This is the Accepted Manuscript version of an article published by Taylor & Francis journals *International Journal of Geographical Information Science* in 2023, which is available at: https://doi.org/10.1080/13658816.2023.2203218
Cite as: Zhao T, Zhengdong H, Tu W, Biljecki F, Long C (2023): Developing a multiview spatiotemporal model based on deep graph neural networks to predict the travel demand by bus. *International Journal of Geographical Information Science*, 1-27.

Developing a multiview spatiotemporal model based on deep graph neural networks to predict the travel demand by bus

Tianhong Zhao[a,b], Zhengdong Huang[a], Wei Tu[a,*], Filip Biljecki[b,c], Long Chen[d]

[a]*School of Architecture and Urban Planning, Shenzhen University, China*
[b]*Department of Architecture, National University of Singapore, Singapore*
[c]*Department of Real Estate, National University of Singapore, Singapore*
[d]*State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, China*

Abstract

The accurate prediction of travel demand by bus is crucial for effective urban mobility demand management. However, most models of travel demand prediction by bus tend to focus on the bus's spatiotemporal dependencies, while ignoring the interactions between buses and other transportation modes, such as metros and taxis. We propose a Multiview Spatiotemporal Graph Neural Network (MSTGNN) model to predict short-term travel demand by bus. It emphasizes the ability to capture the interaction dependencies among the travel demand of buses, metros, and taxis. Firstly, a multiview graph consisting of bus, metro, and taxi views is constructed, with each view containing both a local and global graph. Secondly, a multiview attention-based temporal graph convolution module is developed to capture spatiotemporal and cross-view interaction dependencies among different transport modes. Especially, to address the uneven spatial distributions of features in multiview learning, the cross-view spatial feature consistency loss is introduced as an auxiliary loss. Finally, we conduct intensive experiments using a real-world dataset from Shenzhen, China. The results demonstrate that our proposed MSTGNN model performs better than the existing models. Ablation experiments validate the contributions of various modes of transportation to the improvement of the model's performance.

Keywords: Graph deep learning, multiview learning, travel demand prediction, multimodal transportation, smart card data.

*Corresponding author
Email addresses: zhaotianhong2016@email.szu.edu.cn (Tianhong Zhao), zdhuang@szu.edu.cn (Zhengdong Huang), tuwei@szu.edu.cn (Wei Tu), filip@nus.edu.sg (Filip Biljecki), long.chen@ia.ac.cn (Long Chen)

1. Introduction

The buses serve as one of the most crucial modes of transportation since they contribute to promoting environmental sustainability, offering low-cost and accessible transit, enhancing social and economic connections, and decreasing traffic congestion (May, 2013; McLeod et al., 2017; Sultana et al., 2019). The prediction of travel demand by bus has become an important issue that allows policymakers and transportation authorities to ensure efficient and effective transportation services. The demand for bus travel is affected by the macroscopic built environment and the dynamic factors in the transportation system. In the long term, the spatial layout of the built environment profoundly influences the spatiotemporal distribution of citywide travel demand by bus (McNally, 2007; Ma et al., 2018; Qi et al., 2018). Individual characteristics, such as educational background, income, and family structure, have also been demonstrated to influence long-term travel behaviors, i.e., travel modes, travel times, and travel frequencies (Recker et al., 1986; Wang et al., 2011a). In contrast, in the short term, competition and cooperation among buses, metros, and taxis affect the availability of travel choices, thereby influencing travel demand by bus.

Recently, advances in information and communication technology (ICT) have enabled the collection of massive, timely transportation data, allowing us to capture the dynamic interactions among multiple travel modes. For example, Zhang et al. (2018b) revealed the interaction patterns of buses and taxis by considering the spatial distributions of trips and travel distances. They demonstrated that the competition and cooperation among different transport modes in multimodal trips were spatially correlated. Chen et al. (2020) unraveled latent transfer patterns between buses and metros and reported 21 typical patterns. Their results suggested that the transfer from one mode of travel to another is dynamic across space and time. Wu and Liao (2020) revealed that extreme weather events significantly affect travel behaviors. In experiments conducted on data from Beijing's metro system and a survey, they found that passengers prefer subways or cars to buses and bicycles when the weather is bad.

We summarize the dynamic interactions among buses, metros, and taxis into two types, as shown in Figure 1. (1) Global interaction transfer. For example, a commuter may travel from station *Metro B* to station *Metro A* and then transfer to station *Bus A*. In this case, the travel demand at bus station *Bus A* is affected by the global spatial influence of metro station *Metro B*. (2) Local interaction transfer. For example, if a rainstorm occurs, people who board the bus at *Bus A* may transfer to a taxi service at *Taxi A*, as they will not walk to the bus station in the rain (Wu and Liao, 2020). From the travel records, we can observe that the trips of *Bus A* decrease while those of the neighboring *Taxi A* station increase. Therefore, considering multimodal travel would enable us to better understand and predict travel demand (Li et al., 2021; Ke et al., 2021).

Essentially, predicting the demand for bus travel is a time series problem. Capturing the nonlinear dependencies of travel demand across space and time is a challenge (Zheng et al., 2014). In recent years, many different neural network models have been devel-

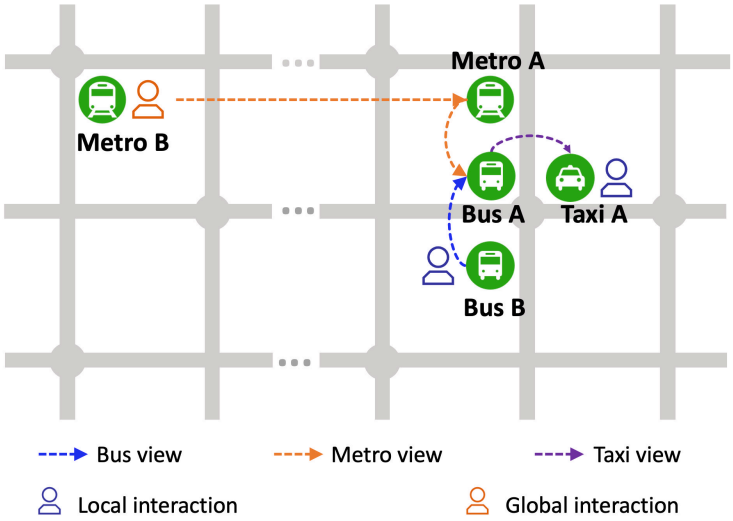

Figure 1: Interactions among the travel demands of buses, metros, and taxis. travel demand by bus may be influenced by other transportation modes in its neighborhood (local) and those that are far away (global).

oped to address this challenge. For instance, graph convolutional networks (GCNs) and recurrent neural networks (RNNs) have been used to capture spatial and temporal dependencies, respectively (Ren et al., 2020; Zhang et al., 2017). Some studies have developed more complex and powerful neural network models to capture specific spatiotemporal dependencies; examples include as long short-term memory (LSTM) (Cui et al., 2019) and graph attention networks (Guo et al., 2019). These models are generally based on historical demand features; they treat each mode of transportation as an isolated and independent environment, ignoring its interaction with other transportation modes. Recently, Liang et al. (2022) proposed a graph-based deep learning approach for bike sharing demand prediction considering the interactions among bike sharing, subway riding, and ride hailing. They developed intermodal similarity and difference modules to capture multimodal transportation interaction dependencies. Their findings demonstrate that graph-based deep learning outperforms existing methods. However, designing an effective structure to capture multimodal dynamic interactions remains a challenge. A multimodal interaction fusion model training strategy is also absent.

To fill these gaps, we propose a new Multiview Spatiotemporal Graph Neural Network (MSTGNN) model to predict short-term travel bus demand. Unlike previously developed approaches, the MTGNN model considers the interactions among different transport modes from both global and local perspectives. Specifically, a multiview graph with bus, metro, and taxi views is constructed according to spatial adjacency and mobility among different regions. A multiview attention-based spatiotemporal graph convolution module is designed to capture spatiotemporal and cross-view interaction de-

pendencies among different transport modes. The cross-view feature spatial consistency (CVSFC) loss is introduced to overcome the issue regarding the uneven spatial distributions of multiview features. Intensive experiments are conducted using a real-world bus, metro, and taxi dataset in Shenzhen. The results demonstrate that the MSTGNN model outperforms the baseline methods.

The remainder of this manuscript is organized as follows. Section 2 reviews related studies. Section 3 defines the studied problem. Section 4 describes the presented MSTGNN model. Section 5 reports the experimental results and discusses them. Finally, Section 6 concludes this study and outlines future work ideas.

2. Related work

2.1. Multimodal urban mobility

Urban mobility, which includes various travel modes, such as walking, cycling, buses, and metros, is typically related to the built environment, demographics, and urban dynamics. (Zhang et al., 2020b; Li et al., 2020b; Tu et al., 2017; Yue et al., 2018; Tu et al., 2020; Gao et al., 2021). As a result, the multimodal demands of different travel types are highly correlated with each other in the long term. Previous studies depended on passenger surveys to investigate the interactions between human travel and multimodal transportation (Barry et al., 2009; Nassir et al., 2015). These studies aimed to investigate the impact of passengers' subjective perceptions and service provisions on multimodal travel in public transportation systems. Cherry and Townsend (2012) used more than 300 surveys and developed a satisfaction analysis and ordinal regression model to analyze the influencing factors of passengers on metro and bus transfers. The results indicated that safety and the distances between metro exits and bus stops were the most critical factors that affected transfers. Hernandez and Monzon (2016) analyzed a transfer satisfaction survey using the principal component analysis method and discovered that minimizing the required waiting time and increasing transfer station comfort are the most important factors that influence the willingness to transfer. Zhao and Li (2017) investigated the interactions between two modes of transportation metros and bicycles in Beijing. It was discovered that trip distance is the most influential element for transfer travel between metro stations and houses or businesses. Furthermore, income, individual views, and the built environment also substantially impact transfer.

With the advancement of ICT, smart cards and GPS data provide comprehensive spatial and temporal travel information, enabling us to explore the spatial scopes and temporal dynamics of multimodal transportation interactions (Fang et al., 2012; Siła-Nowicka et al., 2016; Li et al., 2020a; Zhang et al., 2020b). By combining these big multisource transport data, the state-of-the-art methods can finely identify passengers' transfer behaviors and travel chains in a multimodal transportation system, which acts as a foundation for evaluating multimodal transportation interactions. Zhao et al. (2017)

provided a method for recognizing transfer behaviors using smart card data and constructed various frameworks for analyzing the multifactor interactions involved in multimodal travel, such as built environment associations and spatiotemporal similarity. Tu et al. (2018) explored bus, metro, and taxi ridership; revealed the associations between multimodal ridership and demographic, land use, and transportation factors; and compared the similarities and differences among these associations.

In the short term, the interactions that occur in multimodal transportation are reflected in the travel mode choices made in a dynamic urban environment (Tao et al., 2018; Zhao et al., 2020). Wu et al. (2022) employed a multivariate generalized Poisson regression model to investigate the relationships between bus and subway transfer passenger flows and socioeconomic, built environment, holiday, and weather variables on different days. They found that the transfer ridership of the metro-to-bus mode significantly increased under the high wind, heavy rain, and high-temperature conditions. Kim (2020) examined the impacts of weather and calendar events on subway and bus mode choices. They found that bad conditions, including cold weather, increased subway use, and the magnitudes of the influences of different features varied over the tested periods. In summary, as components of the urban transportation system, buses, metros, and taxis all affect each other. Capturing their dependencies is quite helpful for predicting travel demand.

2.2. Travel demand prediction

Travel demand prediction operates based on historical travel information and additional variables (weather, dates, etc.) to predict long-term or short-term travel demand (Ma et al., 2014; Liu et al., 2020; Karnberger and Antoniou, 2020; Huang et al., 2021). The related methods are mainly divided into two categories: knowledge-driven methods and data-driven methods. Knowledge-driven approaches are generally used in urban research and queuing theory; they predict travel demand by understanding the laws of travel demand and the built environment, as well as by constructing resident behaviors (Ma et al., 2018; Xu et al., 2014). Despite their strong interpretability, these methods cannot effectively capture the dynamic changes in travel demand in a complex realistic context. Typical data-driven methods include the moving average (MA) model and the autoregressive integrated moving average (ARIMA) model (Ahmed and Cook, 1979; Hamed et al., 1995). These models have high computational efficiency, can be applied to various fields and are widely used in transportation prediction. However, such methods can only capture the linear temporal correlations in travel demand data; the more complex nonlinear space-time correlations cannot be captured, limiting the prediction performance of such approaches.

In recent years, deep learning models have achieved great performance in terms of travel demand prediction because they can capture temporal and spatial dependencies (Zhao et al., 2019; Miglani and Kumar, 2019; Kashyap et al., 2022). In terms of temporal dependence, an RNN considers the time series correlations among multiple

input time points, processes the input information from the previous instant, and retains it in the operation used for the current information; thus, RNNs are widely used to capture temporal dependencies (Zaremba et al., 2015; Cho et al., 2014). RNN model variants, such as the LSTM network and gated recurrent unit (GRU) (Hochreiter and Schmidhuber, 1997; Shu et al., 2021), have been developed to overcome the vanishing gradient and overfitting problems and have performed well in transport prediction tasks. In terms of spatial dependence, many studies treat urban space as a standard grid or a non-European graph structure and employ convolutional neural networks (CNNs) and GCNs to capture spatial dependencies (Ren et al., 2020). For example, Ren et al. (2020) divided their study area into a regular two-dimensional grid; captured spatial dependencies through a two-dimensional CNN; used LSTM to capture the proximity, periodicity, and trend features in time series; and predicted the traffic flow. Zhao et al. (2019) built a graph based on a road network and proposed a temporal GCN (T-GCN) model, a combination of a GCN and a GRU, to capture the spatiotemporal characteristics of traffic flows. Many more complex networks have been proposed to better capture the temporal and spatial dependencies in transport prediction tasks, such as graph attention networks (Zheng et al., 2020) and graph WaveNet (GWNET) (Wu et al., 2019).

Most deep learning-based travel demand prediction models focus on complex spatiotemporal dependencies (Yu et al., 2018; Guo and Zhang, 2020), while the interactions of multimodal transportation dependencies receive little attention. Few recent studies have attempted to use a multiview learning approach to fuse multimodal transportation information to achieve improved model performance. Ke et al. (2021) proposed a deep multitask multigraph learning approach for predicting the solo and shared service modes in ride hailing systems. They constructed neighborhood, distance, and functionality graphs and proposed a regularized cross-task learning structure to achieve knowledge sharing between different modes. Liang et al. (2022) proposed a graph-based deep learning approach for bike sharing demand prediction by considering the interactions among bike sharing, subway riding, and ride hailing. They developed intermodal similarity and intermodal difference modules to capture multimodal transportation interaction dependencies. Their results demonstrated the superior performance of their approach over that of existing methods. However, challenges remain with regard to designing an efficient structure for capturing multimodal dynamic interactions, and a model training strategy for the fusion of multimodal interaction features is unavailable. Here, we propose a new MSTGNN model to predict short-term travel demand. Unlike previously developed approaches, this model considers multimodal transportation interactions from both local and global transfer perspectives, developing global and local cross-view multigraph learning structures. Specifically, a CVSFC loss is presented to overcome the uneven distribution issue exhibited by multiview features during model training.

3. Problem statement

Here, we provide the basic definitions used in this study.

Definition 1: Spatial graph

A spatial graph is represented by $G = (V, A, X)$, where $V = v_1, v_2, \cdots, v_N$ is the set of nodes, A is a feature matrix with dimensions of $N \times N$, and X is the feature matrix for the nodes. In this study, each node is a street-level transportation analysis zone (TAZ) (Zhao et al., 2022), and N is the number of TAZs. If there is an edge that goes from node i to node j, then A_{ij}=1; otherwise, A_{ij}=0. The feature matrix X is defined below. Multiple graphs can be constructed by considering the different connections between TAZs, and the detailed graph construction procedure is demonstrated in Section 4.1.

Definition 2: Feature matrix

A feature matrix refers to the historical travel demands of the three transportation modes of each node in the spatial graph G. The feature matrix is denoted by $X \in \mathbb{R}^{M \times F \times N}$. At each time step t, the graph G has a dynamic feature matrix $X_t \in \mathbb{R}^{M \times F \times N}$, where M represents the number of transportation modes, F represents the length of the historical time series, and N is the number of TAZs. X_t^m represents the travel demand for all TAZs at time t under transportation mode m.

Definition 3: Travel demand prediction problem

Given one or more graphs G and its/their feature matrix $X = [X_{t-n}, \cdots, X_{t-1}, X_t]$, the travel demand prediction problem is to learn a function f that can predict the next k steps of graph features belonging to transportation mode m, as shown in Equation 1.

$$\left[X_{t+1}^m, X_{t+2}^m, \cdots, X_{t+k}^m\right] = f\left(G(s); [X_{t-n}, \cdots, X_{t-1}, X_t]\right) \tag{1}$$

4. Methodology

This section introduces the proposed MSTGNN model for travel demand prediction. The schematic structure of the MSTGNN is presented in Figure 2. The MSTGNN model consists of three main components: a multiview graph construction module, a multiview attention-based temporal GCN (MVATGCN) module, and a cross-view spatial-temporal feature fusion module. First, multiview graphs (e.g., bus, metro, and taxi graphs) are constructed. Each view has local and global graphs, which are built according to the historical origin-destination records of trips. Then, the multiview graph features are independently fed into the MVATGCN module to capture both local and global cross-view dependencies. Finally, the predicted travel demand results are obtained by fusing the global and local cross-view spatiotemporal dependencies through a parameter matrix. The details of each module are described below.

4.1. Generating a multiview graph

It is challenging to directly model multimodal travel flow data due to the complex spatial and temporal dependencies of various transportation modes. Multiview learning

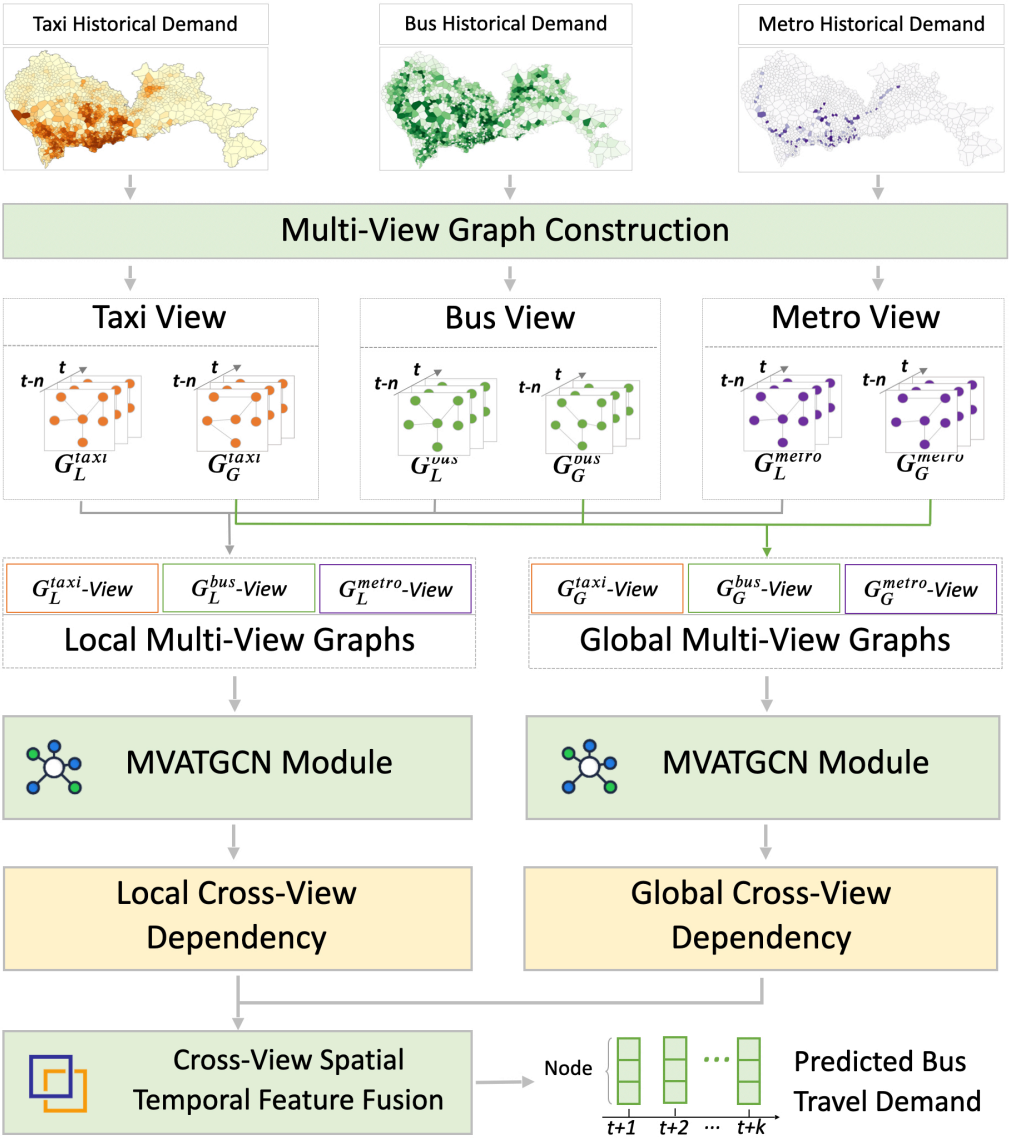

Figure 2: The framework of the proposed MSTGNN model.

can decompose complex interaction patterns into multiple independent simple patterns, separately capture the spatiotemporal dependencies of each view, and then fuse the multiview features(Zhang et al., 2020a; Sun et al., 2022; Wang et al., 2022). On the one hand, this approach reduces the burden of capturing complex nonlinear dependencies. On the other hand, it makes the resulting model more interpretable. The foundation of graph-based multiview learning is to build multiple spatiotemporal graphs. We struc-

ture bus, metro, and taxi views according to the different transportation modes. In each view, we consider two types of spatial relationships among TAZs, spatial adjacency and mobility, which two graphs can represent. (1) The local graph $G_L = (V, A_L, X)$ encodes the spatial adjacency relations. Tobler's first law of geography states that 'everything is related to everything else, but near things are more connected than distant things' (Tobler, 1970). In a transportation system, adjacent TAZs may share similar travel patterns. Therefore, we construct adjacency relationship graphs to consider spatial autocorrelations, and the adjacency matrix A_L can be computed as in Equation 2. (2) The global graph $G_G = (V, A_G, X)$ encodes the mobility among TAZs. Intuitively, geographically distant TAZs with strong mobility have higher travel demand correlations due to transfer. Historical travel demand records provide information on the mobility among different TAZs, which can be used to build a global graph; A_G can be calculated as shown in Equation 3, where v_{ij} denotes the number of travel between TAZs i and j and δ is a threshold. Here, we set the threshold to the count of the top 20% of trips entering a TAZ. This means that we construct edges for each TAZ with other TAZs that have the top 20% of the strongest mobility.

$$A_{L,ij} = \begin{cases} 1, & v_i \text{ and } v_j \text{ are adjacent} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$A_{G,ij} = \begin{cases} 1, & v_{ij} > \delta \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The bus, metro, and taxi views of the graph data models G^{bus}, G^{metro}, and G^{taxi} are constructed for multiview learning, as shown in Equations 4–6. Each view contains two types of data graph models: a local graph G_L and a global graph G_G. Although the G_G have different edges, they share the nodes V, and the travel demand features X belong to the nodes. The local multiview graphs include G_L^{bus}, G_L^{metro} and G_L^{taxi}, and the global multiview graphs include G_G^{bus}, G_G^{metro} and G_G^{taxi}. Finally, the global and local multiview graphs are fed into the MVATGCN module.

$$G^{\text{bus}} = \left\{G_L^{\text{bus}}, G_G^{\text{bus}}\right\} = \left\{\left(V, A_L^{\text{bus}}, X^{\text{bus}}\right), \left(V, A_G^{\text{bus}}, X^{\text{bus}}\right)\right\} \tag{4}$$

$$G^{\text{metro}} = \left\{G_L^{\text{metro}}, G_G^{\text{metro}}\right\} = \left\{\left(V, A_L^{\text{metro}}, X^{\text{metro}}\right), \left(V, A_G^{\text{metro}}, X^{\text{metro}}\right)\right\} \tag{5}$$

$$G^{\text{taxi}} = \left\{G_L^{\text{taxi}}, G_G^{\text{taxi}}\right\} = \left\{\left(V, A_L^{\text{taxi}}, X^{\text{taxi}}\right), \left(V, A_G^{\text{taxi}}, X^{\text{taxi}}\right)\right\} \tag{6}$$

4.2. The MVATGCN

The MVATGCN module captures heterogeneous spatiotemporal and cross-view dependencies. As shown in Figure 3, it includes a multiview TGCN that stacks four temporal convolutional networks (TCNs) and GCNs, as well as a cross-view attention network. The multiview TGCN is used to capture the spatiotemporal dependencies of the

individual transportation views. The cross-view attention network is used to capture the cross-view dependencies of various transportation modes.

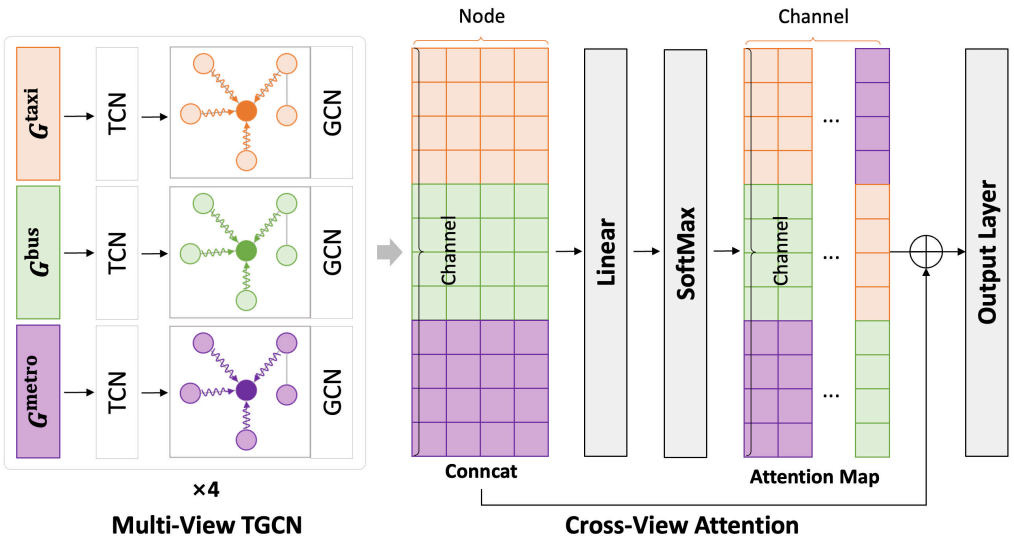

Figure 3: The multiview attention-based temporal graph convolution module.

4.2.1. Multiview TGCN

Convolution is an excellent method for aggregating neighborhood information, and it is commonly used in local feature extraction tasks (Rawat and Wang, 2017; Cao et al., 2020). Inspired by previous studies, we use the convolutional-based method to extract temporal and spatial features. In terms of the temporal aspect, although RNN-based models have become widespread in time series analysis cases, many challenges remain regarding the use of recurrent networks for travel prediction, including their long training time requirements and their difficulty in dealing with long-range sequences. CNN-based models have the advantages of quick training processes, simple structures, and no dependency limitations concerning prior stages. As a result, we adopt dilated causal convolution structures (Yu and Koltun, 2016) as our TCNs to capture the temporal dependency of each zone.

As shown in Figure 4, the dilated causal convolution operation, a special case of standard 1D convolution, slides over inputs by skipping values with a certain step size. The dilated convolution algorithm varies from the traditional CNN because it has a wider kernel, which can more effectively learn connections between data at different time intervals. For each group of view graph data G^v, X^v is the time series feature of view v, and the dilated causal convolution operation of X^v at step s is represented as:

$$\bar{X}^v = \sum_{l=1}^{L} X^v(l) \times K(s - l \times d) \tag{7}$$

where $\bar{X}^v$ denotes the features obtained after the dilated causal convolution, $K \in \mathbb{R}^L$ is the kernel with a size of L and d is the dilation factor that controls the skipping distance. When the network is dilated by a factor of 1, it is similar to a CNN, and a kernel with a dilation factor of 2 has one kernel that skips the input. This structure can obtain a larger receptive field while the convolution kernel size remains the same. The receptive field of the network can be calculated as $r = 2^{(n-1)} \times L$, which means that the r value of the network grows exponentially with the number of network layers n. Gating mechanisms have been demonstrated to be effective for controlling the flow of information through layers in dilated causal convolution networks. We use gated activation units to further capture temporal dependencies, as shown in Equation 8. In the equation, W_f and W_g are the learnable convolution filters, X is the input feature, $\odot$ denotes an elementwise multiplication operator, $\sigma(\cdot)$ is a sigmoid function, and b and c are learnable bias parameters. Finally, the graph feature $\dot{X}^v$ of each view v after temporal convolution is obtained.

$$\dot{X}^v = \tanh\left(W_f * \bar{X}^v + b\right) \odot \sigma\left(W_g * \bar{X}^v + c\right) \tag{8}$$

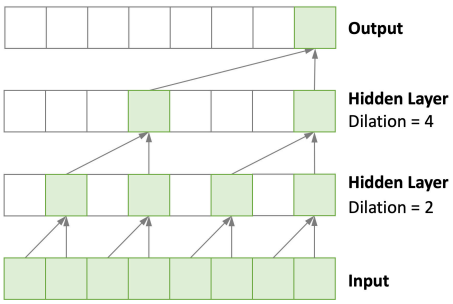

Figure 4: The dilated causal convolution with a kernel size of 2.

GCNs have recently become popular for capturing non-Euclidean spatial dependencies (Yu et al., 2018; Zhang et al., 2020c). GCNs can be roughly categorized into two groups: spectral graph convolution and spatial graph convolution approaches. Spectral graph convolution transfers signals from the graph domain to the Fourier domain through a graph Laplacian. This paper uses the spectral-based graph convolution method to implement the GCN. The input is the graph $\dot{G}^v \in (V, A^v, \dot{X}^v)$ obtained after temporal convolution. $\ddot{X}^v$ denotes the feature matrix after utilizing the GCN. Generally, a graph convolution can be expressed as $\ddot{X}^v = \Theta(L)\dot{X}^v$, i.e., the multiplication of a feature matrix $\dot{X}^v$ with a kernel Θ. Here, $L = I - D^{-1/2}AD^{-1/2}$ is the symmetric normalized Laplacian matrix of G, I is an identity matrix, and D denotes the diagonal degree matrix

of A. However, this normalization approach is time-consuming, with a complexity of $o(n^2)$. We use Chebyshev polynomial approximation to simplify the normalization task instead (Defferrard et al., 2017); formally, the graph convolution can then be rewritten as:

$$\ddot{X}^v = \Theta(L)\dot{X}^v \approx \sum_{k=0}^{K} w_k T_k(\hat{L})\dot{X}^v \tag{9}$$

where $\hat{L} = 2L/\lambda_{\max} - I$ denotes the scaled Laplacian matrix. λ_{max} is the largest eigenvalue of L, and w_k is the Chebyshev coefficient. The Chebyshev polynomials are recursively defined as $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$, with $T_0(x) = 1$ and $T_1(x) = x$. K is the number of consecutive filtering operations or convolutional layers in a model, where node information utilizes the information derived from the $K-1$st-order neighborhood of the central node in a convolution operation.

4.2.2. Cross-view attention network

The mutual influences of different transportation views are diverse; for example, during morning rush hour, the demands for metros and buses are strongly correlated. An attention mechanism is capable of adaptively assigning weights between different views (Vaswani et al., 2017; Guo et al., 2019). We design a cross-view attention module to learn the contribution of each transportation mode to the target zone.

$$a = softmax\Big(U\Big(concat\left(\ddot{X}^{bus}, \ddot{X}^{metro}, \ddot{X}^{taxi}\right)W + b\Big)\Big) \tag{10}$$

where $concat(*)$ concatenates the three view (bus, metro and taxi view) features obtained after the temporal and spatial convolutional networks $\ddot{X}^v \in \mathbb{R}^{C_{in} \times N}$ into a high-dimensional vector. $W \in \mathbb{R}^{N \times C_{out}}$ is a learnable parameter matrix that maps the connected high-dimensional features to a new feature space, C_{out} is the output channel of the high-dimensional features, and $C_{out} = 3 * C_{in}$. $U \in \mathbb{R}^{C_{out} \times C_{out}}$ is a learnable parameter matrix, and b is a trainable bias. The resulting weight matrix $a \in \mathbb{R}^{C_{out} \times C_{out}}$ is the learned attention weight. Finally, the features of the three views are concatenated and multiplied by the attention weight matrix to obtain the cross-view dependency features $\tilde{X}$; this process is formulated as follows:

$$\tilde{X} = concat\left(\ddot{X}^{bus}, \ddot{X}^{metro}, \ddot{X}^{taxi}\right) \cdot a \tag{11}$$

The cross-view attention network has the ability to assign different attention weights to the various transportation views. Intuitively, as shown in Figure 5, the GCN aggregates the local and global information of buses, metros, and taxis from a spatial perspective. The three transportation mode signals are then fed into the attention module, where the bus, metro, and taxi signals are aggregated for each zone and the weights of various transportation modes are allocated by using the attention mechanism.

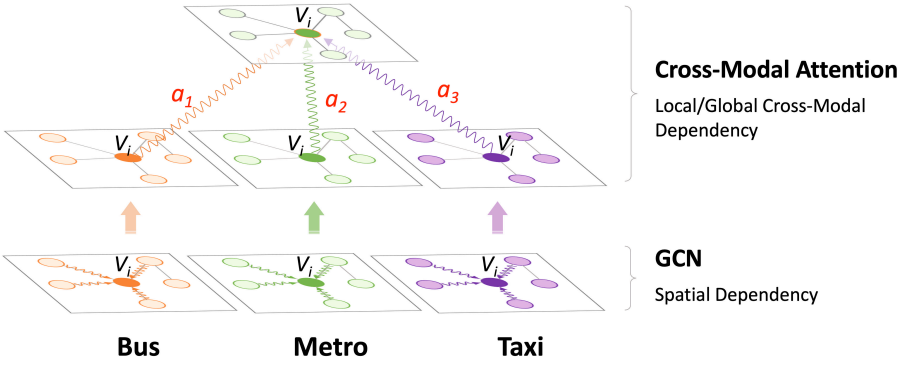

Figure 5: The process of sharing messages between different spaces and multiple views through a GCN and a cross-view attention network.

4.3. Cross-view spatial-temporal feature fusion

Through the MVATGCN module, we obtain local and global cross-view dependency features, which are denoted as $\tilde{X}_l$ and $\tilde{X}_g$, respectively. We propose utilizing a fusion layer to fuse the local and global cross-view spatial-temporal features. Previous multi-view transportation prediction tasks found that various views can have different degrees of influence on the final prediction results (Sun et al., 2022). Inspired by this, the two types of features can be combined using a parametric matrix-based fusion method:

$$\hat{Y} = W_l \odot \tilde{X}_l + W_g \odot \tilde{X}_g \tag{12}$$

where W_l and W_g are the learnable parameters, and $\odot$ represents the Hadamard product.

4.4. Loss functions

To achieve effective model training and address the uneven spatial distribution of multiview features, two kinds of losses are proposed to constrain the network training process, i.e., major and auxiliary losses. The major loss is the mean squared error loss $\mathcal{L}_{L2}$. The auxiliary loss function is the spatial feature distribution consistency loss $\mathcal{L}_{SC}$. The overall loss is formulated as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{L2} + \lambda_2 \mathcal{L}_{SC} \tag{13}$$

where λ_1 and λ_2 control the relative importance levels of the two losses. The best values for λ_1 and λ_2 are found by changing the value of λ_1/λ_2 from 0.01 to 100 while keeping the other parameters fixed. Specifically, $\mathcal{L}_{L2}$ calculates the gaps between the elements in the predicted travel demand by bus $\hat{Y}$ and the ground truth Y. It can be expressed as follows:

$$\mathcal{L}_{L2} = \frac{1}{k}\sum_{t=0}^{M}\sum_{i=0}^{N}\left(\hat{Y}_{t,i} - Y_{t,i}\right)^2 \tag{14}$$

The multimodal travel demand is not uniformly distributed across space. For example, metro travel demand is only observed in limited TAZs, and the taxi travel demand in the central urban area is significantly higher than that in the suburbs. However, the spatial feature distribution imbalance may be considered data noise. To allow the GCN to better learn the interactions of features between different TAZs, we design an auxiliary loss to ensure the consistency of the spatial feature intensity distribution across views, thus avoiding the effect of data noise. In this way, the CVSFC loss is computed by the cosine distance as:

$$\mathcal{L}_{sc} = \sum_{v\in(metro,taxi)}\left(1 - \frac{\ddot{X}^{bus}\cdot\ddot{X}^{v}}{\left\|\ddot{X}^{bus}\right\|\cdot\left\|\ddot{X}^{v}\right\|}\right) \tag{15}$$

where $\ddot{X}^{v}$ represents the feature of the transportation view v calculated by the TCN and GCN.

5. Results and discussion

5.1. Datasets

To demonstrate the effectiveness of the presented MSTGNN, experiments are conducted on multimodal datasets from Shenzhen, China. The collected travel demand data for buses, metros, and taxis extend from December 1, 2018, to December 31, 2018. These data come from the Shenzhen Transportation Bureau; specifically, the dataset has the following three parts.

- Bus: The data of travel demand by bus come from Shenzhen smart card and GPS data. The data cleaning procedures involve adjusting the swipe system time and GPS time and removing some GPS abnormalities. By employing the technique suggested by Wang et al. (2011b), we can estimate the origin stop, destination stop, and travel time of each trip using GPS and smart card data. Finally, approximately 3.1 million bus trips are recorded every day, encompassing 7875 bus stops.

- Metro: The metro travel demand data are collected from the same smart card data, which include the times and station names of each trip when stations are entered and exited. A person's metro trip information, including their boarding/alighting location and time, may be retrieved directly from the smart card database. Anomalies related to trip times and metro stations are removed from the records as part of the data cleaning procedure. The 290 metro stations are distributed across 156 TAZs, and approximately 2.2 million metro trip records are created daily.

- Taxi: The taxi GPS dataset includes record times, taxi locations, and service statuses (0 for idle, 1 for service). We delete GPS records with abnormal passenger load statuses and those that are not in Shenzhen. Consecutive taxi GPS data with statuses changing from 0 to 1 indicate that a taxi trip has started. Utilizing the method, taxi trips are estimated (Tu et al., 2018). Approximately 0.78 million taxi pick-up and drop-off locations and time records per day are created from 20 thousand taxis.

Following Zhao et al. (2022), 1386 TAZs are generated using a Voronoi diagram based on the main roads in Shenzhen, which includes the primary, secondary, and trunk types of the OpenStreetMap (OSM) road classification. This method has the advantage that bus stops on both sides of the same route are grouped together, and the predicted results are better for bus schedules. We measure the travel demands for different modes of transportation in each TAZ with a 5-minute time resolution. This new multimodal transportation travel demand dataset (SZ) is available.

5.2. Preliminary analysis

Before formally presenting the prediction experiments and analytical results, it is necessary to conduct a preliminary analysis to validate the cross-view spatiotemporal dependencies. We use Spearman correlation analysis to evaluate various dependencies. Inspired by Li et al. (2017), the total travel demand for each transportation mode in each region is used as a spatial correlation variable. The sum of all TAZ travel demands every 5 minutes for each transportation mode is used as a temporal correlation variable. As shown in Table 1, from the spatial perspective, the spatial correlation coefficients of bus-metro, bus-taxi, and metro-taxi travel demands are 0.39, 0.55, and 0.53, respectively. This suggests that the travel demands for buses and taxis have the strongest spatial correlation, while those of buses and metros have the weakest correlation due to the lack of metro stations in many TAZs. From a temporal perspective, the correlation coefficient of bus-metro travel is 0.97, which is higher than their spatial correlation. The bus-taxi and metro-taxi travel demands have correlation coefficients of 0.31 and 0.41, respectively. This shows that the temporal correlations between taxi travel demand and other transportation modes are relatively weak. Specifically, Figure 6 shows the temporal distributions of the travel demands for the three modes of transportation. The temporal distributions of bus and metro travel demands are similar, with obvious morning and evening peaks. The travel demand for buses is higher during the morning peak than the metro demand, but the two demands are almost equal during the evening peak. The peak demand for taxis occurs in the early hours (22:00-24:00), and no significant demand fluctuation occurs throughout the day (8:00-20:00). Furthermore, significant differences are observed between the characteristics of taxi and bus operations. Taxis operate throughout the day. However, during the nighttime hours of 0:00-6:00 AM and 22:00-24:00, bus services are reduced or discontinued, and the demand for bus travel is not observed.

To reduce the effect of data noise, we reduce the demand for taxis to zero for these two periods in the dataset.

Table 1: Spatial and temporal correlation coefficients.

	Spatial correlation			Temporal correlation		
	Bus	Metro	Taxi	Bus	Metro	Taxi
Bus	1	0.39	0.55	1	0.97	0.31
Metro	0.39	1	0.53	0.97	1	0.41
Taxi	0.55	0.53	1	0.31	0.41	1

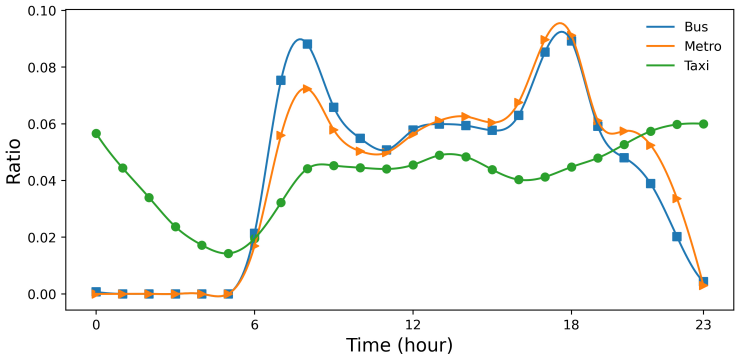

Figure 6: Temporal distributions of the three transportation modes.

5.3. Hyperparameter selection

Parameter tuning is essential for obtaining a deep learning model with optimal performance (Ke et al., 2017; Zhang et al., 2018a). The MSTGNN model contains three important hyperparameters: K_t, K_s, and D_{stg}. K_t is the size of the temporal convolution kernel, K_s is the size of the spatial convolution kernel, and D_{stg} is the number of spatiotemporal convolution layers. The optimal model parameter combination is generally found using a control variable approach (Cheng et al., 2020). To test K_t, K_s and D_{stg} are fixed to 3 and 4, respectively, and K_t varies from 2 to 5. K_s varies from 2 to 5 when K_t and D_{stg} are fixed to 3 and 4, respectively. Correspondingly, D_{stg} varies from 1 to 5 when K_t and K_s are fixed to 3. We use the root mean square error (RMSE) and the running time required for one epoch as evaluation metrics, and Figure 7 presents the obtained results. The RMSE decreases and then increases with increasing K_s and K_t values. The proposed model performs best when $K_s/K_t = 3$; thus, we set K_s/K_t to 3. Even though the performance tends to be better as D_{stg} rises, the training time also significantly increases. Finally, we set D_{stg} to 4.

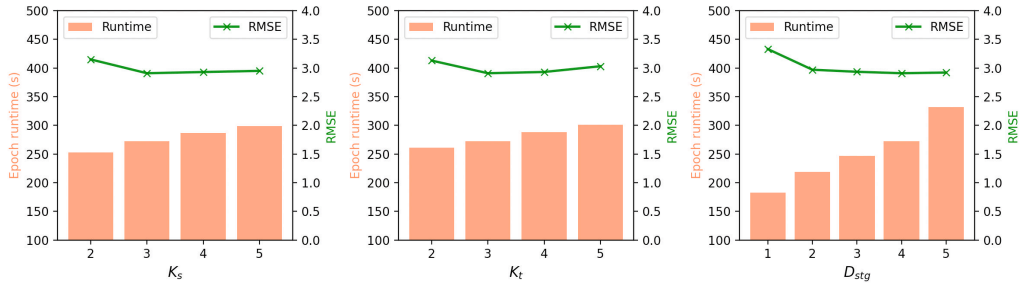

Figure 7: RMSE and epoch runtime changes obtained with different parameters on the SZ dataset.

5.4. Experimental setup

5.4.1. Evaluation metrics

The ground truth is the travel demand by bus within each TAZ. y_i denotes the ground truth of the $i-th$ TAZ. $\hat{y}_i$ is the corresponding predicted value, and $\bar{y}$ is the average of all y_i. We use three evaluation metrics, including the mean absolute error (MAE), RMSE, and coefficient of determination (R^2), which are defined as Equations 16, 17 and 18, respectively. Here, S denotes the set of all TAZs.

$$\mathrm{MAE} = \frac{1}{S}\sum_{i=1}^{S}|y_i - \hat{y}_i| \tag{16}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{S}\sum_{i=1}^{S}(y_i - \hat{y}_i)^2} \tag{17}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{S}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{S}(y_i - \bar{y})^2} \tag{18}$$

5.4.2. Baselines

We compare the MSTGNN model with several baseline methods, including a statistical method (ARIMA), a machine learning method (SVR), and several deep learning methods (LSTM, STGCN, and T-GCN*). In particular, we use the improved T-GCN model to test the advantages of our proposed model in fusing the three modes of transportation. The six baseline methods are described briefly below.

- **ARIMA:** This is a statistical method that is widely used for time series prediction (Box and Pierce, 1970).
- **SVR:** This is a machine learning-based regression approach to modeling the relationship between future demand and historical time series (Lau and Wu, 2008). Here, we use a radial basis function (RBF) as the kernel, and the penalty parameter C is set to 1.

- **LSTM:** This is an RNN-based deep learning model that captures complex relationships and long-term dependencies in time series data, and it has attracted extensive attention in traffic prediction tasks (Hochreiter and Schmidhuber, 1997). Here, we set the number of hidden layers toas 3 and the hidden state size to 128.

- **Spatiotemporal GCN (STGCN):** This is a deep learning model that captures temporal and spatial dependencies by employing gated CNNs and spectral-based GCNs, respectively (Yu et al., 2018). Here, we set the numbers of channels in the three layers of the ST-Conv block to 64, 32, and 128 each, and both the graph and temporal convolution kernel sizes are set to 3.

- **GWNET:** This is a spatiotemporal graph learning approach based on a self-learned adjacency matrix for capturing complex spatial dependencies via node embedding (Wu et al., 2019). Here, we adopt a self-adaptive adjacency matrix and a graph convolution layer with a diffusion step size of 2 in the model.

- **T-GCN*:** This is one of the most widely used transport prediction models, which is combined with a GCN and a GRU (Zhao et al., 2019). Here, the main structure of the T-GCN model remains unchanged. We use the concatenation operation to integrate the features of the three transport modes instead of inputting one transport mode as input. We set the number of hidden units to 100 in our experiments.

5.4.3. Network training

We aggregate the multiview transportation demand data into 5-minute intervals and apply z score normalization to each mode to align the different datasets. The input historical time step is m=12, and the prediction step is n=6, as this setting utilizes the historical observed travel demand of the previous hour to predict the demand for the following half hour. Seventy percent of the data are used for model training, 10% are employed for validation, and 20% are applied for model testing. The adaptive moment estimation (Adam) optimizer is used to train the networks with a learning rate of 0.001 that decays by a factor of 0.8 after 50 epochs. The training batch size is 64. According to the hyperparameter optimization results obtained in Section 5.3, the temporal convolution kernel is set to 3, the spatial convolution kernel is set to 3, the number of convolutional layers is set to 4, and the number of convolution channels is set to 32. The output channel size of the multiview TGCN module is set to 96. Fixed random seeds are used for each experiment. We use LibCity (Wang et al., 2021) to implement the baseline model and tune the parameters based on the performance of the validation set. All experiments are implemented on an NVIDIA DGX A100 server with 512 GB of RAM and eight NVIDIA A100 tensor core GPUs (40 GB).

5.5. Results and analysis

5.5.1. Overall results

The prediction performance of the proposed MSTGNN model is reported in Table 2 and Figure 8. Overall, the model's three metrics (the MAE, RMSE, and R^2) are 1.216, 2.909, and 0.965, respectively, when the time granularity is 15 minutes. The prediction accuracy decreases as the time granularity increases; for example, the MAE increases from 0.807 to 1.283 when the time granularity is increased from 5 minutes to 30 minutes. For visualization analysis, we choose the prediction results obtained for TAZ No. 617 at various time granularities. Figure 8 (a), (b), and (c) show the true travel demand by bus and predicted values for time granularities of 5, 15, and 30 minutes in a day, respectively, as well as green bars indicating their absolute errors. The larger residual occurs between 7:00-9:00 and 17:00-19:00 due to the strong fluctuating travel demand during this period, which is consistent with the results obtained by Guo and Zhang (2020).

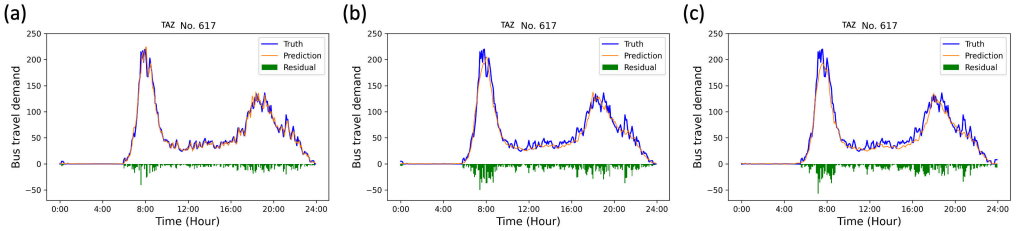

Figure 8: The ground truths and prediction results, as well as an error analysis of the MSTGNN model for TAZ No. 617. (a) 5-minute time granularity; (b) 15-minute time granularity; (c) 30-minute time granularity.

5.5.2. Comparison with the baselines

The performances of various models are summarized in Table 2. Taking a time granularity of 15 minutes as an example, compared to those of the best baseline (T-GCN*), the MAE and RMSE of the proposed model are reduced by 5.59% and 3.51%, respectively. This suggests that the MSTGNN outperforms these baseline methods. The ARIMA performs worse than the other baseline models, with MAE and RMSE values of 2.006 and 5.557, respectively, demonstrating the limitations of traditional statistical models in obtaining nonlinear spatiotemporal relationships. SVR performs better than the ARIMA model, with MAE and RMSE values of 1.66 and 4.956, respectively, due to its ability to handle complex temporal features; however, it performs worse than the deep learning models. LSTM, a deep learning model, can capture nonlinear temporal dependencies and outperform traditional statistical and machine learning models; its MAE and RMSE reach 1.375 and 3.698, respectively. However, among the deep learning models, LSTM performs worst, demonstrating the importance of considering spatial dependencies in travel demand prediction. The MAEs of the STGCN and GWNET are 1.369 and 1.327, respectively; GWNET has a 3.17% advantage. Although GWNET and the STGCN both consider spatial dependencies, the adaptive spatial adjacency matrix used

in GWNET enables it to outperform the STGCN, demonstrating that the utilized graph structure significantly impacts the prediction results. T-GCN* combines the three transportation modes via a simple connection approach, and the MAE improves by 3.35% over that of GWNET. Our model outperforms all other baselines because it considers multiple spatial relationships and the cross-view dependencies of various transportation modes. In addition, we evaluate the prediction results obtained for workdays and weekends, and our proposed MSTGNN model still performs best. The R^2 values produced for workdays and weekends are 0.968 and 0.957, respectively. The R^2 value obtained for weekends is lower than that of workdays. As there are few commuters on weekends, bus travel is highly random and difficult to predict on weekends.

Table 2: Performance comparison among different models.

Model	5 min			15 min			30 min		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
ARIMA	1.235	3.277	0.958	2.006	5.557	0.876	2.847	8.285	0.722
SVR	1.256	4.304	0.925	1.660	4.956	0.900	1.982	5.963	0.854
LSTM	1.124	3.506	0.950	1.375	3.698	0.944	1.641	3.752	0.925
STGCN	1.040	2.447	0.975	1.369	3.314	0.955	1.438	3.504	0.949
GWNET	0.897	2.173	0.983	1.327	3.147	0.963	1.460	3.697	0.942
T-GCN*	0.882	2.019	0.983	1.284	3.011	0.964	1.398	3.503	0.951
MSTGNN	**0.807**	**1.919**	**0.985**	**1.216**	**2.909**	**0.965**	**1.283**	**3.080**	**0.961**

5.5.3. Contributions of different modes of transportation

We construct three simplified versions of the MSTGNN to quantify the contributions of various modes of transportation.

- *MSTGNN-B:* This model only accepts bus transportation data, and the rest of its architecture is consistent with that of the MSTGNN model. The cross-view attention module is removed because only one mode of transportation is considered here.
- *MSTGNN-BM:* The structure of this model is consistent with that of the MSTGNN model, and both local and global cross-view interaction views are considered, but the input data are limited to the combination of bus and metro data.
- *MSTGNN-BT:* Similar to the MSTGNN-BT model, the bus and taxi transportation mode data are combined and fed into this model.

Table 3 summarizes the performance achieved by the MSTGNN model variants with different modes of transportation. The MSTGNN performs best when the bus, metro, and taxi information are considered, while MSTGNN-B performs poorly when only the bus

view is considered. Utilizing the 15-minute time granularity as an example, the MAE and RMSE of the MSTGNN model are improved by 10.06% and 11.51%, respectively, over those of MSTGNN-B. The difference in performance between the MSTGNN-BM and MSTGNN-BT models is minor, especially when the time granularity is 5 minutes, and MSTGNN-BM performs better than MSTGNN-BT when the time granularities are 15 and 30 minutes. From the analysis in Section 5.2, it is clear that the spatial correlation between taxis and buses is stronger. However, the temporal correlation between metros and buses is stronger. Combining the results, it can be suggested that the temporal correlations among multiple modules can play a more important role in achieving model performance improvements than spatial correlations.

Table 3: Performance comparison performed while considering different views.

Model	5 min			15 min			30 min		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
MSTGNN	0.807	1.919	0.985	**1.216**	**2.909**	0.965	**1.283**	**3.080**	**0.961**
MSTGNN-B	0.857	2.019	0.983	1.347	3.249	0.957	1.464	3.561	0.948
MSTGNN-BM	**0.803**	1.922	0.985	1.233	2.947	0.965	1.319	3.117	0.960
MSTGNN-BT	0.807	**1.917**	0.985	1.246	2.983	0.964	1.324	3.166	0.959

Figure 9 illustrates the spatial MAE distributions of the MSTGNN-B, MSTGNN-BM, and MSTGNN-BT models to investigate the contributions of metros and taxis. Darker TAZs have larger MAEs. MSTGNN-BM and MSTGNN-BT, which incorporate additional transportation modes, achieve improved prediction accuracies for most of the TAZs compared to MSTGNN-B. When considering metros, the MAE of MSTGNN-BM increases significantly in the TAZs along the metro lines and in the downtown area. Taking highlighted area ② as an example, MSTGNN-BM reduces the MAEs of the TAZs belonging to and around the metro station. MSTGNN-BT can make some contribution to the prediction performance in the suburbs where there are no metro stations (such as highlighted area ①).

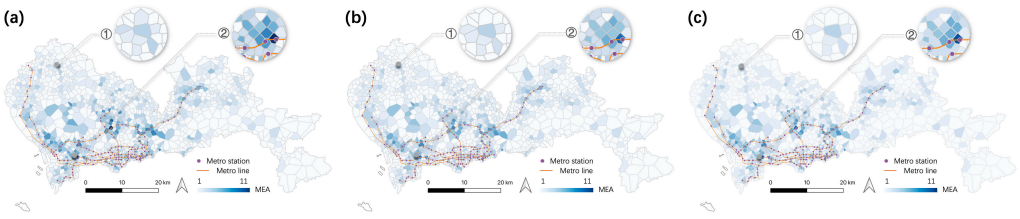

Figure 9: Spatial MAE distributions. (a) MSTGNN-B, (b) MSTGNN-BM, (c) MSTGNN-BT.

5.5.4. Contributions of local and global cross-view dependencies

Multimodal transportation interactions are mainly reflected in the local neighborhood transfer and the global commuter transfer processes. We develop two components in the MSTGNN model to capture local and global cross-view dependencies. To explore the contributions of the global and local cross-view dependencies to the model, we construct two simplified models to conduct ablation experiments.

- *MSTGNN-G:* The architecture is the same as that of the MSTGNN, but the adjacency graph network is abandoned, and the local MVATGCN module is removed.
- *MSTGNN-L:* The connection graph network and global MVATGCN module are removed, similar to MSTGNN-G, and only the local influences of the three views are considered.

Table 4 compares the performance achieved when considering global or local cross-view dependencies in the MSTGNN model. Overall, the MSTGNN, which considers both global and local cross-view dependencies, performs best. Specifically, although the MAE of the MSTGNN is greater than those of the other two models when the time granularity is set to 5 minutes, the RMSE and R^2 are superior to those of the other models. When the time granularity levels are 15 and 30 minutes, the MSTGNN significantly outperforms the other two models in terms of all three metrics. This suggests that the use of multiple graph structures can yield improved model performance by considering more comprehensive spatial relationships. Among the simplified models, MSTGNN-L is slightly better than the MSTGNN-G model considering global cross-view dependencies. This suggests that global dependencies contribute less than local dependencies to the travel demand prediction performance.

Table 4: Performance comparison between local and global cross-view dependencies.

Model	5 min			15 min			30 min		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
MSTGNN	0.807	**1.919**	**0.985**	**1.216**	**2.909**	**0.965**	**1.283**	**3.080**	**0.961**
MSTGNN-L	0.804	1.963	0.984	1.293	3.197	0.958	1.388	3.408	0.952
MSTGNN-G	**0.791**	1.967	0.984	1.318	3.241	0.957	1.435	3.427	0.952

5.5.5. Evaluation of the proposed losses

The travel demands for different transportation modes are unevenly distributed across space. In multiview learning, if the features of different views significantly differ, they may be considered data noise, thus reducing the resulting model performance (Cao et al., 2020). We propose the CVSFC loss to constrain the training process of the network. The auxiliary loss function may cope with missing and abnormal data for a particular view. It

has the ability to learn features of other views via feature consistency rules. Specifically, we design noisy datasets based on the SZ dataset. They are tested in two scenarios to demonstrate the effectiveness of the proposed CVSFC loss. (1) *SZ:* This is the real-world dataset from Shenzhen. (2) *SZ-Missing:* Based on the Shenzhen dataset, the bus data are kept constant, and 20% of the areas are randomly selected for metro and taxi data removal. On these two datasets, we examine the predictive performance achieved with and without the CVSFC loss. The loss weights λ_1 and λ_2 are set to 1 and 1, respectively.

The results are shown in Table 5; overall, the results obtained on these two datasets show that the proposed CVSFC loss can improve the robustness of the model with a very limited negative impact and even a slight performance improvement. Specifically, on the SZ dataset, although the accuracy of the model without the CVSFC loss is slightly higher than that with the CVSFC loss at a time granularity of 5 minutes, the performance achieved with the CVSFC loss improves at other time granularities, with MAE improvements of 3.45% and 3.82%, respectively. On the SZ-Missing dataset, the CVSFC loss enables MAE improvements of 2.34%, 3.90%, and 1.83% for the three time granularities. The CVSFC loss can improve the model performance at all time granularities when the feature distribution inhomogeneities are artificially increased. These results demonstrate that the proposed CVSFC loss is efficient for training our model and beneficial for addressing the uneven distribution of multiview features and enhancing the model's robustness.

Table 5: The results of loss performance comparisons conducted in different scenarios.

Dataset	loss	5 min			15 min			30 min		
		MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
SZ	$\mathcal{L}_{L2}$	**0.805**	**1.904**	0.985	1.258	2.943	0.965	1.332	3.097	0.961
	$\mathcal{L}_{L2} + \mathcal{L}_{sc}$	0.807	1.919	0.985	**1.216**	**2.909**	0.965	**1.283**	**3.080**	0.961
SZ-Missing	$\mathcal{L}_{L2}$	0.832	1.946	0.985	1.279	3.009	0.963	1.333	3.133	0.960
	$\mathcal{L}_{L2} + \mathcal{L}_{sc}$	**0.813**	**1.924**	0.985	**1.231**	**2.947**	**0.965**	**1.309**	**3.117**	0.960

5.5.6. Multiview interaction analysis

To intuitively investigate the roles of the attention mechanisms in our model and the dynamic interactions of buses, metros, and taxis, we plot an attention matrix. The model contains local and global multiview attention matrices. We select three different periods of the day from the two matrices for analysis purposes: morning rush hour (7:00-9:00), evening rush hour (17:00-19:00), and off-rush hour (14:00-17:00). We calculate the average attention weight across all channels of each view to obtain a $3 \times$ three attention matrix. Each row and column of the attention matrix can be divided into three parts, representing the bus, metro, and taxi views, as the input features are concatenated from the three transport views. As shown in Equation 11, the attention-weighted feature $\tilde{X} = concat\left(\ddot{X}^{bus}, \ddot{X}^{metro}, \ddot{X}^{taxi}\right) \cdot a$.

As shown in Figure 10, in the attention matrix, the weights of the columns represent each transportation mode's contribution to the feature $\tilde{X}$. The weights of the rows indicate how a certain transportation mode is affected. For the local attention map, overall, buses play a significant role in feature $\tilde{X}$, with metros making the weakest contribution. The contributions of buses, metros and taxis vary during different periods. During the morning rush hour, metros are more important than taxis, while during the evening rush hour, taxis are more critical than metros; the two contributions are similar during off-rush hours. For the global attention map, buses have the most significant impact on these features. The contribution of buses is most obvious during off-rush periods, but there are also substantial contributions from metros and taxis during the morning and evening rush hours. When comparing the local and global attention maps, the local features appear to be more influenced by buses. At the same time, the other modes (i.e., metros and taxis) also play important roles in the global features.

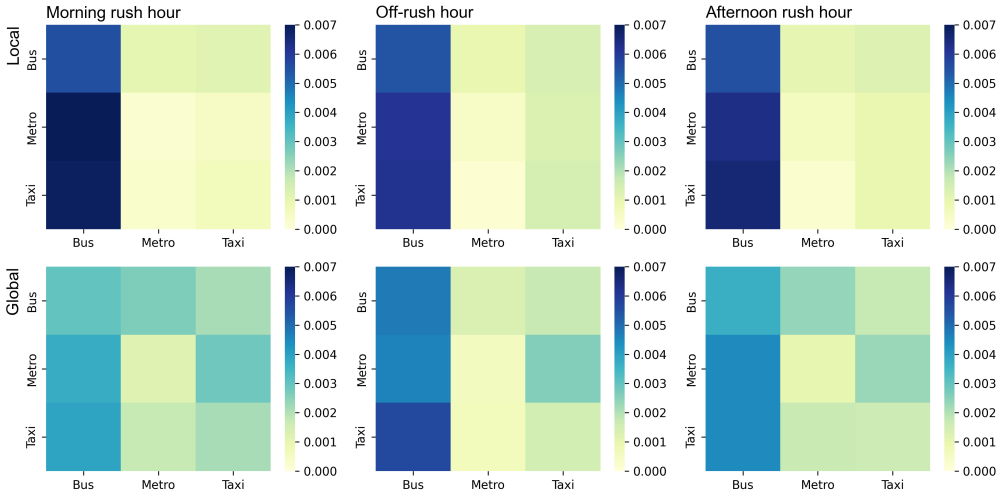

Figure 10: Global and local attention maps produced at various periods throughout the day.

6. Conclusions

This study presents an MSTGNN model that integrates bus, metro, and taxi information to predict travel demand by bus. The TCN and GCN layers are stacked to capture the temporal and spatial dependencies of each transportation mode, and an attention mechanism is developed to capture the cross-view interaction dependencies. The proposed model is evaluated on a real-world dataset from Shenzhen, which includes bus, metro, and taxi travel records. Experiments show that the MSTGNN model outperforms many state-of-the-art baselines according to the RMSE, MAE, and R^2 metrics. The proposed model reduces the MAE and RMSE by 5.59% and 3.51%, respectively. Further

analysis of MSTGNN's attention mechanism demonstrates its good interpretability for understanding the interactions between various transport modes. With the development of smart cities, it is possible to collect a large amount of real-time data on various transport modes, including shared bicycles, walking, and even private cars. The novel model of to capture the interaction dependencies between different modes of transportation can maximize the value of multisource big data.

The proposed model can be further improved in the following ways. First, several other factors that influence travel demand, including the built environment, weather conditions, emergencies, etc., can be further incorporated (Qi et al., 2018; Guo and Zhang, 2020). This research focused on the interactions among different transportation modes in multimodal travel cases to achieve more accurate travel demand prediction based on historical travel demand data. In future studies, multisource data can be examined for data fusion to further improve the prediction performance and robustness of the model. Second, our study was conducted only at the spatial scale of TAZs. The interactions among multiple transportation modes at various spatial resolutions may be worthy of exploration in future studies.

7. Acknowledgment

The authors would like to thank the editors Prof. May Yuan and Prof. Christophe Claramunt, and the anonymous reviewers for their constructive comments to improve the quality of the article. We gratefully acknowledge the Shenzhen Institute of Beidou Applied Technology for providing the raw data. We thank the members of the NUS Urban Analytics Lab for the discussions.

8. Data and code availability statement

The data and codes that support the findings of this study are available at figshare.com with the following link: https://doi.org/10.6084/m9.figshare.20394165.

9. Disclosure statement

No potential conflict of interest was reported by the author(s).

10. Funding

This study is supported and funded by the National Natural Science Foundation of China (No. 42071357, No. 42071360, No. 42001393); Key Project of Natural Science Foundation of Shenzhen (No. JCYJ20220818100200001); the Basic Research Program of Shenzhen Science and Technology Innovation Committee (No. JCYJ20220530152817039); KartoBit Research Network (No. KRN2202GK); Guangdong Science and Technology Strategic Innovation Fund (the Guangdong-Hong Kong-Macau Joint Laboratory Program, No. 2020B1212030009).

References

Ahmed, M.S., Cook, A.R., 1979. Analysis of freeway traffic time-series data by using Box-Jenkins techniques. 722.

Barry, J.J., Freimer, R., Slavin, H., 2009. Use of entry-only automatic fare collection data to estimate linked transit trips in New York City. Transportation research record 2112, 53–61.

Box, G.E., Pierce, D.A., 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. Journal of the American statistical Association 65, 1509–1526.

Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., Zhang, Q., Li, Q., Qiu, G., 2020. Deep learning-based remote and social sensing data fusion for urban region function recognition. ISPRS Journal of Photogrammetry and Remote Sensing 163, 82–97.

Chen, E., Zhang, W., Ye, Z., Yang, M., 2020. Unraveling latent transfer patterns between metro and bus from large-scale smart card data. IEEE Transactions on Intelligent Transportation Systems 23, 3351–3365.

Cheng, S., Peng, P., Lu, F., 2020. A lightweight ensemble spatiotemporal interpolation model for geospatial data. International Journal of Geographical Information Science 34, 1849–1872.

Cherry, T., Townsend, C., 2012. Assessment of potential improvements to metro–bus transfers in bangkok, thailand. Transportation research record 2276, 116–122.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv:1406.1078.

Cui, Z., Henrickson, K., Ke, R., Wang, Y., 2019. Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. IEEE Transactions on Intelligent Transportation Systems 21, 4883–4894.

Defferrard, M., Bresson, X., Vandergheynst, P., 2017. Convolutional neural networks on graphs with fast localized spectral filtering. arXiv:1606.09375.

Fang, Z., Shaw, S.L., Tu, W., Li, Q., Li, Y., 2012. Spatiotemporal analysis of critical transportation links based on time geographic concepts: a case study of critical bridges in wuhan, china. Journal of Transport Geography 23, 44–59.

Gao, F., Li, S., Tan, Z., Wu, Z., Zhang, X., Huang, G., Huang, Z., 2021. Understanding the modifiable areal unit problem in dockless bike sharing usage and exploring the interactive effects of built environment factors. International Journal of Geographical Information Science 35, 1905–1925.

Guo, G., Zhang, T., 2020. A residual spatio-temporal architecture for travel demand forecasting. Transportation Research Part C: Emerging Technologies 115, 102639.

Guo, S., Lin, Y., Feng, N., Song, C., Wan, H., 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting, in: Proceedings of the AAAI conference on artificial intelligence, pp. 922–929.

Hamed, M.M., Al-Masaeid, H.R., Said, Z.M.B., 1995. Short-term prediction of traffic volume in urban arterials. Journal of Transportation Engineering 121, 249–254.

Hernandez, S., Monzon, A., 2016. Key factors for defining an efficient urban transport interchange: Users' perceptions. Cities 50, 158–167.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Huang, L., Liu, X.X., Huang, S.Q., Wang, C.D., Tu, W., Xie, J.M., Tang, S., Xie, W., 2021. Temporal hierarchical graph attention network for traffic prediction. ACM Transactions on Intelligent Systems and Technology (TIST) 12, 1–21.

Karnberger, S., Antoniou, C., 2020. Network–wide prediction of public transportation ridership using spatio–temporal link–level information. Journal of Transport Geography 82, 102549.

Kashyap, A.A., Raviraj, S., Devarakonda, A., Nayak K, S.R., KV, S., Bhat, S.J., 2022. Traffic flow prediction models–a review of deep learning techniques. Cogent Engineering 9, 2010510.

Ke, J., Feng, S., Zhu, Z., Yang, H., Ye, J., 2021. Joint predictions of multi-modal ride-hailing demands:

A deep multi-task multi-graph learning-based approach. Transportation Research Part C: Emerging Technologies 127, 103063.

Ke, J., Zheng, H., Yang, H., Chen, X.M., 2017. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. Transportation research part C: Emerging technologies 85, 591–608.

Kim, K., 2020. Effects of weather and calendar events on mode-choice behaviors for public transportation. Journal of transportation engineering, Part A: Systems 146, 04020056.

Lau, K., Wu, Q., 2008. Local prediction of non-linear time series using support vector regression. Pattern recognition 41, 1539–1547.

Li, C., Bai, L., Liu, W., Yao, L., Waller, S.T., 2021. A multi-task memory network with knowledge adaptation for multimodal demand forecasting. Transportation Research Part C: Emerging Technologies 131, 103352.

Li, W., Batty, M., Goodchild, M.F., 2020a. Real-time GIS for smart cities. International Journal of Geographical Information Science 34, 311–324.

Li, W., Wang, S., Zhang, X., Jia, Q., Tian, Y., 2020b. Understanding intra-urban human mobility through an exploratory spatiotemporal analysis of bike-sharing trajectories. International Journal of Geographical Information Science 34, 2451–2474.

Li, X., Tu, W., Shen, S., Yue, Y., Luo, N., Li, Q., 2017. Revealing spatial variation and correlation of urban travels from big trajectory data. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 42, 53–57.

Liang, Y., Huang, G., Zhao, Z., 2022. Bike sharing demand prediction based on knowledge sharing across modes: A graph-based deep learning approach. arXiv:2203.10961.

Liu, Y., Lyu, C., Liu, X., Liu, Z., 2020. Automatic feature engineering for bus passenger flow prediction based on modular convolutional neural network. IEEE Transactions on Intelligent Transportation Systems 22, 2349–2358.

Ma, X., Zhang, J., Ding, C., Wang, Y., 2018. A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership. Computers, Environment and Urban Systems 70, 113–124.

Ma, Z., Xing, J., Mesbah, M., Ferreira, L., 2014. Predicting short-term bus passenger demand using a pattern hybrid approach. Transportation Research Part C: Emerging Technologies 39, 148–163.

May, A.D., 2013. Urban transport and sustainability: The key challenges. International journal of sustainable transportation 7, 170–185.

McLeod, S., Scheurer, J., Curtis, C., 2017. Urban public transport: planning principles and emerging practice. Journal of Planning Literature 32, 223–239.

McNally, M.G., 2007. The four-step model. Emerald Group Publishing Limited.

Miglani, A., Kumar, N., 2019. Deep learning models for traffic flow prediction in autonomous vehicles: A review, solutions, and challenges. Vehicular Communications 20, 100184.

Nassir, N., Hickman, M., Ma, Z.L., 2015. Activity detection and transfer identification for public transit fare card data. Transportation 42, 683–705.

Qi, G., Huang, A., Guan, W., Fan, L., 2018. Analysis and prediction of regional mobility patterns of bus travellers using smart card data and points of interest data. IEEE Transactions on Intelligent Transportation Systems 20, 1197–1214.

Rawat, W., Wang, Z., 2017. Deep convolutional neural networks for image classification: A comprehensive review. Neural computation 29, 2352–2449.

Recker, W.W., McNally, M.G., Root, G.S., 1986. A model of complex travel behavior: Part I—theoretical development. Transportation Research Part A: General 20, 307–318.

Ren, Y., Chen, H., Han, Y., Cheng, T., Zhang, Y., Chen, G., 2020. A hybrid integrated deep learning model for the prediction of citywide spatio-temporal flow volumes. International Journal of Geographical Information Science 34, 802–823.

Shu, W., Cai, K., Xiong, N.N., 2021. A short-term traffic flow prediction model based on an improved gate recurrent unit neural network. IEEE Transactions on Intelligent Transportation Systems 23, 16654–

16665.
Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J.A., Demšar, U., Fotheringham, A.S., 2016. Analysis of human mobility patterns from GPS trajectories and contextual information. International Journal of Geographical Information Science 30, 881–906.
Sultana, S., Salon, D., Kuby, M., 2019. Transportation sustainability in the urban context: A comprehensive review. Urban geography 40, 279–308.
Sun, J., Zhang, J., Li, Q., Yi, X., Liang, Y., Zheng, Y., 2022. Predicting citywide crowd flows in irregular regions using multi-view graph convolutional networks. IEEE Transactions on Knowledge and Data Engineering 34, 2348–2359.
Tao, S., Corcoran, J., Rowe, F., Hickman, M., 2018. To travel or not to travel:'weather'is the question. modelling the effect of local weather conditions on bus ridership. Transportation research part C: emerging technologies 86, 147–167.
Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. Economic geography 46, 234–240.
Tu, W., Cao, J., Yue, Y., Shaw, S.L., Zhou, M., Wang, Z., Chang, X., Xu, Y., Li, Q., 2017. Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. International Journal of Geographical Information Science 31, 2331–2358.
Tu, W., Cao, R., Yue, Y., Zhou, B., Li, Q., Li, Q., 2018. Spatial variations in urban public ridership derived from gps trajectories and smart card data. Journal of Transport Geography 69, 45–57.
Tu, W., Zhu, T., Xia, J., Zhou, Y., Lai, Y., Jiang, J., Li, Q., 2020. Portraying the spatial dynamics of urban vibrancy using multisource urban big data. Computers, Environment and Urban Systems 80, 101428.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. arXiv:1706.03762.
Wang, D., Chai, Y., Li, F., 2011a. Built environment diversities and activity–travel behaviour variations in Beijing, China. Journal of Transport Geography 19, 1173–1186.
Wang, J., Jiang, J., Jiang, W., Li, C., Zhao, W.X., 2021. Libcity: An open library for traffic prediction, in: Proceedings of the 29th International Conference on Advances in Geographic Information Systems, pp. 145–148.
Wang, P., Zhang, T., Zheng, Y., Hu, T., 2022. A multi-view bidirectional spatiotemporal graph network for urban traffic flow imputation. International Journal of Geographical Information Science 36, 1231–1257.
Wang, W., Attanucci, J.P., Wilson, N.H., 2011b. Bus passenger origin-destination estimation and related analyses using automated data collection systems. Journal of Public Transportation 14, 7.
Wu, J., Liao, H., 2020. Weather, travel mode choice, and impacts on subway ridership in beijing. Transportation research part A: policy and practice 135, 264–279.
Wu, P., Li, J., Pian, Y., Li, X., Huang, Z., Xu, L., Li, G., Li, R., 2022. How determinants affect transfer ridership between metro and bus systems: A multivariate generalized poisson regression analysis method. Sustainability 14, 9666.
Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C., 2019. Graph wavenet for deep spatial-temporal graph modeling. arXiv:1906.00121.
Xu, X.y., Liu, J., Li, H.y., Hu, J.Q., 2014. Analysis of subway station capacity with the use of queueing theory. Transportation research part C: emerging technologies 38, 28–43.
Yu, B., Yin, H., Zhu, Z., 2018. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 3634–3640.
Yu, F., Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122.
Yue, M., Kang, C., Andris, C., Qin, K., Liu, Y., Meng, Q., 2018. Understanding the interplay between bus, metro, and cab ridership dynamics in shenzhen, china. Transactions in GIS 22, 855–871.
Zaremba, W., Sutskever, I., Vinyals, O., 2015. Recurrent neural network regularization. arXiv:1409.2329.
Zhang, J., Chen, F., Guo, Y., Li, X., 2020a. Multi-graph convolutional network for short-term passenger flow forecasting in urban rail transit. IET Intelligent Transport Systems 14, 1210–1217.

Zhang, J., Zheng, Y., Qi, D., 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. arXiv:1610.00081.

Zhang, J., Zheng, Y., Qi, D., Li, R., Yi, X., Li, T., 2018a. Predicting citywide crowd flows using deep spatio-temporal residual networks. Artificial Intelligence 259, 147–166.

Zhang, T., Li, Y., Yang, H., Cui, C., Li, J., Qiao, Q., 2020b. Identifying primary public transit corridors using multi-source big transit data. International Journal of Geographical Information Science 34, 1137–1161.

Zhang, X., Xu, Y., Tu, W., Ratti, C., 2018b. Do different datasets tell the same story about urban mobility—a comparative study of public transit and taxi usage. Journal of Transport Geography 70, 78–90.

Zhang, Y., Cheng, T., Ren, Y., Xie, K., 2020c. A novel residual graph convolution deep learning model for short-term network-based traffic forecasting. International Journal of Geographical Information Science 34, 969–995.

Zhao, D., Wang, W., Woodburn, A., Ryerson, M.S., 2017. Isolating high-priority metro and feeder bus transfers using smart card data. Transportation 44, 1535–1554.

Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., Deng, M., Li, H., 2019. T-GCN: A temporal graph convolutional network for traffic prediction. IEEE Transactions on Intelligent Transportation Systems 21, 3848–3858.

Zhao, P., Li, S., 2017. Bicycle-metro integration in a growing city: The determinants of cycling as a transfer mode in metro station areas in Beijing. Transportation research part A: policy and practice 99, 46–60.

Zhao, T., Huang, Z., Tu, W., He, B., Cao, R., Cao, J., Li, M., 2022. Coupling graph deep learning and spatial-temporal influence of built environment for short-term bus travel demand prediction. Computers, Environment and Urban Systems 94, 101776.

Zhao, X., Yan, X., Yu, A., Van Hentenryck, P., 2020. Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. Travel behaviour and society 20, 22–35.

Zheng, C., Fan, X., Wang, C., Qi, J., 2020. GMAN: A graph multi-attention network for traffic prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1234–1241.

Zheng, Y., Capra, L., Wolfson, O., Yang, H., 2014. Urban computing: concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology (TIST) 5, 1–55.