Dual-branch Spatial-Temporal Self-supervised Representation for Enhanced Road Network Learning

Qinghong Guo[1], Yu Wang[1], Ji Cao[1], Tongya Zheng[1,2,3*], Junshu Dai[1], Bingde Hu[1,4], Shunyu Liu[5], Canghong Jin[2]

[1]State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2]Zhejiang Provincial Engineering Research Center for Real-Time SmartTech in Urban Security Governance, School of Computer and Computing Science, Hangzhou City University
[3]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[4]Bangsun Technology
[5]Nanyang Technological Univerisity
{q.h_guo, yu.wang, caoj25, djs, tonyhu}@zju.edu.cn,
doujiang_zheng@163.com, shunyu.liu@ntu.edu.sg, jinch@hzcu.edu.cn

Abstract

Road network representation learning (RNRL) has attracted increasing attention from both researchers and practitioners as various spatiotemporal tasks are emerging. Recent advanced methods leverage Graph Neural Networks (GNNs) and contrastive learning to characterize the spatial structure of road segments in a self-supervised paradigm. However, spatial heterogeneity and temporal dynamics of road networks raise severe challenges to the neighborhood smoothing mechanism of self-supervised GNNs. To address these issues, we propose a **D**ual-branch **S**patial-**T**emporal self-supervised representation framework for enhanced road representations, termed as DST. On one hand, DST designs a mix-hop transition matrix for graph convolution to incorporate dynamic relations of roads from trajectories. Besides, DST contrasts road representations of the vanilla road network against that of the hypergraph in a spatial self-supervised way. The hypergraph is newly built based on three types of hyperedges to capture long-range relations. On the other hand, DST performs next token prediction as the temporal self-supervised task on the sequences of traffic dynamics based on a causal Transformer, which is further regularized by differentiating traffic modes of weekdays from those of weekends. Extensive experiments against state-of-the-art methods verify the superiority of our proposed framework. Moreover, the comprehensive spatiotemporal modeling facilitates DST to excel in zero-shot learning scenarios.

Code — https://github.com/chaser-gua/DST
Extended version — https://arxiv.org/abs/2511.06633

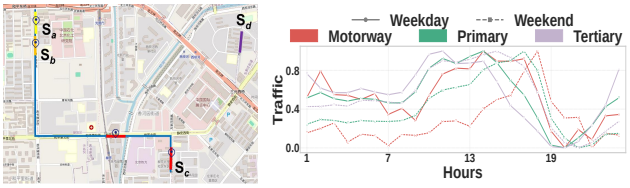
Introduction

The road network is a vital infrastructure in a city, reflecting connectivity and accessibility between road segments, which guide human mobility. With the development of smart traffic systems (Azgomi and Jamshidi 2018; Camero and Alba 2019; Wang et al. 2024c,d), the road network has become a core component supporting various smart city applications, such as traffic inference (Yang et al. 2022; Zou et al. 2023) and destination prediction (Xue et al. 2013; Zong et al. 2019). Road network representation learning (RNRL) aims to learn universal and low-dimensional vector representations for road networks to enhance performance on downstream tasks (Zhou et al. 2024). Consequently, RNRL has gained significant attention from researchers and is emerging as a powerful tool for spatial-temporal management.

Early efforts leverage the natural graph topology of the road network to learn task-specific representations based on Graph Neural Networks (GNNs) (Jepsen et al. 2019; Li et al. 2020; Gharaee et al. 2021). While these methods demonstrate robust performance on their original tasks, their efficacy often diminishes when transferred to different downstream tasks. Recent research on RNRL methods is inspired by powerful self-supervised learning paradigms to explore various geospatial relationships between roads through reconstruction and contrastive learning tasks (Wu et al. 2020; Chen et al. 2021). Additionally, abundant trajectory data are utilized to characterize the dynamic relationships between roads, augmenting road representations with auxiliary information (Mao et al. 2022; Schestakov et al. 2023). Nonetheless, the neighborhood smoothing mechanism inherent in GNNs leads to suboptimal road representations due to the spatial-temporal dynamics of roads, which adversely affects applications in dynamic downstream tasks.

Figure 1 illustrates the representation challenges of spatial heterogeneity and temporal dynamics for RNRL. Firstly, spatial heterogeneity reveals that road similarities can be estimated from several aspects beyond geospatial distance. For example, as illustrated in Figure 1(a), in the blue track, the next hop S_b and S_c after multiple hops from road S_a have configurations similar to S_a. Extending the analysis to the broader road network shows that S_d, although spatially farther from S_a, serves a similar function (i.e., residential). Hence, methods restricted by nearby roads may yield an incomplete understanding of distant roads. Secondly, temporal traffic dynamics at different hours serve as a crucial comple-

*Corresponding author.



(a) Spatial heterogeneity. (b) Temporal dynamics.

Figure 1: An illustration example of spatial heterogeneity and temporal dynamics. (a) Distant roads with similar configurations can be connected by a travel trajectory, whereas nearby roads may not necessarily share similar characteristics. (b) The traffic patterns of roads are characterized not only by road types but also by temporal dynamics.

ment to road representations, which cannot be adequately captured by road networks alone. As shown in Figure 1(b), distinct road types exhibit markedly divergent traffic volumes temporally. Critically, dynamic patterns demonstrate pronounced weekday-weekend divergence even for identical road categories. Therefore, road representation necessitates the consideration of both spatial heterogeneity and temporal dynamics to effectively adapt to dynamic downstream tasks.

To address these issues, we propose a **D**ual-branch **S**patial-**T**emporal self-supervised representation framework (DST) to obtain road representations from dual perspectives. From the spatial perspective, DST overcomes spatial heterogeneity by extracting dynamic relationships among roads using a mix-hop transition matrix, which is employed in the initial stage of graph convolution. Additionally, we design a road hypergraph incorporating three types of hyperedges beyond the natural graph topology and integrate it with the standard road network through contrastive learning. From the temporal perspective, DST utilizes next-token prediction for self-supervised representation based on a causal Transformer, along with a regularization task that differentiates traffic modes on weekdays from those on weekends. To accommodate heterogeneous inputs while optimizing learning efficiency, the dual-branch representations are pre-trained separately and subsequently fused for downstream tasks.

Overall, our contributions can be summarized as below:

- We propose a novel dual-branch road representation framework from both spatial and temporal views to address the challenges of spatial heterogeneity and temporal dynamics, called DST.
- We design a mix-hop transition matrix and hypergraph-based contrastive learning for the spatial branch. Meanwhile, we devise next-token prediction and an auxiliary discrimination task for the temporal branch.
- Extensive experiments on three datasets across three downstream tasks indicate the *state-of-the-art* performance of DST based on superior road representation. Additionally, cross-city experiments demonstrate the strong transferability of our framework in zero-shot learning scenarios.

Preliminaries

In this section, we introduce the definitions of notation and descriptions of variables used in this paper.

Definition 1: ***Road Network***. *A road network can be represented as a directed graph* $\mathcal{G} = \{\mathcal{R}, \mathcal{E}, X_{\mathcal{R}}, X_{\mathcal{E}}\}$*, where* $\mathcal{R}$ *and* $\mathcal{E}$ *denote the set of road segments and their connections. For brevity, we let* $N = |\mathcal{R}|$ *denote the number of road segments and refer to the road segment as "road" in the later sections.* $A_{\mathcal{G}} \in \mathbb{R}^{N \times N}$ *is the adjacency matrix. If* (r_i, r_j) *is reachable, then* $A_{\mathcal{G}}[r_i, r_j] = 1$*; otherwise* $A_{\mathcal{G}}[r_i, r_j] = 0$*.* $X_{\mathcal{R}} \in \mathbb{R}^{N \times C_1}$ *and* $X_{\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}| \times C_2}$ *are the features of* $\mathcal{R}$ *retrieved from OpenStreetMap (OSM) and* $\mathcal{E}$ *computed from* $X_{\mathcal{R}}$*, where* C_1 *and* C_2 *represent the number of features.*

Definition 2: ***Road Hypergraph***. *A road hypergraph* $\mathcal{G}_{\mathcal{H}} = (\mathcal{R}, \mathcal{E}_{\mathcal{H}})$ *consists of roads and hyperedges, where* $\mathcal{E}_{\mathcal{H}} = (e_1, e_2, \ldots, e_K)$ *is the set of hyperedges. A simple graph is inadequate for capturing high-order relationships and functional properties between roads. Therefore, we introduce road hypergraph and utilize hyperedges to achieve these objectives.* $A_{\mathcal{H}} \in \mathbb{R}^{N \times K}$ *is the incidence matrix of a hypergraph. For* r_i *and* e_k*, if* $r_i \in e_k$*, then* $A_{\mathcal{H}}[r_i, e_k] = 1$*; otherwise* $A_{\mathcal{H}}[r_i, e_k] = 0$*.*

Definition 3: ***Trajectory***. *A trajectory* $\tau \in \mathcal{T}$ *is a sequence of continuously reachable* r *in* $\mathcal{G}$*, denoted as* $\tau = (r_1, r_2 \ldots, r_M)$*. Trajectories capture movement patterns and contain rich dynamic information. In the following sections, hop distance represents the number of hops within the trajectory, with* $hop[r_i, r_j] = j - i$*.*

Definition 4: ***Traffic Dynamics***. *The traffic dynamics* $\mathcal{D}_{\mathcal{R}} = [u_i]_{i=1}^{T} \in \mathbb{R}^{N \times T \times C}$ *is the travel mode of* $\mathcal{R}$ *extracted from the trajectory. Let* u_i *represent the number of visits on the road at the* i*-th time. The travel mode is divided by weekdays and weekends, with 24 hours in a day. Thus, we have* $T = 24$ *and* $C = 2$*.*

Problem Statement: Given a road network $\mathcal{G}$, the goal of Road Network Representation Learning is to learn a low-dimensional representation $v_r \in \mathbb{R}^d$ for each road, where d represents the vector dimension. The learned representation can be utilized for various downstream tasks based on roads and trajectories.

Methodology

In this section, we present the proposed DST framework. DST overcomes spatial heterogeneity and captures high-order relationships by the learnable mix-hop transition matrix and hypergraph. Meanwhile, DST employs a Transformer-based traffic dynamic encoder and two self-supervised tasks to model the temporal travel traffic dynamics information of the roads. The overall architecture of the framework is shown in Figure 2.

Mix-hop Transition Matrix Weighting

Before all commencement, each static feature of the road network is embedded through an independent embedding layer. The initial embedding of a road segment is formed

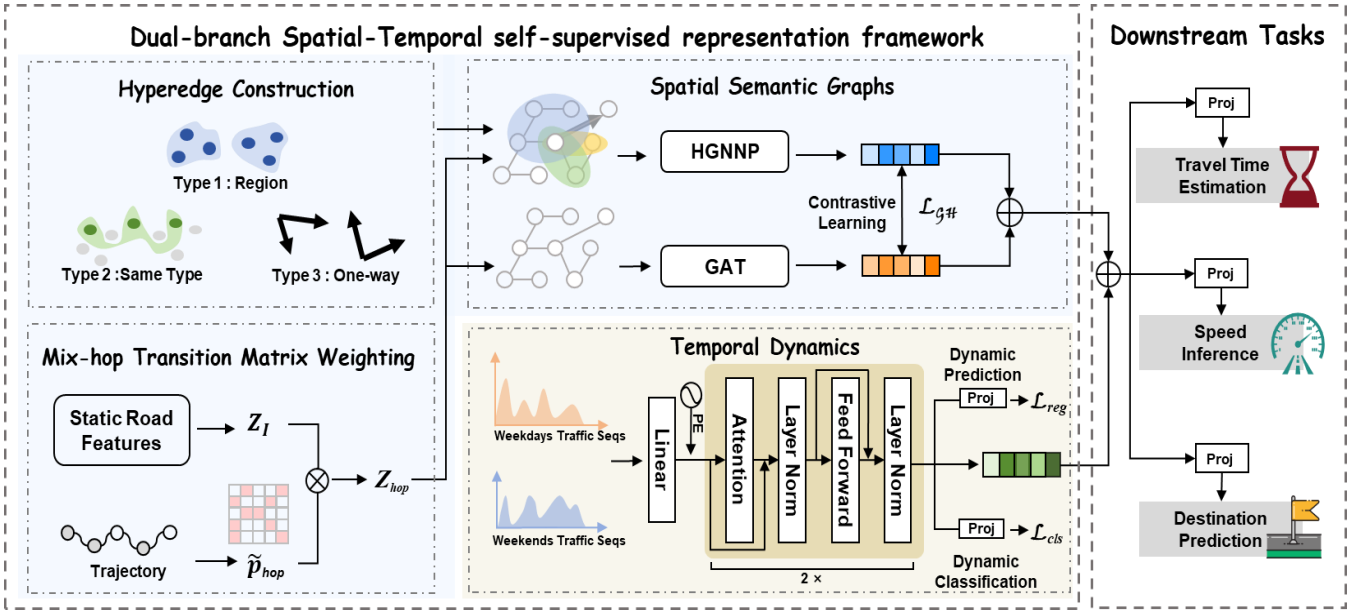


Figure 2: The overview of the proposed DST framework. The high-order relationships are modeled via mix-hop transition matrix weighting and multi-view graph contrastive learning. The temporal travel traffic dynamics are integrated by the Transformer with two specific task-driven updates. Both block co-enhanced representations power downstream tasks jointly.

by concatenating the outputs of these layers:

$$Z_{\mathcal{I}} = \Big\|_{c=1}^{C_1} \text{Emb}\left(X_{\mathcal{R}_c} | X_{\mathcal{R}_c} \in X_{\mathcal{R}}\right). \tag{1}$$

Here, $||$ is the vector concatenation operation and $Z_{\mathcal{I}} \in \mathbb{R}^{N\times d}$ is the feature embedding of all roads.

Next, we consider extracting the mix-hop transition matrix from trajectories on road networks. Unlike existing methods, our approach emphasizes the reachable and functional information of both the next hop and multiple hops of the road within the trajectory:

$$P_{hop}\left[r_i, r_j\right] = \sum_{\tau\in\mathcal{T}} \sum_{1\le i<j\le m} m - (j-i). \tag{2}$$

Here, m denotes the number of roads within the single trajectory. The weights are assigned based on hop distance. In this context, the smaller hop distance receives larger initial weights, and the longer hop distance receives smaller initial weights. This strategy emphasizes adjacent links while incorporating reachable distant links.

To mitigate the impact of magnitude differences, we apply row normalization to the resulting transition matrix:

$$\widetilde{P}_{hop}\left[r_i, r_j\right] = \frac{P_{hop}\left[r_i, r_j\right]}{\sum_{j=1}^{N} P_{hop}\left[r_i, r_j\right]}. \tag{3}$$

Here, $\widetilde{P}_{hop} \in \mathbb{R}^{N\times N}$ is the final mix-hop transition matrix. If $\sum_{j=1}^{N} P_{hop}\left[r_i, r_j\right] = 0$, we set $\widetilde{P}_{hop}\left[r_i, r_i\right] = 1$. Then, we use it to initialize a learnable weight matrix, which is then used to update the representation further:

$$Z_{hop} = \widetilde{P}_{hop} \cdot Z_{\mathcal{I}}. \tag{4}$$

Here, $\widetilde{P}_{hop}$ is the learnable mix-hop transition matrix and $Z_{hop} \in \mathbb{R}^{N\times d}$ is mix-hop weighted feature vectors.

Spatial Semantic Graph Learning

Road networks exhibit heterogeneous characteristics, and the overemphasis of simple graph neural networks on topology may induce the over-smoothing problem(Chen et al. 2021, 2024). To mitigate this, we employ the spatial semantic hypergraph to capture high-order relationships between roads. Inspired by contrastive learning, we model road networks through graph-hypergraph views, updating parameters by maximizing mutual information (MI) between positive node pairs across representations.

Spatial Graph View For the graph view, we employ a multi-layer Graph Attention Network (GAT) (Veličković et al. 2018) to obtain the representation that captures the spatial topology of the road network. At this stage, we also take into account some characteristics of road network connectivity $X_{\mathcal{E}}$. The specific formula for GAT is:

$$Z_{\mathcal{G}} = \Big\|_{h=1}^{H} \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{h} \cdot W^{h} v_j\right), \tag{5}$$

$$\alpha_{ij} = \frac{\exp(\sigma(a^T \cdot \left[Wv_i \,\|\, Wv_j \,\|\, X_{\mathcal{E}_{ij}}\right]))}{\sum_{k\in\mathcal{N}(i)} \exp(\sigma(a^T \cdot \left[Wv_i \,\|\, Wv_k \,\|\, X_{\mathcal{E}_{ik}}\right]))}, \tag{6}$$

where H represents the number of heads in the multi-head attention mechanism, σ is the activation function, W^h is the learnable parameter of h-th head, and $v_j \in Z_{hop}$ is the representation vector of the road r_j. The graph encoder generates the spatial representation $Z_{\mathcal{G}}$ for the graph view.

Semantic HyperGraph View For the hypergraph view, as illustrated in Figure 2, our hypergraph framework models complex semantic relationships through three specialized hyperedge types. $\mathcal{E}_{\mathcal{H}1}$ represents the functional zone

generated through spectral clustering (Von Luxburg 2007) of roads. $\mathcal{E}_{\mathcal{H}2}$ captures long-range dependencies by grouping roads of identical types, regardless of geographical proximity (Zhang and Long 2023). $\mathcal{E}_{\mathcal{H}3}$ operationalizes Tobler's First Law of Geography (Tobler 1970) through hyperedges that connect geographically adjacent unidirectional roads.

We employ a multi-layer General Hypergraph Neural Networks (HGNN$^+$) (Gao et al. 2022) to further capture the functional semantics of the road network. By constructing hyperedges to represent various types of potential high-order and long-range correlations, HGNN$^+$ employs an adaptive hyperedge fusion strategy to effectively combine these correlations. Specifically, the formula for HGNN$^+$ is:

$$Z_{\mathcal{H}} = D\left(\mathcal{R}\right) \cdot A_{\mathcal{H}} \cdot D\left(\mathcal{E}_{\mathcal{H}}\right) \cdot {A_{\mathcal{H}}}^{T} \cdot Z_{hop}, \tag{7}$$

$$\mathcal{E}_{\mathcal{H}} = \mathcal{E}_{\mathcal{H}1} \parallel \mathcal{E}_{\mathcal{H}2} \parallel \mathcal{E}_{\mathcal{H}3}, \tag{8}$$

where the diagonal matrix $D(\mathcal{R})$ represents the degree of the nodes, denoting the number of hyperedges associated with the nodes, while $D(\mathcal{E}_{\mathcal{H}})$ represents the degree of the hyperedges, indicating the number of nodes contained in each hyperedge, $\mathcal{E}_{\mathcal{H}}$ is the set of three types hyperedges. The hypergraph encoder generates the semantic representation matrix $Z_{\mathcal{H}}$ for the hypergraph view.

Contrastive Learning Loss We pursue the generalized rather than specific representation for downstream tasks through self-supervised learning. Therefore, we treat the representations $Z_{\mathcal{G}}$ and $Z_{\mathcal{H}}$ as two views of the road network for contrastive learning. We optimize the following MI maximization objective across these views (Zhu et al. 2021).

$$\mathcal{L}_{\mathcal{GH}} = -\frac{1}{N} \sum_{r_i \in \mathcal{R}} \left[\frac{1}{\left|\mathcal{H}\left(r_i\right)\right|} \sum_{r_j \in \mathcal{H}(r_i)} I\left(v_{r_i}, h_{r_j}\right) \right], \tag{9}$$

where $\mathcal{H}(r_i)$ is the set of nodes corresponding to the hypergraph view of r_i, $v_{r_i} \in Z_{\mathcal{G}}$ and $h_{r_j} \in Z_{\mathcal{H}}$ is the representation vector of r_i and r_j separately, and $I(\cdot)$ is the MI estimator (Mao et al. 2022). The mini-batch strategy (Li et al. 2014) is implemented to optimize memory utilization.

Temporal Dynamics Learning

The coarse-grained road utilization feature fails to integrate into the temporal traffic flow characteristics. To this end, we employ a Transformer-based encoder to model fine-grained traffic dynamics. The joint optimization of the dynamic prediction and classification tasks enables comprehensive learning of 24-hour traffic evolution trends and patterns.

Temporal Travel Dynamics Encoder The Transformer-based encoder ingests the previously defined traffic dynamic sequence, with its final hidden state serving as the compressed sequence representation. This architecture proves particularly effective for dynamic sequence modeling due to its self-attention mechanism, which captures both local fluctuations and global temporal dependencies:

$$Z_{\mathcal{D}} = \text{TransEnc}\left(\text{PosEnc}\left(\mathcal{D}_{\mathcal{R}}\right)\right)[-1]. \tag{10}$$

Here, PosEnc$(\cdot)$ is sinusoidal positional encoding and TransEnc$(\cdot)$ denotes a standard implementation of Transformer Encoder (Waswani et al. 2017). The temporal travel dynamic encoder produces the temporal representation $Z_{\mathcal{D}}$.

Two-task Joint Dynamic Loss We aim to capture the patterns and trends of temporal traffic dynamic sequences through two tasks. Intuitively, the dynamic prediction task leverages the history sequence to predict the value of the sequence at the next time step, thereby capturing the long-term trend of the dynamic sequence. The dynamic classification task requires the model to categorize the input sequence into two types: weekday traffic and weekend traffic to distinguish between different dynamic sequence patterns. The corresponding regression value and classification probability are obtained after the representation passes through the project head, and these form the loss for the dynamic sequence task:

$$\mathcal{L}_{reg} = \frac{1}{N \times C} \sum_{i=1}^{N \times C} \|y_i - \hat{y}_i\|^2, \tag{11}$$

$$\mathcal{L}_{cls} = \frac{1}{N \times C} \sum_{i=1}^{N \times C} \sum_{c=1}^{C} -y_i(c) \log(\hat{y}_i(c)), \tag{12}$$

$$\mathcal{L}_{d} = \lambda_{reg} \cdot \mathcal{L}_{reg} + \lambda_{cls} \cdot \mathcal{L}_{cls}. \tag{13}$$

Here, $\hat{y_i}$ is the predicted output from the fully connected layer and y_i is the ground truth, λ_{reg} and λ_{cls} are the hyperparameters to balance the two tasks.

Representation Fusion for Downstream Tasks

Finally, the spatial semantic representation and temporal dynamic representation are fused to form the final joint representation:

$$Z = \text{Fusion}\left(Z_{\mathcal{G}}, Z_{\mathcal{H}}, Z_{\mathcal{D}}\right), \tag{14}$$

where $Z_{\mathcal{G}}$, $Z_{\mathcal{H}}$, and $Z_{\mathcal{D}}$ are the updated spatial and temporal representations derived from the graph, hypergraph, and sequence encoder. $Z \in \mathbb{R}^{N \times d}$ is the final representation matrix of roads. The fusion methods we explore comprise concatenation along the last dimension, direct summation, and integration of representations via a gating mechanism. In this work, the first is employed for representation fusion.

This fusion method retains all the information of different representations, which is sufficiently effective for downstream tasks. By incorporating temporal travel traffic dynamics into the road network, the final road representation becomes more informative, enhancing its effectiveness and versatility. Furthermore, our approach captures semantic relationships within road networks, which are essential for downstream tasks. For more fusion experiments and further discussion of fusion, please refer to Appendix D.1.

Experiments

In this section, we present and analyze the experimental results across three real-world datasets and downstream tasks. The experimental results validate the effectiveness of leveraging spatiotemporal information from both trajectories and road networks to improve the quality of road network representations. Due to space limitations, additional experimental details and results are provided in Appendices C and D.

Experimental Setup

Datasets We use datasets from three real-world cities: Beijing, Porto, and Xi'an. The initial trajectory data consisted of GPS points collected by vehicles. We employ a map-matching algorithm (Yang and Gidofalvi 2018) to map it to the road sequence trajectory defined in the previous section. Please refer to Appendix C.1 for more details.

Downstream Tasks The effectiveness of road representations is validated on three traffic-related downstream tasks: road speed inference (SI), travel time estimation (TE), and trajectory destination prediction (DP). A more detailed description can be found in Appendix C.2.

Evaluation Metrics Following prior works (Mao et al. 2022; Schestakov et al. 2023), we employ established evaluation metrics for each task: mean absolute error (MAE) and root mean square error (RMSE) for the SI and TE; and accuracy at 1 (ACC@1) and mean reciprocal rank (MRR) for the DP. These metrics are standard for their respective domains.

Baseline Methods We evaluate DST against a series of baselines, including both classic graph learning methods and self-supervised RNRL methods. The former includes Node2Vec (Grover et al. 2016), GCN (Kipf et al. 2017), GAE (Veličković et al. 2018) and T-GCN (Zhao et al. 2019), while the latter comprises SRN2Vec (Wang et al. 2020), Toast (Chen et al. 2021), JCLRNT (Mao et al. 2022), TrajRNE (Schestakov et al. 2023) and DyToast (Chen et al. 2024). See Appendix C.3 for more details.

Overall Performance

Table 1 and 2 present the mean values of the results obtained from five random seeds for DST and baseline models across different downstream tasks. We can find that DST outperforms the baseline models in all tasks, demonstrating its effectiveness in modeling high-order and long-range information about road network function and movement, as well as its superiority in integrating road network dynamic travel traffic. Among the baseline methods, Node2Vec, GCN, and GAE are not specifically designed for road networks. They are limited to modeling simple graph structures and do not capture the intricate relationships between roads, leading to suboptimal performance in tasks. This highlights the importance of unique designs tailored for road networks. Although TGCN models temporal traffic dynamics, its architectural paradigm is not optimized specifically for road networks. Furthermore, it fails to capture the functional semantics of road networks. SRN2Vec and Toast apply random walks for road weighting, thus achieving better performance than the traditional graph learning methods. DyToast augments Toast with trigonometric time features, enhancing performance in destination prediction and travel time estimation tasks. However, its degraded performance in the speed inference task reveals representational robustness limitations. Suboptimal dynamic modeling approaches may degrade performance. In contrast, JCLRNT and TrajRNE leverage trajectory information to enhance road network representations, achieving better performance across tasks than other methods. Nevertheless, they are deficient in modeling temporal dynamics, while DST mines the long-range dependencies of roads in trajectory and models both spatial and temporal traffic representations of the road network, achieving superior performance compared to other methods.

Ablation Studies

DST employs the following core designs for road network representation: the mix-hop pattern extraction strategy initializes the learnable transition matrix to capture road long-range dependencies contained in trajectories; the hypergraph with different hyperedges models higher-order functional information; the dynamic travel traffic integration supplements temporal information. To evaluate the impact of these designs, we conduct ablation studies by removing specific components. (1) *w/o* P_{hop}*:* this variant excludes the learnable transition matrix; (2) *w/o* hg_1*:* this variant eliminates region relationship hyperedges from the hypergraph structure; (3) *w/o* hg_2*:* this variant eliminates the same type relationship hyperedges from the hypergraph structure; (4) *w/o* hg_3*:* this variant eliminates one-way relationship hyperedges from the hypergraph structure; (5) *w/o tm:* this variant omits temporal travel traffic dynamic modeling.

Figure 3 illustrates the performance of the model in the Beijing dataset after the removal of different design components. The results on other datasets are similar. It can be observed that removing any of these components results in inferior performance on downstream tasks compared to the complete DST. Specifically, we make the following analysis. Firstly, the variants *w/o* P_{hop} and *w/o* hg_2 exhibit the most severe performance degradation in the speed inference task, indicating that higher-order functional and motion relationships between roads are essential for learning effective road representations. Secondly, the variant *w/o tm* demonstrates significantly degraded performance across both trajectory-based tasks, underscoring the critical complementary role of dynamic travel traffic information in road network characterization. Finally, the variants *w/o* hg_1 and *w/o* hg_3 exhibit moderate performance degradation across all tasks, indicating that diverse hyperedge types mutually reinforce each other to comprehensively model road network higher-order relationships. Please refer to Appendix D.2 for more details.

Parameter Sensitivity

In this section, we analyze the effect of parameters on DST performance: the mini-batch size, the traffic batch size, and the ratio of λ_{reg} and λ_{cls}. The evaluation is conducted on the destination prediction task for the three datasets. The results are shown in Figure 4. We can find that DST maintains consistent excellent performance across all datasets despite parameter variations. The specific conclusions are as follows. Please refer to Appendix D.5 for more details.

Firstly, the results of *varying mini-batch size* indicate that a smaller size reduces the number of negative samples, potentially causing insufficient sample discrimination and underfitting. Conversely, a larger size increases computational complexity and memory demands while hindering effective learning. Thus, a moderate batch size achieves an optimal balance between learning efficiency and model per-

Methods	Beijing				Porto				Xi'an			
	Destination Prediction		Travel Time Estimation		Destination Prediction		Travel Time Estimation		Destination Prediction		Travel Time Estimation	
	ACC@1 (↑)	MRR (↑)	MAE (↓)	RMSE (↓)	ACC@1 (↑)	MRR (↑)	MAE (↓)	RMSE (↓)	ACC@1 (↑)	MRR (↑)	MAE (↓)	RMSE (↓)
Node2Vec	0.1954	0.2884	253.0633	378.8966	0.2201	0.3364	109.5741	150.4529	0.4088	0.5083	287.3482	430.5943
GCN	0.2189	0.2995	256.2389	381.1153	0.3780	0.4861	108.4090	151.8520	0.4054	0.4904	249.5978	387.8792
GAE	0.2472	0.3254	249.8528	376.1648	0.3997	0.5165	107.6947	150.9522	0.4223	0.5091	283.0110	417.5603
TGCN	0.2080	0.2911	274.5425	402.6056	0.4584	0.5827	111.8000	155.3018	0.3887	0.4823	231.3772	341.2320
SRN2Vec	0.4158	0.4702	242.3774	368.4888	0.6552	0.7745	101.9539	144.1221	0.6966	0.7503	236.8440	378.1186
Toast	0.3031	0.3652	248.4413	374.0255	0.6142	0.7324	101.7503	145.7847	0.7103	0.7673	244.5761	381.7282
JCLRNT	0.4222	0.5528	246.7876	373.1402	0.5133	0.6626	101.6065	145.2937	0.6752	0.7711	240.6833	380.1059
TrajRNE	0.6728	0.7603	237.1361	363.2190	0.6728	0.8063	102.5061	145.2530	0.8260	0.8831	207.6418	314.7823
DyToast	0.4440	0.5164	239.6679	365.3374	0.4887	0.6025	102.3489	145.3401	0.7508	0.8080	238.4887	379.9900
Ours	**0.7288**	**0.8213**	**236.6965**	**363.2039**	**0.6766**	**0.8101**	**101.4223**	**143.6598**	**0.8335**	**0.8950**	**202.8479**	**307.7553**

Table 1: The results on three real-world datasets in terms of the destination prediction task and the travel time estimation task. The best one is denoted by **bold** and the second-best is denoted by underline. ↑ and ↓ denote higher is better and lower is better.

Methods	Beijing		Porto		Xi'an	
	MAE (↓)	RMSE (↓)	MAE (↓)	RMSE (↓)	MAE (↓)	RMSE (↓)
Node2Vec	8.2268	8.8945	9.0367	10.1075	6.8403	8.8547
GCN	8.1926	8.8436	8.9723	10.0485	6.8014	8.8107
GAE	8.1600	8.8154	8.9442	10.0804	6.8040	8.8114
TGCN	5.8026	6.7065	6.5242	7.8042	6.0006	8.0067
SRN2Vec	6.9959	7.8806	7.0657	8.3563	6.0201	7.9983
Toast	6.6538	7.4272	7.2837	8.4783	5.5102	7.5338
JCLRNT	2.8512	3.9013	3.7475	4.9999	4.5138	5.7294
TrajRNE	3.0756	4.3049	4.7854	6.3375	5.1898	7.1767
DyToast	8.0957	8.8124	8.6818	9.8597	6.7461	8.7050
Ours	**2.4595**	**3.2557**	**3.4259**	**4.5538**	**4.4987**	**5.6557**

Table 2: The results on three real-world datasets in terms of the speed inference task. The best one is denoted by **bold** and the second-best is denoted by underline.

formance. Secondly, the results of *varying traffic batch size* indicate that DST is not sensitive to this parameter. Nonetheless, there is a slight performance improvement in small-sized traffic batches, potentially due to the sparsity of the traffic sequence. A larger batch scale may introduce more noise. Finally, the results of *varying the λ ratio* indicate that increasing the next token prediction loss weight enhances model performance. This adjustment compensates for the significant magnitude difference in initial training losses between tasks, thereby improving task balance during training.

Generalization on Zero-Shot Learning Scenarios

Given substantial variations in urban systems, such as network topology, traffic patterns, and human mobility, transferable models significantly reduce deployment costs and resource expenditures across cities. To evaluate the cross-city transferability of DST, we conduct a zero-shot learning experiment. Recognizing dependencies between the mix-hop matrix and urban nodes, we remove this component and train exclusively in Beijing. Evaluation then occurs on the Porto for three downstream tasks. In this context, we compare DST-transfer with GAE and JCLRNT, which support zero-shot learning. Table 3 demonstrates competitive zero-shot performance of DST attributable to comprehensive spatiotemporal representation learning.

Case Study

In this section, we conduct a case study to demonstrate the ability of DST to capture high-order relationships between

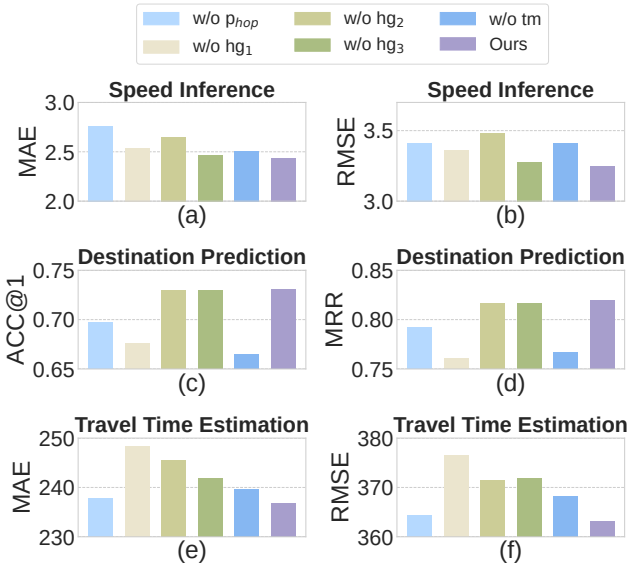


Figure 3: Ablation study on Beijing dataset.

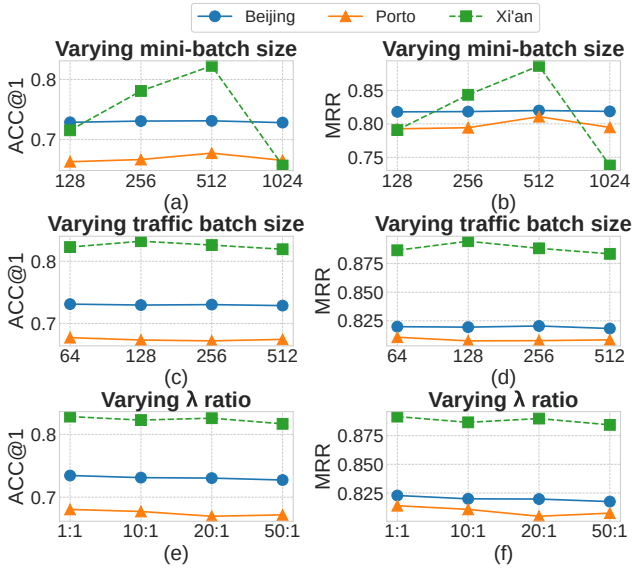
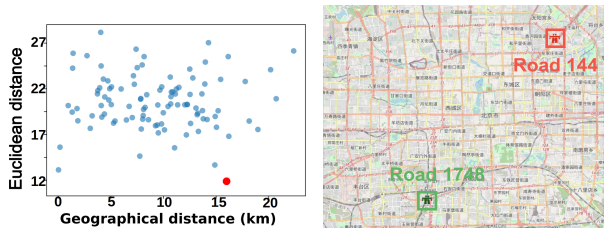


Figure 4: Parameter sensitivity on destination prediction.

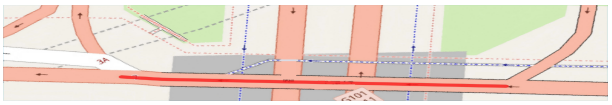
Methods	Speed Inference		Destination Prediction		Travel Time Estimation	
	MAE (↓)	RMSE (↓)	ACC@1 (↑)	MRR (↑)	MAE (↓)	RMSE (↓)
GAE	9.0482	10.1209	0.3119	0.4204	112.1389	154.2606
JCLRNT	4.1047	5.3195	0.0167	0.0338	109.7691	151.6060
DST-transfer	3.5126	4.6181	0.6424	0.7765	108.0329	150.2018

Table 3: Zero-shot learning results (Beijing → Porto).

road segments. We choose the highest-traffic road (ID=144) in Beijing as an anchor road, and compute representation and geographic distances to other roads. Figure 5(a) shows the distance distribution of other roads and the anchor road. We identified a distant road segment (ID=1748) exhibiting close representational proximity. In Figure 5(b), we visualize them in the road network. The anchor road is red, and the selection road is green. The visualization reveals that both the anchor and selected segments lie on the third ring of Beijing. Fine-grained analysis in Figure 5(c) and 5(d) shows they share identical functional roles, unidirectional flow patterns, and dual inflow and outflow characteristics. This confirms DST successfully captures functional and traffic-dynamic relationships beyond spatial proximity.



(a) Distance to the anchor road. (b) Cases in the road network.



(c) The highest traffic road detailed visualization.

(d) The road geographically distant but representationally close.

Figure 5: Case study of the highest traffic road in Beijing.

Related Work

Road Network Representation Learning. Road networks are inherently structured as graphs. Most existing approaches draw inspiration from graph representation learning (Zheng et al. 2022, 2023; Wang et al. 2024a; Yang et al. 2025) and employ graph neural networks (GNNs) for modeling. Numerous GNN-based methods (Wu et al. 2020; Pei et al. 2020; Zhang and Long 2023; Cao et al. 2025) develop specialized graph encoders to capture the spatial topology of road networks. As trajectory data proliferates, trajectory-enhanced methods (Mao et al. 2022; Schestakov et al. 2023) extract movement and transfer patterns to advance road network representation learning. However, these methods inadequately model higher-order information within functional and movement. Our method integrates spatial semantic information from the perspectives of transfer mode and functional structure while incorporating travel traffic dynamics.

Graph and Hypergraph Learning. Early graph learning (Perozzi et al. 2014; Grover et al. 2016) focuses on the topology of graphs and learns the shallow representations of graphs. With the development of deep learning, graph convolutional network (Kipf et al. 2017) and graph attention network (Veličković et al. 2018) achieve significant success in graph learning by aggregating features in different ways. Temporal graph convolutional network considers dynamic changes in graphs (Zhao et al. 2019; Wang et al. 2024b) later. Limited by the adjacency in graphs, these models struggle to represent higher-order relationships between nodes. Therefore, hypergraph learning is employed to model higher-order relationships among two or more entities. Hypergraph Neural Networks (Feng et al. 2019; Gao et al. 2022) adopt different message-passing mechanisms to learn hypergraph representations. However, neither graphs nor hypergraphs can be directly applied to road networks that require tailored designs due to their complex spatiotemporal characteristics.

Spatial-temporal Self-supervised Learning. Graph self-supervised learning methods are categorized into generative and contrastive paradigms. Generative approaches (Bondy and Hemminger 1977) leverage intrinsic graph structure as supervision, focusing on reconstruction tasks. Contrastive methods (You et al. 2020; Liu et al. 2023) operate on augmented graph views, exploiting invariance across views and discriminative information between negative pairs as supervisory signals. Sequence self-supervised learning typically employs predictive pretext tasks, where supervisory signals are autonomously generated through statistical heuristics or domain knowledge. These tasks model implicit data-label relationships to facilitate representation learning (Medina-Salgado et al. 2022). Our framework integrates multi-view graph contrastive learning to capture spatial semantics, while dynamically designed prediction and classification tasks model travel traffic dynamics, jointly enhancing road network representation.

Conclusion

This paper proposes DST, a comprehensive road network representation framework that integrates spatial semantics and temporal travel traffic dynamics. DST simultaneously models the functionality of the road network and the movement provided by trajectories. For spatial semantics, we design a mix-hop transition matrix and hyperedges to capture long-range dependencies between roads. For temporal travel traffic dynamics, we employ two well-designed tasks to model traffic trends and patterns. Extensive experimental results on three real-world datasets across three downstream tasks demonstrate the effectiveness and robustness of DST. Furthermore, DST demonstrates satisfactory transferability. In the future, we will investigate the emergency forecasting problem of road networks in the context of extreme natural disasters, such as typhoons and storms.

Acknowledgments

This work is supported by the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (Grant No. SN-ZJU-SIAS-001), Zhejiang Provincial Natural Science Foundation of China (Grant No. LMS25F020012), the Hangzhou Joint Fund of the Zhejiang Provincial Natural Science Foundation of China under Grant No.LHZSD24F020001, Zhejiang Province High-Level Talents Special Support Program "Leading Talent of Technological Innovation of Ten-Thousands Talents Program" (No.2022R52046), the Fundamental Research Funds for the Central Universities (No.226-2024-00058), and the advanced computing resources provided by the Supercomputing Center of Hangzhou City University.

References

Azgomi, H. F.; and Jamshidi, M. 2018. A brief survey on smart community and smart transportation. In *ICTAI*, 932–939. IEEE.

Bondy, J. A.; and Hemminger, R. L. 1977. Graph reconstruction—a survey. *Journal of Graph Theory*, 1(3): 227–268.

Camero, A.; and Alba, E. 2019. Smart City and information technology: A review. *cities*, 93: 84–94.

Cao, J.; Zheng, T.; Guo, Q.; Wang, Y.; Dai, J.; Liu, S.; Yang, J.; Song, J.; and Song, M. 2025. Holistic Semantic Representation for Navigational Trajectory Generation. *AAAI*.

Chen, Y.; Li, X.; Cong, G.; Bao, Z.; and Long, C. 2024. Semantic-Enhanced Representation Learning for Road Networks with Temporal Dynamics. *arXiv preprint arXiv:2403.11495*.

Chen, Y.; Li, X.; Cong, G.; Bao, Z.; Long, C.; Liu, Y.; Chandran, A. K.; and Ellison, R. 2021. Robust road network representation learning: When traffic patterns meet traveling semantics. In *CIKM*, 211–220.

Feng, Y.; You, H.; Zhang, Z.; Ji, R.; and Gao, Y. 2019. Hypergraph neural networks. In *AAAI*, 3558–3565.

Gao, Y.; Feng, Y.; Ji, S.; and Ji, R. 2022. HGNN+: General hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3181–3199.

Gharaee, Z.; Kowshik, S.; Stromann, O.; and Felsberg, M. 2021. Graph representation learning for road type classification. *Pattern Recognition*, 120: 108174.

Grover, A.; et al. 2016. node2vec: Scalable feature learning for networks. In *SIGKDD*, 855–864.

Jepsen, T. S.; et al. 2019. Graph convolutional networks for road networks. In *SIGSPATIAL*, 460–463.

Kipf, T. N.; et al. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Li, M.; Zhang, T.; Chen, Y.; and Smola, A. J. 2014. Efficient mini-batch training for stochastic optimization. In *SIGKDD*, 661–670.

Li, X.; et al. 2020. Spatial transition learning on road networks with deep probabilistic models. In *ICDE*, 349–360. IEEE.

Liu, S.; Zhou, Y.; Song, J.; Zheng, T.; Chen, K.; Zhu, T.; Feng, Z.; and Song, M. 2023. Contrastive Identity-Aware Learning for Multi-Agent Value Decomposition. In *AAAI*, volume 37, 11595–11603.

Mao, Z.; Li, Z.; Li, D.; Bai, L.; and Zhao, R. 2022. Jointly contrastive representation learning on road network and trajectory. In *CIKM*, 1501–1510.

Medina-Salgado, B.; Sánchez-DelaCruz, E.; Pozos-Parra, P.; and Sierra, J. E. 2022. Urban traffic flow prediction techniques: A review. *Sustainable Computing: Informatics and Systems*, 35: 100739.

Pei, H.; Wei, B.; Chang, K. C.-C.; Lei, Y.; and Yang, B. 2020. Geom-gcn: Geometric graph convolutional networks. In *ICLR*.

Perozzi, B.; et al. 2014. Deepwalk: Online learning of social representations. In *SIGKDD*, 701–710.

Schestakov, S.; et al. 2023. Road network representation learning with vehicle trajectories. In *PAKDD*, 57–69. Springer.

Tobler, W. R. 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46: 234–240.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2018. Graph attention networks. In *ICLR*.

Von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17: 395–416.

Wang, M.-X.; Lee, W.-C.; Fu, T.-Y.; and Yu, G. 2020. On representation learning for road networks. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(1): 1–27.

Wang, Y.; Cao, J.; Huang, W.; Liu, Z.; Zheng, T.; and Song, M. 2024a. Spatiotemporal gated traffic trajectory simulation with semantic-aware graph learning. *Information Fusion*, 108: 102404.

Wang, Y.; Liu, S.; Zheng, T.; Chen, K.; and Song, M. 2024b. Unveiling global interactive patterns across graphs: Towards interpretable graph neural networks. In *SIGKDD*, 3277–3288.

Wang, Y.; Zheng, T.; Liang, Y.; Liu, S.; and Song, M. 2024c. Cola: Cross-city mobility transformer for human trajectory simulation. In *WWW*, 3509–3520.

Wang, Y.; Zheng, T.; Liu, S.; Feng, Z.; Chen, K.; Hao, Y.; and Song, M. 2024d. Spatiotemporal-Augmented Graph Neural Networks for Human Mobility Simulation. *TKDE*, 36(11): 7074–7086.

Waswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Wu, N.; Zhao, X. W.; Wang, J.; and Pan, D. 2020. Learning effective road network representation with hierarchical graph neural networks. In *SIGKDD*, 6–14.

Xue, A. Y.; Zhang, R.; Zheng, Y.; Xie, X.; Huang, J.; and Xu, Z. 2013. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. In *ICDE*, 254–265. IEEE.

Yang, C.; and Gidofalvi, G. 2018. Fast map matching, an algorithm integrating hidden Markov model with precomputation. *International Journal of Geographical Information Science*, 32(3): 547–570.

Yang, J.; Wang, Y.; Chen, K.; Zheng, T.; Zhou, Y.; Xiao, Z.; Cao, J.; Song, M.; and Liu, S. 2025. From GNNs to Trees: Multi-Granular Interpretability for Graph Neural Networks. In *ICLR*.

Yang, Z.; Sun, H.; Huang, J.; He, L.; Jia, X.; Zhao, J.; and Qiao, S. 2022. Robust traffic speed inference with Ensemble Learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(10): 17241–17257.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *NeurIPS*, 33: 5812–5823.

Zhang, L.; and Long, C. 2023. Road network representation learning: A dual graph-based approach. *ACM Transactions on Knowledge Discovery from Data*, 17(9): 1–25.

Zhao, L.; Song, Y.; Zhang, C.; Liu, Y.; Wang, P.; Lin, T.; Deng, M.; and Li, H. 2019. T-GCN: A temporal graph convolutional network for traffic prediction. *IEEE transactions on intelligent transportation systems*, 21(9): 3848–3858.

Zheng, T.; Feng, Z.; Zhang, T.; Hao, Y.; Song, M.; Wang, X.; Wang, X.; Zhao, J.; and Chen, C. 2022. Transition propagation graph neural networks for temporal networks. *TNNLS*, 35(4): 4567–4579.

Zheng, T.; Wang, X.; Feng, Z.; Song, J.; Hao, Y.; Song, M.; Wang, X.; Wang, X.; and Chen, C. 2023. Temporal aggregation and propagation graph neural networks for dynamic representation. *TKDE*, 35(10): 10151–10165.

Zhou, H.; Huang, W.; Chen, Y.; He, T.; Cong, G.; and Ong, Y. S. 2024. Road Network Representation Learning with the Third Law of Geography. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *NeurIPS*, volume 37, 11789–11813. Curran Associates, Inc.

Zhu, Y.; Xu, Y.; Liu, Q.; and Wu, S. 2021. An Empirical Study of Graph Contrastive Learning. In *NeurIPS*.

Zong, F.; Tian, Y.; He, Y.; Tang, J.; and Lv, J. 2019. Trip destination prediction based on multi-day GPS data. *Physica A: Statistical Mechanics and its Applications*, 515: 258–269.

Zou, G.; Lai, Z.; Ma, C.; Li, Y.; and Wang, T. 2023. A novel spatio-temporal generative inference network for predicting the long-term highway traffic speed. *Transportation research part C: emerging technologies*, 154: 104263.