Towards integrated and fine-grained traffic forecasting: A spatial-temporal heterogeneous graph transformer approach

Guangyue Li[a], Zilong Zhao[a,1], Xiaogang Guo[a], Luliang Tang[a,1], Huazu Zhang[a], Jinghan Wang[b]

[a] *State Key Laboratory for Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China*

[b] *School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China*

Abstract

Integrated and fine-grained traffic forecasting plays a crucial role in advancing intelligent transportation systems (ITS). Current traffic forecasting studies primarily focus on predicting the average traffic states of bidirectional road segments. While a few studies have refined the forecasting scale to the turn-level at intersections, they overlook the dependencies between road segments and intersection turns. In real-world traffic networks, roads and turns demonstrate heterogeneous spatial structures and traffic states, yet they are interconnected due to spatial proximity. To achieve integrated traffic forecasting, we propose a novel **S**patial-**T**emporal **H**eterogeneous **G**raph Transformer (STHGFormer) jointly considering roads and turns. For road network representation, we innovatively define a Heterogeneous Road network Graph (HRG), which provides a comprehensive depiction of the complete traffic network and emphasizes its inherent heterogeneity. Spatially, we develop a Heterogeneous Spatial Embedding (HSE) module to encode spatial heterogeneity, including attributes, significance of different nodes, and their relevancy. Based on the spatial information encoded by HSE, our unified SpaFormer simultaneously captures dependencies between roads and turns on an entire traffic network. Temporally, we incorporate an Adaptive Soft Threshold (AST) module with the TempFormer to handle highly fluctuating traffic states. Furthermore, to account for the temporal heterogeneity exhibited by roads and turns, we employ separate TempFormers to learn their characteristics. Our experiments conducted on a real-world dataset from Wuhan, China, demonstrate that STHGFormer outperforms state-of-the-art methods, achieving an 8.1% improvement in road forecasting and a 10.9% improvement in turn forecasting.

Keywords: Integrated and fine-grained traffic forecasting, Heterogeneous road network graph, Spatial-temporal transformer, Heterogeneous spatial embedding

1. Introduction

As the cornerstone of Intelligent Transportation Systems (ITS) (Chauhan et al., 2021; Li et al., 2021), traffic forecasting is critical for route planning (Liebig et al., 2017), travel time estimation (Zhang et al., 2022), and traffic signal control (Wei et al., 2019). Accurate predictions provide essential information for traffic participants to make reasonable decisions, which enhances the safety, security, and efficiency of daily transportation (Yin et al., 2021).

Crowd-sourced data, including trajectory data, has significantly contributed to the refinement and timeliness of traffic perception and forecasting. Currently, the forecasting scale has refined from coarse grids (Zhang et al., 2017) to specific components of the road network, such as road segments and intersection turns. For instance, the studies conducted by (Guo et al., 2020; Zhao et al., 2019a) achieved precise traffic forecasting for bi-directional road segments, enabling accurate short-term citywide traffic prediction. Moreover, as a remarkable advancement, (Fang et al., 2021) refined the forecasting scale to the turn-level, realizing detailed prediction for different turns within intersections.

While previous studies have accomplished fine-grained traffic forecasting to some extent, their depiction of the traffic network remains incomplete. In general, as shown in Fig. 1, the road network can be divided into two main components: unconstrained road segments and intersections controlled by traffic signals. Intersections play a vital role in the traffic network as they connect road segments and facilitate turning movements. The tight physical connections between roads

[1] Corresponding author.

E-mail address: guangyueli@whu.edu.cn (G. Li), zilzhao@whu.edu.cn (Z. Zhao), tll@whu.edu.cn (L. Tang).

and turns are shaped by numerous intersections within the road network. Moreover, the traffic states of both road segments and turns are closely related due to the frequent transmission and feedback of traffic flows between them. For instance, congestion at an intersection often propagates to the surrounding road segments, leading to decreased speeds. Therefore, it is crucial to integrate road segments and turns in order to gain a comprehensive understanding of the spatial dependencies within the traffic network, which ultimately enhances the accuracy and robustness of traffic forecasts.

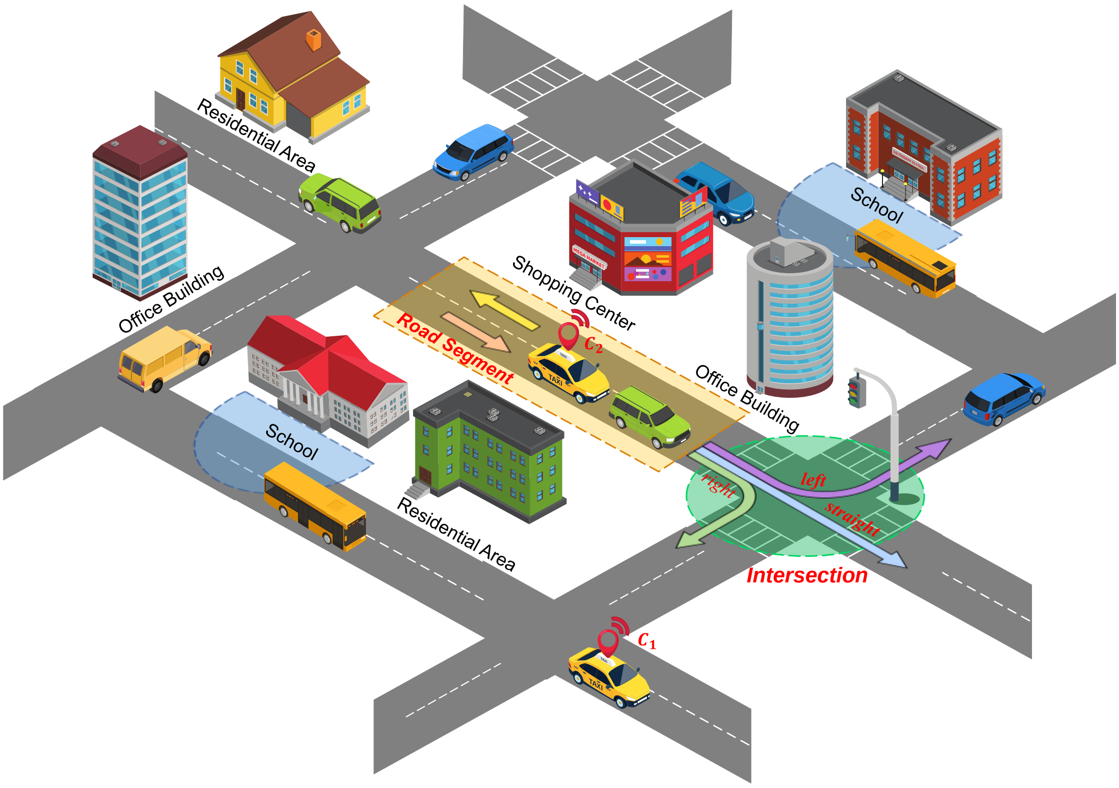
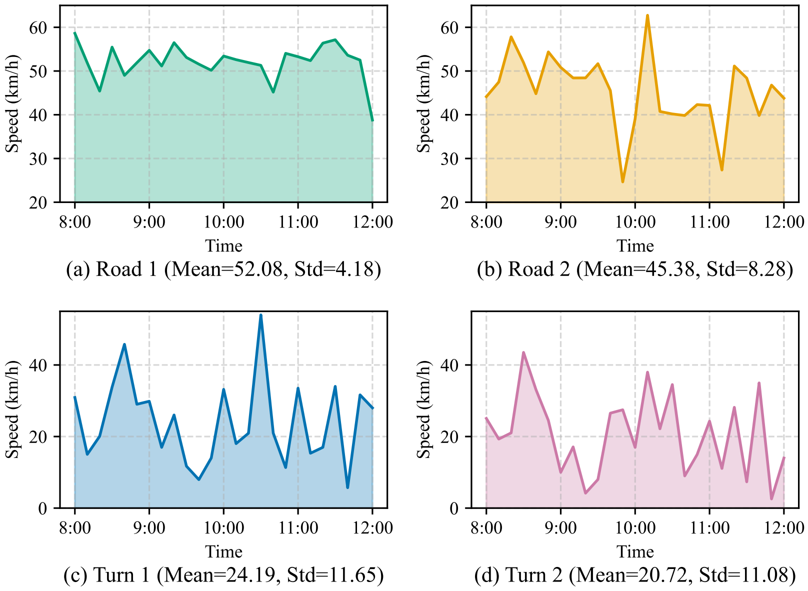


Fig. 1. A real-world traffic system. Floating cars equipped with position sensors can be considered as mobile traffic detectors, as shown in C_1 and C_2. The traffic network consists of road segments (depicted in yellow) and intersections (depicted in green). At intersections, traffic flows exhibit complex turning patterns. The blue semicircles represent areas with similar traffic patterns.



(a) Road 1 (Mean=52.08, Std=4.18)

(b) Road 2 (Mean=45.38, Std=8.28)

(c) Turn 1 (Mean=24.19, Std=11.65)

(d) Turn 2 (Mean=20.72, Std=11.08)

Fig. 2. Traffic states of road segments and intersection turns at the same intersection in a real-world dataset. (a) and (b) depict roads with relatively steady traffic states, characterized by a high mean and low variance. Conversely, (c) and (d) illustrate turns with highly variable traffic conditions, exhibiting a lower mean and higher variance.

In addition, significant spatial-temporal heterogeneity exists between road segments and intersection turns. Regarding spatial structure, the characteristic of traffic flows tends to be stable on road segments, where vehicles can change lanes within safety limits. However, as vehicles enter the intersection region, they display complex turning patterns (Kan et al., 2019), including left-turn, right-turn, and straight-ahead, giving rise to diverse travel times. Regarding temporal state, as depicted in Fig. 2, turns typically exhibit slower speeds but experience more pronounced fluctuations, primarily due to traffic control measures at intersections. Therefore, how to effectively distinguish the spatial-temporal heterogeneity between road segments and turns while exploring their spatial dependencies remains a critical challenge for integrated traffic forecasting.

To overcome these challenges, we propose a Spatial-Temporal Heterogeneous Graph Transformer (STHGFormer)，forecasting road segments and intersection turns integrally. To the best of our knowledge, the proposed STHGFormer pioneers integrated traffic forecasting for different components of traffic networks. Spatially, SpaFormer (namely the spatial module of STHGFormer) simultaneously captures the dependencies between roads and turns within a complete road network. Temporally, due to the heterogeneity of different forecasting elements, we employ separate TempFormers to learn the temporal feature of roads and turns. The key contributions of this paper are summarized as follows.

1) We innovatively define a Heterogeneous Road network Graph (HRG) to comprehensively represent the complex topology in a transportation system. Considering the spatial-temporal heterogeneity, HRG incorporates different types of nodes and edges to depict the characteristics of road segments and intersection turns, as well as their intricate relationships.

2) Utilizing the road network information extracted by the innovative Heterogeneous Spatial Embedding (HSE) module, SpaFormer effectively explores the intricate interdependencies between road segments and turns. Moreover, leveraging its expansive receptive field, SpaFormer dynamically models the correlations present in the traffic patterns.

3) In the temporal dimension, TempFormer addresses the issue of highly fluctuating traffic conditions via an Adaptive Soft Threshold (AST) module. By dynamically adjusting the threshold, TempFormer enhances its ability to capture temporal correlations in the presence of noise, ensuring more accurate and robust forecasting results.

Experiments on a real-world traffic speed dataset in Wuchang, Wuhan, China, demonstrate that the proposed STHGFormer outperforms other baseline models with an improvement of 8.1% in road segment forecasting and an improvement of 10.9% in turn forecasting for a ten-minute traffic prediction.

The article is structured as follows. Section 2 outlines the relevant traffic forecasting studies. Section 3 provides an overview of preliminary definitions, and describes the construction method of the HRG and the proposed STHGFormer in detail. Section 4 presents experiments conducted to validate the proposed model. Finally, Section 5 provides the concluding remarks.

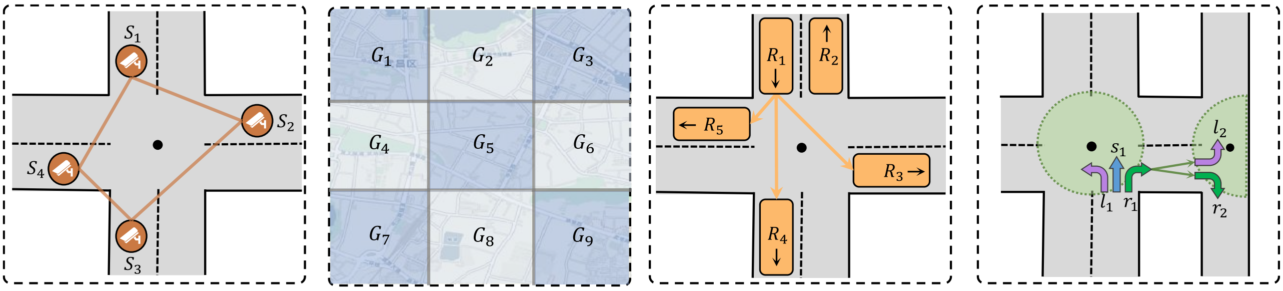
2. Literature Review

In this section, we review previous traffic forecasting studies, considering both forecasting targets and deep learning models they used.

2.1. Forecasting target

Most forecasting studies rely on data collected from fixed sensors, such as PEMS (Chen et al., 2001) and METR-LA (Li et al., 2017) datasets. In such studies, a sensor graph is usually used to model the relationship between adjacent sensors, as illustrated in Fig. 3 (a). Yet, the high costs of installing and maintaining traffic sensors preclude their widespread use in metropolitan transport systems (Fang et al., 2021). When there is enough trajectory data to cover most of the roads, researchers can leverage it to percept traffic states in a cost-effective, high-coverage, and wide-range way (Kan et al., 2019). For instance, some studies(Guo et al., 2021; Zhang et al., 2017) use regular grids to slice the road network and construct a grid graph as Fig. 3 (b). They regard average states of trajectory points within each grid as prediction targets, which provides greater coverage. However, the use of grids disrupts the natural structure of traffic networks (Ye et al., 2022).

Currently, the fine-grained traffic forecasting attracts more research interest. (Guo et al., 2020; Zhang et al., 2019) have successfully predicted traffic states on bidirectional road segments. Their road network graphs, shown in Fig. 3 (c), define nodes as unidirectional road segments and edges as the traffic flow movement between roads. In addition, (Fang et al., 2021) predicted traffic flow states at intersections, including left-turn, right-turn, and straight ahead, leading to a more detailed turn-level prediction. They designed a turn graph, which is a dual graph of the road graph, with turns as nodes and roads as edges, as depicted in Fig. 3 (d). Although these studies have refined the forecasting scale, they are limited to specific parts of the traffic network and do not have integrated forecasting capabilities.



(a) Sensor Graph (b) Grid Graph (c) Road Graph (d) Turn Graph

Fig. 3. Different kinds of traffic road network graph.

2.1. Forecasting model

With the rise of deep learning, forecasting models based on graph neural networks (GNNs) (Wu et al., 2020) are prevalent nowadays (Ye et al., 2022). GNN-based methods can fully leverage the spatial features present in traffic networks (Zhao et al., 2022). Due to this advantage, GNNs have been coupled with temporal neural networks, including recurrent neural networks (RNNs) (Yu et al., 2019), temporal convolutional networks (TCNs) (Wu et al., 2019), and others. For example, Wu et al. (Wu et al., 2019) introduced an adaptive adjacency matrix to reveal the dynamic associations within road networks and utilized GNN to capture the hidden spatial dependencies. In addition, Guo et al. (Guo et al., 2021) constructed traffic network graphs at both road and regional levels and employed a model consisting of GNN and TCN to capture multi-level spatial-temporal relationships.

However, as pointed out by (Kreuzer et al., 2021), GNNs suffer from several issues including over-smoothing and over-squashing. Moreover, using GNNs to mine spatial correlations heavily relies on topological connections of roads, making them unsuitable for detecting dynamic pattern correlations that are independent of road networks (Zhao et al., 2022).

In recent years, the attention mechanism (Vaswani et al., 2017) has emerged as a popular tool in deep learning researches. It enables models to dynamically analyze the correlation between the target sequences and adaptively focus on the desired content. As a result, researchers have been exploring the use of the Transformer and its variants for graph modeling. Graph, as a uniquely non-Euclidean data structure, cannot be processed directly by the attention mechanism, unlike other data structures such as images and texts (Ying et al., 2021). Therefore, many works have employed methods such as positional embedding and attention matrices improving (Min et al., 2022) to incorporate graph information into the vanilla transformer. These studies have demonstrated that successful graph transformers, such as Graphormer (Ying et al., 2021) and GraphiT (Mialon et al., 2021), can even outperform GNNs.

In traffic forecasting, the attention mechanism has been widely applied due to its versatility and advantages. For instance, Liao et al. (Liao et al., 2022) used the attention mechanism to handle long sequences and improve the accuracy of results for long-term forecasting tasks. Regarding potential spatial dependence, Zheng et al. (Zheng et al., 2020) proposed a multi-headed spatial attention mechanism that can adaptively capture multiple types of spatial correlations.

3. Method

3.1. Preliminaries

In this study, we define the traffic forecasting target as a complete road network, which consists of bidirectional road segments and intersection turns. We define $n_i^{rd}, i \in [1, N^{rd}]$ and $n_j^{tr},\ j \in [1, N^{tr}]$ as roads and turns respectively, where N^{rd} and N^{tr} are their total number. Then, we extract the traffic speeds on n_i^{rd} and n_j^{tr} from the trajectory data using a three-step method:

Firstly, we filter out trajectory points with severe deviation and match the remaining points with roads, following the method of Tang et al. (Tang et al., 2017). Then, referring to the method of Fang et al. (Fang et al., 2021), we divide roads and turns by dynamically estimating the queue starting point of each turn with trajectory data. Finally, we extract the speeds of roads by taking the average speed of all floating cars that travel on n_i^{rd} during time *t*, as shown in Eq. (1). Similarly, we collect the speeds of turns in the same way, and define $s_t^{n_j^{tr}}$ as the speed of n_j^{tr} during time *t*.

$$s_t^{n_i^{rd}} = \frac{\sum_{k=1}^{K} avg(v_t^{k,n_i^{rd}})}{K} \tag{1}$$

where $v_t^{k,n_i^{rd}}$ represents the speed of the k-th vehicle at the time t, K is the total number of vehicles passing, and $s_t^{n_i^{rd}}$ is the average speed of n_i^{rd}.

To better articulate traffic forecasting tasks, we define $x_t^{rd} = \left[s_t^{n_1^{rd}}, s_t^{n_2^{rd}}, \dots\right] \in R^{N^{rd}}$ and $x_t^{tr} = [s_t^{n_1^{tr}}, s_t^{n_2^{tr}}, \dots] \in R^{N^{tr}}$ as the traffic states of all road segments and intersection turns during time t, respectively. Further, we use $X_{t-T:t}^{rd} = [x_{t-T+1}^{rd},\ x_{t-T+2}^{rd}, \dots, x_t^{rd}]$ and $X_{t-T:t}^{tr} = [x_{t-T+1}^{tr},\ x_{t-T+2}^{tr}, \dots, x_t^{tr}]$ to represent traffic speeds of them during past T time slots before time t. In the end, the core task of traffic forecasting is to find an appropriate function F that relies on historical traffic speeds and the road network information to forecast future traffic states. The forecasting process can be expressed by Eq. (2).

$$X_{t:t+P}^{rd}; X_{t:t+P}^{tr} = F(X_{t-T:t}^{rd}; X_{t-T:t}^{tr}; G) \tag{2}$$

where P is the length of time slots to be predicted, $X_{t:t+P}^{rd} \in R^{N^{rd} \times P}, X_{t:t+P}^{tr} \in R^{N^{tr} \times P}$ are the traffic speed during this period, and G is the road network information.

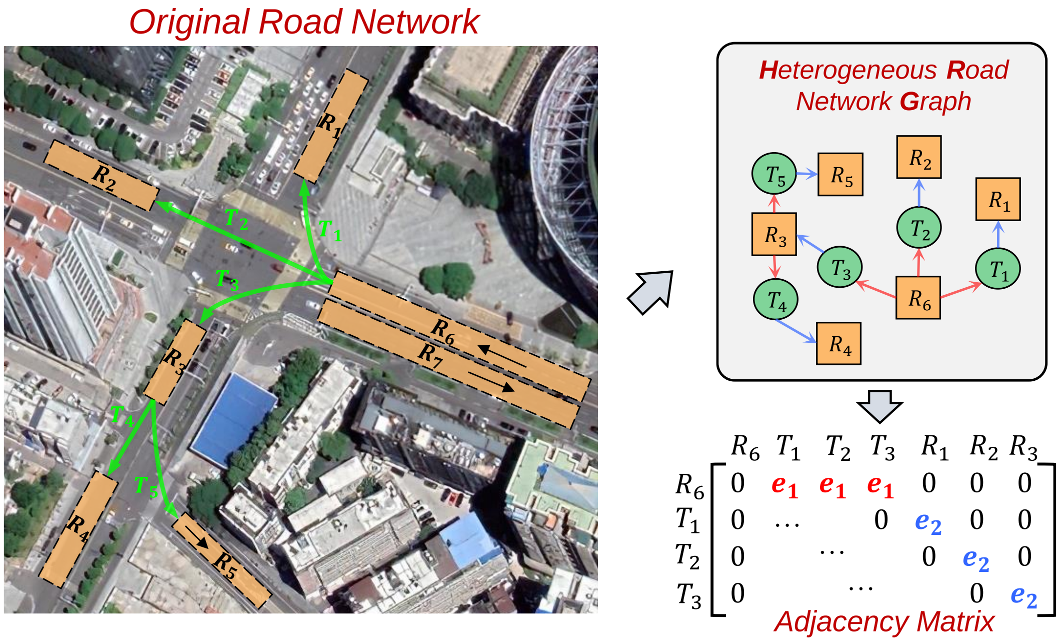
3.2. Heterogeneous Road Network Graph

To provide a comprehensive depiction of the complete traffic network, we present a definition of heterogeneous road network graph (HRG), which represent road segments, intersection turns and their spatial relationships integrally.

Definition 1: Heterogeneous Road network Graph (HRG):

Heterogeneous graphs are graphs with multiple types of nodes and edges (Wang et al., 2019), which can better describe the rich semantic and topological structure of complex road networks. The proposed HRG is defined as $HRG = (V,\ E)$, where nodes $V \in R^N$ represent different parts of the traffic network to be forecasted, e.g., roads and turns. Moreover, edges $E \in R^{N \times N}$ represent various relationships between them, namely different modes of traffic flows when entering and exiting intersections.

By constructing the HRG, we abstracted the topological structure of the heterogeneous road network where traffic flows transmit, providing spatial information for deep learning models to uncover the correlation between roads and turns. Fig. 4 illustrates the construction process of HRG.



$$\begin{array}{c|ccccccc} & R_6 & T_1 & T_2 & T_3 & R_1 & R_2 & R_3 \\ \hline R_6 & 0 & \boldsymbol{e_1} & \boldsymbol{e_1} & \boldsymbol{e_1} & 0 & 0 & 0 \\ T_1 & 0 & \cdots & & 0 & \boldsymbol{e_2} & 0 & 0 \\ T_2 & 0 & & \cdots & & 0 & \boldsymbol{e_2} & 0 \\ T_3 & 0 & & & \cdots & & 0 & \boldsymbol{e_2} \end{array}$$

Fig. 4. Heterogeneous road network graph. In the original road network, the yellow rectangles denote unidirectional roads, while the green arrows represent intersection turns. In the heterogeneous road network graph, yellow nodes represent roads, and green nodes represent turns. The different-colored arrows indicate the distinct relationships between roads and turns. The adjacency matrix includes e_1 and e_2, which refer to distinct edges of different types in the heterogeneous graph.

Definition 2: Heterogeneous nodes: We divide each bi-directional road into two unidirectional road segments with opposite directions and represent them as v_i^{rd}, v_j^{rd}, e.g., R_6 and R_7 in Fig. 4. For one-way road, e.g., R_5, we directly use a single road node v_k^{rd} to represent it. Then, we use turn nodes v^{tr} to describe different types of turns at intersections, including right-turn (T_1), straight-ahead (T_2), and left-turn (T_3).

Definition 3: Heterogeneous edges: Vehicles tend to slow down while they are approaching the intersection due to the traffic control measures, whereas they tend to accelerate when exiting. Thus, we utilize different types of edges to represent such relationships between roads and turns. Specifically, if a vehicle can reach road node v_j^{rd} from v_i^{rd} through a turn node v_p^{tr} at the intersection, then there exists a directed edge of type e_1 between v_i^{rd} and v_p^{tr}, and a directed edge of type e_2 between v_p^{tr} and v_j^{rd}. As shown in Fig. 4, vehicles on road segment R_6 can arrive at R_1 by turning right through T_1. Then, in HRG, there are different kinds of edges between R_6 and T_1, T_1 and R_1. Finally, we create the corresponding adjacency matrix $A = (a_{ij})_{(N \times N)}$ of HRG, which can be defined as follows:

$$a_{v_i v_j} = \begin{cases} e_1, if\ e_{v_i v_j}\ exists, and\ v_i \in v^{rd}, and\ v_j \in v^{tr} \\ e_2, if\ e_{v_i v_j}\ exists, and\ v_i \in v^{tr}, and\ v_j \in v^{rd} \\ \qquad\qquad 0,\ otherwise \end{cases} \tag{3}$$

where v_i, v_j are two different nodes in HRG, and $e_{v_i v_j}$ represent the connection between them.

3.3. Overall Architecture of STHGFormer

This section elaborates on the basic framework of the Spatial-Temporal Heterogeneous Graph Transformer (STHGFormer), as shown in Fig. 5. To integrally predict future traffic states of different parts in the traffic network, STHGFormer utilizes historical traffic speeds of road segments ($X_{t-T:t}^{rd}$) and intersection turns ($X_{t-T:t}^{tr}$) and the heterogeneous road network graph as inputs. At the initial input stage, STHGFormer applies convolution layers to augment the hidden dimension, as indicated in Eq. (4), and uses the resulting H_1^{rd} and H_1^{tr} as inputs for the first layer of the Spatial-Temporal-block (ST-block).

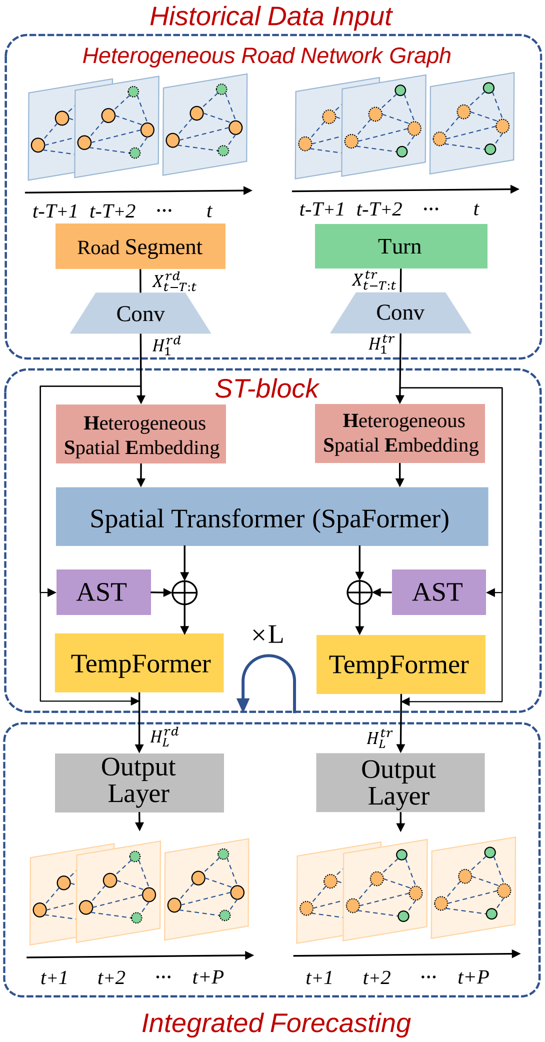


Fig. 5. Overview of STHGFormer. ST-block refers to Spatial-Temporal-block; AST refers to adaptive soft threshold; TempFormer refers to Temporal Transformer.

The ST-block comprises two main components: a Spatial Transformer (SpaFormer) for spatial dependencies, and a Temporal Transformer (TempFormer) for temporal dependencies. Within a complete road network, SpaFormer captures spatial dependencies between road segments and intersection turns simultaneously. Due to the heterogeneity of traffic states in different forecasting elements, two independent TempFormers are deployed to uncover the time series correlations.

STHGFormer stacks L layers of ST-blocks, featuring rich residual connections that allow for the exploration of deep spatial-temporal dependencies. Eventually, H_L^{rd} and H_L^{tr}, outputs of the final ST-block, are transferred through the output layer, thus generating the ultimate forecasting results, as exemplified in Eq. (5). The output layer of STHGFormer comprises a single convolutional layer and a fully connected layer. The purpose of the convolutional layer is to reduce dimensionality, while the fully connected layer is used to produce the final output.

$$H_1^{rd} = Conv(X_{t-T:t}^{rd}), H_1^{tr} = Conv(X_{t-T:t}^{tr}) \tag{4}$$

$$X_{t:t+P}^{rd} = FC\left(Conv(H_L^{rd})\right), X_{t:t+P}^{tr} = FC\left(Conv(H_L^{tr})\right) \quad (5)$$

where $H_1^{rd} \in R^{N^{rd} \times T \times d}$ and $H_1^{tr} \in R^{N^{tr} \times T \times d}$ are the inputs for the first ST-block. $FC()$ denotes the vanilla fully connected layer. $Conv(H_L^{rd}) \in R^{N^{rd} \times T}$ and $Conv(H_L^{tr}) \in R^{N^{tr} \times T}$ are temporary outputs of the convolution layer when generating the final results.

3.4. Heterogeneous Spatial Embedding

Existing graph transformers primarily focus on homogeneous graphs (Min et al., 2022), where node and edge types are treated as the same. To our knowledge, there is no clear method to extract information from heterogeneous graphs and incorporate it into transformers. To tackle this problem, we propose the Heterogeneous Spatial Embedding (HSE) approach, as illustrated in Fig. 5. HSE extracts heterogeneous spatial information from the HRG as a complement to the transformer. This is done from three perspectives: the attributes of roads and turns, their significance in the road space, and their relevancy within the traffic network.

1) **Attribute.** According to Shao et al. (Shao et al., 2022), one crucial factor influencing the accuracy of forecasting models is their ability to distinguish between different forecasting elements. As mentioned in section 1, roads and turns are two distinct areas with significant differences in their spatial-temporal attributes. Therefore, it is essential to categorize them accordingly. To accomplish this, we convert the type identities of each element (i.e., road or turn) into embedding matrices, which serve as prior knowledge for the model, as illustrated in Eq. (6). Furthermore, we address the heterogeneity within the same type of forecasting elements by transforming their identification numbers into embedding matrices. By capturing the correlation between traffic state and attributes, the model can enhance its understanding of the traffic network's heterogeneity.

$$Att = E(tp_V) + E(id_V), Att \in R^{N \times d} \quad (6)$$

where $tp_V \in R^N$ represents the type information of roads and turns; $id_V \in R^N$, from 1 to N, represents the identification number; The function $E()$ is the learnable embedding layer; d is the hidden dimension of embedding matrices.

2) **Significance.** Degree centrality provides a potential method for measuring the significance of forecasting elements by quantifying the number of edges connected to a node. In daily life, traffic accidents occurring at intersections with a higher degree centrality are more likely to cause congestion (Kan et al., 2019). Thus, we integrate degree centrality as a metric to assess the importance of forecasting elements. To achieve this, we employ an embedding layer to transform degree centrality, as illustrated in Eq. (7). By encoding degree centrality, our model can identify which traffic features are more influential and emphasize their effects.

$$Sig = E(deg_V^-) + E(deg_V^+) \quad (7)$$

where Sig denotes a matrix $\in R^{N \times d}$. deg_V^- and deg_V^+ represent the in-degree and out-degree, respectively.

3) **Relevancy.** In transportation systems, closely situated areas often exhibit similar traffic states due to their dense proximity. Additionally, it is important to note that traffic flows have distinct patterns when entering and exiting intersections, and the edges of HRG possess multiple properties. To capture these relationships, we introduce relevancy encoding to highlight edge heterogeneity while representing neighboring connections. We encode the adjacency matrix A of HRG using an embedding layer and a fully connected layer, as illustrated in Eq. (8). Subsequently, it is incorporated as a mask in the calculation of the attention mechanism. The relevancy encoding empowers our model to uncover the associations between diverse forecasting elements.

$$Rel = FC(E(A)), Rel \in R^{N \times N} \quad (8)$$

where $E(A) \in R^{N \times N}$ represent the embedded adjacency matrix.

3.5. Spatial Transformer

Traffic forecasting differs significantly from other time-series forecasting tasks due to the transmission and feedback of traffic flows, which result in interdependencies among prediction elements. It is essential to capture the complex spatial correlations for accurate forecasting, especially when dealing with heterogeneous elements. Given that road segments and intersection turns coexist on the same road network, we employ a unified module called Spatial Transformer (SpaFomer) to reveal the spatial dependencies between them.

As shown in Fig. 6, SpaFomer receives the corresponding hidden states of roads and turns as its inputs. For the l-th SpaFomer inputs, we denote them as $H_l^{rd} \in R^{N^{rd} \times T \times d}$ and $H_l^{tr} \in R^{N^{tr} \times T \times d}$, respectively. By concatenating H_l^{rd} and H_l^{tr}, SpaFomer generates a unified input, denoted as $H_l = concat(H_l^{rd}, H_l^{tr})$, where $H_l \in R^{N \times T \times d}$ and N is the sum of N^{rd} and N^{tr}. To discern the various attributes of distinct forecasting targets and their significance, SpaFomer incorporates the structural information outputted by HSE into the primary input, as illustrated in Eq. (9).

$$H_{l,s}^{e} = H_l + Att + Sig, H_{l,s}^{e} \in R^{N \times T \times d} \tag{9}$$

where $H_{l,s}^{e}$ is the hidden state of forecasting elements embedded with HSE.

Referring to the multi-head attention mechanism in the transformer (Vaswani et al., 2017), we employ trainable fully connected neural networks to linearly transform and generate multiple subspaces of Query (Q), Key (K), and Value (V), as described in Eq. (10). The utilization of multiple heads enables individual attention computations to be conducted within each subspace, thereby capturing distinct spatial correlations.

$$Q_{l,i}^{s} = H_{l,s}^{e} W_{l,Q_i}^{s}, K_{l,i}^{s} = H_{l,s}^{e} W_{l,K_i}^{s}, V_{l,i}^{s} = H_{l,s}^{e} W_{l,V_i}^{s}, where\ i \in [1, h] \tag{10}$$

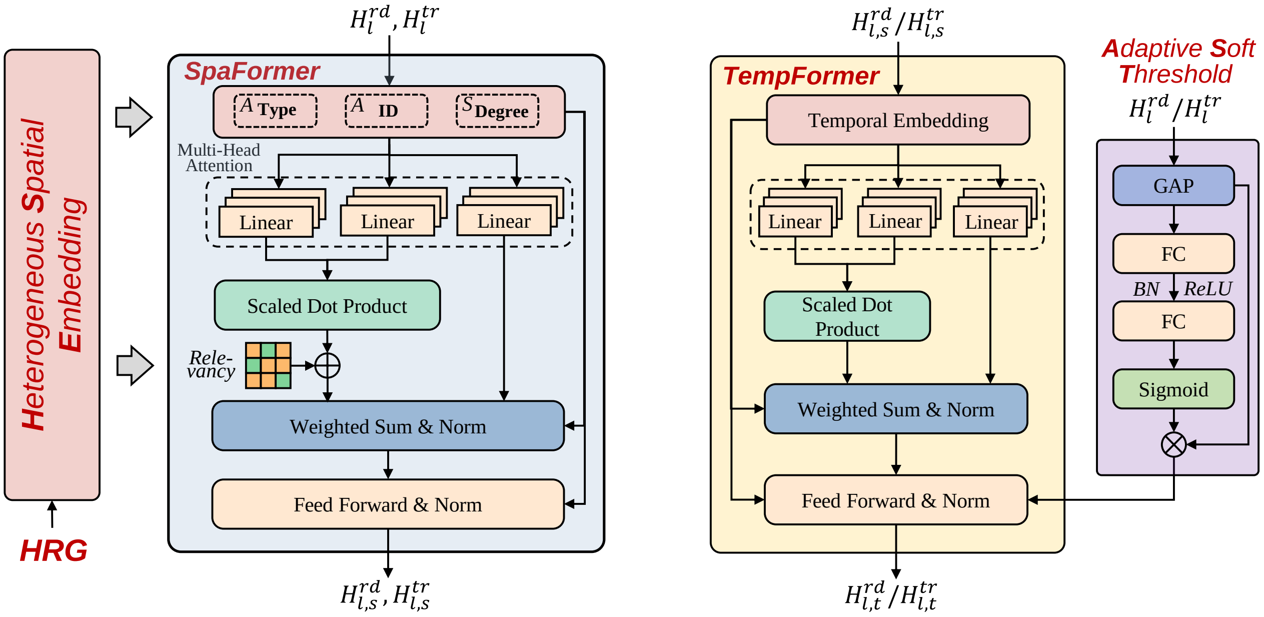
where h refers to the number of heads and i represents an individual head. $W_{l,Q_i}^{s}, W_{l,K_i}^{s} \in R^{d \times d'}$ and $W_{l,V_i}^{s} \in R^{d \times d}$ are matrices of trainable weights. $Q_{l,i}^{s}, K_{l,i}^{s} \in R^{T \times N \times d'}$ and $V_{l,i}^{s} \in R^{T \times N \times d}$ are subspaces tensors that undergo linear transformation.

For correlations between each forecasting element, we establish an attention weight matrix between $Q_{l,i}^{s}$ and $K_{l,i}^{s}$ through dot product and normalize it using Softmax. Furthermore, to capture the heterogeneous topological relevance in the traffic network, we incorporate the output of HSE, denoted as Rel, as a mask in the attention weight matrix, as depicted in Eq. (11). The output of the i-th head is then obtained by updating $V_{l,i}^{s}$ with $A_{l,i}^{s}$. Subsequently, the outputs of each head are concatenated using a trainable linear network to generate the final result, as illustrated in Eq. (12).

$$A_{l,i}^{s} = Softmax(\frac{Q_{l,i}^{s}(K_{l,i}^{s})^{T}}{\sqrt{d'}}) + Rel, A_{l,i}^{s} \in R^{T \times N \times N} \tag{11}$$

$$MultiHeadS(H_{l,s}^{e}) = FC(concat\left(A_{l,1}^{s}V_{l,1}^{s}, A_{l,2}^{s}V_{l,2}^{s}, \ldots, A_{l,h}^{s}V_{l,h}^{s}\right)) \tag{12}$$

where the dot product between Q and K is represented as $Q_{l,i}^{s}(K_{l,i}^{s})^{T}$. $\sqrt{d'}$ acts as the normalization constant. $MultiHeadS(H_{l,s}^{e}) \in R^{N \times T \times d}$ represents the output of multi-head attention.



(a) SpaFormer

(b) TempFormer

Fig. 6. Structure of SpaFomer and TempFormer.

In contrast to traditional GNNs that have a limited receptive field restricted to nearby nodes, SpaFomer captures spatial dependencies using global information. As a result, each node can assess the similarity of traffic patterns by considering all other nodes, without relying solely on the road network. Furthermore, SpaFomer incorporates the Rel (Relevancy) as a mask to the attention weights, allowing the forecasting elements to dynamically focus on other roads or turns based on the structure of the traffic network. For example, if Rel learns the differences in adjacency relations between distinct elements, the model will pay greater attention to the heterogeneity in traffic flow transmission, rather than solely focusing on connectivity.

In addition, to address the problems of disappearing and exploding gradients, a skip connection is introduced by incorporating the input H_l. Furthermore, LayerNorm (LN) (Vaswani et al., 2017) is applied to normalize the value, which helps accelerate the convergence of the model. Finally, the output of the l-th SpaFomer is generated using the feedforward layer, with the hidden states of roads and turns being separated. In summary, the output process of SpaFomer can be described as follows.

$$H_l^s = LN(MultiHeadS(H_{l,s}^e) + H_l), H_l^s \in R^{N\times T\times d} \tag{13}$$

$$H_{l,s}^{rd}, H_{l,s}^{tr} = split(FC(H_l^s) + H_l^s) \tag{14}$$

where $H_{l,s}^{rd} \in R^{N^{rd}\times T\times d}$ and $H_{l,s}^{tr} \in R^{N^{tr}\times T\times d}$ denote the hidden states of roads and turns, respectively, after the SpaFomer calculation.

3.6. Adaptive Soft Threshold and Temporal Transformer

Traffic states are heavily influenced by past historical states, and this correlation exhibits a non-linear variation over time (Zheng et al., 2020). As a solution, we propose the utilization of a module called Temporal Transformer (TempFormer) to extract the non-linear temporal correlations, as depicted in Fig. 6. Given the substantial heterogeneity between road segments and turns, we employ two separate TempFormers to handle each type independently. Additionally, as we refine the prediction targets, external factors such as traffic control exert a more significant impact on speed. To mitigate these external effects, we introduce an additional module called Adaptive Soft Threshold (AST) into TempFormer.

3.6.1. Adaptive Soft Threshold

Excessive noise in traffic states can significantly impact the attention mechanism's learning ability. The presence of external outliers affects the mechanism's global receptive field, resulting in inaccuracies in forecasting results. To generate more consistent predictions that align with the overall trend of traffic states, we incorporate the Adaptive Soft Threshold (AST) into TempFormer, inspired by (Zhao et al., 2019b).

The soft threshold plays a crucial role in many signal denoising algorithms. By defining a threshold, we can eliminate features with absolute values below it and compress features with absolute values exceeding it. Furthermore, noise levels often vary across traffic state samples. For instance, during peak times, fluctuations in traffic speeds become more pronounced. To tackle this problem, AST dynamically assigns thresholds based on the characteristics of the time series.

In the AST module, the original inputs (H_l^{rd}, H_l^{tr}) of the l-th SpaFormer undergo global average pooling (GAP) and dimensionality reduction, generating the feature vector $gap = GAP(H_l^{\epsilon}) \in R^{N' \times 1 \times 1}$.Here, $\epsilon \in [rd, tr]$ represents the identification of roads and turns, and N' denotes their respective numbers. Next, the gap is fed into a sub-network with two fully connected layers, where the output is normalized to the range of 0-1 using the Sigmoid function. The sub-network produces a dynamic coefficient α, as shown in Eq. (15). Subsequently, the outputs of AST $(H_{l,ast}^{\epsilon})$ are generated by utilizing the $\alpha \times gap$ as the threshold, as illustrated in Eq. (16).

$$\alpha = Sigmoid\big(FC(H_l^{\epsilon})\big), \alpha \in R^{N' \times 1 \times 1} \tag{15}$$

$$H_{l,ast}^{\epsilon} = \begin{cases} H_l^{\epsilon} - \alpha \times gap, H_l^{\epsilon} > \alpha \times gap \\ \qquad 0, H_l^{\epsilon} \leq \alpha \times gap \end{cases} \tag{16}$$

where α is the learnable weight factor coefficient. FC represents the sub-network consisting of two fully connected layers. $H_{l,ast}^{\epsilon} \in R^{N' \times T \times d}$ is the final output of AST.

3.6.2 Temporal Transformer

Unlike recurrent neural networks, the attention mechanism processes time series in parallel, which can result in the loss of sequence order information. Therefore, we use positional encoding (PE) (Vaswani et al., 2017) to transform the sequence of temporal series into vectors. In addition, there is often a correlation between traffic states and the specific time they are collected. For example, traffic speeds typically slow down during rush hours. To account for this, we divide a day into twelve time periods and encode the day-of-week and the time-of-day for each time step with one-hot encoding. Then, we concatenate them into time information embedding ($TIE \in R^{7+12}$), as shown in Eq. (17).

$$H_{l,t}^{e} = H_{l,s}^{\epsilon} + PE + E(TIE) \tag{17}$$

where $H_{l,s}^{\epsilon} \in R^{N' \times T \times d}$ represents the hidden state of roads or turns generated by SpaFormer. $PE \in R^{T \times d}$ represents the position encoding and $E(TIE) \in R^{T \times d}$ represents the time information embedding.

After embedding temporal features, TempFormer utilizes the multi-head attention mechanism to generate multiple subspaces of Q, K, and V, capturing diverse time-related correlations. Unlike SpaFormer, TempFormer focuses solely on the temporal dimension and does not apply a mask during attention weight calculation. The process is described in more detail below.

$$Q_{l,i}^{t} = H_{l,t}^{e} W_{l,Q_i}^{t}, K_{l,i}^{t} = H_{l,t}^{e} W_{l,K_i}^{t}, V_{l,i}^{t} = H_{l,t}^{e} W_{l,V_i}^{t}, where\ i \in [1, h] \tag{18}$$

$$A_{l,i}^{t} = Softmax\Big(\frac{Q_{l,i}^{t} (K_{l,i}^{t})^{T}}{\sqrt{d'}}\Big), A_{l,i}^{t} \in R^{N' \times T \times T} \tag{19}$$

$$MultiHeadT(H_{l,t}^{e}) = FC(concat(A_{l,1}^{t}V_{l,1}^{t}, A_{l,2}^{t}V_{l,2}^{t}, \ldots, A_{l,h}^{t}V_{l,h}^{t})) \tag{20}$$

where $Q_{l,i}^{t}, K_{l,i}^{t} \in R^{N' \times T \times d'}$ and $V_{l,i}^{t} \in R^{N' \times T \times d}$ represent the Q, K, and V of the i-th head, respectively. $A_{l,i}^{t}$ is the attention weight, and $MultiHeadT(H_{l,t}^{e}) \in R^{N' \times T \times d}$ represents the output of the multi-head attention.

The outputs of the attention mechanism and the AST are combined and then passed through a feed-forward layer. This step aims to integrate the information extracted from both sources and produce a comprehensive representation.

$$H_{l}^{t} = LN(MultiHeadT(H_{l,t}^{e}) + H_{l,ast}^{\epsilon}), H_{l}^{t} \in R^{N' \times T \times d} \tag{21}$$

$$H_{l,t}^{\epsilon} = FC(H_{l}^{t}) + H_{l,s}^{\epsilon}, H_{l,t}^{\epsilon} \in R^{N' \times T \times d} \tag{22}$$

where $H_{l,t}^{\epsilon}$ represents the hidden states of roads or turns through TempFormer, which is also the output of the ST-block.

4. Experiments

4.1 Dataset

The study area chosen for this research is the Wuchang district, which is a heavily populated urban area in Wuhan, China. The area features a high volume of taxis and various amenities including hospitals, schools, and parks. The region of interest is illustrated in Fig. 7, spanning from longitude 114.353° E to 114.295° E and latitude 30.562° N to 30.520° N, using the coordinate reference system EPSG:4326 - WGS 84.

Traffic speed data are collected from approximately 4,000 taxis operating in the Wuchang district between July 1 and July 31, 2018. We construct the heterogeneous road network graph following the methodology described in section 3.2. It contains 277 road segments and 269 turns in 43 intersections. Also, we extract and record the average speed for each road and turn at 10-minute intervals. In total, we gathered 2,437,344 speeds during the 31-day period, which are then divided into training and test sets, with 70% and 30% allocation, respectively.

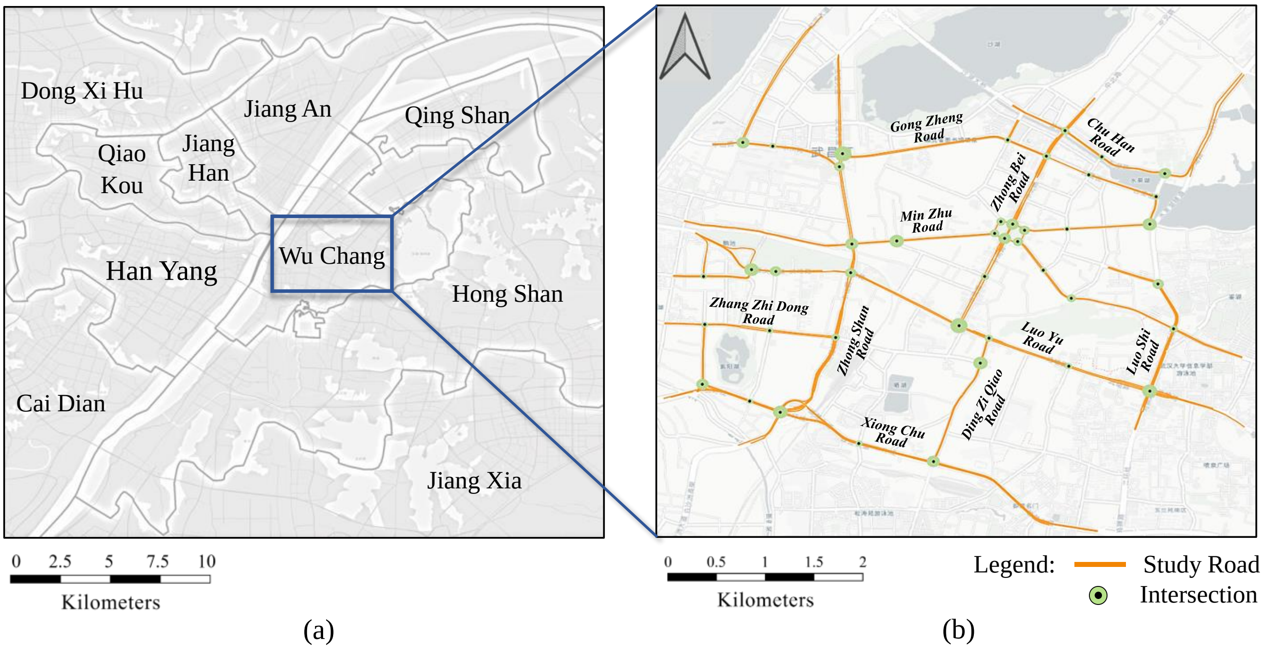


Fig. 7. Road network of the study area. (a) Central districts of Wuhan, China; (b) Road network.

4.2 Experimental Settings

The STHGFormer model is constructed via the PyTorch library with the CUDA version being 10.1. The experiments are conducted on Intel Core i9-10900X @ 3.70GHz CPUs and NVIDIA GeForce RTX 3090 GPUs. The AdamW optimizer is utilized to train the STHGFormer for 200 epochs, and the model employs MSELoss as its loss function. The learning rate is set to 10^{-3} initially, and it is scaled down to 10^{-4} by the final epoch. The batch size and the hidden dimension d are both 32. STHGFormer sets T to 12 and P to 3 for forecasting. It predicts traffic speeds of next 10, 20, and 30 minutes with traffic state data recorded over the past two hours. Unless stated otherwise, STHGFormer is composed of two ST-blocks, using four heads in both SpaFormer and TempFormer.

To comprehensively evaluate the forecasting accuracy of STHGFormer, the mean absolute error (MAE) and the root mean square error (RMSE) are selected as the metrics, which are defined as

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \widehat{y_i}|}{N} \tag{22}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}{N}} \tag{23}$$

where y_i is the true value, $\widehat{y_i}$ is the predicted value, and N is the number of forecasting elements.

In this paper, we present a comparative analysis between STHGFormer and nine other widely utilized deep learning models, including:

· **Long Short-Term Memory (LSTM)** (Yu et al., 2019): A deep learning model that belongs to the recurrent neural network family and is commonly employed in sequence data processing.

· **Spatial-Temporal Graph Convolution Network (STGCN)** (Yu et al., 2017): A deep learning model designed for processing spatial-temporal data, utilizing convolutional operations to learn spatial-temporal dependencies.

· **Attention-based Spatial-Temporal Graph Convolutional Network (ASTGCN)** (Guo et al., 2019): A deep learning model that uses a multi-branch structure to model three different temporal properties of traffic flow. It integrates the attention mechanism with spatial-temporal convolution to enhance model performance.

· **Graph-WaveNet (GWNet)** (Wu et al., 2019): A CNN-based model that mixes adaptive graphs with dilated convolutional operations to capture latent spatial-temporal connections.

· **Spatial-Temporal Transformer Network (STTN)** (Xu et al., 2020): An attention-based model that utilizes spatial transformers, GCNs, and temporal transformers. The model employs a multi-head attention mechanism.

· **Optimized Graph Convolution Recurrent Neural Network (OGCRNN)** (Guo et al., 2020): OGCRNN combines the architectures of GCN and RNN in its design. Unlike conventional graph-based neural networks, OGCRNN utilizes a residual graph model that aims to substitute empirical and fixed graphs.

· **Hierarchical Graph Convolution Network (HGCN)** (Guo et al., 2021): HGCN is designed to take advantage of the innate hierarchical structure present in traffic systems. This network operates on both micro and macro traffic graphs, leading to enhanced performance.

· **Dual Dynamic Spatial-Temporal Graph Convolution Network (DDSTGCN)** (Sun et al., 2022): DDSTGCN captures the dynamic spatial-temporal feature of graph edges in traffic flow data by transforming the data into a dual hypergraph.

· **Spatial-Temporal Identity Network (STID)** (Shao et al., 2022): STID is a simple, yet effective model for Multivariate Time Series (MTS) forecasting. The model utilizes simple MLPs to solve the previously challenging indistinguishability issue, achieving superior performance.

4.3 Comparison of forecasting performance

The predictive performance of STHGFormer is evaluated by comparing it to baseline models using a real-world traffic speed dataset. The training epochs for all models are set to 200. Baseline models are trained with the parameters defined in their original papers. We use the heterogeneous road network graph (HRG) as input for models that require a predefined graph. In contrast to STHGFormer, all baseline models consider road segments and intersection turns as identical forecasting targets. Table 1 presents the forecast performances of all models for both roads and turns in 10-minute (1 step), 20-minute (2 steps), and 30-minute (3 steps) forecasting tasks.

Table 1 Performance comparison of different models

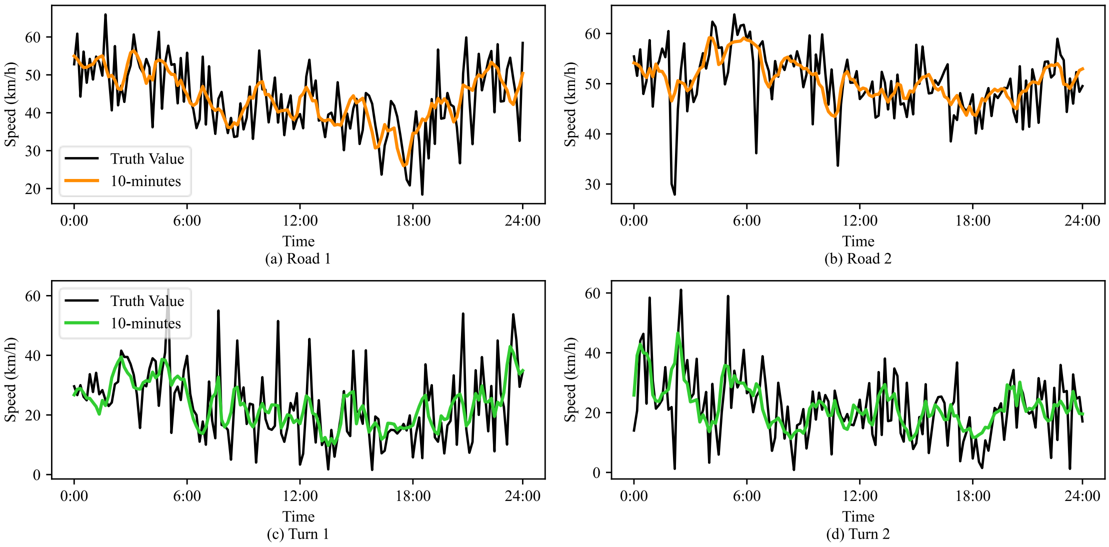
Target	Model	10min		20min		30min	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
Road	LSTM	5.12	6.75	5.26	6.93	5.37	7.07
	STGCN	5.06	6.68	5.18	6.83	5.28	6.95
	ASTGCN	4.90	6.49	4.94	6.55	4.99	6.60
	OGCRNN	4.90	6.49	4.95	6.57	5.00	6.62
	GWNet	4.80	6.40	4.88	6.49	4.93	6.56
	HGCN	4.87	6.47	4.95	6.57	5.01	6.64
	STTN	4.86	6.42	4.87	6.44	4.94	6.54
	DDSTGCN	4.80	6.39	4.90	6.50	4.93	6.56
	STID	4.84	6.41	4.87	6.45	4.93	6.53
	STHGFormer	**4.41**	**5.81**	**4.29**	**5.65**	**4.50**	**5.95**
	Improvements	8.12%	9.08%	11.91%	12.26%	8.72%	9.30%
Turn	LSTM	8.76	11.73	8.88	11.84	8.97	11.92
	STGCN	8.69	11.64	8.79	11.75	8.87	11.81
	ASTGCN	8.46	11.43	8.49	11.46	8.52	11.49
	OGCRNN	8.45	11.42	8.50	11.47	8.53	11.50
	GWNet	8.22	11.28	8.21	11.21	8.22	11.19
	HGCN	8.43	11.42	8.49	11.49	8.54	11.54
	STTN	8.37	11.21	8.36	11.20	8.42	11.28
	DDSTGCN	8.28	11.27	8.29	11.24	8.27	11.23
	STID	8.40	11.24	8.44	11.30	8.45	11.31
	STHGFormer	**7.32**	**9.70**	**7.22**	**9.58**	**7.84**	**10.44**
	Improvements	10.95%	13.93%	12.06%	14.46%	4.62%	6.70%

Predicting traffic speeds in intersection turns is more challenging due to the greater speed fluctuation compared to road segments. The MAE and RMSE for turn forecasting are considerably larger than those for road forecasting across all models. Nonetheless, our proposed model, STHGFormer, outperforms all baseline models in every evaluation metric, demonstrating its suitability for integrated and fine-grained traffic forecasting.

In comparison to LSTM, which only captures temporal dependencies, models utilizing GNNs such as STGCN and OGCRNN emphasize the importance of the road network and demonstrate improved performance. In addition, dynamic spatial-temporal analysis models like GWNet and DDSTGCN outperform static models that only consider fixed road network information. Furthermore, STID achieves relatively good results with its simple structure by effectively distinguishing the spatial-temporal heterogeneity of forecasting elements.

The proposed STHGFormer achieves state-of-the-art performance in integrated traffic forecasting for both roads and turns. Specifically, when predicting traffic speeds for the next 10 minutes, STHGFormer demonstrates an 8.1% improvement in MAE for road forecasting and a 10.9% improvement in MAE for turn forecasting. In the presence of forecasting targets with heterogeneity, STHGFormer highlights the significance of diverse spatial-temporal attributes and relationships, generating superior performance. Moreover, when confronted with fluctuating traffic states in turns, STHGFormer demonstrates an even more substantial improvement in accuracy, emphasizing the importance of adaptive soft thresholding.

In order to visually illustrate the prediction performance of STHGFormer, we present the forecasting results using line graphs. Fig. 8 compares the actual speeds with the forecasted results for two roads and two turns at the intersection of Zhongbei Road and Chuhan Road throughout the entire day of July 23rd.



(a) Road 1

(b) Road 2

(c) Turn 1

(d) Turn 2

Fig. 8. Visualization of traffic forecasting results. The forecasting targets shown are also form roads and turns near the intersection of Zhong Bei Road and Chu Han Road.

STHGFormer accurately predicts the overall trends in speed for both roads and turns within a day. It can be observed that turns exhibit more unpredictable fluctuations compared to roads. This variability is likely influenced by factors such as traffic control, which significantly affect vehicle speed at intersections. In STHGFormer, the proposed AST efficiently filters out excessive noise, enabling our model to accommodate speed variation in turns.

4.4. Ablation Study

We construct four variations of STHGFormer by removing its main modules. STHGFormer-SF excludes the SpaFormer module, while STHGFormer-TF removes the TempFormer module. Similarly, STHGFormer-HSE does not use heterogeneous spatial embedding (HSE) for representing the HRG. Finally, STHGFormer-AST is designed without the adaptive soft threshold (AST) for noise suppression.

For a more in-depth analysis of the impact of HSE, we conduct additional ablation experiments focusing on its attribute encoding (-- Att), significance encoding (-- Sig), and relevancy encoding (-- Rel). The models used in these experiments maintained the same settings as the STHGFormer, with the exception of the components under investigation. Then, we assessed the forecasting performance of each variation to demonstrate their respective influences, as shown in Table 2.

Table 2 demonstrates that all incomplete STHGFormer models exhibit varying levels of accuracy degradation, highlighting the significance of each module in capturing spatial-temporal correlation. Notably, STHGFormers that exclude the temporal modules (STHGFormer-TF and STHGFormer-AST) display greater accuracy degradation compared to those that exclude spatial modules (STHGFormer-SF and STHGFormer-HSE). This suggests that temporal correlations have a more pronounced influence on traffic forecasting than spatial correlations. Additionally, the exclusion of the AST module leads to a more substantial decline in accuracy for turn speed forecasting. Specifically, when compared to the 10-minute road forecasting, the mean absolute error (MAE) gap of turn forecasting between STHGFormer-AST and the complete model (STHGFormer) increased by 1.4%, highlighting the significance of the soft threshold.

Table 2 Ablation analysis of different STHGFormer components

Target	Model	10min		20min		30min	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
Road	STHGFormer	**4.41**	**5.81**	**4.29**	**5.65**	**4.5**	**5.95**
	STHGFormer-SF	4.48	5.92	4.38	5.77	4.55	5.60
	STHGFormer-TF	4.81	6.37	4.82	6.38	4.88	6.46
	STHGFormer-HSE	4.59	6.07	4.46	5.89	4.60	6.09
	-- Att	4.43	5.84	4.30	5.66	4.51	5.96
	-- Sig	4.69	6.19	4.59	6.06	4.68	6.18
	-- Rel	4.52	5.97	4.39	5.78	4.56	6.02
	STHGFormer-AST	4.78	6.34	4.78	6.34	4.84	6.43
Turn	STHGFormer	**7.32**	**9.70**	**7.22**	**9.58**	**7.84**	**10.44**
	STHGFormer-SF	7.43	9.85	7.28	9.66	7.88	10.49
	STHGFormer-TF	8.03	10.70	8.04	10.73	8.08	10.78
	STHGFormer-HSE	7.63	10.11	7.42	9.83	7.91	10.52
	--- Att	7.41	9.83	7.24	9.61	7.85	10.45
	-- Sig	7.93	10.53	7.76	10.34	8.03	10.70
	-- Rel	7.47	9.89	7.29	9.67	7.86	10.46
	STHGFormer-AST	8.11	10.86	8.02	10.73	8.16	10.92

It is important to note that the STHGFormer-HSE, which only remove the heterogeneous spatial embedding, shows a more significant reduction in accuracy compared to the STHGFormer-SF model that completely eliminates the SpaFormer module. We speculate that the lesser accuracy of STHGFormer-HSE is due to the lack of information regarding heterogeneous road networks. The lack of road network information makes it difficult for the attention mechanism to properly identify spatial dependencies within complex traffic flows, consequently reducing the forecasting capabilities of the model. Additional ablation experiments indicate that significance encoding has a more significant impact on forecasting results within the HSE when compared to attribute and relevancy encoding.

4.6. Comparison of integrated and disintegrated STHGFormer

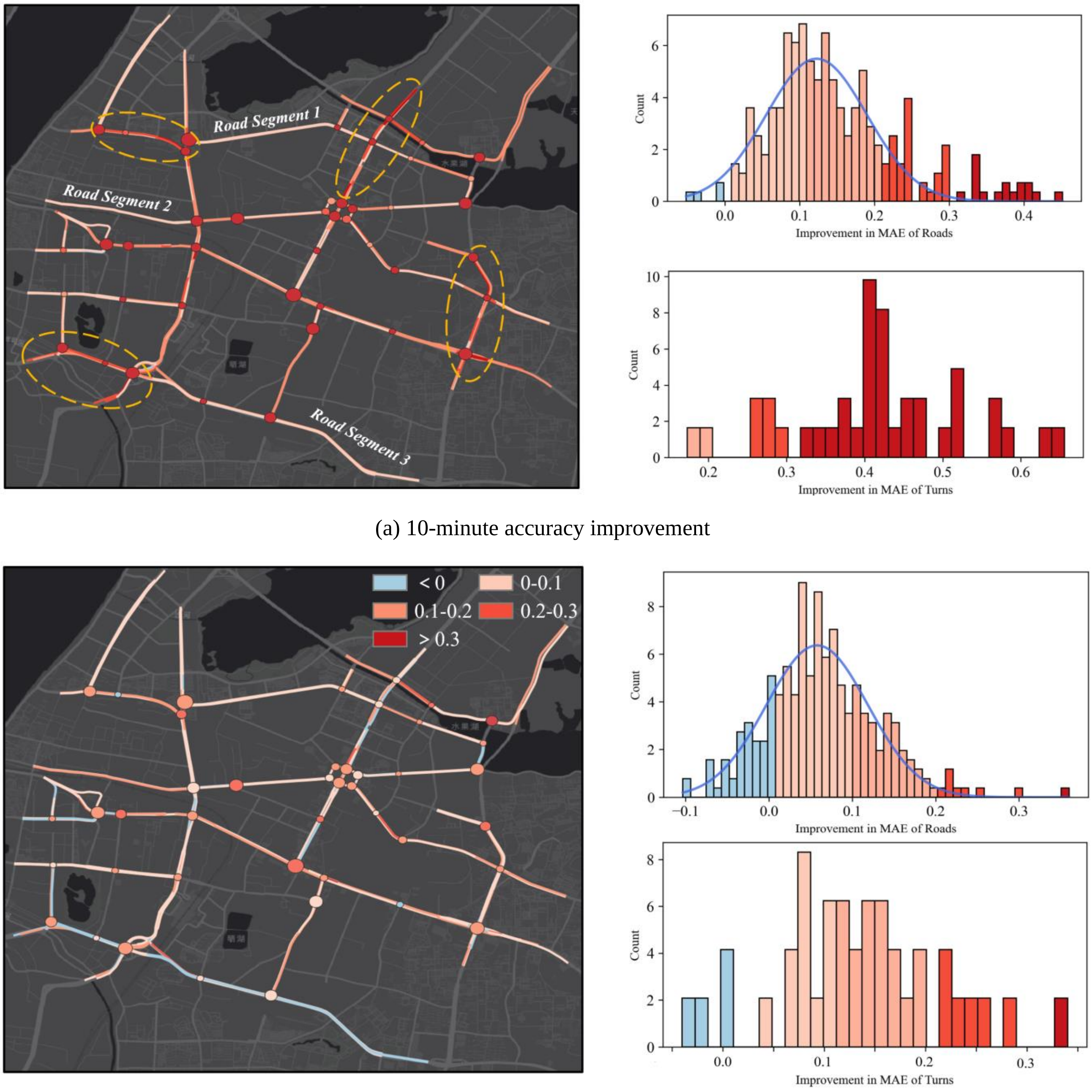
To examine the influence of simultaneously forecasting, we compare the integrated model, STHGFormer, against two disintegrated models: STHGFormer (R), which solely forecasts road segments, and STHGFormer (T), which solely forecasts turns. To ensure comparison consistency, we replace the HRG used in integrated forecasting with adjacent graphs when training STHGFormer (R) and STHGFormer (T). Specifically, we consider that connected road segments are linked and turns within the same intersection are linked. We present the comparison of predictive accuracy achieved by STHGFormer, STHGFormer (R), and STHGFormer (T) in Table 3.

Table 3 Performance comparison of integrated and disintegrated STHGFormer

Target	Model	10min		20min		30min	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
Road	STHGFormer (R)	4.53	5.99	4.40	5.81	4.55	6.03
	STHGFormer	**4.41**	**5.81**	**4.29**	**5.65**	**4.50**	**5.95**
	Improvements	2.65%	3.01%	2.50%	2.75%	1.10%	1.33%
Turn	STHGFormer (T)	7.73	10.25	7.56	10.02	7.99	10.63
	STHGFormer	**7.32**	**9.70**	**7.22**	**9.58**	**7.84**	**10.44**
	Improvements	5.30%	5.37%	4.50%	4.39%	1.88%	1.79%

The data presented in Table 3 indicates that integrated forecasting achieves higher accuracy for both road and turn prediction across various time intervals. There is a more notable improvement in turn forecasting compared to road forecasting. This could be attributed to the fact that turns are situated in intersections and have closer connections to other turns and roads. As a result, the advantages of integrated forecasting are more pronounced for turns.

Fig. 9 illustrates the accuracy enhancement of roads and turns in two different prediction intervals (10 and 30 minutes). To offer a more visual representation, we compute the average improvement of turns at the same intersection and use the accuracy enhancement of intersections to portray the enhancement of turns. As shown in Fig. 9, most roads and turns experience an improvement in accuracy for both prediction intervals. This result indicates that integrated forecasting significantly enhances accuracy by capturing dependencies between road segments and intersection turns.



(a) 10-minute accuracy improvement

(b) 30-minute accuracy improvement

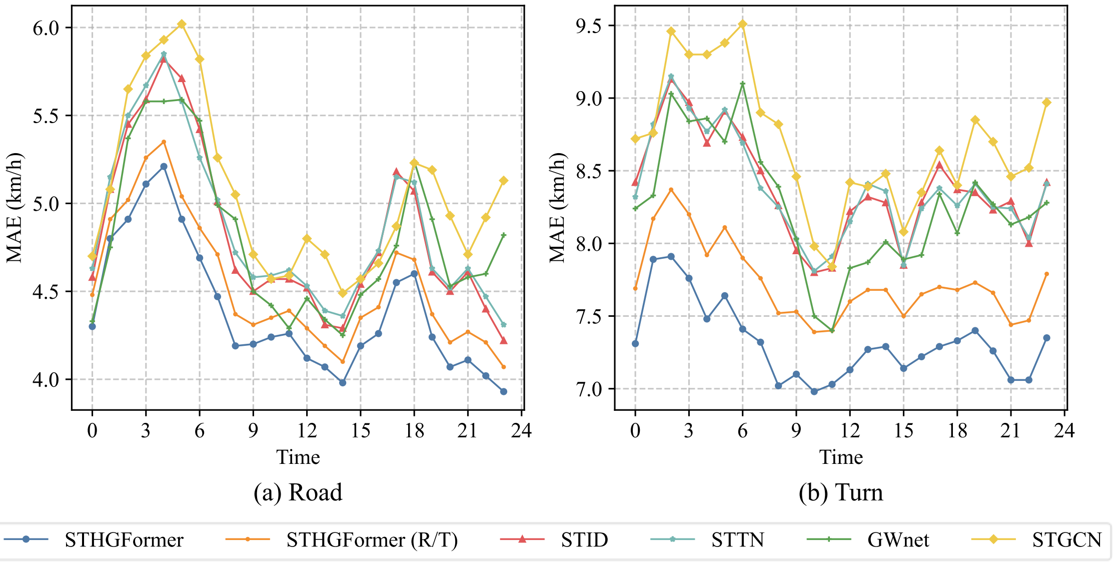
Fig. 9. The distribution of MAE improvement of integrated STHGFormer against disintegrated STHGFormer.

In the 10-minute forecast, road segments close to intersections exhibit a higher improvement in accuracy. For example, roads circled in yellow exhibit a significantly greater enhancement in accuracy compared to those located farther from intersections, such as road segments 1, 2, and 3. Moreover, almost all intersections demonstrate significant improvements in accuracy. To summarize, regions with more complex topological structures show more noticeable improvements in accuracy. By examining the spatial relationships within a complete traffic network, both roads and turns obtain more accurate forecasting results as they benefit from each other's traffic states.

Ultimately, by comparing the results of the 10-minute prediction with the 30-minute prediction in Fig. 9(a) and Fig. 9(b), it is observed that the accuracy of long-term prediction decreases throughout the overall traffic road network. As the time scale becomes longer, the spatial relationship between roads and turns becomes increasingly complex, posing a greater challenge for accurate prediction. In particular, the decline in accuracy is more significant at the edge of the study area, where the topology is relatively simple.

4.7. Time-varying Accuracy of Models

To investigate the predictive stability over a day, we compared STHGFormer with five other models, as well as the disintegrated versions of STHGFormer, namely STHGFormer (R) and STHGFormer (T). We computed the mean absolute error (MAE) of the 10-minute predictions for each hour and illustrated the results of roads and turns in Fig. 10(a) and Fig. 10(b) respectively.



(a) Road

(b) Turn

Fig. 10. The average MAE (km/h) with respect to time of day of different models.

The results demonstrate that STHGFormer consistently outperforms the other models in terms of predictive accuracy. Notably, STHGFormer showcases the narrowest range of fluctuating errors compared to the alternative models when forecasting both roads and turns. This finding highlights the exceptional capability of STHGFormer in effectively capturing intricate spatial-temporal dependencies.

5. Conclusion and future works

This paper proposes a novel approach called Spatial-Temporal Heterogeneous Graph Transformer (STHGFormer) for integrated traffic forecasting. The proposed method addresses the challenges posed by the complex relationships between roads and turns and their heterogeneous spatial-temporal attributes. By constructing the Heterogeneous Road network Graph (HRG), we completely depict the traffic network where roads and turns coexist. Moreover, the proposed Heterogeneous Spatial Embedding (HSE) module extracts complex spatial information within HRG from multiple dimensions. Through the combination of HSE and the multi-headed attention mechanism, the SpaFormer efficiently uncovers the spatial correlations between roads and turns. To handle the highly fluctuating time series, the TempFormer incorporates an Adaptive Soft Threshold (AST) module, which leverages a learnable threshold to mitigate the impact of noise.

Experimental results on a real-world urban traffic dataset demonstrated that STHGFormer outperforms other baselines, achieving outstanding forecasting accuracy for both road segments and intersection turns. Ablation studies provided further evidence of the indispensability of all proposed modules within STHGFormer. In addition, the proposed method exhibited consistent and reliable forecasting stability across various times of the day.

Future research directions will focus on incorporating additional factors such as weather conditions, traffic accidents, and other variables which influence traffic states. Besides, we aim to leverage the forecasting results of roads and turns to enhance the accuracy and granularity of route planning and travel time estimation. Furthermore, we will further explore effective methods to integrate diverse data sources, such as data collected by fixed sensor and trajectory data, to achieve a more comprehensive traffic forecasting.

References

Chauhan, R., Dhamaniya, A., Arkatkar, S., 2021. Driving behavior at signalized intersections operating under disordered traffic conditions. *Transportation research record* 2675(12), 1356-1378.

Chen, C., Petty, K., Skabardonis, A., Varaiya, P., Jia, Z., 2001. Freeway performance measurement system: mining loop detector data. *Transportation Research Record* 1748(1), 96-102.

Fang, M., Tang, L., Yang, X., Chen, Y., Li, C., Li, Q., 2021. FTPG: A fine-grained traffic prediction method with graph attention network using big trace data. *IEEE Transactions on Intelligent Transportation Systems*.

Guo, K., Hu, Y., Qian, Z., Liu, H., Zhang, K., Sun, Y., Gao, J., Yin, B., 2020. Optimized graph convolution recurrent neural network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems* 22(2), 1138-1149.

Guo, K., Hu, Y., Sun, Y., Qian, S., Gao, J., Yin, B., 2021. Hierarchical Graph Convolution Network for Traffic Forecasting, *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 151-159.

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

Kan, Z., Tang, L., Kwan, M.-P., Ren, C., Liu, D., Li, Q., 2019. Traffic congestion analysis at the turn level using Taxis' GPS trajectory data. *Computers, Environment and Urban Systems* 74, 229-243.

Kreuzer, D., Beaini, D., Hamilton, W., Létourneau, V., Tossou, P., 2021. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems* 34, 21618-21629.

Li, F., Feng, J., Yan, H., Jin, G., Yang, F., Sun, F., Jin, D., Li, Y., 2021. Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution. *ACM Transactions on Knowledge Discovery from Data*.

Li, Y., Yu, R., Shahabi, C., Liu, Y., 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint* arXiv:.01926.

Liao, L., Hu, Z., Zheng, Y., Bi, S., Zou, F., Qiu, H., Zhang, M., 2022. An improved dynamic Chebyshev graph convolution network for traffic flow prediction with spatial-temporal attention. *Applied Intelligence*, 1-13.

Liebig, T., Piatkowski, N., Bockermann, C., Morik, K., 2017. Dynamic route planning with real-time traffic predictions. *Information Systems* 64, 258-265.

Mialon, G., Chen, D., Selosse, M., Mairal, J., 2021. Graphit: Encoding graph structure in transformers. *arXiv preprint arXiv:.05667*.

Min, E., Chen, R., Bian, Y., Xu, T., Zhao, K., Huang, W., Zhao, P., Huang, J., Ananiadou, S., Rong, Y., 2022. Transformer for graphs: An overview from architecture perspective. *arXiv preprint arXiv:.08455*.

Shao, Z., Zhang, Z., Wang, F., Wei, W., Xu, Y., 2022. Spatial-Temporal Identity: A Simple yet Effective Baseline for Multivariate Time Series Forecasting, *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 4454-4458.

Sun, Y., Jiang, X., Hu, Y., Duan, F., Guo, K., Wang, B., Gao, J., Yin, B., 2022. Dual Dynamic Spatial-Temporal Graph Convolution Network for Traffic Prediction. *IEEE Transactions on Intelligent Transportation Systems* 23(12), 23680-23693.

Tang, L., Ren, C., Liu, Z., Li, Q., 2017. A road map refinement method using delaunay triangulation for big trace data. *ISPRS International Journal of Geo-Information* 6(2), 45.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., Yu, P.S., 2019. Heterogeneous graph attention network, *The world wide web conference*, pp. 2022-2032.

Wei, H., Zheng, G., Gayah, V., Li, Z., 2019. A survey on traffic signal control methods. *arXiv preprint arXiv:.08117*.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y.J.I.t.o.n.n., 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* 32(1), 4-24.

Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C.J.a.p.a., 2019. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint* arXiv:.00121.

Xu, M., Dai, W., Liu, C., Gao, X., Lin, W., Qi, G.-J., Xiong, H., 2020. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint* arXiv:.02908.

Ye, J., Zhao, J., Ye, K., Xu, C., 2022. How to Build a Graph-Based Deep Learning Architecture in Traffic Domain: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 23(5), 3904-3924.

Yin, X., Wu, G., Wei, J., Shen, Y., Qi, H., Yin, B., 2021. Multi-stage attention spatial-temporal graph networks for traffic prediction. *Neurocomputing* 428, 42-53.

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., Liu, T.-Y., 2021. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems* 34, 28877-28888.

Yu, B., Yin, H., Zhu, Z., 2017. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:.04875*.

Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation* 31(7), 1235-1270.

Zhang, J., Zheng, Y., Qi, D., 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction, *Thirty-first AAAI conference on artificial intelligence*.

Zhang, W., Zhu, F., Lv, Y., Tan, C., Liu, W., Zhang, X., Wang, F.-Y., 2022. AdapGL: An adaptive graph learning algorithm for traffic prediction based on spatiotemporal neural networks. *Transportation Research Part C: Emerging Technologies* 139, 103659.

Zhang, Y., Cheng, T., Ren, Y., 2019. A graph deep learning method for short-term traffic forecasting on large road networks. *Computer-Aided Civil and Infrastructure Engineering* 34(10), 877-896.

Zhao, J., Chen, C., Liao, C., Huang, H., Ma, J., Pu, H., Luo, J., Zhu, T., Wang, S., 2022. 2F-TP: Learning Flexible Spatiotemporal Dependency for Flexible Traffic Prediction. *IEEE Transactions on Intelligent Transportation Systems*.

Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., Deng, M., Li, H., 2019a. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems* 21(9), 3848-3858.

Zhao, M., Zhong, S., Fu, X., Tang, B., Pecht, M., 2019b. Deep residual shrinkage networks for fault diagnosis. *IEEE Transactions on Industrial Informatics* 16(7), 4681-4690.

Zheng, C., Fan, X., Wang, C., Qi, J., 2020. Gman: A graph multi-attention network for traffic prediction, *Proceedings of the AAAI conference on artificial intelligence*, pp. 1234-1241.