When Will We Arrive? A Novel Multi-Task Spatio-Temporal Attention Network Based on Individual Preference for Estimating Travel Time

Guojian Zou, Ziliang Lai, Changxi Ma, Meiting Tu, Jing Fan, and Ye Li

***Abstract*— Predicting how long a trip will take may allow travelers plan ahead, save money, and avoid traffic congestion. The journey time estimation model should take into account three crucial factors: (1) individual travel preference, (2) dynamic spatio-temporal correlations, and (3) the association between long-term speed forecast and travel time estimate. In order to overcome these challenges, this study proposes a unique parallel architecture called the multi-task spatio-temporal attention network (MT-STAN) to estimate journey times. To extract the dynamic spatio-temporal correlations of the road network, we first develop a traffic speed prediction model based on spatio-temporal block and bridge transformer networks, combining the road, timestamp, and traffic speed information into hidden states. Second, we offer a personalized model for estimating journey times that makes use of cross-network, holistic attention, and semantic transformer. In this approach, travel preferences extraction through cross-network, holistic attention permits correlations between the dynamic road network's hidden states and individual journey characteristics, which are subsequently transformed into global semantics by the semantic transformer; preferences and semantics are integrated during the estimate phase. Finally, a multi-task learning component is included, which combines both traffic speed prediction and individual journey time estimate, via the sharing of underlying network parameters and the improvement of the contextual semantic knowledge of the latter job. Evaluation experiments are carried out using a highway dataset collected in Yinchuan City, Ningxia Province, China. The proposed prediction model outperforms state-of-the-art baseline approaches in experiments.**

***Index Terms*— Dynamic spatio-temporal correlations, individual travel preference, holistic attention, multi-task learning, traffic speed prediction, travel time estimation.**

I. Introduction

INTELLIGENT transportation systems (ITS) [1], [2] provide drivers with real-time departure and arrival timings through smartphone applications to aid in route planning, making precise travel time estimates a crucial service, as shown in Fig. 1. Various persons have different travel characteristics, hence individual travel time estimation (ITTE) has emerged as a new method in traffic information services [3]. The ITTE is crucial to the ITS because it keeps drivers from becoming stopped in traffic, cuts down on gas money, and eases stress on the road [4].

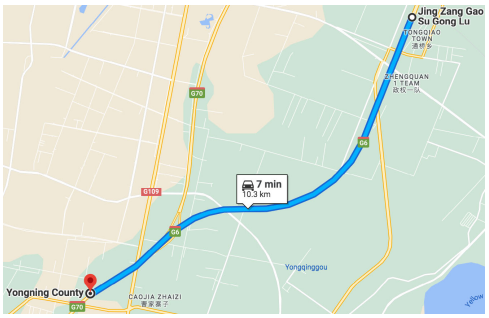


Fig. 1. Estimated time of arrival refers to the estimated travel time between a pair of origin and destination along the given route.

As one travel route is composed of several road segments (or local paths), there are two different perspectives for travel time estimation (TTE) to the given route: (1) For local TTE, first, the whole route is broken up into a few smaller ones; next, the travel time of each of those smaller paths is modeled; and finally, the travel time of all of those smaller paths is added together to get the total travel time for the route. (2) For collective TTE, it makes direct use of traffic data to calculate the total travel time along a given route. So far, many machine-learning and deep-learning methods have been developed from these two perspectives.

Researchers have utilized machine learning techniques to identify nonlinear correlations from trajectory data [5], [6], [7], [8], [9]. One such technique is support vector regression (SVR), whose performance exceeds that of the historical average predictor (HA) [5]. When compared to conventional statistical techniques, machine learning approaches excel in areas of high dimension and nonlinear characteristics, where they are able to address the nonlinear issue that has previously plagued these areas. Since journey time is modified by both spatial and temporal dimensions, TTE is a dynamic spatio-temporal prediction challenge. However, machine learning approaches can only be used to extract shallow nonlinear characteristics, and they cannot be used to simulate the deep dynamic spatio-temporal correlations of traffic data [10], [11].

Manuscript received 4 January 2023; revised 13 April 2023 and 6 May 2023; accepted 12 May 2023. This work was supported in part by the Project of the National Key Research and Development Program of China under Grant 2018YFB1601301, in part by the National Natural Science Foundation of China under Grant 71961137006, in part by the Natural Science Foundation of China under Grant 52062027, in part by the Key Research and Development Project of Gansu Province under Grant 22YF7GA142, in part by the Gansu Provincial Science and Technology Major Special Project—Enterprise Innovation Consortium Project under Grant 22ZD6GA010, and in part by the Lanzhou Jiaotong University Basic Research Top Talents Training Program under Grant 2022JC02. The Associate Editor for this article was Y. Kamarianakis. *(Corresponding authors: Jing Fan; Ye Li.)*

Please see the Acknowledgment section of this article for the author affiliations.

Digital Object Identifier 10.1109/TITS.2023.3276916

Recently, deep learning techniques like a magic box have been effectively used in a variety of contexts. It uses complex neural networks with many layers to automatically pull out features from input variables and then match the predicted and observed values. As a result, numerous fields of study, including traffic demand prediction [12], traffic speed prediction [13], [14], traffic flow prediction [15], and trip time estimate [3], [10], [11], [16], [17], [18], [19], [20], have been reexamined in light of this novel approach. Local- and collective-TTEs are performed using recurrent neural networks (RNNs) [3], [16], [17], [18], [21], convolutional neural networks (CNNs) [20], and graph neural networks (GNNs) [10], [11], [19], all of which provide superior performance than machine learning approaches. Thus, we might draw motivation from deep learning techniques to inform the development of more effective model architectures.

The ITTE in existing studies is based on the trajectory dataset with only taxi records [3], [10], [11], [16], [17], [18], [19]. However, the travel features, including the vehicle ID, location, road length, and departure time of all types of vehicles, are recorded by ETC intelligent monitoring sensors on highways, which can be used to get the driving path of each vehicle on the highway network. In addition, Individual Travel Time Estimation (ITTE) indicates that the travel time is estimated for the specific driver at the departure time after a given target route, the accuracy of which is easily affected by the two aspects of (1) traffic speed, the traffic speed of the road network is dynamic and has a different impact on individual travel time during different periods of time; (2) individual travel preference, the complex individual travel features indirectly reflect individual travel preference, such as vehicle ID, destination, and other complex features. Therefore, future traffic speeds can be utilized to provide vital information embedded within the ITTE. The validity of the traffic velocities is a prerequisite, and the idea is to couple the traffic speed prediction with the ITTE model. Furthermore, individual travel time on the local path is closely related to the entire route, and it is significant to estimate the travel time for each segment and whole route simultaneously. Using a single model to accomplish three tasks, an ideal inspiration emerges according to the preceding analysis.

A unique multi-task spatio-temporal attention network (MT-STAN) is suggested. The first essential component of MT-STAN is the traffic speed prediction (TSP) model, which includes the spatio-temporal block (ST-Block) and bridge transformer (BridgeTrans). ST-Block uses spatio-temporal attention and a fusion gate network to extract the dynamic spatio-temporal correlations of past road networks, whereas BridgeTrans uses multi-head attention mechanism to generate the future road network's hidden states. Second, we develop an ITTE model based on cross-network, holistic attention, and semantic transformer; the cross-network is used to zero in on interactions between travel features like vehicle-ID and -Type, while the holistic attention computes correlations between the individual travel features and the holistic sequence states (past and future) of each local path to update the individual travel representation, and then all local path's representations are translated into a global semantic by the semantic transformer. Third, multi-task learning is performed to exchange underlying network parameters, which increases ITTE task semantic knowledge by linking it with TSP. The contributions of this paper are summarized as follows:

1) We developed a TSP model using ST-Block and BridgeTrans to extract the dynamic spatio-temporal correlations of the road network across time. Specifically, ST-Block takes into account the dynamic influence of the road network on the traffic states of the target route, and BridgeTrans creates future traffic states on the target route, supplying crucial traffic information to the ITTE phase.
2) High individual travel preference representation is desired, hence the ITTE model based on cross-network, holistic attention, and semantic transformer is suggested. To simulate the interplay of separate journey characteristics, we use a cross-network. To further account for the influence of traffic state on the driver's trip duration, we establish a novel individual travel representation by calculating the correlations between the hidden states of a dynamic road network and the individual travel characteristics of each local route. Then, a semantic transformer is employed to convert all route representations into a global semantic.
3) Contextual semantic information for the ITTE task is improved by MT-STAN, a multi-task learning approach that takes advantage of the relationship between tasks and shares underlying network parameters. Speed prediction provides the traffic states of the highway network to the ITTE model, and the individual journey time predictions on each local path and entire route are promoted mutually. To prevent feature loss and internal covariate shift, residual connection [22] and batch normalization (BN) [23] are added to the network.
4) The highway traffic dataset serves as the basis for a number of investigations. The experimental results compare the suggested MT-STAN model favorably to its baseline counterparts, demonstrating its superiority.

The remaining sections of this work are structured as follows. Previous studies in this field are reviewed in Section II. A number of preliminary materials and a statement of the research topic are presented in Section III. The MT-STAN model is described in Section IV. Section V details the experiments conducted and their results. Lastly, Section VI concludes the paper and makes recommendations for further research.

II. Related Work

The function of TTE has been incorporated into an online map service that plays a crucial part in intelligent transportation systems (ITS) [24] and is extensively used in logistics planning [25] and emergency response planning. Recent efforts on TTE are extensive, but we concentrate primarily on two categories: overall and individual trip time calculations.

A. Overall-Based Travel Time Estimation

In order to accurately predict the total travel time of a route, traditional TTE algorithms relied heavily on macro factors (traffic speed, origin-destination travel time, traffic statuses, etc.). In TTE, the linear model is utilized since

it is a straightforward and common technique [26], [27]. For example, dynamic linear model (DLM) is proposed by Kwak and Geroliminis [26] to represent the non-linear traffic conditions. Machine learning techniques, in contrast to linear models, guarantee a global minimum of error for a given set of training data [5], [6], [7], [8], [9]. For example, to better represent the traffic pattern, unlike the linear model-historical average predictor (HA), the support vector regression (SVR) method used for trip time prediction by Wu et al. [5] is more accurate. In order to predict future traffic conditions and trip durations, Chiabaut and Faitout [6] suggest clustering the previous dataset of comparable traffic circumstances, for example using k-means, and calculating the similarity between the incoming data and the current cluster. A mean link travel time estimate methodology based on path recovery (PR-LTTE) is proposed by Sun et al. [7], which involves inferring the most probable route and then calculating the link travel time using the least-squares estimation using the inferred path. While these techniques may be used to discover things like nonlinear correlations or superficial traffic distributions in a historical dataset, they struggle when it comes to extracting deep, complicated spatial and temporal relationships.

Numerous studies have shown the superior performance of deep learning approaches to machine learning for TTE [19], [21], [28], [29], [30], [31], [32], [33], [34], [35]. **For temporal correlation**, for example, He et al. [34] offer a traffic pattern-centric segment coalescing framework (TP-SCF) that uses non-negative matrix factorization (NMF) and long short-term memory (LSTM) to learn diverse patterns of traffic situations and temporal correlations along various bus line segments. Ting et al. [33] present a deep hybrid model that combines the gated recurrent unit (GRU) with XGBoost through a linear layer to enable the temporal correlations extraction. **For spatio-temporal correlation**, for example, Petersen et al. [29] use a deep learning network that blends CNNs with LSTM layers to extract spatio-temporal correlations of bus route data, which is useful for studying temporal relationships between events. Graph neural networks (GNNs) can handle arbitrary issues in non-Euclidean space and are often used to extract the spatial correlations, whereas CNNs have limits in non-Euclidean space [13], [14], [19], [32]. Recent studies have shown that GNNs can be effectively applied to the TTE task; for instance, Ma et al. [19] propose a multi-attention graph neural network for city-wide bus travel time estimation (MAGTTE), which employs multi-head graph neural network (GAT) to model the dynamic spatial correlations and LSTM and Transformer [36] to extract temporal correlations.

B. Individual-Based Travel Time Estimation

Individual micro features (vehicle ID, vehicle Type, departure time, road index, traffic speed, etc.) are primarily used by individual-based TTE methods to represent the individual's travel preference, and deep learning methods are used as a key technique to achieve ITTE [2], [3], [4], [10], [11], [16], [17], [18], [37]. For example, Wang et al. [18] offer a multi-task deep learning model for travel time estimation (DeepTTE) that utilises a combination of spatio-temporal correlations and individual travel variables to provide estimates for the total journey duration and for each local path duration. Using a deep recurrent neural network, GPS trajectories, smartphone inertial data, and road networks are fused to provide individualized trip time estimates (CTTE; Gao et al. [17]). Using both Euclidean and non-Euclidean information, Han et al. [2] offer a multi-semantic path representation technique to boost estimate accuracy. Using high-order spatial and temporal dependence encoded in complex representations, Fu et al. [10] offer an estimated time of arrival (ETA) model called CompactETA, which also uses positional encoding to record the sequential information of the journey path. To address this issue, Jin et al. [11] present a unique spatial-temporal graph neural network (STGNN-TTE) that integrates individual journey time estimates with traffic conditions of road networks. These approaches, however, did not take into account how the interplay between input factors affected ITTE.

We believe that interactions between various individual travel characteristics have a favourable influence on our ITTE and must be accounted for in our prediction models. The factorization-machine (FM) algorithm imitates the interactions between variables and even solves the data sparsity problem in ITTE [38]. It has been employed as a crucial component of prediction algorithms to simulate the relationships between various individual travel characteristics [3], [16]. For example, Sun et al. [16] and Wang et al. [3] employ wide-deep-recurrent (WDR) to derive the individual travel preference since many drivers lack personalized data. Alleviating data sparsity concerns in estimated time of arrival (ADS-ETA) is a novel methodology proposed by Sun et al. [4] to address road network and driver sparsity issues. Chen et al. [37] offer a hybrid deep learning framework (HSETA) to estimate journey time by using FM and multi-layer perceptron (MLP) to extract global features, gate recurrent unit (GRU) to model route features, and an attention mechanism to combine the two.

In the category of overall-based TTE, these approaches lack individualised service; that is, it is impossible to deliver a correct TTE to a single individual driver for a given route due to the vast differences in travel preferences. In the category of individual-based TTE, these techniques do not account for dynamic traffic conditions on the road network, the association between individual travel features and dynamic traffic conditions on each local path, and the relationship between TSP, local path ITTE, and total route ITTE. In addition, all groups restrict the vehicle type, ignoring the impact of other vehicle types, such as coaches and maintenance trucks, on the target study. In this research, which was inspired by previous works on TTE [3], [4], [11], [16], a unique multi-task spatio-temporal attention network based on individual preference, known as MT-STAN, is presented and implemented in the highway network with all sorts of vehicles.

III. Preliminary

In this section, we discuss various preliminary considerations and define our problem formally.

Definition 1 (Road Network):
The road network is a linked graph in which the correlation between individual roads is intricate and changes depending on the speed of traffic and the passage of time. Giving the road network contains target study area, and then some properties

are added on the roads at time step t, including traffic speed $\mathrm{X}_{speed,t} \in \mathrm{R}^{\mathrm{N}\times 1}$, road index $\mathrm{X}_{index,t} \in \mathrm{R}^{\mathrm{N}\times 1}$, and timestamp $\mathrm{X}_{timestamp,t} \in \mathrm{R}^{\mathrm{N}\times 4}$ (week, day, hour, and minute). Before entering into the TSP model, the speed $\mathrm{X}_{speed,t} \in \mathrm{R}^{\mathrm{N}\times 1}$ is transformed to $\mathrm{XE}_{speed,t} \in \mathrm{R}^{\mathrm{N}\times d}$ using fully-connected layers. The timestamp and index are mapped to the dense matrixes $\mathrm{XE}_{index,t} \in \mathrm{R}^{\mathrm{N}\times d}$ and $\mathrm{XE}_{timestamp,t} \in \mathrm{R}^{\mathrm{N}\times 4\times d}$ through one-hot, and the converted method is similar to the embedded way in the BERT [39]. As an additional data input to the TSP model, we fed timestamp and road index embeddings via a two-layer fully connected network, resulting in spatio-temporal embeddings $\mathrm{STE} \in \mathrm{R}^{\mathrm{N}\times d}$. The number of road segments of the highway network is N; to facilitate multi-head attention internal divide exactly and consider computation cost, the dimension of d is set to 64 based on prior knowledge.

Definition 2 (Route):
This research uses the route, which is a sequence link with additional individual travel features on the intended origin-destination path, to study the individual journey time prediction. Giving travel features $\mathrm{T}^k \in \mathrm{R}^{d_\mathrm{T}+\mathrm{L}*2}$ of driver k on entire path, five types of properties are contained in local path v_i, including vehicle ID $\mathrm{T}^k_{vehicle} \in \mathrm{R}$, vehicle types $\mathrm{T}^k_{type} \in \mathrm{R}$, driver's departure time $\mathrm{T}^k_{departure} \in \mathrm{R}^5$ (week, day, hour, minute, and second), local path distance $\mathrm{T}^k_{distance,v_i} \in \mathrm{R}$, and local path index $\mathrm{T}^k_{index,v_i} \in \mathrm{R}$ on which road segment locates. Note, the vehicle ID, vehicle types, driver's departure time, and local path index are first mapped into one-hot vector, respectively, and local path distance maintain original value; new travel features obtained, $\acute{\mathrm{T}}^k = \left\{\acute{\mathrm{T}}^k_{vehicle}, \acute{\mathrm{T}}^k_{type}, \acute{\mathrm{T}}^k_{departure}, \acute{\mathrm{T}}^k_{distance}, \acute{\mathrm{T}}^k_{index}\right\} \in \mathrm{R}^n$, and n denotes the dimension of $\acute{\mathrm{T}}^k$. A sharing matrix space $\mathrm{ET} \in \mathrm{R}^{n\times d}$ is then defined, and map new travel features $\acute{\mathrm{T}}^k \in \mathrm{R}^n$ to the sharing matrix and obtain a dense representation $\hat{\mathrm{T}}^k \in \mathrm{R}^{(d_\mathrm{T}+\mathrm{L}*2)\times d}$. The segment count of the target route is L, $max\,(\mathrm{L}) = \mathrm{N}$; $d_\mathrm{T} = 7$ denotes the number of input variables (except local path distance and local path index) to the real-time ITTE model.

Definition 3 (Problem Statement):
As part of the training process, we learn to (1) extract the past and the future dynamic spatio-temporal correlations of the road network, (2) extract the individual travel preference through drivers' route and model interaction between features, (3) compute the correlations between individual travel features and the dynamic spatio-temporal correlations on each local path, and (4) associate different prediction tasks, share underlying network parameters. During the test phase, given the historical road network observations $\mathrm{X} = \left\{\mathrm{X}_{t_1}, \cdots, \mathrm{X}_{t_\mathrm{P}}\right\} \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d_\mathrm{X}}$ and drivers' route $\mathrm{T}^k \in \mathrm{R}^{d_\mathrm{T}+\mathrm{L}*2}$, we aim to predict the traffic speed of road network, and evaluate the entire- and local- path travel time of driver k, expressed as $\hat{\mathrm{Y}}^1 = \left\{\hat{\mathrm{Y}}^1_{t_{\mathrm{P}+1}}, \cdots, \hat{\mathrm{Y}}^1_{t_{\mathrm{P}+\mathrm{Q}}}\right\} \in \mathrm{R}^{\mathrm{Q}\times\mathrm{L}\times 1}$, $\hat{\mathrm{Y}}^2 \in \mathrm{R}$, and $\hat{\mathrm{Y}}^3 = \left\{\hat{\mathrm{Y}}^3_{v_1}, \cdots, \hat{\mathrm{Y}}^3_{v_\mathrm{L}}\right\} \in \mathrm{R}^{\mathrm{L}\times 1}$, respectively. Note, 3 denotes the number of tasks; $d_\mathrm{X} = 6$ presents the number of input variables to the TSP model.

Remark: The traffic speed and route data in the aforementioned sets (training, validation, and test) must first be aligned. The study does not focus on the topic of route optimization.

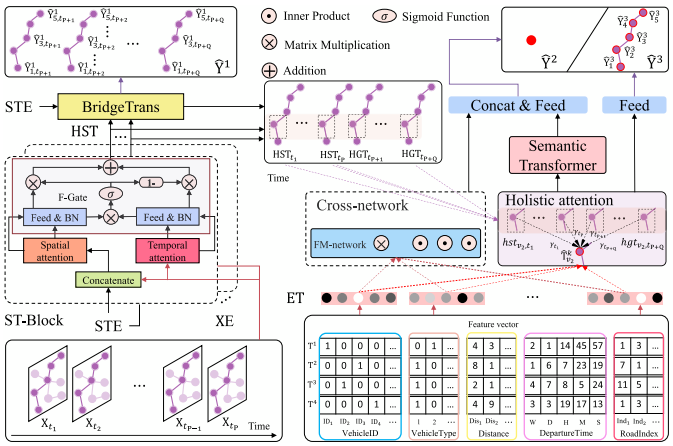


Fig. 2. Architecture of MT-STAN. **Left**: traffic speed prediction model. **Right**: travel time estimation model.

IV. Proposed Approach

A. Framework Overview

The proposed MT-STAN model is shown in Fig. 2 as a parallel structure consisting of traffic speed prediction-, journey time estimation-, and multi-task learning models. For starters, the highway system is a dynamic graph, with traffic speeds being a major influence on how long it takes to get from one place to another.

A traffic speed prediction model built on ST-Block and BridgeTrans is intended to extract the spatio-temporal correlations of the highway system, allowing for more accurate predictions of travel times. A ITTE model based on cross-network, holistic attention, and semantic transformer is proposed to directly model the interaction between driver's route features, compute the correlations between individual travel features and dynamic spatio-temporal correlations and then form a global semantic. This is because the driver's route features reflect the individual travel preference and play a significant role in estimating the travel time for a given trip. Third, there is a strong relationship between TSP and ITTE; a multi-task learning approach is utilized to merge underlying both TSP- and ITTE- models, sharing underlying network parameters, and enriching the contextual semantic information for ITTE. The next section expands upon every facet of the aforementioned strategy.

B. Traffic Speed Prediction Model

1) ST-Block: Assume that the input of ST-Block is $\mathrm{XE} \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d}$ and $\mathrm{STE} \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d}$, and output is $\mathrm{HST} \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d}$, in which the hidden output state of road segment v_i at time step t_j is $hst_{v_i,t_j} \in \mathrm{R}^d$. The outputs of spatial attention, temporal attention, and fusion gate network in the l^{th} layer are $\mathrm{HDS}^l \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d}$, $\mathrm{HDT}^l \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d}$, and $\mathrm{HST}^l \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d}$, respectively, while the dynamic spatial correlation, dynamic temporal correlation, and spatio-temporal correlation of road segment v_i at time step t_j are $hds^l_{v_i,t_j} \in \mathrm{R}^d$, $hdt^l_{v_i,t_j} \in \mathrm{R}^d$, and $hst^l_{v_i,t_j} \in \mathrm{R}^d$, where P denotes the input series length of ST-Block.

Since this study uses a nonlinear transformation function at high frequencies, it is first defined as:

$$f\,(x) = \mathrm{ReLU}\,(x\mathrm{W} + b) \tag{1}$$

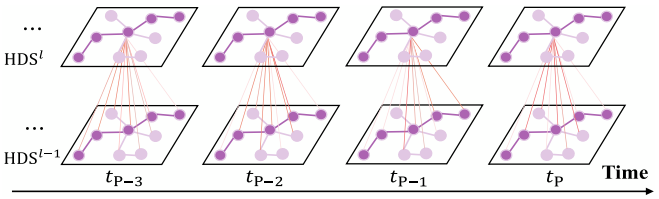


Fig. 3. Spatial attention model of the spatial correlations between different road segments.

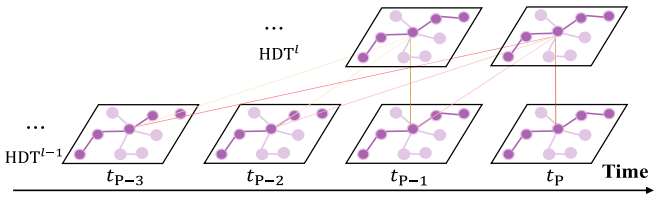


Fig. 4. Temporal attention model of the temporal correlations between different time steps.

where x represents the input data, W and b denote the weight matrices and bias, respectively, and ReLU is the nonlinear activation function.

a) Spatial attention: The term "dynamic spatial correlation" describes the way in which the effect weights by others on a certain road segment in the road network varies over time, and hence affects the traffic speed of the target road segment. For instance, as shown in Fig. 3, the upstream of shallow purple roads may have a negative impact on the traffic speeds of the dark purple road during rush hour, and this impact may lessen when the congestion is alleviated. To model the dynamic spatial correlation, we designed a spatial attention network, i.e., a multi-head graph attention network (multi-head GAT) to adaptively model the correlations between distinct road segments.

For road segment v_i, at time step t_j, the correlation coefficient $\alpha^m_{v_i,v}$ between road segments v_i and v is,

$$\alpha^m_{v_i,v} = \frac{exp\left(\mu^m_{v_i,v}\right)}{\sum_{v_r}^{\mathrm{V}} exp\left(\mu^m_{v_i,v_r}\right)} \tag{2}$$

where $\mu^m_{v_i,v}$ denotes the relevance between v_i and v, V represents the whole road segments.

The relevance $\mu^m_{v_i,v}$ can be obtained by the inner product of the query vector of road segment v_j and the key vector of road segment v,

$$\mu^m_{v_i,v} = \frac{\left\langle f^m_q\left(\left[hds^{l-1}_{v,t_j} ste_{v,t_j}\right]\right), f^m_k\left(\left[hds^{l-1}_{v,t_j}, ste_{v,t_j}\right]\right)\right\rangle}{\sqrt{d}} \tag{3}$$

where f^m_q and f^m_k are respectively the nonlinear transformation functions in the m^{th} head attention of the query vector and the key vector, $\langle *, * \rangle$ represents the inner product operator, and $[*, *]$ represents the binary concatenation.

After obtaining the correlation coefficient $\mu^m_{v_i,v}$ between road segments v_i and v in the m^{th} head attention, the l^{th} layer dynamic spatial correlation $hds^l_{v_i,t_j} \in \mathrm{R}^d$ of road segment v_i at time step t_j can be formulated as,

$$hds^{l,m}_{v_i,t_j} = \sum_{v_r}^{\mathrm{V}} \alpha^m_{v_i,v_r} f^m_v\left(\left[hds^{l-1}_{v_r,t_j}, ste_{v_r,t_j}\right]\right) \tag{4}$$

$$hds^l_{v_i,t_j} = \mathrm{BN}\left(\|^{\mathrm{M}}_{m=1} hds^{l,m}_{v_i,t_j} \mathrm{W}_{ds} + hds^{l-1}_{v_i,t_j}\right) \tag{5}$$

where f^m_v is the nonlinear transformation function in the m^{th} head attention of the value vector, M is the number of heads, and $\|$ represents the concatenation. The final dynamic spatial correlation $hds_{v_i,t_j} \in \mathrm{R}^d$ of road segment v_i can be calculated using Equations (2)-(5) at time step t_j. The initial input is $\mathrm{XE} \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d}$ and $\mathrm{STE}[:\mathrm{P}] \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d}$, and the output is $\mathrm{HDS} \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d}$.

b) Temporal attention: With dynamic temporal correlation, the effect weight of past traffic speeds on future speeds at any given road segment in a traffic network varies continuously over time. For instance, congestion during morning rush hour may be influenced by the previous traffic pace, and this effect may build up over time until easing. To adaptively describe the relationships between multiple time steps, we develop a temporal attention technique based on Transformer [36], as illustrated in Fig. 4.

For road segment v_i, at time step t_j, the correlation coefficient $\beta^m_{t_j,t}$ between time steps t_j and t is,

$$\beta^m_{t_j,t} = \frac{exp\left(\xi^m_{t_j,t}\right)}{\sum_{t_r}^{\mathcal{N}_{t_\mathrm{P}}} exp\left(\xi^m_{t_j,t_r}\right)} \tag{6}$$

where $\xi^m_{t_j,t}$ denotes the relevance between t_j and t, $\mathcal{N}_{t_\mathrm{P}}$ denotes a set of time steps before t_P.

The relevance can be obtained by the inner product of the query vector of road segment v_i at time step t_j and the key vector of road segment v_i at time step t,

$$\xi^m_{t_j,t} = \frac{\left\langle f^m_q\left(hdt^{l-1}_{v_i,t_j}\right), f^m_k\left(hdt^{l-1}_{v_i,t}\right)\right\rangle}{\sqrt{d}} \tag{7}$$

Once the correlation coefficient $\beta^m_{t_j,t}$ in the m^{th} head attention is obtained, the l^{th} layer temporal correlation $hdt^l_{v_i,t_j} \in \mathrm{R}^d$ of road segment v_i at time step t_j can be formulated as,

$$hdt^{l,m}_{v_i,t_j} = \sum_{t_r}^{\mathcal{N}_{t_\mathrm{P}}} \beta^m_{t_j,t_r} f^m_v\left(hdt^{l-1}_{v_i,t_r}\right) \tag{8}$$

$$hdt^l_{v_i,t_j} = \mathrm{BN}\left(\|^{\mathrm{M}}_{m=1} hdt^{l,m}_{v_i,t_j} \mathrm{W}_{dt} + hdt^{l-1}_{v_i,t_j}\right) \tag{9}$$

The final temporal correlation $hdt_{v_i,t_j} \in \mathrm{R}^d$ of road segment v_i can be calculated using Equations (6)-(9) at time step t_j. The initial input of temporal attention is just $\mathrm{XE} \in \mathrm{R}^{\mathrm{Q}\times\mathrm{N}\times d}$, and the output is $\mathrm{HDT} \in \mathrm{R}^{\mathrm{Q}\times\mathrm{N}\times d}$.

c) Fusion gate network (F-Gate): The traffic speed on a given road segment at a given time step is connected to its speed at earlier and subsequent time steps, as well as to the speeds on other roads. In order to successfully fuse spatial and temporal representations, we create a new basic network, as shown in Fig. 2 (left), and the output of F-Gate is $\mathrm{HST}^l \in \mathrm{R}^{\mathrm{P}\times\mathrm{L}\times d}$ without any further parameters.

The working process of F-Gate is,

$$\mathrm{HST}^l = \mathcal{Z} \odot \mathrm{HDS}^l + (1 - \mathcal{Z}) \odot \mathrm{HDT}^l \tag{10}$$

with

$$\mathcal{Z} = \sigma\left(\mathrm{HDS}^l \odot \mathrm{HDT}^l\right) \tag{11}$$

where σ represents sigmoid activation, and $\mathcal{Z} \in \mathrm{R}^{\mathrm{Q}\times\mathrm{L}\times d}$ is weight vector that controls the flow of spatial and temporal representations at each time step.

2) BridgeTrans: BridgeTrans based on Transformer is used to generate target sequence instead of step-by-step decoding way [36], [40]. There are two differences between BridgeTrans and temporal attention: (1) we just consider the correlation between historical inputs and future spatio-temporal embeddings, because the future traffic speeds are unknown, and (2) the generative inference decoding way is used that can avoid the inference error propagation in spatial and temporal dimensions [40]. The initial inputs are $\mathrm{HST} \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d}$ and $\mathrm{STE[P:P+Q]} \in \mathrm{R}^{\mathrm{Q}\times\mathrm{N}\times d}$, and output is $\mathrm{HGT} \in \mathrm{R}^{\mathrm{Q}\times\mathrm{L}\times d}$. Note, for the generative inference decoding way, the hidden representations of road network at time step $t_{\mathrm{P}+j}$ are not limited by previous generative results from time steps $t_{\mathrm{P}+1}$ to $t_{\mathrm{P}+j}$ and just have relationship with historical spatio-temporal correlations $\mathrm{HST} \in \mathrm{R}^{\mathrm{P}\times\mathrm{N}\times d}$ and spatio-temporal embeddings $\mathrm{STE[P:P+Q]} \in \mathrm{R}^{\mathrm{Q}\times\mathrm{N}\times d}$.

C. Individual Travel Time Prediction Model

1) Cross-Network: Personal preferences for many aspects of transportation are reflected in the data collected by ETC's intelligent monitoring sensors. The travel attributes of a car and a van, for example, are different from those of a bus or a truck. However, the current literature on individual trip time prediction does not take into account these variations. In addition, the interaction between travel features reflects the individual travel regularity more. For example, the feature pair <vehicle ID, departure time>: <F-17502, 2021-05-07 08:15:23> represents that driver of vehicle F-17502 may be a worker and often depart from home to the office at 08:15. Therefore, the interaction between features is essential for individual travel time estimation because different feature pairs maybe express different aspects of travel preference. To model these properties, we proposed a cross-network based on factorization machine (FM) network [38]; the interaction between features is deeply extracted using FM network to obtain high-level semantics which breaks the independence of features and can better mine the correlation between features. The process of FM network is defined as,

$$\mathrm{HC}^k = \sum_{i}^{n} \mathrm{W}_i \acute{\mathrm{T}}_i^k + \sum_{i}^{n} \sum_{j=i+1}^{n} \left\langle \mathrm{ET}_i, \mathrm{ET}_j \right\rangle \acute{\mathrm{T}}_i^k \acute{\mathrm{T}}_j^k + b \tag{12}$$

where W_i is a weight coefficient, which is used to model the strength of $\acute{\mathrm{T}}_i^k$, and b denotes a bias. When the $\acute{\mathrm{T}}_i^k \acute{\mathrm{T}}_j^k$ is 'one', there is an interaction between i^{th}- and j^{th}- features; on the contrary, if $\acute{\mathrm{T}}_i^k \acute{\mathrm{T}}_j^k$ is 'zero', there is no interaction between i^{th}- and j^{th}- features. $\langle \mathrm{ET}_i, \mathrm{ET}_j \rangle$ represents the interaction between i^{th}- and j^{th}- features when $\acute{\mathrm{T}}_i^k \acute{\mathrm{T}}_j^k$ is 'one'. The initial inputs are $\acute{\mathrm{T}}^k$ and ET, and the output is $\mathrm{HC}^k \in \mathrm{R}$.

2) Holistic Attention: An individual's time spent in transit is influenced by the congestion levels of the road system as a whole. For instance, if we decide to take a trip during rush hour, it might end up costing us more time on the road than usual since the speed of each individual car is limited by the congestion on the roads. In this study, we propose a holistic attention network to link past and projected traffic speeds of a road network with specific travel characteristics, and then combine the findings of holistic attention with cross-network to estimate individual journey times, as shown in Fig. 2. (right). Assume that the output hidden state of holistic attention is $hdh_{\hat{\mathrm{T}}_{v_i}^k} \in \mathrm{R}^d$ for local path v_i, and the correlation coefficient $\gamma^m_{\hat{\mathrm{T}}_{v_i}^k, t_j}$ between travel representation $\hat{\mathrm{T}}_{v_i}^k$ of driver k and time step t_j is,

$$\gamma^m_{\hat{\mathrm{T}}_{v_i}^k, t_j} = \frac{exp\left(\tau^m_{\hat{\mathrm{T}}_{v_i}^k, t_j}\right)}{\sum_{t_r}^{\mathcal{N}_{t_{\mathrm{P}+\mathrm{Q}}}} exp\left(\tau^m_{\hat{\mathrm{T}}_{v_i}^k, t_r}\right)} \tag{13}$$

where $\tau^m_{\hat{\mathrm{T}}_{v_i}^k, t_j}$ denotes the relevance between travel representation $\hat{\mathrm{T}}_{v_i}^k$ of driver k and time step t_j, and $\mathcal{N}_{t_{\mathrm{P}+\mathrm{Q}}}$ denotes a set of time steps before $t_{\mathrm{P}+\mathrm{Q}}$.

The relevance can be obtained by the inner product of the query vector of driver k at local path v_i and the key vector of local path v_i at time step t_j,

$$\tau^m_{\hat{\mathrm{T}}_{v_i}^k, t_j} = \frac{\left\langle f_q^m\left(hdh^{l-1}_{\hat{\mathrm{T}}_{v_i}^k}\right), f_k^m\left(hgt_{v_i,t_j} \; if \;\; (t_j > t_{\mathrm{P}}) \;\; else \; hst_{v_i,t_j}\right)\right\rangle}{\sqrt{d}} \tag{14}$$

Once the correlation coefficient $\gamma^m_{\hat{\mathrm{T}}_{v_i}^k, t_j}$ in the m^{th} head attention is obtained, the l^{th} layer holistic hidden state $hdh^l_{\hat{\mathrm{T}}_{v_i}^k} \in \mathrm{R}^d$ of driver k at local path v_i can be formulated as,

$$hdh^{l,m}_{\hat{\mathrm{T}}_{v_i}^k} = \sum_{t_r}^{\mathcal{N}_{t_{\mathrm{P}+\mathrm{Q}}}} \gamma^m_{\hat{\mathrm{T}}_{v_i}^k, t_r} f_v^m\left(hgt_{v_i,t_r} \; if \;\; (t_r > t_{\mathrm{P}}) \;\; else \; hst_{v_i,t_r}\right) \tag{15}$$

$$hdh^l_{\hat{\mathrm{T}}_{v_i}^k} = \mathrm{BN}\left(\|_{m=1}^{\mathrm{M}} \, hdh^{l,m}_{\hat{\mathrm{T}}_{v_i}^k} \mathrm{W}_{dh} + hdh^{l-1}_{\hat{\mathrm{T}}_{v_i}^k}\right) \tag{16}$$

The final holistic hidden state $hdh_{\hat{\mathrm{T}}_{v_i}^k} \in \mathrm{R}^d$ of driver k at local path v_i can be calculated using Equations (13)-(16). The initial input of holistic attention is $\hat{\mathrm{T}}^k \in \mathrm{R}^{(d_{\mathrm{T}}+\mathrm{L}*2)\times d}$, $\mathrm{HST} \in \mathrm{R}^{\mathrm{P}\times\mathrm{L}\times d}$, and $\mathrm{HGT} \in \mathrm{R}^{\mathrm{Q}\times\mathrm{L}\times d}$, and the output is $\mathrm{HDH}^k \in \mathrm{R}^{\mathrm{L}\times d}$.

3) Semantic Transformer: After the holistic attention, we can get new presentation of driver k at each local path v_i, that is, hidden state $hdh_{\hat{\mathrm{T}}_{v_i}^k} \in \mathrm{R}^d$. To combine all the hidden states on the entire route with length L, a semantic transformer based on 1-D convolutional neural networks (1-D CNNs) is used to translate the hidden conditions into a global semantic and then combined with travel preference on travel time estimation phase [41]. For individual travel time

estimation on the entire path, we use 1-D CNNs to extract the global semantic of $\mathrm{HDH}^k \in \mathrm{R}^{\mathrm{L}\times d}$ between different road segments,

$$\mathrm{H}^k = \sum_i^{\mathrm{L}} f\left(w * \mathrm{HDH}^k\right)_{v_i} \tag{17}$$

where $'*'$ denotes convolution calculation, w represents convolutional kernel. The initial input of 1-D CNN is $\mathrm{HDH}^k \in \mathrm{R}^{\mathrm{L}\times d}$, and the output is $\mathrm{H}^k \in \mathrm{R}^d$.

D. Multi-Task Learning

We use traffic speed and travel time models to obtain four different final hidden representations HST, H^k, HC^k, and HDH^k, and feeding to different subtask layers to generate target prediction values,

$$\widehat{\mathrm{Y}} = \begin{cases} \widehat{\mathrm{Y}}^1 = (\mathrm{HST})\,\mathrm{W}_1, \widehat{\mathrm{Y}}^1 \in \mathrm{R}^{\mathrm{Q}\times\mathrm{L}\times 1} \\ \widehat{\mathrm{Y}}^2 = \left(\mathrm{H}^k\right)\mathrm{W}_2 + \mathrm{HC}^k, \widehat{\mathrm{Y}}^2 \in \mathrm{R}^1 \\ \widehat{\mathrm{Y}}^3 = \left(\mathrm{HDH}^k\right)\mathrm{W}_3, \widehat{\mathrm{Y}}^3 \in \mathrm{R}^{\mathrm{L}\times 1} \end{cases} \tag{18}$$

where $\mathrm{W}_1 \in \mathrm{R}^{d\times 1}$, $\mathrm{W}_2 \in \mathrm{R}^{d\times 1}$, and $\mathrm{W}_3 \in \mathrm{R}^{d\times 1}$ represent the weight matrices of the three different fully connected layers. $\hat{\mathrm{Y}}^1$ represents predicted traffic speeds, $\hat{\mathrm{Y}}^2$ represents the travel time of driver k on the entire path, and $\hat{\mathrm{Y}}^3$ represents local paths' travel times of driver k.

The loss function of MT-STAN corresponding to the multi-task layer is defined as the mean absolute error (MAE) between observed values Y and predicted values $\hat{\mathrm{Y}}$,

$$\begin{aligned} L(\theta) = &\frac{\lambda}{\mathrm{Q}\times\mathrm{L}}\sum_j^{\mathrm{Q}}\sum_i^{\mathrm{L}}\left|\mathrm{Y}^1_{v_i,t_j} - \widehat{\mathrm{Y}}^1_{v_i,t_j}\right| + \eta\left|\mathrm{Y}^2 - \widehat{\mathrm{Y}}^2\right| \\ &+ \frac{(1-\lambda-\eta)}{\mathrm{L}}\sum_i^{\mathrm{L}}\left|\mathrm{Y}^3_{v_i} - \widehat{\mathrm{Y}}^3_{v_i}\right| \end{aligned} \tag{19}$$

where θ denotes all the learnable parameters in MT-STAN, λ and η represent loss weights.

V. Experiments

A. Data Description

The ETC intelligent monitoring sensors installed at the gantries and toll stations along the highway in Yinchuan City, Ningxia Province, China, provide the traffic data utilized in this analysis. As can be seen in Fig. 5, the highway network under investigation is broken up into 108 individual road segments by the monitoring devices. Driving paths of each vehicle on the highway network are produced using data from ETC intelligent monitoring sensors that capture the real-time location of the vehicle. As our ITTE research objectives, we've settled on the **Fuyin Expressway** (G70), **Jingzang Expressway** (G6), **YinKun Expressway** (G85), and **Qingyin Expressway** (G20).

Five road segments totaling 47.959 km are chosen for G70, with the gantry between Yinchuan North Hub and Helan Mountain Road Toll Station serving as the beginning and the gantry between Yongning- and Yesheng- Toll Stations serving as the destination. Five road segments are chosen with a total length of 50.882 km for G6, with the gantry between Shilidian Hub and Jingui Toll Station serving as the starting point and the gantry between Yongning- and Yesheng Toll Stations serving as the ending point. For G85, three road segments totaling 29.827 km are chosen between the origin at the gantry between Lingwu North- and Airport South- Toll Station and the destination at the gantry between Shiba- and Zhangzheng- Hubs. Five road segments totaling 29.636 km in length are chosen for G20, with the gantry starting at the Yinchuan North Hub and ending at the Helan Mountain Road Toll Station serving as the origin and the gantry ending at the Shuidonggou Toll Station serving as the destination.

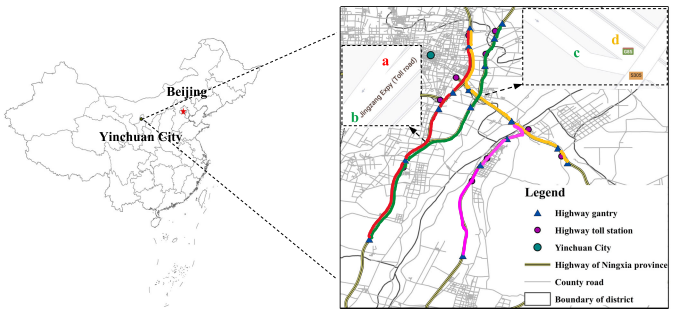


Fig. 5. Study area.

Our research region has a total of 28.090 million driving recordings from all of the monitoring devices. However, full route data between origin and destination are few, and there are 45.986k records in G70, 43.525k in G6, 1.683k in G85, and 19.233k in G20. Because highways have numerous ramps, from the main road to the exit toll station and the main road to the other main road through the Hub. Given a long route from the origin (O) to the destination (D), multiple ramps exist between O and D, and a high ratio of vehicles from O to the exit toll station or other main roads via Hub. As a result, only a few vehicles travel directly from O to D, while numerous vehicles travel from O to other locations before D. To prevent traffic speed data sparsity, the time series form of the traffic speed is produced by periodically measuring the speed along each road segment, in this case, every 15 minutes. The start date is June 1, 2021, and the end date is August 31, 2021.[1]

B. Baselines and Metrics

1) Individual Travel Time Estimation: DNN [31], which uses a deep multi-layer perceptron to estimate individual travel time in this paper. CoDriver ETA [16], which introduces an auxiliary task to learn an embedding of the personalized historical driving preference under multi-task learning. DeepTTE [18], a multi-task deep learning model, and it combines spatio-temporal correlation with individual travel preference to estimate the travel time of both the entire- and local-paths. WDR [3], explore the personalization information and solve the data sparsity problem because many drivers are short of personalized data. CompactETA [10], which encodes high order spatial and temporal dependency into sophisticated

[1] https://github.com/zouguojian/Travel-time-prediction/tree/main/data

representations, and further encode the sequential information of the travel route by positional encoding. CTTE [17], which uses deep recurrent neural networks to fuse features from phone and road networks, using multi-task learning that combines individual travel time estimation with traffic speed prediction. However, for the highway network, there is no phone data in this paper. MT-STAN, proposed in this paper, is based on the ITTE model, TSP model, and multi-task learning. MT-STAN-1, which is variant of MT-STAN that does not contain the TSP model and multi-task learning, and merely using ITTE model based on cross-network and semantic transformer for individual travel time estimation on entire path. Note that we use two nonlinear transformation layers instead of holistic attention to model each local path's representation. MT-STAN-2, which is variant of MT-STAN that does not contains holistic attention component in the ITTE model, using the addition operation to combine the historical spatio-temporal correlations of road network with to replace the holistic attention. MT-STAN-3, which is variant of MT-STAN that does not contains multi-task learning component, merely one task, that is, individual travel time estimation on entire path.

2) Travel Speed Prediction: LSTM [42], which is used to capture the nonlinear traffic dynamic characteristics. T-GCN [43], which combines the GCN and GRU to model the spatio-temporal correlations. DCRNN [44], which captures spatial correlation using bidirectional random walks on the graph, using the diffusion convolutional GRU to model the temporal correlation. ST-GRAT [14], which designs an encoder-decoder architecture using spatial and temporal attention based on the transformer to capture the spatio-temporal dynamics in road networks [36].

In order to evaluate the prediction performance of the MT-STAN model, three metrics are used to determine the difference between the observed values and the prediction values: the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Note that low RMSE, MAE, and MAPE values indicate a more accurate prediction performance.

C. Experimental Settings

Our MT-STAN model's hyperparameters and baselines are optimised during training by picking the model with the lowest mean absolute error (MAE) from the validation set. Therefore, the model error determined by the validation set is used to choose the best model to use. The procedure entails the following details: 50 epochs are used in each experiment. The model is put to the test on the validation set once it has been trained for an epoch. We revise and save the model parameters if the MAE of the prediction model drops on the validation set. Once the prediction impact of the prediction model on the validation set is maximized after extensive experimentation with various parameter settings, the training procedure is complete. The test set samples are then iterated until a prediction is reached. All of our studies use an early-stop mechanism, with parameters of 300 early-stop rounds and 10 maximum epochs. To consist with existing studies, we decided on a value of 6 for the objective time step Q and a value of 12 for the historical time step P. Seventy percent of the information was used as a training set, 15 percent as a validation set, and 15 percent as a test set in the experiment.

TABLE I
MODEL PARAMETERS

Component Name	Parameters	Values
Spatial Attention	hidden nodes	64
	blocks	1
	number of heads	2
Temporal Attention	hidden nodes	64
	blocks	1
	number of heads	4
BridgeTrans	hidden nodes	64
	blocks	1
	number of heads	4
Holistic Attention	hidden nodes	64
	blocks	1
	number of heads	2
Semantic Transformer	is padding	True
	stride	1
	filter size	3
Multi-task Layer	hidden nodes	[128, 64, 1]
		[128, 64, 1]
		[128, 64, 1]
-	Batch size	64
-	Dropout	0.2
-	Decay rate	0.99
-	Learning rate	0.0005
-	Epoch	50
-	Training method	Adam optimizer
-	λ	0.3
-	η	0.4

The final model framework parameters are established after many training stages. The MT-STAN model's layer count, node count, and other relevant hyperparameters are listed in Table I. The MT-STAN and baselines are implemented in TensorFlow and PyTorch. The server's 4 NVIDIA Tesla V100S-PCIE-32GB GPUs and 24 CPU cores are used for model training and testing. It is worth noting that both the suggested MT-STAN model and the baseline models' **implementation codes** are freely accessible on the author's **GitHub**.[2]

D. Experimental Results

1) The Convergence of MT-STAN: To demonstrate the stability of our proposed method, we have plotted the convergence of MT-STAN on a real-world dataset, specifically the G70 dataset. As depicted in Fig. 6 (left), MT-STAN converged quickly during the training phase, with a total loss of 2.318 after only 100 iterations. In Fig. 6 (right), we can observe that the training loss only experiences slight fluctuations when the number of training rounds reaches 200, and each subtask's validation loss is close to optimal and exhibits smooth changes. Based on these results, we can draw the following conclusions: (1) our proposed model can converge rapidly in less than 100 rounds; and (2) MT-STAN exhibits highly stable performance for each subtask, ensuring that all tasks converge simultaneously with low error rates and do not fluctuate on the validation dataset.

[2]https://github.com/zouguojian/Travel-time-prediction

TABLE II
PERFORMANCE COMPARISON FOR INDIVIDUAL TRAVEL TIME ESTIMATION ON G70, G6, G85, AND G20

Model	Fuyin Expressway (G70)			Jingzang Expressway (G6)			Yinkun Expressway (G85)			Qingyin Expressway (G20)		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
DNN [31]	5.187	28.663	9.376%	12.089	53.375	14.569%	1.256	1.679	5.434%	4.205	28.636	9.158%
CoDriver ETA [16]	5.200	28.630	9.414%	12.104	53.209	15.145%	1.380	1.817	5.980%	4.236	28.724	9.077%
DeepTTE [18]	6.108	29.718	12.104%	13.106	53.394	17.635%	3.716	5.944	13.182%	5.524	**26.512**	18.333%
WDR [3]	5.172	28.666	**9.250%**	12.059	53.126	14.720%	1.406	1.827	6.171%	4.212	28.687	8.989%
CompactETA [10]	**5.154**	28.641	9.285%	12.075	**53.082**	15.010%	**1.236**	**1.620**	5.341%	**4.193**	28.714	**8.833%**
CTTE [17]	5.191	**28.617**	9.384%	**11.963**	53.182	**14.335%**	1.271	1.960	**5.195%**	4.197	28.486	8.942%
MT-STAN-1	5.181	28.649	9.341%	12.060	**53.237**	14.725%	1.306	1.698	5.730%	4.222	28.702	9.065%
MT-STAN-2	**5.146**	**28.618**	**9.239%**	12.032	53.301	14.425%	1.250	1.683	5.484%	**4.179**	**28.685**	8.842%
MT-STAN-3	5.159	28.625	9.325%	**12.008**	53.388	**14.285%**	**1.156**	**1.538**	**4.995%**	4.181	28.705	**8.765%**
MT-STAN	**5.130**	**28.615**	**9.190%**	**11.978**	**53.087**	**14.592%**	**1.229**	**1.656**	**5.386%**	**4.167**	**28.672**	**8.774%**
MT-STAN (no-type)	5.884	29.060	11.583%	12.918	53.808	17.461%	1.484	2.035	6.329%	4.942	28.971	12.357%

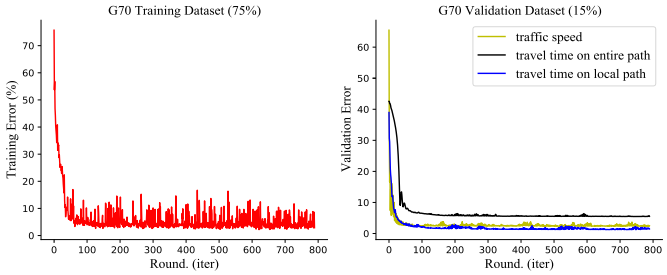


Fig. 6. Training error (**left**) and validation error (**right**) in the G70. The error measure function is the MAE.

2) Performance Comparison for Individual Travel Time Estimation on Entire Path: Table II shows the experimental outcomes of the proposed MT-STAN model and baselines, respectively. There are four different freeways where the models have been put into action. Some of the most intriguing conclusions are highlighted in black boldface below, which we derived from the data in the table.

DeepTTE was initially suggested in the baselines for individual travel time estimation, and its superior performance compared to conventional approaches has been verified by extensive experimental data [18]. Recently, WDR based on FM, Deep neural network, and LSTM is used to extract the individual travel preference, and the performance on the three criteria is superior to DeepTTE's. On the G70, G6, G85, and G20 expressways, for instance, WDR lowered the MAE by 15.324%, 8.044%, 62.164%, and 23.751%; on the same roads, WDR reduced the MAPE by 23.579%, 16.530%, 53.186%, and 50.968% compared with DeepTTE. What's more, DNN's performance in this area is far higher than that of DeepTTE. Compared to WDR and DNN, DeepTTE's performance is subpar because of two main factors: first, DeepTTE doesn't take into account interactions between features; in contrast, WDR uses FM to model this property generally; and second, traffic speed is a directed factor that affects individual travel time; however, DeepTTE doesn't take this into account.

According to Table II, many novel techniques, such as CoDriver ETA, CompactETA, and CTTE, have been presented that are based on WDR. Since CoDriver ETA employs an auxiliary triplet loss function, which may restrict the vehicle ID embedding learning under the driver speed similarity, its performance is inferior than that of WDR. CompactETA's usage of the GAT mechanism rather than the FM-, deep-, and recurrent- modules to model the spatial and temporal dependence for the road segments results in a 0.348%, 12.091%, and 0.451% reduction in MAE compared to WDR on the G70, G85, and G20 expressways, respectively. Also, under multi-task learning, the traffic speed prediction model supplies the spatio-temporal correlations on the road segments for individual trip time estimate, resulting in improved performance of CTTE over WDR on the G6, G85, and G20 expressways, respectively (0.796%, 9.602%, and 0.356% in terms of MAE; 2.615%, 15.816%, and 0.523% in terms of MAPE). Insightful new directions for this paper's investigation have emerged from the aforementioned comparative findings, which show how crucial multi-task learning and spatio-temporal correlations on road segments are for accurate individual journey time estimates.

TSP model, ITTE model, and multi-task learning are the three parts that make up the suggested model. Although ST-Block represents the dynamic spatio-temporal correlations of the past road network for the TSP model, BridgeTrans produces unknown future road network states. The ITTE model considers the interaction between individual travel characteristics and the effect of past and future traffic conditions on individual travel times. When learning across several tasks, it's helpful to have the ability to share the parameters of the underlying network and to collaborate on individual subtasks while assigning distinct loss values to each. As can be seen from the findings, our MT-STAN performs better than all baselines on all expressways across all three criteria. In the case of the G70, for instance, MT-STAN improved MAE by 1.346% compared to CoDriverETA, 0.812% compared to WDR, 0.466% compared to CompactETA, and 1.175% compared to CTTE; reduced MAPE by 2.379% compared to CoDriverETA, 0.649% compared to WDR, 1.023% compared to CompactETA, 2.067% compared to CTTE. The benefits of MT-STAN's prediction performance are most seen when used to estimate journey times for an individual.

3) The Importance of Vehicle Type in the Individual Travel Time Prediction: As shown in Table II, no-type indicates that the prediction model does not take the vehicle type into account during prediction processing. Compared to the condition that incorporated the vehicle type into prediction, the MAE, RMSE, and MAPE of MT-STAN (no-type) increased by 12. 812%, 1.531%, and 20.660%, respectively, in the G70;

TABLE III
LOCAL PATH TRAVEL TIME ON FUYIN (G70)-, JIANGZANG (G6)-, YINKUN (G85)-, AND QINGYIN (G20)- EXPRESSWAYS

Road Segment Index	MAE	RMSE	MAPE	Length (km)	Data Size (k)	MAE	RMSE	MAPE	Length (km)	Data Size (k)
	Fuyin Expressway (G70)					Jingzang Expressway (G6)				
Road Segment 1	0.192	0.253	9.974%	3.210		0.168	0.221	10.360%	2.984	
Road Segment 2	0.505	0.658	7.451%	8.856		0.221	0.313	7.668%	5.247	
Road Segment 3	0.968	8.449	13.624%	3.318	45.986	0.367	0.527	7.984%	7.631	43.525
Road Segment 4	0.503	0.709	8.037%	10.381		0.670	0.921	8.084%	12.827	
Road Segment 5	3.789	27.042	11.233%	22.194		11.431	53.054	22.000%	22.194	
Road Segment Index	Yinkun Expressway (G85)					Qingyin Expressway (G20)				
Road Segment 1	1.081	1.495	6.083%	21.221		0.158	0.221	7.771%	3.210	
Road Segment 2	0.102	0.151	5.079%	4.363		0.445	0.631	6.381%	7.814	
Road Segment 3	0.260	0.333	7.895%	4.243	1.683	3.296	28.537	11.506%	8.100	19.233
Road Segment 4	-	-	-	-		0.418	1.034	6.978%	7.900	
Road Segment 5	-	-	-	-		0.126	0.357	7.860%	2.612	

TABLE IV
PERFORMANCE COMPARISON FOR TRAFFIC SPEED PREDICTION ON G70, G6, G85, AND G20

Model	Fuyin Expressway (G70)			Jingzang Expressway (G6)			Yinkun Expressway (G85)			Qingyin Expressway (G20)		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
LSTM [42]	3.965	5.887	4.312%	4.357	6.659	4.263%	5.641	9.210	5.288%	4.003	6.084	4.658%
T-GCN [43]	3.606	5.531	4.000%	4.116	6.189	4.077%	4.993	8.502	4.778%	3.860	5.876	4.588%
DCRNN [44]	3.937	6.001	4.379%	4.142	6.265	4.076%	5.029	8.774	4.924%	3.831	5.823	4.564%
ST-GRAT [14]	3.555	5.596	3.984%	3.894	6.073	3.889%	4.902	8.833	4.834%	3.501	5.566	4.225%
MT-STAN	**3.293**	**5.223**	**3.662%**	**3.532**	**5.708**	**3.485%**	**4.683**	**8.279**	**4.501%**	**3.325**	**5.186**	**3.923%**

by 7.277%, 1.340%, and 16.431%, in the G6; by 17.183%, 18.624%, and 14.900%, in the G85; by 15.500%, 1.032%, and 28.996%, in the G20. The comparison results demonstrate that vehicle type is an indispensable input variable for estimating travel time. In addition, the following analyses can be summed up: first, the vehicle type variable reflects the similarity in travel speed among vehicles of the same type, as well as the heterogeneity between types of vehicles, i.e., travel time is closely related to speed; second, if the vehicle ID does not appear in the training dataset, it isn't easy to obtain the travel preference in the test dataset without the vehicle type feature. Fortunately, this drawback in baselines has been addressed in this paper, as all baselines' inputs contain a variable of vehicle type.

4) Detailed the Individual Travel Time Estimation on Local Path: In this novel approach, MT-STAN, Tables III shows the results of estimating the individual trip time on each local path (or road segment) using the multi-task learning model. Except for a few of road segments, the prediction error can be kept under control when the length of the local route is less than around 22 kilometres. The MAE, RMSE, and MAPE on section four of road G70 at a length of 10.381 km are 0.503, 0.709, and 8.037%, respectively; on section four of road G6 at a length of 12.827 km, these values are 0.670, 0.921, and 8.084%, respectively. The following also provides some interpretable analysis and some other intriguing findings.

Table III shows that for the overlapping road segment 5 (as Fig. 5), the estimation performance of MT-STAN on G70 is better than on G6 by 66.853%, 49.029%, and 48.941% in terms of MAE, RMSE, and MAPE, and that the total vehicles of G6 are high than G70 when the individual travel time is greater than 20 minutes, as shown in Fig. 7. The MAE of road section five on road G6 is higher than G75 for one reason: there is a conjunction area near before and the flow from G75 to G6; these phenomena are also founded on road segment 3

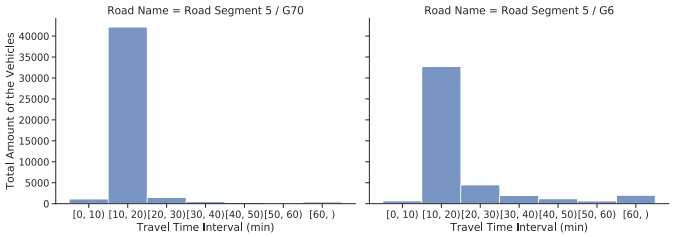


Fig. 7. The total amount of vehicles on the overlaping road segment for G70 and G6 expressways.

(as Fig. 5) of road G20. The foregoing shows that vehicles entering from a ramp or another road always have a negative effect on the followed vehicles from an main road, and that this effect is reflected in the travel times of the latter; the proposed MT-STAN can accurately estimate individual travel times on the local path without a conjunction area, even if the local path is too long; for instance, in the road G85, MT-STAN achieves high performance on road segment one despite the road segment's 21.221 km.

5) Performance Comparison for Traffic Speed Prediction: Table IV shows that when compared to all available baselines, the suggested technique, MT-STAN, provides superior performance across all measures on all four expressways. Due to the importance of spatial correlation in traffic prediction, the MAE was lowered by 16.948%, 18.935%, 16.983%, and 16.937% using the spatio-temporal model T-GCN compared to the temporal model LSTM across all expressways. MT-STAN outperforms other spatio-temporal models such as T-GCN, DCRNN, and ST-GRAT on all expressways with respect to MAE, respectively by 7.370%, 9.296%, 4.468%, and 5.027%. The findings show that ST-Block takes into account dynamic spatio-temporal correlations, whereas BridgeTrans prevents the transmission of prediction error in both space and time, leading to more accurate predictions of traffic speeds. The

accuracy of traffic speed forecasts is critical for accurately depicting conditions on the roads. Therefore, one of the most important parts of the MT-STAN is the novel TSP model based on ST-Block and BridgeTrans, which feeds the ITTE model the dynamic spatio-temporal correlation information of the road network in the past and the future. This helps to prevent the evaluation shift of individual travel times.

6) Influence of Each Component: MT-STAN-1: Tables II shows that MT-STAN-1 surpasses partial baselines on all expressways without taking traffic speed into account, and its performance is comparable to that of WDR. For instance, on the road G70, MT-STAN-1 achieved an MAE reduction of 0.116% compared to DNN, 0.366% compared to Codriver ETA, and 17.872% compared to DeepTTE, and on the road G6, an MAE reduction of 0.240% compared to DNN, 0.364% compared to DeepTTE, and 7.981% compared to DeepTTE. The findings show that the MT-STAN-1's based on cross-network and semantic transformer guidance leads to a good performance and positively impacts individual journey time prediction, even when road traffic speeds are ignored. In addition, the performance of MT-STAN-1 is low than MT-STAN-2 regarding the three metrics. For example, compared with MT-STAN-2, MT-STAN-1 increased MAE by 0.676% on the road G70; 4.288% on the road G85. These results also demonstrate that the traffic condition information of the highway network and multi-task learning are critical for ITTE.

MT-STAN-2: According above comparisons, the performance of MT-STAN-2 gets significantly improved compared with MT-STAN-1. However, compared with MT-STAN, MT-STAN-2 does not fare as well, as shown in Tables II. For instance, on the road G20, MT-STAN-2 increases upon MT-STAN by 0.287%, 0.045%, and 0.769% in terms of MAE, RMSE, and MAPE, respectively. The root reason is a neglect of the dissimilar impact of past and future spatio-temporal correlations of the road network on individual travel time estimates. In addition, the CTTE has better performance than MT-STAN-2 on the road G6, while the CompactETA has better performance than MT-STAN-2 on the road G85; maybe MT-STAN-2 is impacted by the addition operation. The results justify the decision to adopt a holistic focus in MT-STAN instead of a more conventional approach, like adding operations in MT-STAN-2.

MT-STAN-3: It is a variant of MT-STAN that uses the same TSP model and ITTE model, but only cares about estimating trip times throughout the whole route for individuals, rather than focusing on predicting traffic speeds or estimating times at local path. As shown in Tables II, when comparing MAE on the road for both G85 and G20, the MT-STAN-3 is superior to all baselines. Nonetheless, with the exception of G85, MT-STAN-3 performs worse than the planned MT-STAN on all expressways for all three criteria. The MAE on the road G70 was 0.562% higher with MT-STAN-3 compared to MT-STAN, while the MAE on the road G20 was 0.335% higher. The results show that both the traffic speed prediction and local individual travel time estimation tasks improve the accuracy of whole-path individual travel time estimates. In light of this, this work treats multi-task learning as an essential part of MT-STAN.

TABLE V
COMPUTATION TIME DURING THE TRAINING AND INFERENCE PHASES ('Y' MEANS CONTAIN RNNS ARCHITECTURE, 'N' MEANS NOT CONTAIN RNNS)

Model	Training (seconds)/(100 rounds)	Inference (seconds)
DNN (N)	32.013	0.008
CoDriver ETA (Y)	31.187	0.012
DeepTTE (Y)	29.172	0.012
WDR (Y)	66.403	0.011
CompactETA (N)	33.139	0.010
CTTE (Y)	29.595	0.014
LSTM (Y)	15.866	0.024
T-GCN (Y)	12.963	0.026
DCRNN (Y)	20.115	0.152
ST-GRAT (N)	7.493	0.077
MT-STAN (N)	31.374	0.029

7) Computation Time: Our proposed model is a multi-task learning architecture that ensures high performance across three tasks while keeping computation costs to a minimum. Both the proposed method and baselines were trained and evaluated using the same GPUs and CPU. Table V presents several interesting findings: first, using recurrent neural networks (RNNs) to model temporal dependency requires a more computational cost to train the prediction model. For instance, WDR requires 66.403s every 100 rounds because RNNs cannot achieve parallel computing on GPU in the temporal dimension. Second, MT-STAN abandons RNNs and multilayer feed-forward networks, instead adopting various multi-head attention mechanisms to extract spatial, temporal, and holistic correlations. The training cost of MT-STAN is equal to or even lower than baselines on the GPU. Third, the inference cost of MT-STAN is greater than that of some baselines due to its complexity being greater than theirs and the inferior performance of parallel computing on the CPU compared to the GPU. These results demonstrate that the proposed method has a significantly more effective training procedure than those based on RNNs or multilayer feed-forward networks. Moreover, MT-STAN is a multi-task learning model that can complete all three tasks with a single trained model. Therefore, we can tolerate these minute differences between the proposed method and baselines during the inference phase.

Furthermore, our proposed MT-STAN method has a significant potential advantage in that we can transfer the pre-trained traffic speed prediction (TSP) model to the proposed individual travel time estimating (ITTE) model without having to go through the training process again. In other words, we only need to provide the traffic condition information to the ITTE model. Therefore, our proposed method can not only achieve better performance than the baselines but also has the potential for more efficient and scalable applications in real-world scenarios

8) Case Study: Using the expressway as an example, we choose two ideal baselines (CompactETA and CTTE) and propose MT-STAN, and we display the fitting performance

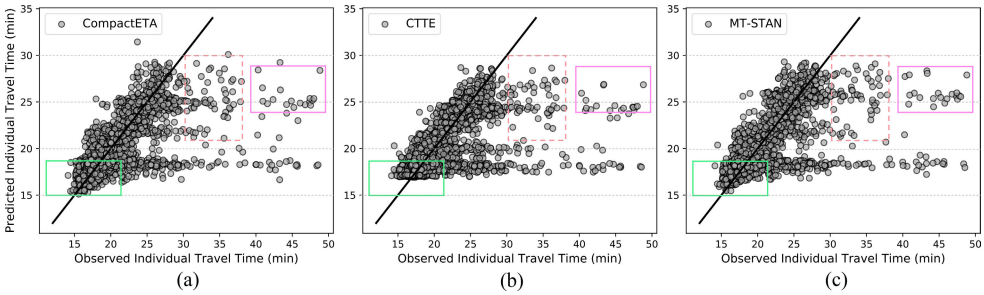


Fig. 8. Fitting degree between the observed and predicted ITT on entire path. The **black line** indicates $y = x$, and the black dots indicate the degree of deviation between the observed and predicted values.

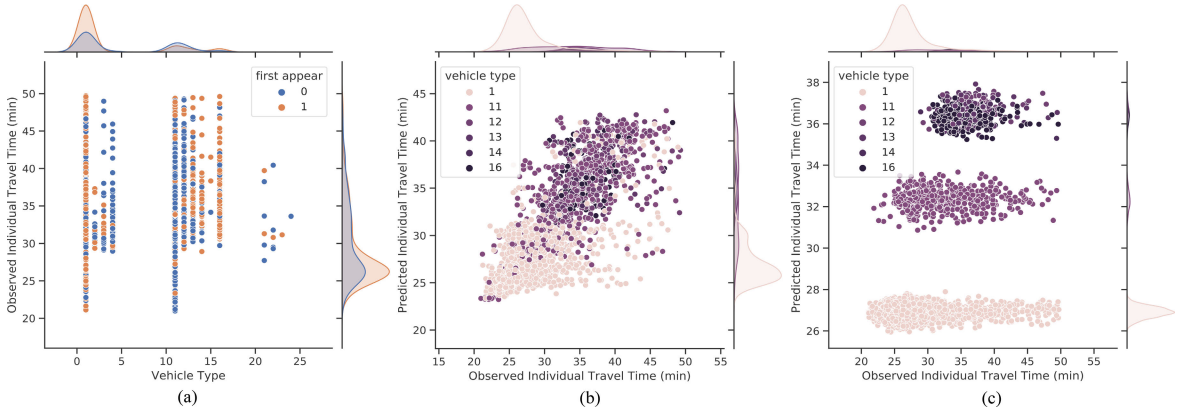


Fig. 9. The performance of MT-STAN on the entire path. (a) denotes the distribution of different types of vehicles, '1' indicates first appear, but '0' is not; (b) denotes the fitting performance under different types of vehicles; and (c) denotes the fitting performance under new vehicles.

of each model, for example, on the G20. As can be shown in Fig. 8, MT-STAN outperforms the two state-of-the-art baselines, CompactETA and CTTE, in terms of its capacity to fit data. It's interesting to note that when the individual trip time is between 0 and 20 minutes, the black dots included in MT-STAN are closer to the black line than CompactETA and CTTE, particularly beneath the line in the green box. Figure 8 shows that when the individual journey duration is more than 30 minutes, the prediction performance is marginally inadequate for all techniques, but MT-STAN has a lower discrete degree of black dots than the other two baselines. More so in the pink and purple boxes, MT-STAN's black dots are clustered nearer to the black line than either of the other baselines. These comparative findings show that MT-STAN is more suited for estimating travel times for individuals, and that even individual drivers may have extended journey times due to unique characteristics such as junction areas.

To further illustrate MT-STAN's prediction performance, we plot the fitting performance based on vehicle type and whether or not the vehicle exists in the training set. An identifier for a vehicle is represented as a 1 if it occurs exclusively in the test set, and a 0 otherwise. Figure 9 (a) displays the distribution of the various vehicle types on the G70, and the top six vehicle types-1, 11, 12, 13, 14, and 16-were chosen for further research (the full vehicle types can be referenced[3]). The difference between the anticipated and actual travel times for individuals in the test set is shown in Fig. 9 (b), and note that these vehicles have appeared in the training set. Various vehicle types have different travel times on the same freeway, as seen in the image. In addition, MT-STAN can make accurate estimates of individual travel time across the board, and even if the individual journey time is more than 35 minutes, indicating that the MT-STAN's prediction function is unaffected by vehicle types. Figure 9 (c) displays the estimated performance of individuals that occur exclusively in the test set. According to the findings, the MT-STAN can handle the data sparsity difficulties when the vehicle ID does not exist in the prior period, allowing for reliable estimates of individual trip times throughout the full route for new vehicles.

Estimates of how long it will take a person to complete the full route are directly affected by how long it takes them to complete the local route. Accordingly, in Fig. 10, we provide a visual representation of the predicted outcomes of MT-STAN over the continued local paths on the G70. Table III shows how study of expressway travel times is greatly impacted by road design. When the individual journey time on local path is more than 5 minutes in Fig. 10 (c), for instance, the dots fall into a critical discrete distribution, and when the individual travel time on local path is greater than 20 minutes in Fig. 10 (e), the same thing happens. However, the suggested MT-STAN model still retains good performance, despite the fact that the individual journey time calculation of these local pathways becomes problematic due to the negative effect of conjunction area (as shown in the Fig. 5). As can be shown

[3]https://www.mot.gov.cn/zhengcejiedu/sfglcltxfcxfl/

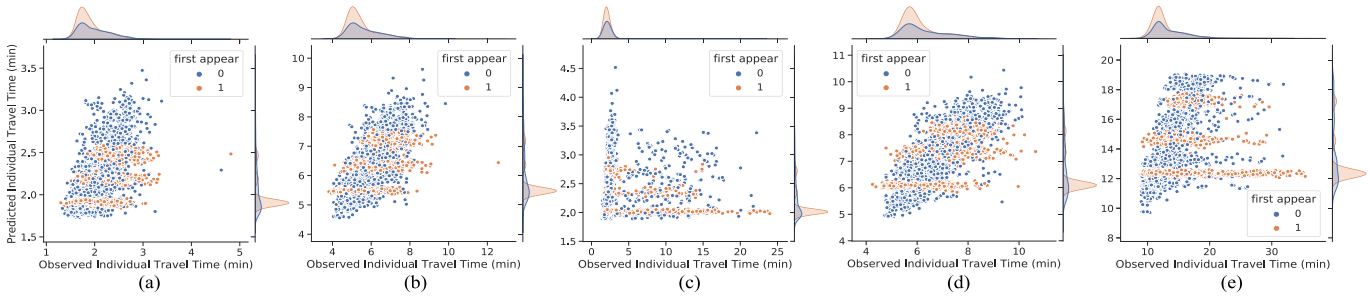


Fig. 10. The performance of MT-STAN on different local paths. (a)-(e) represent the road segments one to five.

in Fig. 10 (a), (b), and (d), MT-STAN is able to provide reliable estimates of travel times, even for conjunction areas and new vehicles, and this performance is independent of the road segments of the previous conjunction region. In addition, comparing local paths three and five with one, two, and four in Fig. 5 and Fig. 10, flow from the main road ramp has a substantial negative effect than the entrance ramp for the local path' individual trip time estimate. Therefore, the local route with conjunction area is one of the essential elements that need to be addressed for the individual trip time estimate in the highway systems, especially for some specific conjunction areas, and it provides us with a new avenue along which to continue advancing our study.

VI. Conclusion

This research proposes a unique multi-task spatio-temporal attention network (MT-STAN) for predicting journey time in the highway network. More specifically, we initially suggested a TSP model, using the ST-Block to extract historical spatio-temporal correlations and BridgeTrans to generate future road networks' hidden states. Following this, we developed an ITTE model, which takes into account the impact of traffic conditions on segment-specific trip times, and concentrates on individual preference by paying comprehensive attention to the interplay between travel features. Furthermore, we presented a multi-task learning, with the underlying networks' parameters shared and the subtasks' collaboration weighted differently for loss.

Experiments on real-world datasets reveal that MT-STAN obtains state-of-the-art performance compared with the baseline model for both traffic speed prediction and individual trip time estimates, with the benefit being more pronounced in multi-task learning. While the proposed MT-STAN model shows superiority in absolute prediction for both TSP and ITTE tasks, the latter is hindered by variables such as road construction and data sparsity, as shown by reflections of ITTE findings at the local route level. Furthermore, comparing the prediction performance of all variants with proposed MT-STAN and baselines, the contributions of each part of the proposed method are highlighted. To further improve ITTE's forecast ability, we anticipate expanding the suggested MT-STAN model with more data pertaining to the conjunction region, such as the history interaction traffic flow in the conjunction area. In future work, we plan to collect and integrate diverse external data, such as traffic accidents, and extend our model's capabilities to cover longer routes with sparse data. This will significantly enhance the model's adaptability to complex traffic environments.

Acknowledgment

The authors would like to express their gratitude to EditSprings (https://www.editsprings.com) for the expert linguistic services provided.

Guojian Zou, Ziliang Lai, Meiting Tu, and Ye Li are with the Key Laboratory of Road and Traffic Engineering, Ministry of Education, and the College of Transportation Engineering, Tongji University, Shanghai 201804, China (e-mail: 2010768@tongji.edu.cn; 2033402@tongji.edu.cn; meitingtu@tongji.edu.cn; JamesLI@tongji.edu.cn).

Changxi Ma is with the School of Traffic and Transportation, Lanzhou Jiaotong University, Lanzhou 730070, China (e-mail: machangxi@mail.lzjtu.cn).

Jing Fan is with China Railway First Survey and Design Institute Group Company Ltd., Xi'an 710043, China, and also with the Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai 201804, China (e-mail: jing.fan@tongji.edu.cn).

References

[1] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big data analytics in intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 383–398, Jan. 2019.

[2] L. Han, B. Du, J. Lin, L. Sun, X. Li, and Y. Peng, "Multi-semantic path representation learning for travel time estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 13108–13117, Aug. 2022.

[3] Z. Wang, K. Fu, and J. Ye, "Learning to estimate the travel time," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 858–866.

[4] Y. Sun et al., "Alleviating data sparsity problems in estimated time of arrival via auxiliary metric learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 23231–23243, Dec. 2022.

[5] 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.

[6] N. Chiabaut and R. Faitout, "Traffic congestion and travel time prediction based on historical congestion maps and identification of consensual days," *Transp. Res. C, Emerg. Technol.*, vol. 124, Mar. 2021, Art. no. 102920.

[7] T. Sun, K. Zhao, C. Zhang, M. Chen, and X. Yu, "PR-LTTE: Link travel time estimation based on path recovery from large-scale incomplete trip data," *Inf. Sci.*, vol. 589, pp. 34–45, Apr. 2022.

[8] E. Jenelius and H. N. Koutsopoulos, "Urban network travel time prediction based on a probabilistic principal component analysis model of probe data," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 2, pp. 436–445, Feb. 2018.

[9] A. Prokhorchuk, J. Dauwels, and P. Jaillet, "Estimating travel time distributions by Bayesian network inference," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 1867–1876, May 2020.

[10] K. Fu, F. Meng, J. Ye, and Z. Wang, "CompactETA: A fast inference system for travel time prediction," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 3337–3345.

[11] G. Jin, M. Wang, J. Zhang, H. Sha, and J. Huang, "STGNN-TTE: Travel time estimation via spatial–temporal graph neural network," *Future Gener. Comput. Syst.*, vol. 126, pp. 70–81, Jan. 2022.

[12] X. Zou, S. Zhang, C. Zhang, J. J. Q. Yu, and E. Chung, "Long-term origin-destination demand prediction with graph deep learning," *IEEE Trans. Big Data*, vol. 8, no. 6, pp. 1481–1495, Dec. 2022.

[13] C. Zheng, X. Fan, C. Wang, and J. Qi, "Gman: A graph multi-attention network for traffic prediction," in *Proc. AAAI*, vol. 34, no. 1, 2020, pp. 1234–1241.

[14] C. Park et al., "ST-GRAT: A novel spatio-temporal graph attention networks for accurately forecasting dynamically changing road speed," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 1215–1224.

[15] C. Ma, G. Dai, and J. Zhou, "Short-term traffic flow prediction for urban road sections based on time series analysis and LSTM_BILSTM method," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5615–5624, Feb. 2021.

[16] Y. Sun et al., "CoDriver ETA: Combine driver information in estimated time of arrival by driving style learning auxiliary task," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4037–4048, May 2022.

[17] R. Gao, F. Sun, W. Xing, D. Tao, J. Fang, and H. Chai, "CTTE: Customized travel time estimation via mobile crowdsensing," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19335–19347, Oct. 2022.

[18] D. Wang, J. Zhang, W. Cao, J. Li, and Y. Zheng, "When will you arrive? Estimating travel time based on deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–8.

[19] J. Ma, J. Chan, S. Rajasegarar, and C. Leckie, "Multi-attention graph neural networks for city-wide bus travel time estimation using limited data," *Exp. Syst. Appl.*, vol. 202, Sep. 2022, Art. no. 117057.

[20] J. Guo, W. Wang, Y. Tang, Y. Zhang, and H. Zhuge, "A CNN-Bi_LSTM parallel network approach for train travel time prediction," *Knowl.-Based Syst.*, vol. 256, Nov. 2022, Art. no. 109796.

[21] J. Pang, J. Huang, Y. Du, H. Yu, Q. Huang, and B. Yin, "Learning to predict bus arrival time from heterogeneous measurements via recurrent neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 9, pp. 3283–3293, Sep. 2019.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[24] L. Li, X. Chen, Z. Li, and L. Zhang, "Freeway travel-time estimation based on temporal–spatial queueing model," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1536–1541, Sep. 2013.

[25] A. Simroth and H. Zähle, "Travel time prediction using floating car data applied to logistics planning," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 243–253, Mar. 2011.

[26] S. Kwak and N. Geroliminis, "Travel time prediction for congested freeways with a dynamic linear model," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 12, pp. 7667–7677, Dec. 2021.

[27] H. Wang, X. Tang, Y.-H. Kuo, D. Kifer, and Z. Li, "A simple baseline for travel time estimation using large-scale trip data," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–22, Mar. 2019.

[28] Z. Wang, M. Liang, and D. Delahaye, "A hybrid machine learning model for short-term estimated time of arrival prediction in terminal manoeuvring area," *Transp. Res. C, Emerg. Technol.*, vol. 95, pp. 280–294, Oct. 2018.

[29] N. C. Petersen, F. Rodrigues, and F. C. Pereira, "Multi-output bus travel time prediction with convolutional LSTM neural network," *Exp. Syst. Appl.*, vol. 120, pp. 426–435, Apr. 2019.

[30] J. J. Q. Yu, "Citywide estimation of travel time distributions with Bayesian deep graph learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 2366–2378, Mar. 2023.

[31] M. Abdollahi, T. Khaleghi, and K. Yang, "An integrated feature learning approach using deep learning for travel time prediction," *Exp. Syst. Appl.*, vol. 139, Jan. 2020, Art. no. 112864.

[32] R. Dai, S. Xu, Q. Gu, C. Ji, and K. Liu, "Hybrid spatio-temporal graph convolutional network: Improving traffic prediction with navigation data," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 3074–3082.

[33] P. Ting et al., "Freeway travel time prediction using deep hybrid model—Taking sun Yat–Sen freeway as an example," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8257–8266, Aug. 2020.

[34] P. He, G. Jiang, S.-K. Lam, and Y. Sun, "Learning heterogeneous traffic patterns for travel time prediction of bus journeys," *Inf. Sci.*, vol. 512, pp. 1394–1406, 2020.

[35] J. Xu, S. Xu, R. Zhou, C. Liu, A. Liu, and L. Zhao, "TAML: A traffic-aware multi-task learning model for estimating travel time," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 6, pp. 1–14, Dec. 2021.

[36] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[37] K. Chen, G. Chu, X. Yang, Y. Shi, K. Lei, and M. Deng, "HSETA: A heterogeneous and sparse data learning hybrid framework for estimating time of arrival," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21873–21884, Nov. 2022.

[38] S. Rendle, "Factorization machines," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 995–1000.

[39] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[40] H. Zhou et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 12, 2021, pp. 11106–11115.

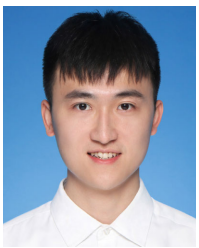
[41] S. Kiranyaz, T. Ince, O. Abdeljaber, O. Avci, and M. Gabbouj, "1-D convolutional neural networks for signal processing applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8360–8364.

[42] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transp. Res. C, Emerg. Technol.*, vol. 54, pp. 187–197, May 2015.

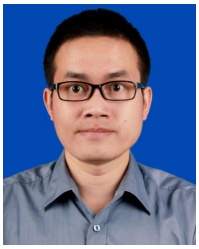
[43] L. Zhao et al., "T-GCN: A temporal graph convolutional network for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3848–3858, Sep. 2020.

[44] 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. 1–11.

Guojian Zou received the M.S. degree from the College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China, in 2020. He is currently pursuing the Ph.D. degree with the College of Transportation Engineering, Tongji University, China. He is the author of more than 15 articles. His research interests include deep learning, natural language processing, computer vision, urban computing, and intelligent transportation systems.



Ziliang Lai received the B.S. degree in transportation engineering from Beijing Jiaotong University, Beijing, China, in 2020. He is currently pursuing the M.S. degree with the College of Transportation Engineering, Tongji University, China. His research interests include autonomous vehicle ridesharing, vehicle path optimization, and deep learning.



Changxi Ma received the B.S. degree in traffic engineering from the Huazhong University of Science and Technology in 2002 and the Ph.D. degree in transportation planning and management from Lanzhou Jiaotong University in 2013. He is currently a Professor with Lanzhou Jiaotong University. He is the author of three books and more than 100 articles. His research interests include ITS, traffic safety, and hazardous materials transportation.

Meiting Tu received the joint Ph.D. degree from Paris Saclay University and Tongji University, Shanghai, in 2022. She is currently a Distinguished Researcher and a Doctoral Supervisor with the College of Transportation Engineering, Tongji University. She was selected by the Shanghai Overseas Leading Talents Program in 2022. Her research interests include shared mobility management, low-carbon traffic planning, spatio-temporal big data mining, and artificial intelligence algorithms.

Jing Fan received the Ph.D. degree in transportation engineering from Tongji University in 2023. As a Joint Ph.D. Student, she also studied at The University of Tokyo from 2019 to 2020. Currently, she is working as a Post-Doctoral Fellow at China Railway First Survey and Design Institute Group Company Ltd., and Tongji University. Her research interests include spatiotemporal data mining, transport planning, transport policy, transportation economic and geography.

Ye Li received the B.S., M.S., and Ph.D. degrees in transportation engineering from Tongji University, Shanghai, China, in 1995, 2000, and 2003, respectively. He has been a Professor with Shanghai Normal University and the Transportation Engineering Department, Tongji University. In 2011, he was a Visiting Scholar with the University of California at Berkeley, Berkeley, USA. He is the author of one book and more than 60 articles. His research interests include public transportation planning, low-carbon transportation system planning, and transportation service pattern innovation with big data. His awards and honors include the State Science and Technology Prize, the China Navigation Technology Award, and the New Century Talents Award.