Information Fusion 102 (2024) 102063

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Towards integrated and fine-grained traffic forecasting: A Spatio-Temporal Heterogeneous Graph Transformer approach

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

[a] State Key Laboratory of 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

ARTICLE INFO

Keywords:
Integrated traffic forecasting
Spatio-temporal heterogeneity
Heterogeneous road network graph
Spatio-temporal transformer
Trajectory data

ABSTRACT

Fine-grained traffic forecasting is crucial for the management of urban transportation systems. Road segments and intersection turns, as vital elements of road networks, exhibit heterogeneous spatial structures, yet their traffic states are interconnected due to spatial proximity. The heterogeneity and interrelationships arising from different road network elements pose major challenges to accurate traffic forecasting. However, existing forecasting studies focus solely on bidirectional road segments, disregarding the relationships between roads and turns. To achieve integrated traffic forecasting that considers both road segments and intersection turns, we propose a novel Spatio-Temporal Heterogeneous Graph Transformer (STHGFormer). 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. Then, we propose a Heterogeneous Spatial Embedding (HSE) module to encode road network information, including heterogeneous attributes and interactions in the HRG. Based on the spatial information encoded by HSE, a unified SpaFormer, serving as the spatial module of STHGFormer, captures the interdependencies between roads and turns across the entire traffic network. To mitigate the impact of high temporal fluctuation, we embed the Adaptive Soft Threshold (AST) module into TempFormer, which dynamically adjusts the threshold to enhance the analysis capability of complex temporal correlations. Experiments conducted on a real-world dataset from Wuhan, China, demonstrate that STHGFormer outperforms state-of-the-art methods, achieving a 6.1 % improvement in road forecasting and an 8.5 % improvement in turn forecasting.

1. Introduction

As the cornerstone of Intelligent Transportation Systems (ITS) [1,2], traffic forecasting is critical for route planning [3], travel time estimation [4], and traffic signal control [5]. Accurate predictions provide essential information for traffic participants to make reasonable decisions, which enhances the safety, security, and efficiency of daily transportation [6].

Crowd-sourced data, such as trajectory data, has significantly contributed to the refinement and timeliness of traffic perception and forecasting [7]. Currently, the forecasting scale has been refined from coarse grids [8] to specific components of the road network, such as road segments and intersection turns. Guo et al. [9] and Zhang et al. [10] achieved accurate traffic forecasting for bi-directional road segments, enabling accurate short-term citywide traffic prediction. Furthermore, Fang et al. [11] refined the forecasting scale to turn-level and realize detailed predictions for different turns within intersections. While previous studies have accomplished fine-grained traffic forecasting to some extent, their depictions of the traffic network remain incomplete, neglecting the interrelationships and heterogeneity arising from different road network elements.

In general, as shown in Fig. 1, a road network can be divided into two main components: road segment (unconstrained) and intersection (controlled by traffic signals). Intersections act as crucial areas where traffic flows interact and change directions [12], connecting road segments and carrying turning movements. Moreover, due to the transmission and feedback of traffic flows, states of roads and turns are closely related. For instance, traffic congestion at an intersection often propagates to surrounding road segments, leading to decreased speeds of roads. Therefore, it is pivotal to integrate road segments and intersection turns to capture the complete spatial dependencies within the traffic network, which can enhance the integrity, fineness, and accuracy

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

https://doi.org/10.1016/j.inffus.2023.102063
Received 13 July 2023; Received in revised form 22 September 2023; Accepted 2 October 2023
Available online 8 October 2023

of traffic forecasts.

Meanwhile, complex spatio-temporal heterogeneity exists between road segments and intersection turns. Regarding spatial structure, the characteristic of traffic flows tends to be homogeneous on road segments, where vehicles can change lanes flexibly within security restrictions. However, when vehicles enter intersections, they display complex turning patterns [12], including left-turn, right-turn, and straight-ahead, resulting in diverse travel times. Regarding temporal state, as shown in Fig. 2, turns typically exhibit slower speeds but demonstrate more pronounced fluctuations, primarily due to traffic control measures (e.g., traffic signals, crosswalks) at intersections. Therefore, how to effectively distinguish the spatio-temporal heterogeneity between road segments and turns while exploring their interrelationships remains a critical challenge for integrated traffic forecasting.

To overcome these challenges, we propose a Spatio-Temporal Heterogeneous Graph Transformer (STHGFormer) to comprehensively forecast road segments and intersection turns. To the best of our knowledge, the proposed STHGFormer pioneers integrated traffic forecasting for different components of traffic networks. For spatial correlations, SpaFormer (i.e., the spatial module of STHGFormer) simultaneously captures the dependencies between roads and turns within a complete road network. For temporal correlations, considering the heterogeneity of different forecasting elements, we employ two TempFormers to learn the dynamic features of roads and turns, respectively. The main contributions of this paper are summarized as follows.

(1) We innovatively define a Heterogeneous Road network Graph (HRG) to comprehensively represent the topological structure of the complete traffic network. Considering the spatio-temporal heterogeneity, HRG incorporates different types of nodes and edges to depict the characteristics of road segments and intersection turns, as well as their synergistic relationships.
(2) We develop a Heterogeneous Spatial Embedding (HSE) module to characterize the heterogeneous road network information from three dimensions: attributes, significance, and relevance. Leveraging the spatial information encoded by the HSE, SpaFormer effectively explores the intricate interdependencies between road segments and turns.
(3) We propose an Adaptive Soft Threshold (AST) module to alleviate the influence of high temporal fluctuation. Integrated with the AST, the proposed TempFormer enhances its capacity to capture intricate temporal correlations in the presence of noise and ensure more accurate and robust forecasting results.

Experiments on a real-world traffic speed dataset in Wuhan, China, demonstrate that the proposed STHGFormer outperforms other baseline models with an improvement of 6.1 % in road forecasting and an improvement of 8.1 % in turn forecasting for ten-minute traffic predictions.

The article is structured as follows. Section 2 outlines the relevant traffic forecasting studies. Section 3 describes the construction method of the innovative HRG and the proposed STHGFormer in detail. Section 4 presents experiments conducted to validate the proposed model. Finally, Section 5 provides the conclusion and future works.

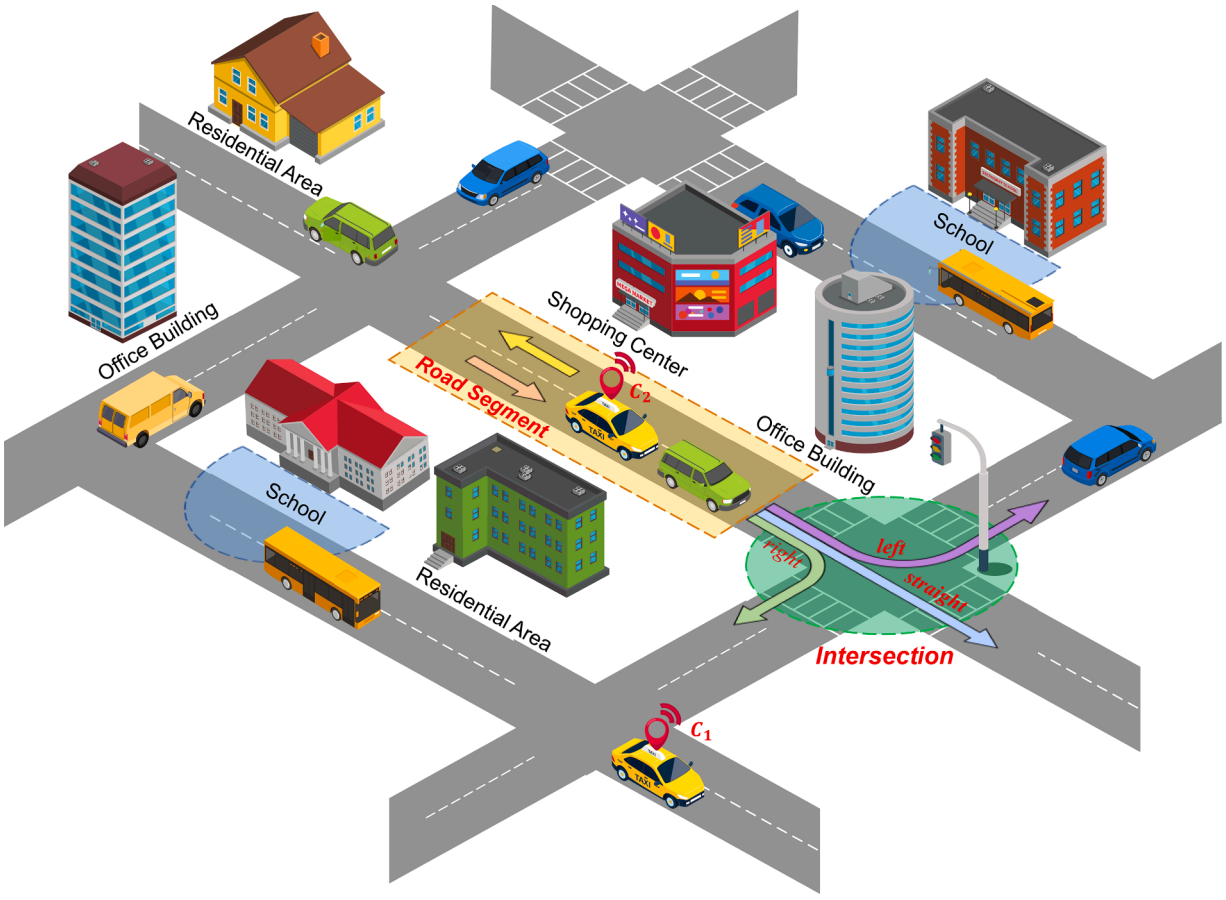


Fig. 1. A real-world traffic system. Floating cars equipped with position sensors can be considered as mobile traffic detectors (e.g., C_1 and C_2). The traffic network consists of road segments (depicted in yellow) and intersections (depicted in green). At intersections, traffic flows exhibit complex turning patterns, including left-turn, right-turn, and straight-ahead. The blue semicircles represent areas with comparable traffic patterns. Similar points of interest (POIs) and citizen traveling behaviors lead to similar traffic patterns. Even two non-adjacent areas can exhibit interrelated traffic states due to their analogous traffic patterns.

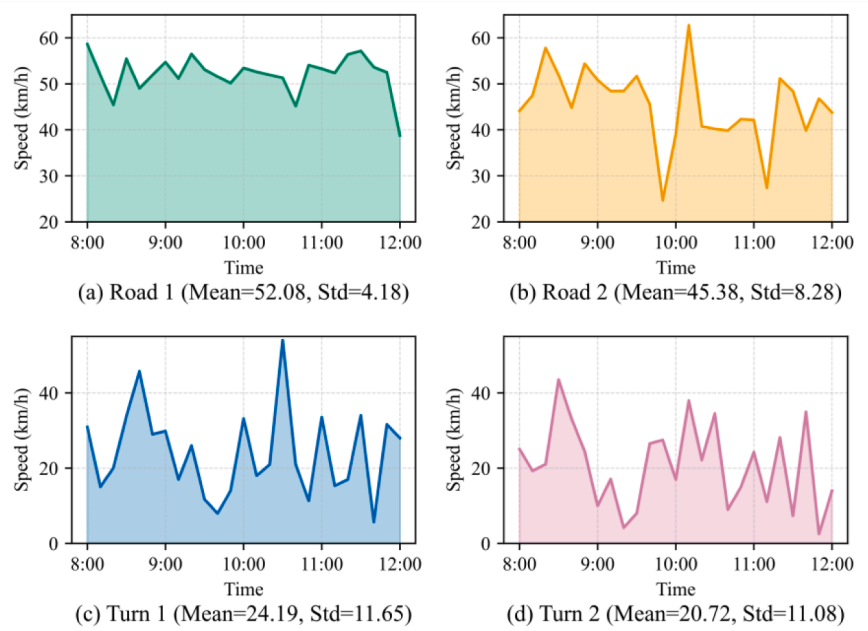


Fig. 2. Traffic states of road segments and intersection turns at the same intersection in a real-world dataset. (a) and (b) depict roads with relatively steady traffic states, characterized by high mean and low variance. Conversely, (c) and (d) illustrate turns with highly variable traffic conditions,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

2. Literature review

2.1. Traffic forecasting target

Most forecasting studies rely on data collected from fixed sensors, such as PEMS [13] and METR-LA [14] 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 [11]. Trajectory data, due to its cost-effectiveness, wide-coverage, and extensive-range, has garnered considerable attention from researchers [12]. For instance, some studies [8,15] 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 provide greater coverage. However, the use of grids disrupts the natural structure of traffic networks [16].

Currently, fine-grained traffic forecasting attracts more research interest. Guo et al. [9] and Zhang et al. [10] have effectively 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. [11] focus their forecasting target on intersections, as shown in Fig. 3(d). By constructing a dual graph and extracting traffic states for different turns (including left-turn, right-turn, and straight ahead), they accomplish turn-level traffic forecasting. Although these studies have refined the forecasting scale to some, they are limited to specific parts of the traffic network and have not achieved integrated forecasting for roads and turns simultaneously.

2.2. Traffic forecasting model

Graph neural networks (GNNs) [17] can effectively capture and comprehend intricate correlations within dynamic interacting systems [18]. As a result, they have found extensive application in various domains, including recommendation systems [19–22], knowledge graphs [23], and mobility prediction [24]. In terms of traffic prediction, GNNs have been synergistically combined with temporal neural networks, including recurrent neural networks (RNNs) [25], temporal convolutional networks (TCNs) [26], and others. For example, Wu et al. [26] introduced an adaptive adjacency matrix to reveal the dynamic associations within road networks and utilized GNN to capture the hidden spatial dependencies. Guo et al. [15] constructed traffic network graphs at both road and regional levels and employed a model consisting of GNN and TCN to capture multi-level spatio-temporal relationships.

However, as pointed out by Zhao et al. [27], the performance of GNNs is still far from satisfactory since their limited representation capacity facing complex traffic networks. Moreover, using GNNs to capture spatial correlations heavily relies on topological connections of

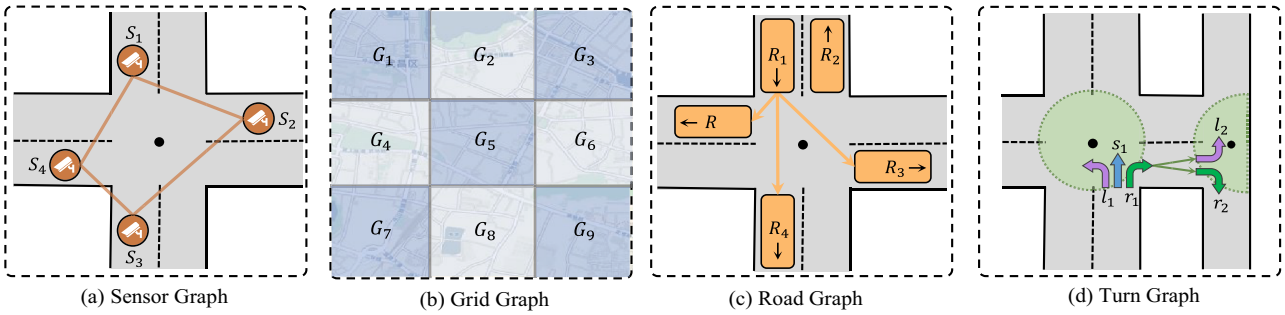


Fig. 3. Different kinds of traffic road network graph.

roads, making them unsuitable for detecting dynamic pattern correlations that are independent of road networks [28].

In recent years, the attention mechanism [29] has emerged as a popular tool in deep learning research. It enables models to dynamically analyze the correlation between the target sequences and adaptively focus on the desired content. In traffic forecasting, the attention mechanism has been widely applied due to its versatility and advantages. For instance, Liao et al. [30] used the attention mechanism to handle long sequences and improve the accuracy of results for long-term forecasting tasks. Regarding potential spatial dependencies, Zheng et al. [31] proposed a multi-headed spatial attention mechanism that can adaptively capture multiple types of spatial correlations.

Meanwhile, researchers have been exploring the use of the transformer and its variants for graph modeling. Graph, as a unique non-Euclidean data structure, cannot be processed directly by the attention mechanism, unlike other data structures such as images and texts [32]. Therefore, many works have employed methods such as positional embedding and attention matrix improvement [33] to incorporate graph information into the vanilla transformer. These studies have demonstrated that successful graph transformers, such as Graphormer [32] and GraphiT [34], can even outperform GNNs.

Nevertheless, existing graph transformers primarily focus on homogeneous graphs [33], where node and edge types are treated as the same. To our knowledge, no methods have clearly illustrated how to extract information from heterogeneous graphs and incorporate it into transformers.

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. Beside, the notations used in this paper are shown in Table 1 for clarity.

- **Fine-grained speed extraction.** In terms of traffic speed extraction, we filter out trajectory points with severe deviation and match the remaining points with roads, following the method of Tang et al. [35]. Then, referring to Fang et al. [11], 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).

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

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

- **Problem formulation.** For traffic forecasting, we define $x_t^{rd} = [s_t^{n_1^{rd}}, s_t^{n_2^{rd}}, \ldots] \in R^{N^{rd}}$ and $x_t^{tr} = [s_t^{n_1^{tr}}, s_t^{n_2^{tr}}, \ldots] \in R^{N^{tr}}$ as the traffic states of all road segments and intersection turns at time t, respectively. Further, we use $X_{t-T:t}^{rd} = [x_{t-T+1}^{rd}, x_{t-T+2}^{rd}, \ldots, x_t^{rd}]$ and $X_{t-T:t}^{tr} = [x_{t-T+1}^{tr}, x_{t-T+2}^{tr}, \ldots, x_t^{tr}]$ to represent traffic speeds of them during past T time slots before time t. 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 as Eq. (2).

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

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 traffic speeds during this period, and G is the road network.

3.2. Heterogeneous road network graph

To provide a comprehensive depiction of the traffic network, we define a heterogeneous road network graph (HRG), which represents road segments, intersection turns, and their synergistic relationships integrally.

Definition 1. Heterogeneous road network graph (HRG).

A heterogeneous graph is a graph with multiple types of nodes and edges [36], 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 elements of the traffic network to be forecasted, i.e., roads and turns. Edges $E \in R^{N \times N}$ represent various relationships between them, i.e., different modes of traffic flows when entering and exiting intersections.

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

Definition 2. Heterogeneous nodes. In the HRG, we employ heterogeneous nodes to represent roads and turns, respectively. Specifically, for bi-directional roads, we divide each of them 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); we directly use a single road node v_k^{rd} to represent a one-way road (e.g., R_5). For intersection turns, 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 perform different driving behaviors when entering and exiting the intersection, which leads to the heterogeneity of edges. Due to the traffic control measures, vehicles tend to slow down while they are approaching the intersection, whereas accelerate when they are exiting. Thus, we utilize different types of edges, i.e., heterogeneous edges, to represent such relationships

Table 1
Summary of notations and descriptions.

Notations	Descriptions	Notations	Descriptions
rd	Road segment	tr	Intersection turn
T	The length of observations	P	The length of predictions
X	Traffic state tensor	$HRG = (V, E)$	Heterogeneous road network graph
$A = (a_{ij})_{(N \times N)}$	The adjacency matrix of HRG	H	Hidden state in STHGFormer
Att	Embedding of attribute	Sig	Embedding of significance
Rel	Embedding of relevancy	l	Layer of ST-block
H_l^s	Outputs of l-th SpaFomer	ϵ	The identification of road or turn
$H_{l,ast}^{\epsilon}$	Outputs of l-th AST	$H_{l,t}^{\epsilon}$	Outputs of l-th TempFormer

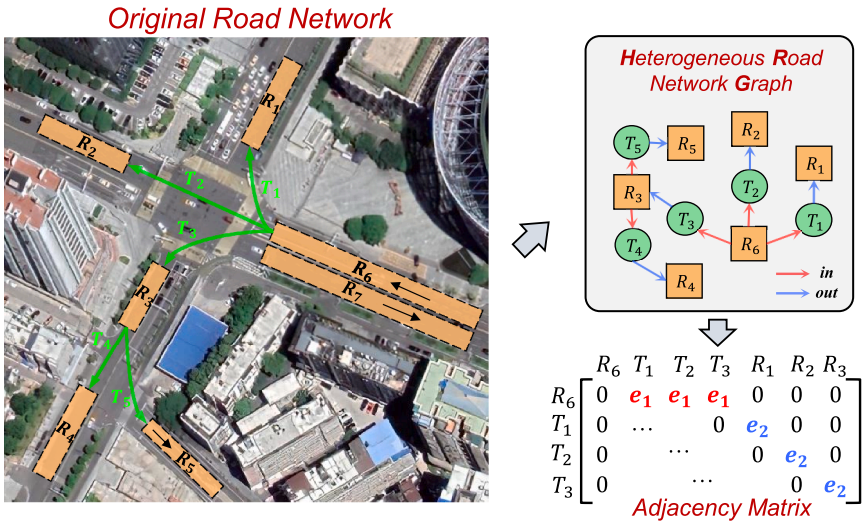


Fig. 4. Heterogeneous road network graph. In the original road network, the yellow rectangles denote unidirectional roads, while the green arrows represent intersection turns. In the heterogeneous road network graph, yellow nodes represent roads, and green nodes represent turns. The different-colored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

between roads and turns. Specifically, if a vehicle can reach the road node v_j^{rd} from v_i^{rd} through a turn node v_p^{tr} at the intersection, 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 the road segment R_6 can arrive at R_1 by turning right through T_1. Then, in HRG, there are two 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:

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

where v_i, v_j are two types of 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 framework of the Spatio-Temporal Heterogeneous Graph Transformer (STHGFormer), as shown in Fig. 5. To integrally predict traffic states of different elements in the traffic network, STHGFormer utilizes historical traffic speeds of road segments ($X_{t-T:t}^{rd}$), 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 and uses H_1^{rd} and H_1^{tr} as inputs for the first layer of the Spatio-Temporal-block (ST-block), as indicated in Eq. (4).

The ST-block comprises two main components: Spatial Transformer (SpaFormer) for spatial dependencies, and Temporal Transformer (TempFormer) for temporal dependencies. Within a complete road network, SpaFormer captures spatial dependencies between road segments and intersection turns simultaneously. As the key part of SpaFormer, Heterogeneous Spatial Embedding (HSE) extracts road network information from HRG and embeds it into the vanilla transformer. Due to the heterogeneity from 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 spatio-temporal dependencies. Finally, H_{L+1}^{rd} and H_{L+1}^{tr}, outputs of the final ST-block, are transferred through the output layer, thus generating the ultimate

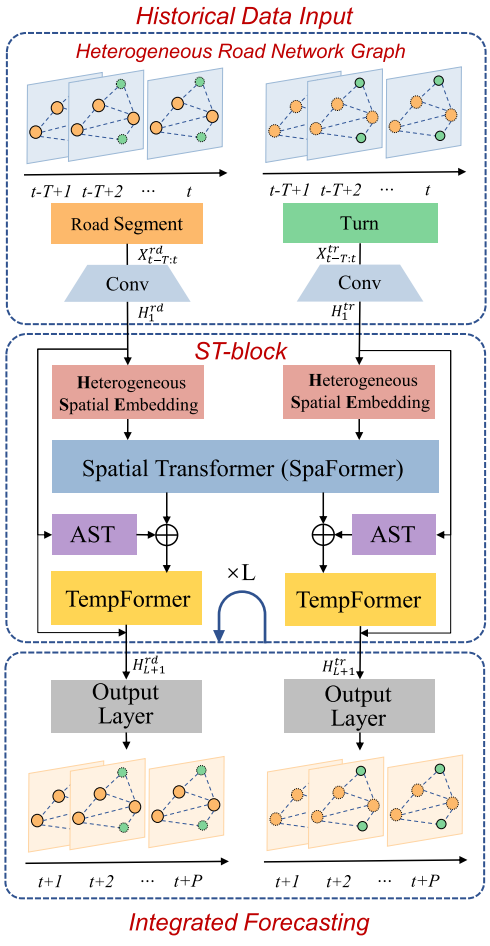


Fig. 5. Overview of STHGFormer. ST-block refers to Spatio-Temporal-block; AST refers to adaptive soft threshold; TempFormer refers to Temporal Transformer.

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 generate the final output.

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

$$X_{t:t+P}^{rd} = FC\left(Conv\left(H_{L+1}^{rd}\right)\right), X_{t:t+P}^{tr} = FC\left(Conv\left(H_{L+1}^{tr}\right)\right) \tag{5}$$

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. $Conv(H_{L+1}^{rd}) \in R^{N^{rd} \times T}$ and $Conv(H_{L+1}^{tr}) \in R^{N^{tr} \times T}$ are temporary outputs of the convolution layer when generating the final results.

3.4. Heterogeneous Spatial Embedding

To incorporate information from the heterogeneous road network graph into transformers, we propose the Heterogeneous Spatial Embedding (HSE) module. HSE extracts heterogeneous spatial information from the HRG and serves as a key component embedded into SpaFormer. The characteristics of heterogeneous nodes and edges are encoded from three perspectives: node attributes, significance, and relevancy within the traffic network.

(1) **Attribute.** Distinguishing different elements is a significant ability for models to improve their forecasting accuracy [37]. In the HRG, road segments and intersection turns are two areas with significant differences in their spatio-temporal attributes. Therefore, in this study, we categorize them by endowing them different attributes. Specifically, we convert the type identity of each element (i.e., road or turn) into the embedding matrix of type ($E(tp_V)$). Furthermore, forecasting elements of the same type can also exhibit heterogeneous traffic states. Thus, we transform the exclusive identification numbers of forecasting elements into the embedding matrix of ID ($E(id_V)$). By capturing the impact of attributes, the model can enhance its understanding of the traffic network's heterogeneity.

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

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; N is the sum of N^{rd} and N^{tr}; The function $E()$ is the learnable embedding layer; d is the hidden dimension of embedding matrices.

(2) **Significance.** Degree centrality provides a potential way for measuring the significance of forecasting elements by quantifying the number of edges connected to a node. Intuitively, intersections with high degree centralities have more complex traffic flow patterns and are more likely to occur traffic congestion [12]. Thus, we integrate degree centrality as a metric to assess the importance of forecasting elements. Specifically, we employ an embedding layer to transform the degree centrality of HRG into the embedding matrix of significance, as illustrated in Eq. (7). By encoding degree centrality, STHGFormer can identify which forecasting elements are more influential.

$$Sig = E\left(deg_V^-\right) + E\left(deg_V^+\right),\ Sig \in R^{N \times d} \tag{7}$$

where Sig denotes the embedding matrix of significance. The notations deg_V^- and deg_V^+ represent the in-degree and out-degree, respectively.

(3) **Relevancy.** Traffic flows exhibit distinct patterns when entering and exiting intersections. Furthermore, closely interconnected areas within the road network tend to display similar traffic states. To comprehensively represent the synergistic relationships between roads and turns, we introduce the definition of a heterogeneous road network graph (HRG), where nodes and edges possess multiple properties. The adjacency matrix of the HRG, denoted as A, encapsulates both the heterogeneity of edges and the first-order neighborhood information in road networks. Thus, we employ relevancy encoding for the adjacency matrix to emphasize edge heterogeneity and represent neighboring connections. Specifically, we encode the matrix A of the HRG using an embedding layer and a fully connected layer, as illustrated in Eq. (8). Subsequently, this encoded matrix is utilized as a mask during the calculation of the attention mechanism. The incorporation of relevancy encoding empowers STHGFormer to perceive the heterogeneous and interconnected associations present in complex road networks.

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

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 complex spatial correlations for accurate forecasting, especially when dealing with heterogeneous elements. For the dependencies between road segments and intersection turns, we employ a unified module called Spatial Transformer (SpaFomer) to capture the synergistic relationship between them and enhance the performance of integrated forecasting.

As shown in Fig. 6(a), 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 $H_l = concat(H_l^{rd}, H_l^{tr}) \in R^{N \times T \times d}$, where 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).

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

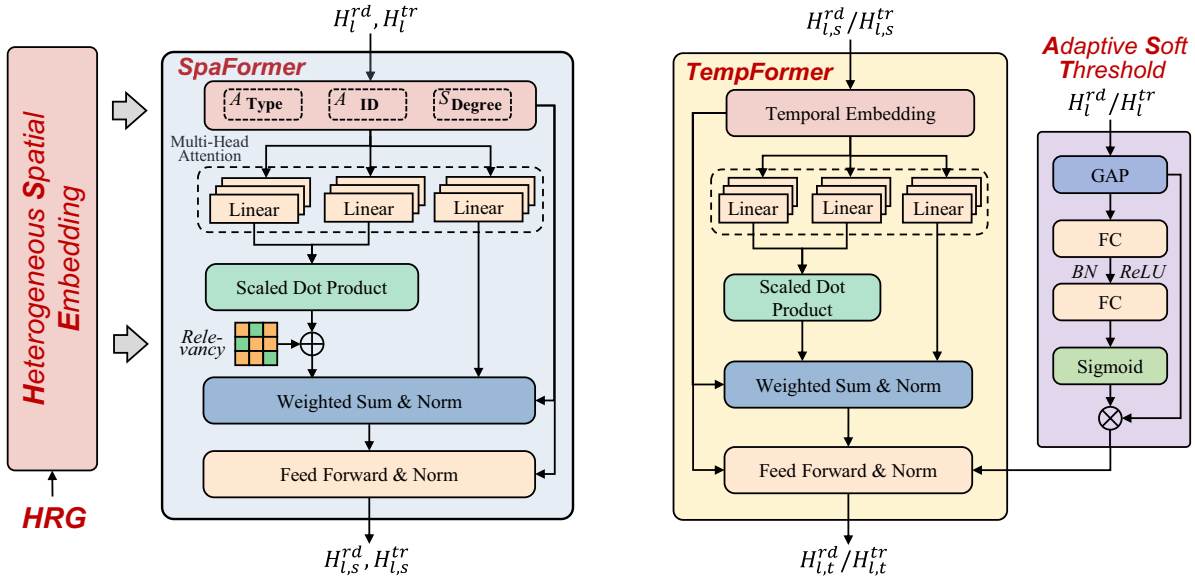
where $H_{l,s}^e$ is the hidden state of forecasting elements embedded with HSE.

Referring to the multi-head attention mechanism in the transformer [29], 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 denotes the index of attention head. $W_{l,Q_i}^s,\ W_{l,K_i}^s \in R^{d \times \dot{d}}$ 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.

To reveal correlations between each forecasting element, we construct 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



(a) SpaFormer (b) TempFormer

Fig. 6. Structure of SpaFomer and TempFormer.

heterogeneous topological relevance in the traffic network, we incorporate the output of HSE, denoted as *Rel*, as a mask into the attention weight matrix, as depicted in Eq. (11). The output of the i-th head is then obtained by updating $V^s_{l,i}$ with $A^s_{l,i}$. Subsequently, the results of each head are concatenated using a trainable linear network to generate the final output, as illustrated in Eq. (12).

$$A^s_{l,i} = Softmax\left(\frac{Q^s_{l,i}\left(K^s_{l,i}\right)^T}{\sqrt{d}}\right) + Rel,\ A^s_{l,i} \in R^{T\times N\times N} \tag{11}$$

$$MultiHeadS\left(H^e_{l,s}\right) = FC\left(concat\left(A^s_{l,1}V^s_{l,1}, A^s_{l,2}V^s_{l,2}, ..., A^s_{l,h}V^s_{l,h}\right)\right) \tag{12}$$

where the dot product between Q and K is represented as $Q^s_{l,i}(K^s_{l,i})^T$. $\sqrt{d}$ acts as the normalization constant. $MultiHeadS(H^e_{l,s}) \in R^{N\times T\times d}$ represents the output of multi-head attention.

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 the other nodes, instead of 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 forecasting elements 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) [29] 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 feed forward layer, with the hidden states of roads and turns being separated. In summary, the output process of SpaFomer can be described as follows.

$$H^s_l = LN\left(MultiHeadS\left(H^e_{l,s}\right) + H_l\right),\ H^s_l \in R^{N\times T\times d} \tag{13}$$

$$H^{rd}_{l,s}, H^{tr}_{l,s} = split\left(FC\left(H^s_l\right) + H^s_l\right) \tag{14}$$

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

3.6. Adaptive Soft Threshold and Temporal Transformer

Traffic states, which heavily depend on historical states, exhibit a non-linear variation over time [31]. Thus, we utilize a module called Temporal Transformer (TempFormer) to extract temporal correlations, as shown in Fig. 6(b). Considering heterogeneity between road segments and turns, we employ two separate TempFormers to handle each type independently. To mitigate the effect of highly fluctuating traffic states, we incorporate an additional module called Adaptive Soft Threshold (AST) into TempFormer.

3.6.1. Adaptive Soft Threshold

External factors, such as traffic signals and pedestrians, cause evident vibration and fluctuation in the traffic state, which can significantly impact the attention mechanism's learning ability. To alleviate the impact of high temporal fluctuation, we incorporate the Adaptive Soft Threshold (AST) into TempFormer [38].

By setting a threshold, we can filter out features with absolute values below it and compress features with absolute values exceeding it. Furthermore, the level of noise varies across different forecasting samples. For instance, during peak times, fluctuations in traffic speeds become more pronounced. To tackle this problem, AST dynamically defines thresholds based on the characteristics of forecasting elements.

In the AST module, the original inputs (H^{rd}_l, H^{tr}_l) of the l-th ST-block undergo global average pooling (GAP) and dimensionality reduction, resulting in the feature vector $gap = GAP(H^e_l) \in R^{N^e\times 1\times 1}$. Here, $e \in [rd,\ tr]$ represents the identification of roads and turns, and N^e denotes their respective sums. Next, H^e_l is fed into a sub-network with two fully connected layers. The sub-network produces a dynamic coefficient α, as shown in Eq. (15). Subsequently, the outputs of AST ($H^e_{l,ast}$) are generated by utilizing the $\alpha \times gap$ as the threshold, as illustrated in Eq. (16).

$$\alpha = Sigmoid\left(FC\left(H^e_l\right)\right),\ \alpha \in R^{N^e\times 1\times 1} \tag{15}$$

$$H^e_{l,ast} = \begin{cases} H^e_l - \alpha \times gap, H^e_l > \alpha \times gap \\ 0, H^e_l \le \alpha \times gap \end{cases} \tag{16}$$

where α is a learnable weight factor coefficient. FC represents the sub-network consisting of two fully connected layers. $H^{e}_{l,ast} \in R^{N^{e} \times T \times d}$ is the final output of AST.

3.6.2. Temporal Transformer

Unlike RNNs, the attention mechanism processes time series in parallel, which results in the loss of sequence order information. Therefore, we use positional encoding (PE) [29] to transform the sequence of temporal series into vectors. In addition, traffic states vary with the specific time of collection. For example, congestion is more likely to occur during the morning and evening rush hours, resulting in slower traffic speeds. To this end, 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).

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

where $H^{e}_{l,s} \in R^{N \times T \times d}$ 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, and captures 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 of TempFormer can be described as follows.

$$Q^{t}_{l,i} = H^{e}_{l,t} W^{t}_{l,Q_i},\ K^{t}_{l,i} = H^{e}_{l,t} W^{t}_{l,K_i},\ V^{t}_{l,i} = H^{e}_{l,t} W^{t}_{l,V_i},\ where\ i \in [1,h] \tag{18}$$

$$A^{t}_{l,i} = Softmax\left(\frac{Q^{t}_{l,i}\left(K^{t}_{l,i}\right)^{T}}{\sqrt{d}}\right),\ A^{t}_{l,i} \in R^{N^{e} \times T \times T} \tag{19}$$

$$MultiHeadT\left(H^{e}_{l,t}\right) = FC\left(concat\left(A^{t}_{l,1}V^{t}_{l,1}, A^{t}_{l,2}V^{t}_{l,2}, ..., A^{t}_{l,h}V^{t}_{l,h}\right)\right) \tag{20}$$

where $Q^{t}_{l,i}, K^{t}_{l,i} \in R^{N^{e} \times T \times d}$ and $V^{t}_{l,i} \in R^{N^{e} \times T \times d}$ represent the Q, K, and V of the i-th head, respectively. $A^{t}_{l,i}$ is the attention weight, and $MultiHeadT(H^{e}_{l,t}) \in R^{N^{e} \times T \times d}$ 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 and produce comprehensive results.

$$H^{t}_{l} = LN\left(MultiHeadT\left(H^{e}_{l,t}\right) + H^{e}_{l,ast}\right),\ H^{t}_{l} \in R^{N^{e} \times T \times d} \tag{21}$$

$$H^{e}_{l,t} = FC\left(H^{t}_{l}\right) + H^{e}_{l,s},\ H^{e}_{l,t} \in R^{N^{e} \times T \times d} \tag{22}$$

where $H^{e}_{l,t}$ represents hidden states of roads or turns through TempFormer, which are also outputs of the i-th ST-block.

4. Experiments

4.1. Dataset

Experiments are performed in Wuchang district, Wuhan, China. This area features a high volume of taxis and various POIs including hospitals, schools, and parks. The region of interest is illustrated in Fig. 7, spanning from longitude 114.295° E to 114.353° E and latitude 30.520° N to 30.562° N under the coordinate reference system EPSG:4326-WGS 84.

We construct the heterogeneous road network graph according to the method described in Section 3.2. It contains 277 road segments and 269 turns at 43 intersections. For each turning node, there exist an entering-type edge (e_1) and a leaving-type edge (e_2) connecting it to road nodes. Consequently, the constructed HRG encompasses a total of 538 heterogeneous edges. The trajectory data are collected from approximately 4000 taxis within the Wuchang district between July 1 and July 31, 2018. We extract and record the average speed for each road and turn at 10 min intervals. In total, we gathered 2437,344 speeds during the 31-day period, which are then divided into training and test sets, with 70 % and 30 % allocation, respectively.

4.2. Experimental settings

4.2.1. Evaluation metrics

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

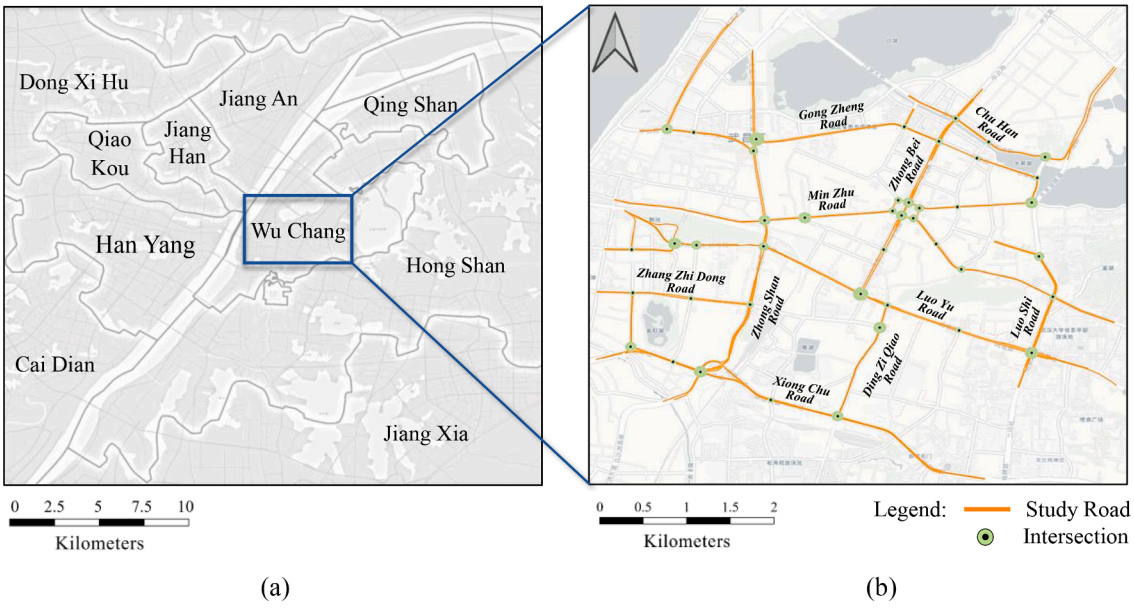


Fig. 7. Road network of the study area. (a) Wuchang District, Wuhan, China; (b) Road network.

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \widehat{y_p}|}{N} \tag{23}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \widehat{y_p})^2}{N}} \tag{24}$$

$$MAPE = \frac{100\%}{N} \left| \frac{\widehat{y_p} - y_i}{y_i} \right| \tag{25}$$

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

4.2.2. Hyperparameter setting

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 @ 3.70 GHz 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. In forecasting, the STHGFormer adopts observation length $T = 12$ and prediction length $P = 3$, enabling predictions of traffic speeds for the upcoming 10, 20, and 30 min based on traffic state data accumulated over the past two hours.

Given that our proposed STHGFormer involves multiple hyper-parameters, this section assesses how different hyper-parameter configurations influence the predictive performance of the model. We investigate the impact of the ST-block count, the number of attention heads, the hidden dimension (d), and the number of training epochs on the accuracy of traffic speed predictions for 10 min, as depicted in Fig. 8.

From Fig. 8(a), the accuracy of STHGFormer does not exhibit a gradual improvement as the number of ST-blocks increases. The results indicate a noticeable enhancement when the number of blocks is set to 2, reaching its optimal state at 3. Considering the computational burden during training, we apply two ST-blocks in subsequent experiments. In addition, Fig. 8(b) and (c) reveal that the model achieves its optimal predictive performance when the number of attention heads is set to 4 and the hidden dimension (d) is set to 32. As these two hyperparameters increase, the model's accuracy tends to decrease. In terms of training epoch, Fig. 8(d) shows that the model's accuracy experiences rapid improvement during the initial 200 training epochs, while its advancement becomes marginal beyond this epoch count. As a result, we set the training epoch to 200.

4.2.3. Baseline models

In this paper, we present comparative analyses between STHGFormer and eleven deep learning models, including:

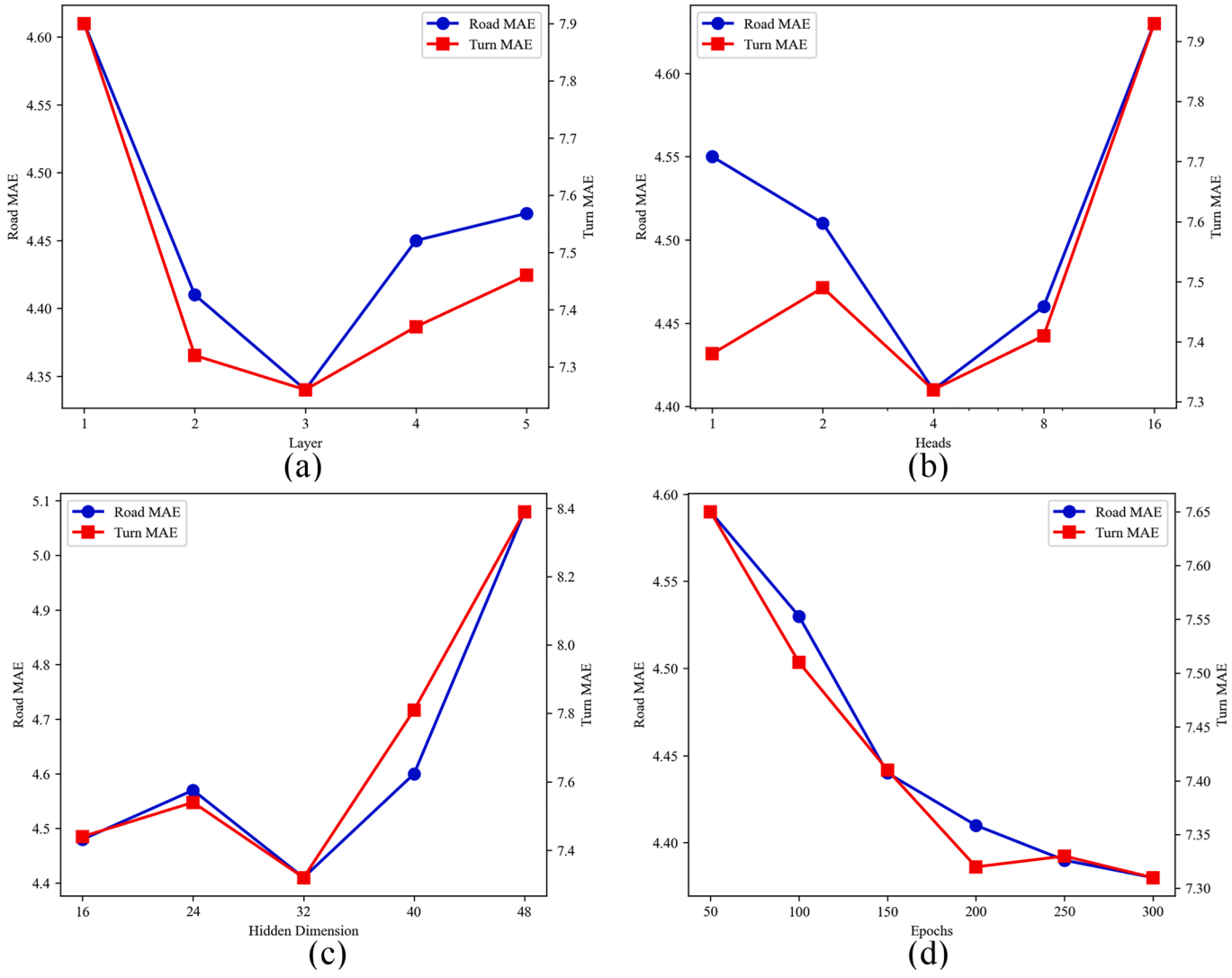


Fig. 8. Hyperparameter analysis. (a) Number of ST-blocks; (c) Number of heads; (b) Hidden dimension size; (d) Number of training epochs.

- **Long Short-Term Memory (LSTM)** [25]: 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)** [39]: A deep learning model designed for processing spatio-temporal data, utilizing convolutional operations to learn spatio-temporal dependencies.
- **Attention-based Spatial-Temporal Graph Convolutional Network (ASTGCN)** [40]: ASTGCN uses a multi-branch structure to model three different temporal properties of traffic flow, which integrates the attention mechanism with spatial-temporal convolution to enhance model performance.
- **Graph-WaveNet (GWNet)** [26]: A CNN-based model that mixes adaptive graphs with dilated convolutional operations to capture latent spatio-temporal connections.
- **Spatial-Temporal Transformer Network (STTN)** [41]: 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)** [9]: 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)** [15]: HGCN takes advantage of the innate hierarchical structure present in traffic systems, which operates on both micro and macro traffic graphs.
- **Dual Dynamic Spatial-Temporal Graph Convolution Network (DDSTGCN)** [42]: DDSTGCN captures the dynamic spatio-temporal feature of graph edges by transforming the data into a dual hypergraph.
- **Spatial-Temporal Identity Network (STID)** [37]: STID utilizes simple MLPs to solve the previously challenging indistinguishability issue, achieving superior performance. It is a simple, yet effective model for time series forecasting.
- **Principal Graph Embedding Convolutional Recurrent Network (PGECRN)** [43]: PGECRN integrates gated recurrent units (GRU) and the adjacency matrix graph embedding (AMGE) for traffic forecasting. The proposed AMGE can solve the data drift problem caused by road network structure changes.
- **Decomposition Dynamic Graph Convolutional Recurrent Network (DDGCRN)** [44]: DDGCRN fuses dynamic graph convolutional recurrent network with RNN-based dynamic graph model for time-varying traffic signals.

4.3. Comparison of forecasting performance

The performance of STHGFormer is evaluated by comparing it with baseline models. 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 pre-defined graph. In contrast to STHGFormer, all baseline models consider road segments and intersection turns as identical forecasting targets. Table 2 presents the forecast performances of all models for both roads and turns in 10 min (1 step), 20 min (2 steps), and 30 min (3 steps) tasks.

Predicting traffic speeds in intersection turns is more challenging due to the greater speed fluctuation compared to road segments. Thus, the MAE and RMSE for turns are larger than those for roads across all models. Please note that STHGFormer does not achieve the best result in terms of MAPE results for turns. We attribute this phenomenon to the unique characteristics inherent in intersection turns. Due to traffic control measures, traffic speeds near intersections always approach zero, which makes the denominator of MAPE converge to zero. In such scenarios, MAPE may not provide an accurate measure of error magnitude. Nonetheless, the proposed STHGFormer outperforms all baseline models across most of the evaluation metrics, which demonstrates its superiority for integrated and fine-grained traffic forecasting.

In comparison to LSTM that only captures temporal dependencies, models utilizing GNNs (e.g., STGCN and OGCRNN) can better capture the spatial features of road networks, thus performing better prediction results. In addition, dynamic spatio-temporal analysis models like GWNet, DDSTGCN and DDGCRN outperform static models that only consider fixed road network information (e.g., STGCN). Furthermore, STID achieves relatively decent results with its simple structure by effectively distinguishing the spatio-temporal heterogeneity of forecasting elements.

Experiment results show that 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

Table 2
Performance comparison of different models.

Target	Model	10min			20min			30min		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Road	LSTM	5.12	6.75	17.89	5.26	6.93	18.35	5.37	7.07	18.87
	STGCN	5.06	6.68	17.61	5.18	6.83	18.04	5.28	6.95	18.46
	ASTGCN	4.90	6.49	16.96	4.94	6.55	17.11	4.99	6.60	17.31
	OGCRNN	4.90	6.49	17.20	4.95	6.57	17.37	5.00	6.62	17.57
	GWNet	4.80	6.40	17.08	4.88	6.49	17.39	4.93	6.56	17.79
	HGCN	4.87	6.47	16.99	4.95	6.57	17.20	5.01	6.64	17.36
	STTN	4.86	6.42	17.59	4.87	6.44	17.85	4.94	6.54	17.97
	DDSTGCN	4.80	6.39	17.17	4.90	6.50	17.55	4.93	6.56	17.85
	STID	4.84	6.41	17.61	4.87	6.45	17.99	4.93	6.53	17.94
	PGECRN	4.85	6.45	17.62	4.94	6.56	18.04	5.03	6.64	18.08
	DDGCRN	4.70	6.29	16.83	4.71	6.32	16.96	4.73	6.34	16.93
	STHGFormer	**4.41**	**5.81**	**15.62**	**4.29**	**5.65**	**15.30**	**4.50**	**5.95**	**16.00**
Turn	LSTM	8.76	11.73	82.23	8.88	11.84	83.33	8.97	11.92	84.17
	STGCN	8.69	11.64	80.40	8.79	11.75	81.27	8.87	11.81	81.64
	ASTGCN	8.46	11.43	77.40	8.49	11.46	78.28	8.52	11.49	78.47
	OGCRNN	8.45	11.42	78.80	8.50	11.47	79.22	8.53	11.50	79.59
	GWNet	8.22	11.28	82.25	8.21	11.21	73.27	8.22	11.19	**72.77**
	HGCN	8.43	11.42	76.84	8.49	11.49	76.78	8.54	11.54	76.79
	STTN	8.37	11.21	85.31	8.36	11.20	89.41	8.42	11.28	88.06
	DDSTGCN	8.28	11.27	91.62	8.29	11.24	74.88	8.27	11.23	74.80
	STID	8.40	11.24	86.94	8.44	11.30	90.76	8.45	11.31	90.15
	PGECRN	8.44	11.45	80.40	8.53	11.56	75.19	8.59	11.59	79.73
	DDGCRN	8.00	11.00	**69.59**	7.93	10.92	**70.18**	8.04	11.07	79.21
	STHGFormer	**7.32**	**9.70**	69.92	**7.22**	**9.58**	70.82	**7.84**	**10.44**	75.89

1 step, STHGFormer achieves 6.1 % and 8.5 % improvements in MAE for road and turn forecasting, respectively. In the presence of forecasting targets with heterogeneity, STHGFormer highlights the significance of diverse spatio-temporal attributes and relationships, generating superior performance. Moreover, when confronted with fluctuating traffic states in turns, STHGFormer achieves even more substantial improvements in accuracy.

For a better presentation of the forecasting performance of STHGFormer, Fig. 9 compares the actual speeds with the forecasting results of two roads and two turns throughout the entire day of July 23.

It can be observed that turns exhibit more unpredictable fluctuations compared to roads. Even though, STHGFormer accurately predicts the overall trends in speeds for both roads and turns within a day. In STHGFormer, the proposed AST efficiently filters out excessive noise, enabling our model to accommodate speed variation in turns. Regardless of the speed of traffic or the degree of fluctuation, our model consistently maintains a relatively high level of accuracy.

4.4. Ablation study

To verify the validity of different modules, we construct four variations of STHGFormer by removing its main modules, respectively:

- **STHGFormer-SF:** STHGFormer that excludes the SpaFormer module.
- **STHGFormer-TF:** STHGFormer that excludes the TempFormer module.
- **STHGFormer-HSE:** STHGFormer that excludes the heterogeneous spatial embedding (HSE) module.
- **STHGFormer-AST:** STHGFormer that excludes the adaptive soft threshold (AST) module.

For more in-depth analyses 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 maintain the same settings as the STHGFormer, except for the components under investigation. Then, we assess the forecasting performance of each variation to demonstrate their respective influences, as shown in Table 3.

Table 3 demonstrates that all incomplete STHGFormer models exhibit varying levels of accuracy degradation, which indicates the significance of each module in capturing spatio-temporal correlations. Notably, STHGFormers that exclude the temporal modules (i.e., STHGFormer-TF and STHGFormer-AST) display greater accuracy degradation compared to those that exclude spatial modules (i.e., 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 significant declines in accuracy for turn speed forecasting. Specifically, for the 10 min (1-step) turn forecasting, the MAE of STHGFormer-AST decreased by 0.79 km/h compared to the full STHGFormer, emphasizing the importance of the soft threshold.

Table 3
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 removes the heterogeneous spatial embedding, shows a more significant reduction in accuracy compared to the STHGFormer-SF model which eliminates the SpaFormer module. We speculate that the lower accuracy of STHGFormer-HSE is due to the lack of heterogeneous graph information, which makes it difficult for the attention mechanism to properly reveal spatial dependencies within complex traffic networks. Additional ablation experiments indicate that significance encoding (– Sig) has a more significant impact on forecasting results within the HSE when compared to attribute (– Att) and relevancy encoding (– Rel).

The above experiments demonstrate that our proposed STHGFormer, which utilizes HRG and HSE to comprehensively represent traffic networks, can effectively explore spatial correlations and achieve accurate prediction results. The diverse nodes and edges within the HRG depict

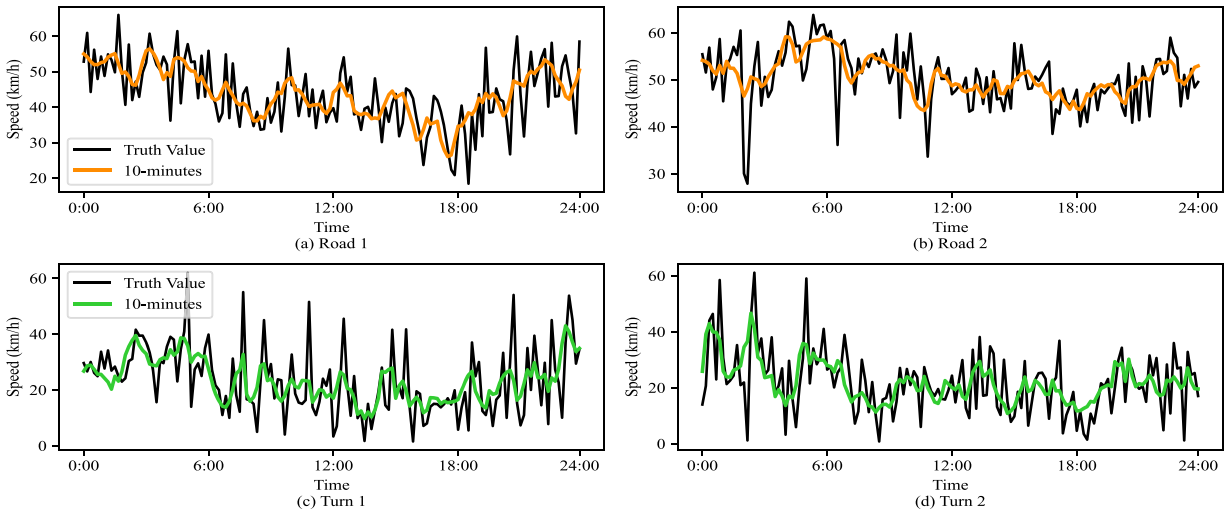


Fig. 9. Visualization of traffic forecasting results near the intersection of Zhong Bei Road and Chu Han Road.

the complex road network structure in detail, which is crucial for integrated traffic forecasting. Specifically, the incorporation of heterogeneity facilitates the distinct perception ability of the transformer for various types of elements. Furthermore, HRG realizes a complete representation of roads and turns, empowering SpaFormer to possess a receptive field that encompasses the entire road network. On the whole, the combined utilization of HRG and HSE allows STHGFormer to implement a comprehensive understanding of the road network, which promotes distinct perception and analysis of fine-grained forecasting elements.

4.5. Comparison of integrated and disintegrated STHGFormer

To demonstrate the advantages of STHGFormer in integrated traffic forecasting, we further compare the integrated model against two disintegrated models:

- **STHGFormer (R):** STHGFormer that solely forecasts road segments.
- **STHGFormer (T):** STHGFormer that solely forecasts intersection turns.

To ensure comparison consistency, we replace the HRG used in integrated forecasting with adjacent graphs when training STHGFormer (R) and STHGFormer (T). Specifically, for STHGFormer (R), we consider that connected road segments are linked. For STHGFormer (T), we consider turns within the same intersection are linked. We present the comparison of predictive accuracy by STHGFormer, STHGFormer (R), and STHGFormer (T) in Table 4.

The results in Table 4 indicate that integrated forecasting achieves higher accuracy for both road and turn forecasting across various time intervals. Note that prediction results of turns achieve higher improvement. This could be attributed to the fact that turns are located at intersections with complex topological relationships and are more closely related to other forecasting elements. As a result, the advantages of integrated forecasting are more pronounced for turns.

Fig. 10 illustrates the accuracy enhancement of roads and turns in two different prediction intervals (10 and 30 min). To visually demonstrate the accuracy improvement of turns, we compute the average improvement of turns at the same intersection and use intersection improvements to portray turn improvements.

As shown in Fig. 10, most roads and turns have better accuracy in different prediction intervals. This result indicates that integrated forecasting significantly enhances accuracy by capturing synergistic relationships between road segments and intersection turns.

In the 10 min forecast, as shown in the histograms of Fig. 10(a), approximately 95 % of the roads demonstrate better accuracy. Moreover, all turns exhibit significant improvements, with a general reduction in MAE of 0.4 km/h. Furthermore, regions characterized by intricate topological structures exhibit more pronounced enhancements. Fig. 10(a) illustrates that roads near intersections (highlighted in yellow) experience greater accuracy improvements compared to those situated farther away, such as Road Segments 1, 2, and 3. Intersection turns featuring complex structures consistently demonstrate noteworthy improvements.

In the 30 min forecast, as depicted in Fig 10(b), it can be observed that the accuracy of long-term prediction decreases across the entire traffic road network. As the time scale increases, spatial correlations between roads and turns become more intricate, presenting greater challenges for accurate forecasting. However, histograms in Fig. 10(b) illustrate that STHGFormer still achieves considerable improvements in accuracy for the majority of roads and turns.

Table 4
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 %

4.6. Time-varying forecasting accuracy

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 compute the mean absolute errors (MAE) of the 10 min prediction for each hour and illustrate the results of roads and turns in Fig. 11(a) and Fig. 11(b), respectively.

The results demonstrate that STHGFormer consistently outperforms the other models in terms of forecasting accuracy. Notably, STHGFormer shows the narrowest range of fluctuation compared to the alternative models for both road and turn forecasting. This finding highlights the exceptional capability of STHGFormer in effectively capturing intricate spatio-temporal dependencies. By capturing complete spatial correlations and suppressing temporal fluctuations, STHGFormer improves the accuracy and stability of traffic forecasting.

4.7. Complexity analysis

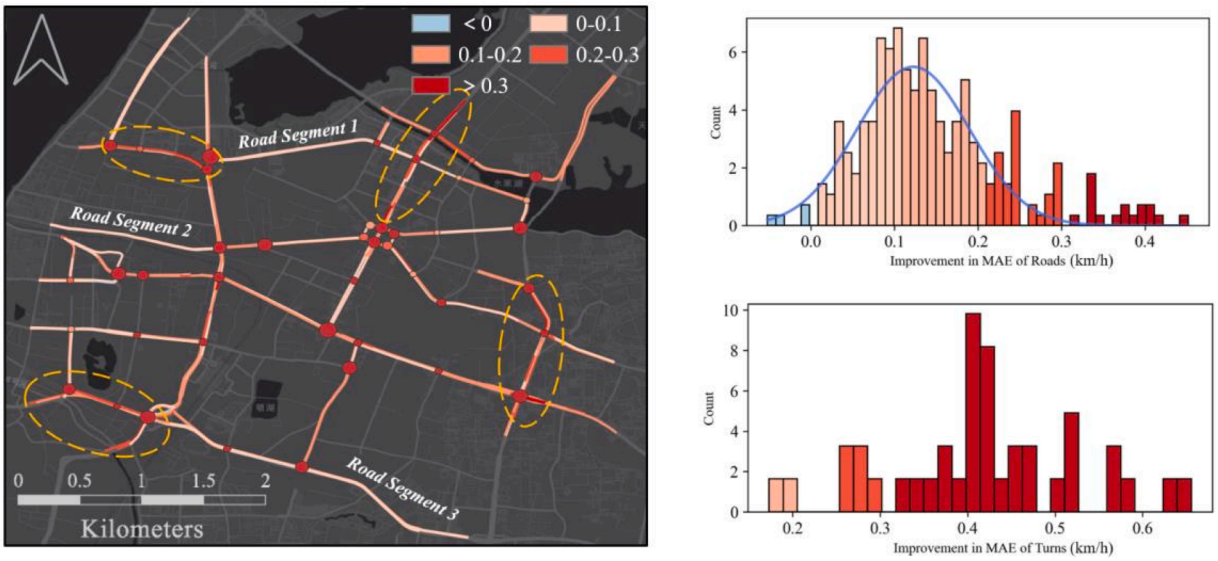
Efficient traffic forecasting models are essential for maintaining real-time performance and practicality within dynamic urban traffic systems. To assess the computational complexity of STHGFormer, we conduct a thorough comparison with four high-performing baseline models. Table 5 shows the parameters, training time and inference time of different models. Additionally, we also investigate the influence of key modules within STHGFormer on complexity.

STTN excels in parameter performance among the factors examined. However, this achievement comes at the cost of the longest training and inference times due to the intricate fusion of Transformer and GNN with gate mechanisms. The inherent computational demands of the Transformer architecture further contribute to this overhead. In contrast, GWNet stands out as the fastest baseline model, requiring only 13.03 s for training and 2.95 s for inference, owing to its streamlined spatio-temporal convolution operations. In terms of training time, DDSTGCN, DDGCRN, and STHGFormer exhibit similar speeds. While STHGFormer's inference speed is slightly slower compared to the other two baseline models, it remains suitable for real-time forecasting needs.

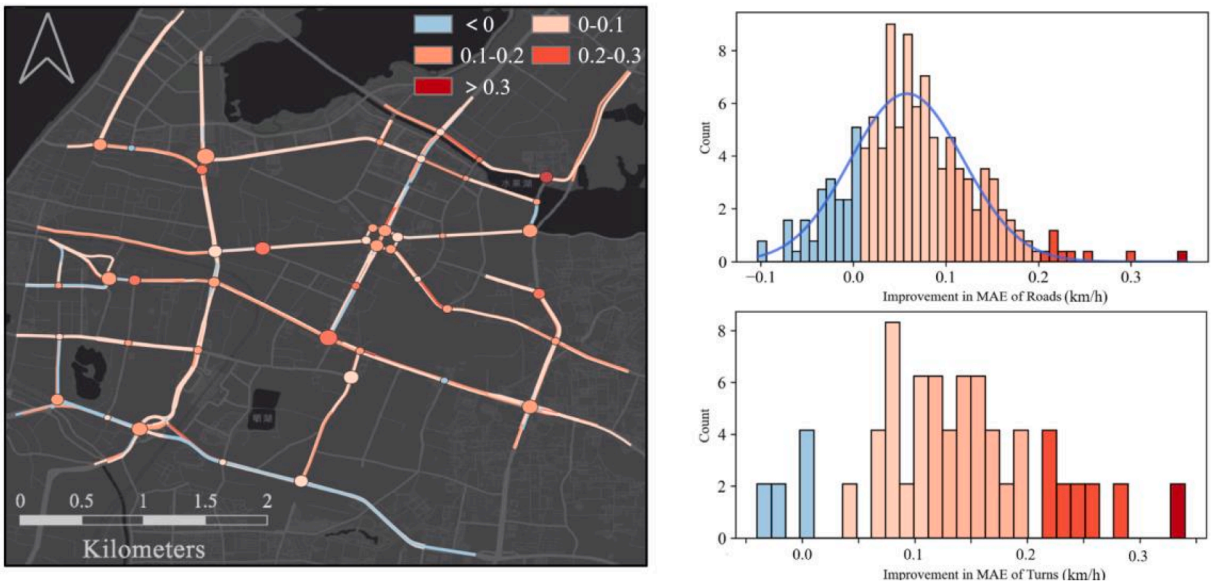
Among the variations of STHGFormer, excluding SpaFormer (STHGFormer-SF) results in the shortest training and inference times. Instead, removing the HSE module (STHGFormer-HSE) only marginally affects complexity. This implies that the attention mechanism responsible for spatial dependency incurs the highest computational cost. Integrated traffic forecasting encompasses intricate spatial interrelationships among numerous road segments and intersection turns, leading to increased computational demands and complexity. Additionally, the proposed AST module enhances prediction accuracy while maintaining quick training times, efficiently mitigating traffic state fluctuations.

5. Conclusion and future works

This paper develops a novel approach called Spatio-Temporal Heterogeneous Graph Transformer (STHGFormer) for integrated and fine-grained traffic forecasting. The proposed method addresses the challenges posed by the complex relationships between roads and turns and



(a) 10-minute accuracy improvement



(b) 30-minute accuracy improvement

Fig. 10. The distribution of MAE improvement of integrated STHGFormer against disintegrated STHGFormer. The histograms illustrate the number of roads or turns with different accuracy improvements.

their heterogeneous spatio-temporal attributes. Specifically, by constructing the Heterogeneous Road network Graph (HRG), we completely represent the topology and heterogeneity of the entire traffic network. Moreover, through the combination of the Heterogeneous Spatial Embedding (HSE) module and the multi-headed attention mechanism, the SpaFormer efficiently reveals the complex spatial correlations between roads and turns. To handle the highly fluctuating time series, an Adaptive Soft Threshold (AST) module is incorporated into the TempFormer, which leverages learnable thresholds to mitigate the impact of fluctuation.

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 modules within STHGFormer. In addition, the proposed method exhibited consistent and reliable forecasting stability across various times of the day. Furthermore, experiments assessing algorithm complexity have demonstrated STHGFormer's capability for real-time forecasting in complex urban traffic scenarios.

In the future, we aim to leverage the forecasting results of roads and turns to enhance the accuracy and granularity of route planning and travel time estimation. Beside, we will also attempt to figure out how to incorporate additional factors such as weather conditions, traffic accidents, and other variables into traffic forecasting. Finally, additional research is warranted to explore the potential of heterogeneous GNNs [36,45] in modeling the intricate spatial relationships within the road network.

CRediT authorship contribution statement

Guangyue Li: Conceptualization, Methodology, Software, Visualization, Writing – original draft. **Zilong Zhao:** Conceptualization,

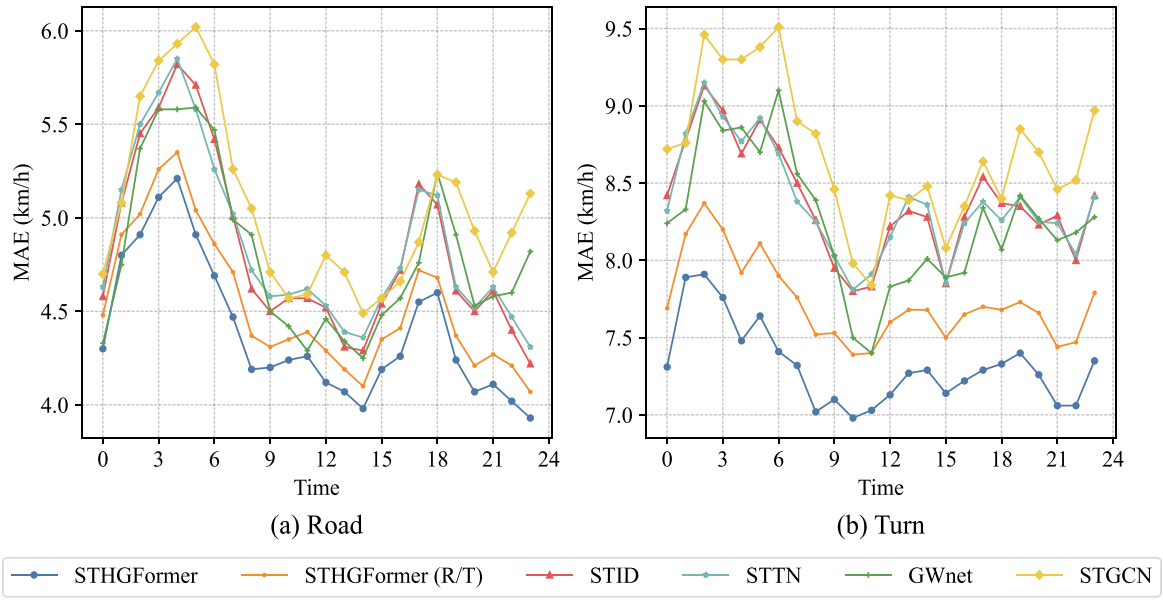


Fig. 11. The average MAE (km/h) of different models at different times.

Table 5
Complexity analysis. Bold font indicates the best results, while the second-best results are underlined.

Model	Total Parameters (M)	Training (s/epoch)	Inference (s)
GWNet	0.30	13.03	**2.95**
STTN	**0.07**	44.32	15.75
DDSTGCN	0.29	23.32	4.14
DDGCRN	0.57	18.56	4.74
STHGFormer	0.35	26.48	8.08
STHGFormer-SF	0.22	**8.26**	3.35
STHGFormer-TF	0.13	18.38	5.55
STHGFormer-HSE	0.25	21.59	5.19
STHGFormer-AST	0.16	25.92	6.16

Software, Formal analysis, Writing – review & editing. **Xiaogang Guo:** Methodology, Formal analysis, Writing – review & editing. **Luliang Tang:** Conceptualization, Writing – review & editing, Funding acquisition. **Huazu Zhang:** Resources, Software, Data curation. **Jinghan Wang:** Visualization, Formal analysis, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] F. Li, J. Feng, H. Yan, G. Jin, F. Yang, F. Sun, D. Jin, Y. Li, Dynamic graph convolutional recurrent network for traffic prediction: benchmark and solution, ACM Trans. Knowl. Discov. Data (2021).
[2] R. Chauhan, A. Dhamaniya, S. Arkatkar, Driving behavior at signalized intersections operating under disordered traffic conditions, Transp. Res. Rec. 2675 (2021) 1356–1378.
[3] T. Liebig, N. Piatkowski, C. Bockermann, K. Morik, Dynamic route planning with real-time traffic predictions, Inf. Syst. 64 (2017) 258–265.
[4] W. Zhang, F. Zhu, Y. Lv, C. Tan, W. Liu, X. Zhang, F.Y. Wang, AdapGL: an adaptive graph learning algorithm for traffic prediction based on spatiotemporal neural networks, Transp. Res. Part C Emerg. Technol. 139 (2022), 103659.
[5] H. Wei, G. Zheng, V. Gayah, Z. Li, A survey on traffic signal control methods, arXiv preprint (2019).
[6] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, B. Yin, Multi-stage attention spatial-temporal graph networks for traffic prediction, Neurocomputing 428 (2021) 42–53.
[7] Z. Zhao, L. Tang, M. Fang, X. Yang, C. Li, Q. Li, Toward urban traffic scenarios and more: a spatio-temporal analysis empowered low-rank tensor completion method for data imputation, Int. J. Geogr. Inf. Sci. (2023) 1–34.
[8] J. Zhang, Y. Zheng, D. Qi, Deep spatio-temporal residual networks for citywide crowd flows prediction, in: Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence, 2017.
[9] K. Guo, Y. Hu, Z. Qian, H. Liu, K. Zhang, Y. Sun, J. Gao, B. Yin, Optimized graph convolution recurrent neural network for traffic prediction, IEEE Trans. Intell. Transp. Syst. 22 (2020) 1138–1149.
[10] Y. Zhang, T. Cheng, Y. Ren, A graph deep learning method for short-term traffic forecasting on large road networks, Comput. Aided Civ. Infrastruct. Eng. 34 (2019) 877–896.
[11] M. Fang, L. Tang, X. Yang, Y. Chen, C. Li, Q. Li, FTPG: a fine-grained traffic prediction method with graph attention network using big trace data, IEEE Trans. Intell. Transp. Syst. (2021).
[12] Z. Kan, L. Tang, M.P. Kwan, C. Ren, D. Liu, Q. Li, Traffic congestion analysis at the turn level using Taxis' GPS trajectory data, Comput. Environ. Urban Syst. 74 (2019) 229–243.
[13] C. Chen, K. Petty, A. Skabardonis, P. Varaiya, Z. Jia, Freeway performance measurement system: mining loop detector data, Transp. Res. Rec. 1748 (2001) 96–102.
[14] Y. Li, R. Yu, C. Shahabi, Y. Liu, Diffusion convolutional recurrent neural network: data-driven traffic forecasting, arXiv preprint (2017).
[15] K. Guo, Y. Hu, Y. Sun, S. Qian, J. Gao, B. Yin, Hierarchical Graph convolution network for traffic forecasting, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 151–159.
[16] J. Ye, J. Zhao, K. Ye, C. Xu, How to build a graph-based deep learning architecture in traffic domain: a survey, IEEE Trans. Intell. Transp. Syst. 23 (2022) 3904–3924.
[17] W. Ju, Z. Fang, Y. Gu, Z. Liu, Q. Long, Z. Qiao, Y. Qin, J. Shen, F. Sun, Z. Xiao, A comprehensive survey on deep graph representation learning, arXiv preprint (2023).
[18] X. Luo, J. Yuan, Z. Huang, H. Jiang, Y. Qin, W. Ju, M. Zhang, Y. Sun, HOPE: high-order graph ODE for modeling interacting dynamics, K. Andreas, B. Emma, C. Kyunghyun, E. Barbara, S. Sivan, S. Jonathan (Eds.), in: Proceedings of the 40th International Conference on Machine Learning, PMLR, 2023, pp. 23124–23139. Proceedings of Machine Learning Research.
[19] Y. Qin, W. Ju, H. Wu, X. Luo, M. Zhang, Learning graph ODE for continuous-time sequential recommendation, arXiv preprint, (2023).
[20] Y. Wang, Y. Li, S. Li, W. Song, J. Fan, S. Gao, L. Ma, B. Cheng, X. Cai, S. Wang, Deep graph mutual learning for cross-domain recommendation, in: Proceedings of the International Conference on Database Systems for Advanced Applications, Springer, 2022, pp. 298–305.
[21] Y. Wang, Y. Qin, F. Sun, B. Zhang, X. Hou, K. Hu, J. Cheng, J. Lei, M. Zhang, DisenCTR: dynamic graph-based disentangled representation for click-through rate prediction, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2314–2318.
[22] Y. Wang, J. Shen, Y. Song, S. Wang, M. Zhang, HE-SNE: heterogeneous event sequence-based streaming network embedding for dynamic behaviors, in: Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), IEEE, 2022, pp. 1–8.

[23] Y. Wang, Y. Song, S. Li, C. Cheng, W. Ju, M. Zhang, S. Wang, DisenCite: graph-based disentangled representation learning for context-specific citation generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 11449–11458.
[24] W. Ju, Y. Qin, Z. Qiao, X. Luo, Y. Wang, Y. Fu, M. Zhang, Kernel-based substructure exploration for next POI recommendation, in: Proceedings of the 2022 IEEE International Conference on Data Mining (ICDM), IEEE, 2022, pp. 221–230.
[25] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: LSTM cells and network architectures, Neural Comput. 31 (2019) 1235–1270.
[26] Z. Wu, S. Pan, G. Long, J. Jiang, C.J.A.P.A. Zhang, Graph wavenet for deep spatial-temporal graph modeling, arXiv preprint, (2019).
[27] Y. Zhao, X. Luo, W. Ju, C. Chen, X.S. Hua, M. Zhang, Dynamic hypergraph structure learning for traffic flow forecasting, in: Proceedings of the International Conference On Data Engineering (ICDE), 2023, pp. 2303–2316.
[28] J. Zhao, C. Chen, C. Liao, H. Huang, J. Ma, H. Pu, J. Luo, T. Zhu, S. Wang, 2F-TP: learning Flexible spatiotemporal dependency for flexible traffic prediction, IEEE Trans. Intell. Transp. Syst. (2022).
[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[30] L. Liao, Z. Hu, Y. Zheng, S. Bi, F. Zou, H. Qiu, M. Zhang, An improved dynamic Chebyshev graph convolution network for traffic flow prediction with spatial-temporal attention, Appl. Intell. (2022) 1–13.
[31] C. Zheng, X. Fan, C. Wang, J. Qi, Gman: a graph multi-attention network for traffic prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 1234–1241.
[32] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, T.Y. Liu, Do transformers really perform badly for graph representation? Adv. Neural Inf. Process. Syst. 34 (2021) 28877–28888.
[33] E. Min, R. Chen, Y. Bian, T. Xu, K. Zhao, W. Huang, P. Zhao, J. Huang, S. Ananiadou, Y. Rong, Transformer for graphs: an overview from architecture perspective, arXiv preprint, (2022).
[34] G. Mialon, D. Chen, M. Selosse, J. Mairal, Graphit: encoding graph structure in transformers, arXiv preprint, (2021).
[35] L. Tang, C. Ren, Z. Liu, Q. Li, A road map refinement method using delaunay triangulation for big trace data, ISPRS Int. J. Geo Inf. 6 (2017) 45.
[36] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, P.S. Yu, Heterogeneous graph attention network, in: Proceedings of the World Wide Web Conference, 2019, pp. 2022–2032.
[37] Z. Shao, Z. Zhang, F. Wang, W. Wei, Y. Xu, Spatial-temporal identity: a simple yet effective baseline for multivariate time series forecasting, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 4454–4458.
[38] M. Zhao, S. Zhong, X. Fu, B. Tang, M. Pecht, Deep residual shrinkage networks for fault diagnosis, IEEE Trans. Ind. Inf. 16 (2019) 4681–4690.
[39] B. Yu, H. Yin, Z. Zhu, Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting, arXiv preprint, (2017).
[40] S. Guo, Y. Lin, N. Feng, C. Song, H. Wan, Attention based spatial-temporal graph convolutional networks for traffic flow forecasting, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 922–929.
[41] M. Xu, W. Dai, C. Liu, X. Gao, W. Lin, G.J. Qi, H. Xiong, Spatial-temporal transformer networks for traffic flow forecasting, arXiv preprint (2020).
[42] Y. Sun, X. Jiang, Y. Hu, F. Duan, K. Guo, B. Wang, J. Gao, B. Yin, Dual dynamic spatial-temporal graph convolution network for traffic prediction, IEEE Trans. Intell. Transp. Syst. 23 (2022) 23680–23693.
[43] Y. Han, S. Zhao, H. Deng, W. Jia, Principal graph embedding convolutional recurrent network for traffic flow prediction, Appl. Intell. (2023) 1–15.
[44] W. Weng, J. Fan, H. Wu, Y. Hu, H. Tian, F. Zhu, J. Wu, A decomposition dynamic graph convolutional recurrent network for traffic forecasting, Pattern Recognit. 142 (2023), 109670.
[45] Y. Wang, S. Tang, Y. Lei, W. Song, S. Wang, M. Zhang, Disenhan: disentangled heterogeneous graph attention network for recommendation, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 1605–1614.