction via spatial temporal graph neural network," in *Proc. Web Conf.* New York, NY, USA: Association for Computing Machinery, 2020, pp. 1082–1092.

[46] D. Yao, C. Zhang, J. Huang, and J. Bi, "SERM: A recurrent model for next location prediction in semantic trajectories," in *Proc. ACM Conf. Inf. Knowl. Manag.*, Nov. 2017, pp. 2411–2414.

[47] L. Zhang et al., "Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 3656–3663.

[48] Z. Xiao, Y. Chen, M. Alazab, and H. Chen, "Trajectory data acquisition via private car positioning based on tightly-coupled GPS/OBD integration in urban environments," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9680–9691, Jul. 2022.

[49] V. Havyarimana, Z. Xiao, A. Sibomana, D. Wu, and J. Bai, "A fusion framework based on sparse Gaussian–Wigner prediction for vehicle localization using GDOP of GPS satellites," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 2, pp. 680–689, Feb. 2020.

[50] Z. Xiao et al., "Toward accurate vehicle state estimation under non-Gaussian noises," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10652–10664, Dec. 2019.

[51] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1655–1661.

[52] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 933–941.

[53] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," 2015, *arXiv:1506.05163*.

[54] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.

[55] W. Long et al., "Unified spatial–temporal neighbor attention network for dynamic traffic prediction," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 1515–1529, Feb. 2023.

[56] H. Yao, X. Tang, H. Wei, G. Zheng, and Z. Li, "Revisiting spatial–temporal similarity: A deep learning framework for traffic prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 5668–5675.

Zhu Xiao (Senior Member, IEEE) received the M.S. and Ph.D. degrees in communication and information systems from Xidian University, China, in 2007 and 2009, respectively.

From 2010 to 2012, he was a Research Fellow with the Department of Computer Science and Technology, University of Bedfordshire, U.K. He is currently a Full Professor with the College of Computer Science and Electronic Engineering, Hunan University, China. His research interests include wireless localization, the Internet of Vehicles, and intelligent transportation systems. He is currently an Associate Editor of the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

Hao Li received the B.S. degree from Zhejiang Sci-Tech University, China, in 2021. He is currently pursuing the M.S. degree in information and communication engineering with Hunan University, Changsha, China.

His research interests include the Internet of Vehicles and trajectory big data mining.

Hongbo Jiang (Senior Member, IEEE) received the Ph.D. degree from Case Western Reserve University in 2008.

He is currently a Full Professor with the College of Computer Science and Electronic Engineering, Hunan University. He was a Professor with the Huazhong University of Science and Technology. His research interests include computer networking, especially algorithms and protocols for wireless and mobile networks. He is an elected member of the Academia Europaea and a fellow of IET, BCS, and AAIA. He was an Editor of IEEE/ACM TRANSACTIONS ON NETWORKING, an Associate Editor of IEEE TRANSACTIONS ON MOBILE COMPUTING, and an Associate Technical Editor of *IEEE Communications Magazine*.

You Li received the Ph.D. degree in geographic information science from the Wuhan University of China in 2017. He is currently an Associate Researcher with the Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China. His research interests include LiDAR processing, 3D GIS, and 3D road environment modeling.

Mamoun Alazab (Senior Member, IEEE) received the Ph.D. degree in computer science from the School of Science, Information Technology and Engineering, Federation University Australia. He is currently an Associate Professor with the College of Engineering, IT and Environment, Charles Darwin University, Australia. He is also a cyber security researcher and a practitioner with industry and academic experience. He has more than 200 research papers in many international journals and conferences. His research interests include cyber security and digital forensics of computer systems with a focus on cybercrime detection and prevention. He is the Founding Chair of the IEEE Northern Territory (NT) Subsection.

Yongdong Zhu received the bachelor's degree from Xi'an Jiaotong University, Xi'an, China, in 1997, and the Ph.D. degree from the University of Essex, Colchester, U.K., in 2007.

He is currently a Professor with the Institution of Intelligent Network, Zhejiang Lab, Zhejiang, China. His current research interests include next-generation mobile communications, mobile edge network architecture, vehicular communication networks, and the Internet of Things.

Schahram Dustdar (Fellow, IEEE) received the Ph.D. degree in business informatics from the University of Linz, Austria, in 1992. He is currently a Full Professor of computer science (informatics) with a focus on internet technologies heading the Distributed Systems Group, TU Wien, Wein, Austria. He has been a member of the IEEE Conference Activities Committee (CAC) since 2016, the Section Committee of Informatics of the Academia Europaea since 2015, and the Academia Europaea (The Academy of Europe) of Informatics Section since 2013. He was a recipient of the ACM Distinguished Scientist Award in 2009 and the IBM Faculty Award in 2012. He has been the Chairperson of the Informatics Section of the Academia Europaea since December 2016. He is an Associate Editor of the IEEE TRANSACTIONS ON SERVICES COMPUTING, *ACM Transactions on the Web*, and *ACM Transactions on Internet Technology*. He is on the editorial board of IEEE.