UniTR: A Unified Framework for Joint Representation Learning of Trajectories and Road Networks

Jie Zhao[1*], **Chao Chen**[1*†], **Yuanshao Zhu**[2], **Mingyu Deng**[1], **Yuxuan Liang**[2]

[1]College of Computer Science, Chongqing University, China
[2]The Hong Kong University of Science and Technology (Guangzhou), China
{csjiezhao, cschaochen, dmy}@cqu.edu.cn, yuanshao@ieee.org, yuxliang@outlook.com

Abstract

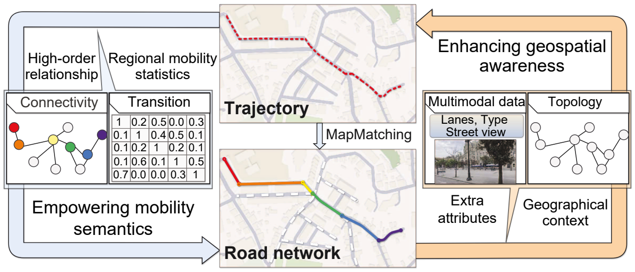
Representation learning of urban spatial-temporal data is fundamental and critical, serving a wide range of intelligent applications. Given that road networks and trajectories are inherently interrelated, their joint representation learning can significantly enhance the accuracy and utility of these applications. However, effectively learning joint representations for these two types of data remains challenging, particularly due to the complexities of interaction modeling and cross-scale optimization. To this end, we propose a unified framework, named UniTR, for joint representation learning of road networks and trajectories. Specifically, we first design a hierarchical propagation mechanism to model the complex many-to-many interactions between road networks and trajectories, thereby generating informative embeddings. Then, a triple-level contrastive optimization module is incorporated to systematically select valid positive and negative samples, further refining the embeddings. Experiments conducted on real-world datasets from two cities clearly demonstrate the effectiveness and superiority of UniTR.

Code — https://github.com/csjiezhao/UniTR

Introduction

Urban spatial-temporal data, particularly road networks and trajectories, are pivotal in understanding and modeling urban dynamics (Wang, Cao, and Philip 2020). The process of representation learning for this data is essential, as it converts spatial-temporal objects into compact and informative vector representations (Chen et al. 2024a). In practice, these representations serve as the foundation for a wide range of intelligent applications, such as traffic prediction (Zhao et al. 2023), route planning (Liu et al. 2023), and mobility analysis (Luca et al. 2021).

Road networks and trajectories are inherently interconnected, with road networks providing the foundational spatial structure and trajectories capturing the dynamic mobility patterns that occur within this structure (Mao et al. 2022). Given this interplay, their respective representations should be learned jointly, rather than in isolation. Integrating these two data types into a unified learning framework allows for mutual enhancement, leading to more effective representations. Figure 1 illustrates how the road network's topology constrains trajectory movements, while multi-modal information from road segments supplements their geographic context. Meanwhile, trajectories reveal the high-order relationships between road segments, with transition frequencies highlighting the different significance of segments.



1	0.2	0.5	0.0	0.3
0.1	1	0.4	0.5	0.1
0.1	0.2	1	0.2	0.1
0.1	0.6	0.1	1	0.5
0.7	0.0	0.0	0.3	1

Figure 1: Illustration of mutual enhancement in joint representation of road networks and trajectories.

A few studies (Chen et al. 2021; Mao et al. 2022; Ma et al. 2024), have sought to jointly characterize road networks and trajectories. These approaches typically use separate branches to extract features from road networks and trajectories, allowing the representations to complement each other in a self-supervised manner. However, this paradigm of separate modeling may not adequately account for the interactions between spatial structure and human mobility. Furthermore, effective joint representation learning of road networks and trajectories requires addressing two key aspects:

First, interactions between road networks and trajectories. Trajectories interact with multiple road segments, and road segments are traversed by numerous trajectories. Modeling these complex many-to-many relationships is crucial for learning effective representations. *Second, joint optimization across different representations*. The embedding spaces for road networks and trajectories exist on different scales. Both same-scale and cross-scale embedding optimization are vital for achieving meaningful representations.

To address the above challenges, we propose a unified contrastive learning framework named **UniTR** for joint representation learning of road networks and trajectories.

[*]These authors contributed equally.

[†]Corresponding author: cschaochen@cqu.edu.cn

Specifically, we design a hierarchical propagation mechanism that models the complex interactions between road networks and trajectories, facilitating their simultaneous representation learning. Additionally, we introduce a triple-level contrastive loss that optimizes embeddings across different scales, enhancing the overall performance of the learned representations. This comprehensive approach ensures that UniTR not only effectively models the intricate relationships within urban spatial-temporal data but also achieves superior performance across various downstream tasks. Our contributions can be summarized as follows:

- **Hierarchical Propagation Mechanism:** We design a hierarchical propagation mechanism that simultaneously learns the representations of road networks and trajectories. This mechanism effectively models the many-to-many interactions between road segments and trajectories, enhancing the overall representational quality.
- **Triple Level Contrastive Optimization:** We introduce a triple-level contrastive optimization that facilitates cross-scale embedding learning. Through carefully designed sampling strategies and contrastive loss functions, our approach ensures that representations are effectively optimized both within and across scales.
- **Comprehensive Evaluation:** We conduct extensive experiments on two real-world datasets and four downstream tasks, demonstrating that UniTR outperforms existing methods in terms of representation quality and downstream task performance, thus validating the effectiveness of the proposed framework.

Related Work

Representation for Trajectory and Road Network. Representation learning for spatial-temporal data, particularly trajectories and road networks, has been extensively explored due to its importance in urban analysis (Chen et al. 2024b; Deng et al. 2024b). Trajectory Representation methods primarily focus on capturing movement behaviors. For instance, Traj2Vec (Yao et al. 2017) uses spatial and temporal windows to convert GPS trajectories into feature vectors. Subsequent works have refined this approach with tailored windows and data augmentation strategies, such as those in (Yao et al. 2019; Li et al. 2018; Yang et al. 2021a; Chang et al. 2023), while TrajFormer (Liang et al. 2022b) introduces continuous point embedding for transformer-based models. On the other hand, Road Network Representation methods leverage graph-based techniques to capture structural and geospatial properties. Early methods like DeepWalk (Perozzi, Al-Rfou, and Skiena 2014) and Node2Vec (Grover and Leskovec 2016) focused on topological embeddings, while more recent approaches utilize graph neural networks (Kipf and Welling 2017) to integrate node and edge features, addressing complex spatial dependencies (Wang et al. 2019; Wu et al. 2020; Zhang and Long 2023). However, these methods typically treat trajectories and road networks as separate entities, failing to account for the interactions between them.

Joint Representation Learning. Acknowledging the interdependence between road networks and trajectories, recent research has increasingly adopted joint representation learning approaches. For example, Toast (Chen et al. 2021) and GTS (Han et al. 2021) integrate auxiliary traffic context and point-of-interest embeddings, respectively, to enhance trajectory representations. Recently, successful progress in contrastive learning for time series (Luo et al. 2023; Lai et al. 2024) and graph learning (Lee and Shin 2023; Xiao et al. 2024) has provided new opportunities for representation learning (Deng et al. 2024a). Base on that, some methods apply contrastive learning and other self-supervised techniques for jointly embedding, employing multi-task learning frameworks with supervision signals like contrastive, autoregressive, and masking losses (Fu and Lee 2020; Liang et al. 2022a; Yang et al. 2021b; Jiang et al. 2023; Mao et al. 2022; Ma et al. 2024). Despite these advancements, effectively managing the many-to-many interactions between trajectories and road networks remains a challenge.

Preliminaries

Definition 1 (Road network) *A road network is denoted as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$. Each vertex $v \in \mathcal{V}$ corresponds to a road segment, typically associated with attributes such as length and coordinates. Each edge $e_{i,j} \in \mathcal{E}$ indicates a direct connection between the road segments v_i and v_j. $\mathbf{A}$ is a binary adjacency matrix that indicates whether an edge exists between road segments.*

Definition 2 (Trajectory) *A trajectory $\tau = \langle p_1, p_2, \ldots, p_{|\tau|}\rangle$ is a sequence of GPS points recorded by a moving object. It can be mapped onto the road network $\mathcal{G}$, resulting in a road segment-based trajectory $\langle v_1, v_2, \ldots, v_n\rangle$.*

Definition 3 (Street view image) *A street view image is a photograph captured at a specific location on the road network, offering a detailed visual snapshot of the surroundings. Specifically, we collect multiple images on each road segment and create a 360-degree view, denoted as $v.img$.*

Problem Statement Given a road network $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ and a set of historical trajectories $\mathcal{T} = \{\tau_i\}_{i=1}^{|\mathcal{T}|}$, the objective is to learn the mapping functions, $\mathcal{F}_v : \mathcal{V} \to \mathbb{R}^d$ and $\mathcal{F}_\tau : \mathcal{T} \to \mathbb{R}^d$, which embed each road segment v_i and each trajectory τ_i into two d-dimensional representation vectors.

Methodology

As depicted in Figure 2, UniTR contains four sequential stages: *Data Preparation, Data Augmentation, Joint Representation Learning, and Joint Contrastive Optimization.*

Data Preparation and Augmentation

Preparation The data preparation stage involves two key processes: segment feature extraction and trajectory preprocessing. First, road segment features are extracted by encoding road attributes such as segment ID and length using linear layers, while street view images are processed through SwinTransformers (Liu et al. 2021). The resulting features are concatenated to form the multi-modal feature matrix (i.e., $\mathbf{X}$). Meanwhile, GPS trajectories are mapped onto the

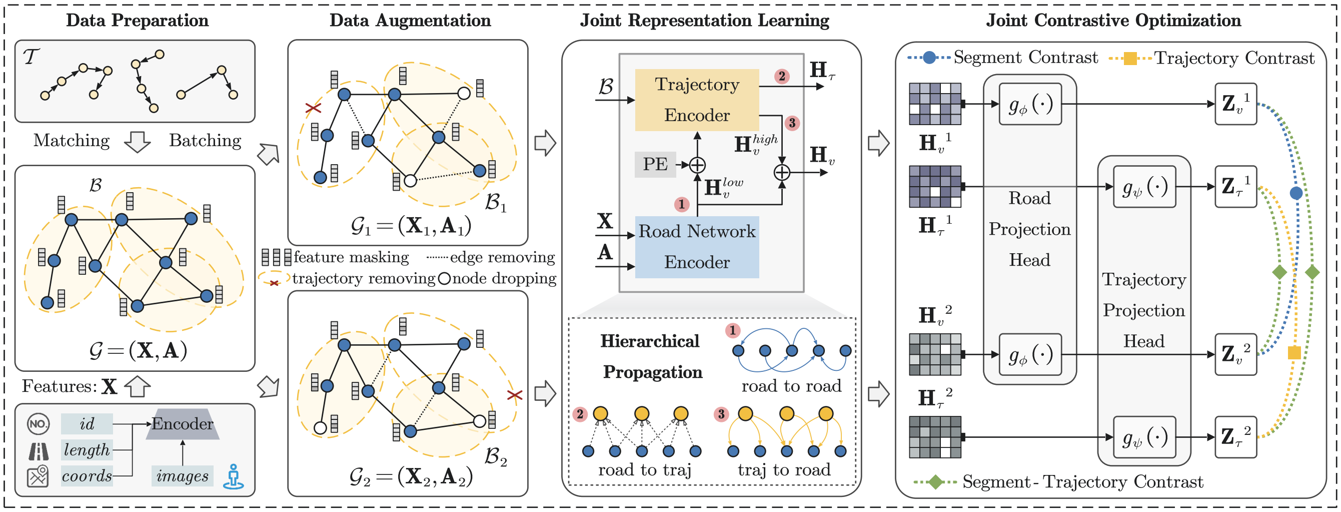


Figure 2: The framework of UniTR, which includes modules for data preparation and augmentation, follows up with joint representation learning and contrastive optimization.

road network by map matching algorithm (Yang and Gidofalvi 2018). These matched trajectories are then organized into batches (i.e., $\mathcal{B}$), to facilitate efficient post-processing.

However, both road networks and trajectory data are subject to various inconsistencies and uncertainties in real-world urban environments, such as missing attributes, incomplete street view images, or temporal changes due to roadworks. To build robust representations that can handle these challenges, we implement a series of data augmentation strategies. These strategies are designed to introduce controlled perturbations into the data, helping the model learn to generalize and maintain performance in the face of incomplete or noisy information. Specifically, we apply Road Network Augmentation and Trajectory Augmentation techniques, which aim to simulate potential variations and uncertainties in road networks and trajectory data.

Road Network Augmentation To address the inherent challenges in road network data, such as missing attributes and dynamic topology changes, we implement two tailored augmentation techniques:

- *Multi-modal feature masking* randomly masks either attribute or visual features from road segment data. By setting a selected modality to a zero vector, the model is encouraged to learn more robust representations that do not rely on complete information. Given attribute feature $\mathbf{x}_i^{attr}$ and visual feature $\mathbf{x}_i^{vis}$ of road segment v_i, the masked features can be expressed as:

$$\mathbf{m}_i^{attr} \odot \mathbf{x}_i^{attr} \,||\, \mathbf{m}_i^{vis} \odot \mathbf{x}_i^{vis}, \tag{1}$$

  where $\odot$ denotes element-wise product, $\mathbf{m}_i^{attr}$ and $\mathbf{m}_i^{vis}$ are two binary mask vectors.
- *Mobility-based edge removing* introduces structural perturbations into the road network by selectively removing edges based on their mobility importance. Unlike random edge modifications, this approach uses historical trajectory data to determine which connections are less crucial for the network functioning. For an edge $e_{i,j}$ from v_i to v_j, the transition probability $p(v_j \mid v_i)$ is calculated as:

$$p(v_j \mid v_i) = \frac{\#trans(v_i \rightarrow v_j)}{\#visit(v_i)}. \tag{2}$$

  The removing probability of $e_{i,j}$ is then defined as:

$$pr(e_{i,j}) = \delta_\epsilon(1 - p(v_j \mid v_i)), \tag{3}$$

  where δ_ϵ is a linear transformation mapping to $[\epsilon, 1-\epsilon]$, ensuring the removing probability never equals 1, even if there is no transition between the segments.

Trajectory Augmentation Trajectory data is also prone to various inconsistencies, such as varying sampling rates and missing records, which can significantly affect the accuracy of movement pattern modeling. To address these issues, we employ two augmentation strategies:

- *Road segment dropping* involves randomly removing road segments from a trajectory. This technique simulates scenarios where trajectory data are corrupted or missing, which encourages the model to learn consistent movement patterns despite the incomplete data.
- *Trajectory removing* adds another layer of variability by randomly excluding entire trajectories from the batch $\mathcal{B}$. This strategy prevents the model from becoming overly reliant on specific trajectory sequences, promoting its ability to generalize to unseen movement patterns.

Joint Representation Learning

Given the intricate and deeply intertwined nature of road networks and trajectory data, effectively capturing the interactions between these two types of data is pivotal for joint representation learning. In this work, we design a Road Network Encoder and a Trajectory Encoder that work together through a hierarchical propagation mechanism, enabling the modeling of *many-to-many* interactions between road networks and trajectories.

Road Network Encoder This encoder is designed to capture the low-order structural representations $\mathbf{H}_v^{low}$ of road segments, focusing on *road-to-road propagation*. Specifically, it utilizes Chebyshev graph convolution (Defferrard, Bresson, and Vandergheynst 2016) to encode the structural information inherent in the road network:

$$\mathbf{H}_v^{low} = \sum_{k=0}^{K-1} T_k(\hat{\mathbf{L}})\mathbf{X}\mathbf{\Theta}_k, \tag{4}$$

where $\hat{\mathbf{L}} = \frac{2\mathbf{L}}{\lambda_{max}} - \mathbf{I}$ represents the normalized Laplacian matrix derived from the adjacency matrix $\mathbf{A}$, with λ_{max} as its largest eigenvalue. T_k is the Chebyshev polynomials.

Trajectory Encoder This encoder aims to: 1) learn trajectory representations by aggregating consecutive road segment representations through *road-to-trajectory propagation*, and 2) supplement road segment representations by incorporating high-order relationships discovered in the trajectories through *trajectory-to-road propagation*.

For a given trajectory τ_j, the encoder first processes the low-order representations and positional encoding of the road segments within the trajectory, and computes the trajectory representation as:

$$\mathbf{h}_{\tau_j} = \sigma\left(\sum_{v_i \in \tau_j} \frac{(\mathbf{h}_{v_i}^{low} + \mathbf{PE}(i))\mathbf{\Theta}_{\tau_j}}{|\tau_j|} + \mathbf{b}_{\tau_j}\right), \tag{5}$$

where $\mathbf{\Theta}_{\tau_j}$ and $\mathbf{b}_{\tau_j}$ are learnable parameters, σ is the activation function, and $\mathbf{PE}(i)$ is the positional encoding function. Subsequently, the learned trajectory representations $\mathbf{h}_{\tau_j}$ are feedback to the corresponding road segments through a convolutional layer in the opposite direction, resulting in high-order representations for each road segment:

$$\mathbf{h}_{v_i}^{high} = \sigma\left(\sum_{\tau_j : v_i \in \tau_j} \frac{\mathbf{h}_{\tau_j}\mathbf{\Theta}_{v_i}}{degree(v_i)} + \mathbf{b}_{v_i}\right). \tag{6}$$

Finally, a comprehensive and rich representation $\mathbf{H}_v = \mathbf{H}_v^{low} + \mathbf{H}_v^{high}$ of road segments is formed, which used together with the batched trajectory data representations $\mathbf{H}_\tau$ for follow-up joint optimization.

Joint Contrastive Optimization

As shown in Figure 2, prior to applying contrastive learning, we employ two projection heads, $g_v(\cdot)$ and $g_\tau(\cdot)$ to map the representations into a new latent space where the loss computation occurs. Specifically, the representations $(\mathbf{H}_v^1, \mathbf{H}_v^2)$ and $(\mathbf{H}_\tau^1, \mathbf{H}_\tau^2)$ are projected individually, yielding two sets of projected representations: $(\mathbf{Z}_v^1, \mathbf{Z}_\tau^1)$ and $(\mathbf{Z}_v^2, \mathbf{Z}_\tau^2)$. However, jointly optimizing their representations is challenging due to the difficulty in positive/negative sample mining and the misalignment of their embedding spaces. Motivated by previous study (Lee and Shin 2023), we design: 1) a proximity and mobility-aware sampling strategy to accurately identify valid contrast pairs; and 2) a triple-level contrastive loss that optimizes representations both within individual scales and across the road network and trajectory scales, ensuring more cohesive and meaningful representations.

Proximity and Mobility-Aware Sampling Identifying valid positive and negative samples for road networks and trajectories is a non-trivial task due to the inherent complex spatial and temporal dependencies. To this end, we introduce a strategy that leverages geographic proximity and mobility semantics to ensure that the selected samples are both spatially and semantically relevant.

(1) *Sample Selection for Road Network*. For each road segment v_i in the road network graph $\mathcal{G}_1$, we identify positive samples $\mathbf{z}_{v_i}^{1,+}$ from two sources: neighboring road segments within a specific geographic proximity and the corresponding road segment in the alternate view. Formally:

$$\mathbf{z}_{v_i}^{1,+} = \{\mathbf{z}_{v_j}^1 \mid v_j \in \mathcal{N}_q^1(v_i)\} \cup \{\mathbf{z}_{v_i}^2\}, \tag{7}$$

where $\mathcal{N}_q^1(v_i)$ denotes the q-hop neighbors of v_i in $\mathcal{G}_1$. Accordingly, negative samples $\mathbf{z}_{v_i}^{1,-}$ are selected with two criteria: 1) they lie outside the q-hop neighborhood of the target node, and 2) they do not co-occur with the target node in the current trajectory batch:

$$\begin{aligned}\mathbf{z}_{v_i}^{1,-} = &\{\mathbf{z}_{v_j}^1 \mid v_j \notin \mathcal{N}_q^1(v_i) \wedge v_j \notin \mathcal{B}_1(v_i)\} \cup \\ &\{\mathbf{z}_{v_j}^2 \mid v_j \notin \mathcal{N}_q^2(v_i) \wedge v_j \notin \mathcal{B}_2(v_i)\}.\end{aligned} \tag{8}$$

(2) *Sample Selection for Trajectories*. Similarly, for each trajectory $\tau_i \in \mathcal{B}_1$, we select positive samples $\mathbf{z}_{\tau_i}^{1,+}$ based on the similarity of their movement paths to τ_i. This involves identifying trajectories with high path similarity within the same view and the corresponding trajectory in the other view $\mathcal{B}_2$. Negative samples $\mathbf{z}_{\tau_i}^{1,-}$ are selected from trajectories that are disjoint, ensuring they offer a clear contrast to the positive examples. The above process can be described as:

$$\begin{aligned}\mathbf{z}_{\tau_i}^{1,+} &= \{\mathbf{z}_{\tau_j}^1 \mid \text{Jac}(\tau_i, \tau_j) \geq \theta\} \cup \{\mathbf{z}_{\tau_i}^2\}, \\ \mathbf{z}_{\tau_i}^{1,-} &= \{\mathbf{z}_{\tau_j}^1 \mid \text{Jac}(\tau_i, \tau_j) = 0 \wedge \tau_j \in \mathcal{B}_1\} \cup \\ &\quad \{\mathbf{z}_{\tau_j}^2 \mid \text{Jac}(\tau_i, \tau_j) = 0 \wedge \tau_j \in \mathcal{B}_2\},\end{aligned} \tag{9}$$

where Jac$(\cdot)$ denotes the measure of Jaccard similarity coefficient, θ is a predefined threshold.

Triple Level Contrastive Loss To align representations cross scales, we design the contrastive loss at segment, trajectory, and cross level. Each level is designed to optimize the representations by focusing on specific aspects of the spatial-temporal data, ensuring that the learned embeddings are both cohesive and discriminative across different scales. For clarity, we first present the similarity measure:

$$s(\mathbf{z}_a, \mathbf{z}_b) = e^{cos(\mathbf{z}_a, \mathbf{z}_b)/x}, \tag{10}$$

where $cos(\cdot)$ represents the cosine similarity between two embeddings, x is the temperature parameter.

(1) *Segment Level Contrast*. At this level, the goal is to enhance the ability to distinguish between different road segments based on their contextual relationships. For each road segment v_i in view $\mathcal{G}_1$, the contrastive loss is designed to maximize the similarity between the anchor embedding $\mathbf{z}_{v_i}^1$ and its corresponding positive samples $\mathbf{z}_{v_i}^{1,+}$, while minimizing the similarity with the negative samples $\mathbf{z}_{v_i}^{1,-}$, that is:

$$l_v(\mathbf{z}_{v,i}^1) = -\log \frac{\sum_{\mathbf{z} \in \mathbf{z}_{v,i}^{1,+}} s(\mathbf{z}_{v,i}^1, \mathbf{z})}{\sum_{\mathbf{z} \in \mathbf{z}_{v,i}^{1,+}} s(\mathbf{z}_{v,i}^1, \mathbf{z}) + \sum_{\mathbf{z} \in \mathbf{z}_{v,i}^{1,-}} s(\mathbf{z}_{v,i}^1, \mathbf{z})}. \tag{11}$$

The segment level loss is then averaged across all segments in both road network views:

$$\mathcal{L}_v = \frac{1}{2|V|} \sum_{i=1}^{|V|} l_v(\mathbf{z}_{v,i}^1) + l_v(\mathbf{z}_{v,i}^2) \tag{12}$$

(2) *Trajectory Level Contrast*. The trajectory level contrast focuses on refining the understanding of movement patterns by optimizing the relationship between trajectories. For each trajectory τ_i in view $\mathcal{B}_1$, the contrastive loss is designed to enhance the agreements between the anchor embedding $\mathbf{z}_{\tau_i}^1$ and its corresponding positive samples in $\mathbf{z}_{\tau_i}^{1,+}$, while pushing away negative samples in $\mathbf{z}_{\tau_i}^{1,-}$:

$$l_v(\mathbf{z}_{\tau,i}^1) = -\log \frac{\sum_{\mathbf{z} \in \mathbf{z}_{\tau,i}^{1,+}} \boldsymbol{s}(\mathbf{z}_{\tau,i}^1, \mathbf{z})}{\sum_{\mathbf{z} \in \mathbf{z}_{\tau,i}^{1,+}} \boldsymbol{s}(\mathbf{z}_{\tau,i}^1, \mathbf{z}) + \sum_{\mathbf{z} \in \mathbf{z}_{\tau,i}^{1,-}} \boldsymbol{s}(\mathbf{z}_{\tau,i}^1, \mathbf{z})}. \tag{13}$$

The trajectory level contrastive loss is obtained by averaging across all trajectories in both trajectory batch views:

$$\mathcal{L}_\tau = \frac{1}{2|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} l_\tau(\mathbf{z}_{\tau,i}^1) + l_\tau(\mathbf{z}_{\tau,i}^2) \tag{14}$$

(3) *Cross Level Contrast*. This type of contrast aims to narrow the gap between road segments and the trajectories that traverse them. The loss maximizes similarity between road segments in $\mathcal{G}_1$ and trajectories in $\mathcal{B}_2$ when they are related, and minimizes it otherwise, and vice versa for trajectories in $\mathcal{B}_2$ and road segments in $\mathcal{G}_1$. Formally:

$$\begin{aligned} l_c(\mathbf{z}_{v,i}^1, \mathbf{z}_{\tau,j}^2) = &-\log \frac{\mathcal{D}(\mathbf{z}_{v,i}^1, \mathbf{z}_{\tau,j}^2)}{\mathcal{D}(\mathbf{z}_{v,i}^1, \mathbf{z}_{\tau,j}^2) + \sum_{k: i \notin k} \mathcal{D}(\mathbf{z}_{v,i}^1, \mathbf{z}_{\tau,k}^2)} \\ &-\log \frac{\mathcal{D}(\mathbf{z}_{\tau,j}^2, \mathbf{z}_{v,i}^1)}{\mathcal{D}(\mathbf{z}_{\tau,j}^2, \mathbf{z}_{v,i}^1)) + \sum_{k: k \notin j} \mathcal{D}(\mathbf{z}_{\tau,j}^2, \mathbf{z}_{v,k}^1)}, \end{aligned} \tag{15}$$

where $\mathcal{D}$ is a discriminator implemented by a combination of a bilinear layer and a sigmoid activation, which quantifies the similarity probability of road segment-trajectory pairs. The overall cross level contrastive loss is computed by summing the losses from both views, and averaging across all road segment-trajectory pairs that co-occur:

$$\mathcal{L}_c = \frac{1}{2|N|} \sum_{i=1}^{|\mathcal{V}|} \sum_{j=1}^{|\mathcal{B}|} \mathbb{I}_{v_i \in \tau_j} (l_c(\mathbf{z}_{v,i}^1, \mathbf{z}_{\tau,j}^2) + l_c(\mathbf{z}_{v,i}^2, \mathbf{z}_{\tau,j}^1)), \tag{16}$$

where $N = \sum_{i=1}^{|\mathcal{V}|} \sum_{j=1}^{|\mathcal{B}|} \mathbb{I}_{v_i \in \tau_j}$ represents the total number of road segment-trajectory pairs that co-occur.

As a result, the final objective function combines the segment, trajectory, and cross level contrastive losses to ensure comprehensive representation learning:

$$\mathcal{L} = \lambda_v \mathcal{L}_v + \lambda_\tau \mathcal{L}_\tau + \lambda_c \mathcal{L}_c, \tag{17}$$

where λ_v, λ_τ and λ_c are weighting factors that balance the contributions of each loss.

Experiments

Datasets We conduct the experiment on two real-world datasets from two distinct sources: a public dataset for Porto, acquired from a Kaggle competition (Kaggle 2015), and a private one for Chengdu, provided by Didi Chuxing. Each dataset comprises road network data, GPS trajectories, and street view images. The road networks for both cities are sourced from OpenStreetMap (OpenStreetMap Contributors 2024), and street-view panoramic images are obtained via the APIs of Google Maps and Baidu Maps. The statistics of these datasets are summarized in Table 1.

Datasets	Porto	Chengdu
# road segments	10,780	6,786
# road connections	24,980	17,542
# trajectories	1,710,670	5,819,383
# street view images	10,780	6,786

Table 1: Statistics of the datasets.

Downstream Tasks and Metrics To assess the effectiveness of our representation learning model, we employ four downstream tasks: two related to road segments and two focused on trajectories (Ma et al. 2024; Mao et al. 2022).

- *Road Type Classification*. Classifies each road segment by type, using a fully connected layer with a Softmax layer. Performance is measured by Micro-F1 (Mi-F1) and Macro-F1 (Ma-F1) scores.
- *Traffic Speed Inference*. Estimates average traffic speed per road segment with a linear regression model, evaluated by Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
- *Travel Time Estimation*. Predicts travel time for a trajectory using a 3-layer Multi-Layer Perceptron, with MAE and RMSE as metrics.
- *Trajectory Similarity Search*. Retrieves top-k similar trajectories for a query trajectory, using HR@10 and Mean Rank (MR) for evaluation.

Baselines We compare our model with 6 baselines, including two random walk-based methods: Node2Vec (Grover and Leskovec 2016) and SRN2Vec (Wang et al. 2020), two GNN-based methods: RFN (Jepsen, Jensen, and Nielsen 2020) and HRNR (Wu et al. 2020), a graph contrastive learning-based method: SARN (Chang et al. 2023), and a joint representation learning method: JCLRNT (Mao et al. 2022). The first five baselines focus on learning road network representations, while the last one jointly learns representations for both road networks and trajectories.

Performance on Downstream Tasks

This section presents a comparison of our model, UniTR, against baseline methods across four downstream tasks. The overall results are given in Table 2, we can find that:

- In the **road segment-based tasks**, UniTR significantly outperforms all baselines. Specifically, UniTR achieves

Task	Road Type Classification		Traffic Speed Inference		Travel Time Estimation		Similar Trajectory Search	
Dataset	Porto	Chengdu	Porto	Chengdu	Porto	Chengdu	Porto	Chengdu
Metric	Mi-F1/Ma-F1	Mi-F1/Ma-F1	MAE/RMSE	MAE/RMSE	MAE/RMSE	MAE/RMSE	MR/HR@10	MR/HR@10
Node2Vec	0.430/0.377	0.467/0.455	3.69/4.89	7.30/8.28	202.4/293.1	137.2/203.2	25.265/0.712	43.062/0.659
SRN2Vec	0.511/0.286	0.442/0.391	2.45/3.23	3.42/4.57	202.0/294.3	226.5/312.2	11.822/0.914	14.410/0.861
RFN	0.501/0.228	0.431/0.327	3.25/4.55	6.79/7.78	163.0/256.7	162.7/235.4	31.654/0.724	54.181/0.639
HRNR	0.409/0.296	0.410/0.370	3.56/4.80	7.24/8.22	205.8/295.7	144.8/212.5	24.743/0.716	46.780/0.653
SARN	0.510/0.305	0.507/0.404	3.50/4.76	7.12/8.11	167.5/263.7	162.1/234.0	28.356/0.703	52.356/0.638
JCLRNT	0.593/0.402	0.634/0.620	2.39/3.29	3.76/4.83	159.3/251.3	120.6/180.8	5.003/0.939	7.457/0.912
UniTR	**0.638/0.564**	**0.668/0.664**	**2.10/2.91**	**3.20/4.26**	**144.6/209.9**	**110.5/162.7**	**1.740/0.984**	**3.002/0.958**

Table 2: Performance comparison on downstream tasks.

Task	Road Type Classification		Traffic Speed Inference		Travel Time Estimation		Similar Trajectory Search	
Metric	Mi-F1	Ma-F1	MAE	RMSE	MAE	RMSE	MR	HR@10
UniTR	0.638	0.564	2.101	2.909	144.6	209.9	1.740	0.984
w/o vis	0.613 ↓	0.497 ↓	1.978 ↑	3.026 ↓	142.7 ↑	211.9 ↑	2.049 ↓	0.946 ↓
w/o t2r	0.600 ↓	0.472 ↓	2.144 ↓	2.973 ↓	145.9 ↑	224.1 ↓	1.964 ↓	0.976 ↓
w/o $\mathcal{L}_\tau$	0.613 ↓	0.505 ↓	2.221 ↓	3.108 ↓	146.6 ↑	214.1 ↓	1.742 ↓	0.983 ↓
w/o $\mathcal{L}_c$	0.633 ↓	0.532 ↓	2.177 ↓	3.046 ↓	142.4 ↑	209.3 ↑	1.756 ↓	0.983 ↓

Table 3: Ablation study on four downstream tasks in Porto dataset. ↓ denotes a decrease in performance, and ↑ an improvement.

the highest Mi-F1 and Ma-F1 scores in road type classification and the lowest MAE and RMSE in traffic speed inference, underscoring the robustness of its joint learning framework. SRN2Vec enhance the Node2Vec model with real trajectory data but remains inferior to UniTR. GNN-based methods like RFN, HRNR and SARN, perform reasonably well but are still limited by under-utilizing trajectory information. The superior performance of UniTR highlights the advantages of joint representation learning that fully exploits both road network structure and mobility patterns.

- For the **trajectory-based tasks**, the first five baselines, which are not designed to specifically learn trajectory representations, underperform as expected. In contrast, joint learning models like JCLRNT and UniTR, which both employ joint representation learning, significantly surpass the others. Notably, UniTR outperforms JCLRNT in all trajectory-based metrics, highlighting its superior ability to model the complex interactions between roads and trajectories. This advantage is largely attributed to the hierarchical propagation mechanism and cross-scale comparative learning of UniTR, which enables it to capture more fine-grained relationships between road network and trajectories.

To summarize, the strong performance across all tasks and datasets underscores the effectiveness of UniTR. By simultaneously capturing the structure of road networks and the mobility patterns of trajectories, UniTR generates robust and generic representations that excel in diverse applications. In addition, most of the metrics are better on the Chengdu dataset with smaller road network scale and high trajectory sampling rate. It means that smaller road network with finer-grained trajectory data appear to enhance the performance by providing detailed spatial-temporal information.

Ablation Study

To assess the contribution of various components in UniTR, we conducted ablation experiments on four variants of the model: 1) **w/o vis**: removing visual information from the road segment features; 2) **w/o t2r**: excluding the trajectory-to-road propagation in the trajectory encoder; 3) **w/o** $\mathcal{L}_\tau$: omitting the trajectory level contrastive loss; and 4) **w/o** $\mathcal{L}_c$: removing the cross level contrastive loss.

The results of ablation experiments on the Porto dataset are presented in Table 3. We can find that each component of UniTR plays a critical role in its overall performance. The removal of visual features (**w/o vis**) leads to a noticeable performance decline in road type classification, suggesting that street view images are particularly valuable for this task. Excluding the trajectory-to-road propagation (**w/o t2r**) results in significant performance drops in traffic speed inference and trajectory similarity search, underscoring the importance of dynamic interaction modeling between trajectories and road segments. The information of road network and trajectory data is indeed complementary, enhancing representation quality and task performance. Moreover, The omission of the trajectory level contrastive loss (**w/o** $\mathcal{L}_\tau$) primarily affects trajectory-based tasks, indicating its crucial role in accurately capturing movement patterns. Similarly, removing the cross level contrastive loss (**w/o** $\mathcal{L}_c$) causes moderate performance declines across all tasks, demonstrating its contribution to aligning road segment and trajectory representations. Notably, none of the ablated variants outperforms the UniTR model across all tasks, highlighting the

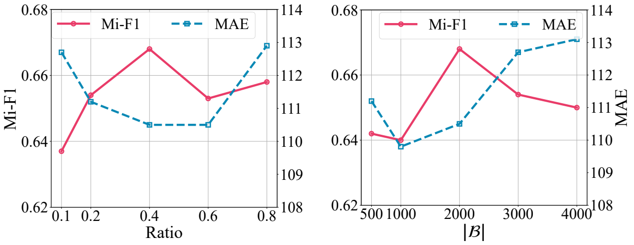


Figure 3: Effects of road segment dropping ratio and trajectory batch size ($|\mathcal{B}|$).

necessity of integrating all components to achieve balanced and robust performance.

Parameter Analysis

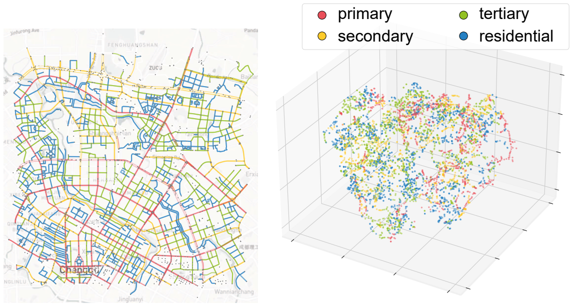
We analyze the impact of two crucial parameters of UniTR: the dropping ratio of road segments and the size of trajectory batch. The evaluation is conducted on the Chengdu dataset, with performance measured by Mi-F1 for road type classification and MAE for travel time estimation. The results are illustrated in Figure 3. For the dropping ratio, we observe a moderate dropping ratio around 0.4, which yields the best performance for both road type classification and travel time estimation. This finding suggests that strategically dropping a portion of road segments encourages the model to learn more robust and effective representations. As the ratio exceeds this optimal point, a noticeable decline in performance occurs. This decline likely results from the excessive loss of critical information, which impairs the ability of UniTR to make accurate predictions. Regarding the trajectory batch size, the result reveals that a mid-sized trajectory batch (around 2000 trajectories) strikes a well balance between learning efficiency and model performance. Smaller size may lack sufficient diversity in trajectory patterns, leading to underfitting. Conversely, larger batch size might introduce excessive noise or complexity, causing the model to struggle with efficient learning. Thus, the trajectory batch size is a key factor in determining the capacity to effectively capture the different movement patterns.

Computational Complexity Analysis

The proposed UniTR incurs computational costs in the following aspects: 1) $\mathcal{O}\left(|\mathcal{V}|F + |\mathcal{E}|\right)$ for road network augmentation, $\mathcal{O}\left(|\mathcal{B}| + |\mathcal{B}| \cdot |\tau|\right)$ for trajectory augmentation; 2) $\mathcal{O}\left((|\mathcal{E}| + |\mathcal{V}|)Fd^2\right)$ for road network encoder, $\mathcal{O}\left((|\mathcal{B}| + |\mathcal{V}|)Fd^2\right)$ for trajectory encoder; 3) $\mathcal{O}\left(|\mathcal{V}|^2 d^2\right)$ for road level contrast, $\mathcal{O}\left(|\mathcal{B}|^2 d^2\right)$ for trajectory level contrast and $\mathcal{O}\left(|\mathcal{V}| \cdot |\mathcal{E}| d^2\right)$ for cross level contrast. Herein, F and d denotes the dimension of hidden state and embedding vector, respectively. Overall, the model complexity is mainly influenced by road network scale and trajectory batch size.

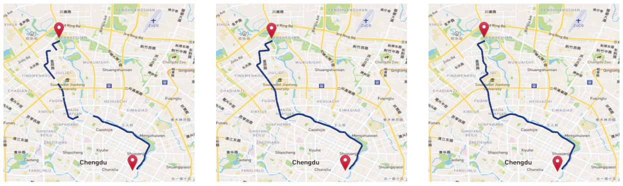
Case Study

Embedding Visualization We visualize the embeddings learned by our model, and observe them in the context of road type classification on Chengdu dataset. As depicted in Figure 4, the embeddings show clear clustering according to road types, with primary and secondary roads distinctly separated from residential and tertiary roads. This demonstrates that UniTR is able to learn meaningful representations from multi-modal data for differentiating road types.



(a) Road map of Chengdu. (b) 3D visualization of embeddings.

Figure 4: Road types and embedding visualization



(a) Query trajectory. (b) 1st retrieved one. (c) 3rd retrieved one.

Figure 5: Illustration of trajectory search.

Similar Trajectory Search Figure 5 shows the query trajectory and the top three retrieved trajectories. The retrieved trajectories closely align with the query, demonstrating that the embeddings effectively capture the underlying mobility patterns and spatial structures, resulting in accurate similarity retrieval. This confirms our model's capability to learn robust trajectory representations that generalize well to similar trajectory identification tasks.

Conclusion

In this paper, we propose a unified framework entitled UniTR for joint representation learning for road network and trajectory data. Specifically, we design a Road Network Encoder and a Trajectory Encoder that work together through a hierarchical propagation mechanism, modeling many-to-many interactions. In addition, we propose a proximity and mobility-aware sampling strategy to identify valid contrast pairs and design a triple-level contrastive loss for representation optimization. Experimental results show that the representation learned by UniTR outperforms existing methods in terms of downstream task performance. In future work, we plan to conduct further exploration from the perspectives of downstream task extension and cross-city generalization.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant 62322601, 62172066, and 62402414, the Independent Research Project of State Key Laboratory of Mechanical Transmission for Advanced Equipment under Grant SKLMT-ZZKT-2024R07, the Fundamental Research Funds for the Central Universities 2024IAIS-QN017. This paper was done when the first author was a research intern at HKUST(GZ).

References

Chang, Y.; Qi, J.; Liang, Y.; and Tanin, E. 2023. Contrastive trajectory similarity learning with dual-feature attention. In *IEEE International Conference on Data Engineering*, 2933–2945.

Chen, W.; Liang, Y.; Zhu, Y.; Chang, Y.; Luo, K.; Wen, H.; Li, L.; Yu, Y.; Wen, Q.; Chen, C.; et al. 2024a. Deep learning for trajectory data management and mining: A survey and beyond. *arXiv preprint arXiv:2403.14151*.

Chen, Y.; Huang, W.; Zhao, K.; Jiang, Y.; and Cong, G. 2024b. Self-supervised Learning for Geospatial AI: A Survey. *arXiv preprint arXiv:2408.12133*.

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 *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 211–220.

Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3844–3852.

Deng, M.; Chen, C.; Zhang, W.; Zhao, J.; Yang, W.; Guo, S.; Pu, H.; and Luo, J. 2024a. HyperRegion: Integrating Graph and Hypergraph Contrastive Learning for Region Embeddings. *IEEE Transactions on Mobile Computing*.

Deng, M.; Zhang, W.; Zhao, J.; Wang, Z.; Zhou, M.; Luo, J.; and Chen, C. 2024b. A Novel Framework for Joint Learning of City Region Partition and Representation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(7): 1–23.

Fu, T.-Y.; and Lee, W.-C. 2020. Trembr: Exploring road networks for trajectory representation learning. *ACM Transactions on Intelligent Systems and Technology*, 11(1): 1–25.

Grover, A.; and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 855–864.

Han, P.; Wang, J.; Yao, D.; Shang, S.; and Zhang, X. 2021. A graph-based approach for trajectory similarity computation in spatial networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 556–564.

Jepsen, T. S.; Jensen, C. S.; and Nielsen, T. D. 2020. Relational fusion networks: Graph convolutional networks for road networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(1): 418–429.

Jiang, J.; Pan, D.; Ren, H.; Jiang, X.; Li, C.; and Wang, J. 2023. Self-supervised trajectory representation learning with temporal regularities and travel semantics. In *IEEE International Conference on Data Engineering*, 843–855.

Kaggle. 2015. ECML/PKDD 15 Taxi Trajectory Prediction. https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

Lai, Z.; Li, H.; Zhang, D.; Zhao, Y.; Qian, W.; and Jensen, C. S. 2024. E2USD: Efficient-yet-effective Unsupervised State Detection for Multivariate Time Series. In *Proceedings of the ACM on Web Conference 2024*, 3010–3021.

Lee, D.; and Shin, K. 2023. I'm me, we're us, and i'm us: Tri-directional contrastive learning on hypergraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8456–8464.

Li, X.; Zhao, K.; Cong, G.; Jensen, C. S.; and Wei, W. 2018. Deep representation learning for trajectory similarity computation. In *IEEE International Conference on Data Engineering*, 617–628.

Liang, X.; Zhu, F.; Zhu, Y.; Lin, B.; Wang, B.; and Liang, X. 2022a. Contrastive instruction-trajectory learning for vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1592–1600.

Liang, Y.; Ouyang, K.; Wang, Y.; Liu, X.; Chen, H.; Zhang, J.; Zheng, Y.; and Zimmermann, R. 2022b. TrajFormer: Efficient trajectory classification with transformers. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 1229–1237.

Liu, H.; Han, J.; Fu, Y.; Li, Y.; Chen, K.; and Xiong, H. 2023. Unified route representation learning for multi-modal transportation recommendation with spatiotemporal pre-training. *The VLDB Journal*, 32(2): 325–342.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Luca, M.; Barlacchi, G.; Lepri, B.; and Pappalardo, L. 2021. A survey on deep learning for human mobility. *ACM Computing Surveys*, 55(1): 1–44.

Luo, D.; Cheng, W.; Wang, Y.; Xu, D.; Ni, J.; Yu, W.; Zhang, X.; Liu, Y.; Chen, Y.; Chen, H.; et al. 2023. Time series contrastive learning with information-aware augmentations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4534–4542.

Ma, Z.; Tu, Z.; Chen, X.; Zhang, Y.; Xia, D.; Zhou, G.; Chen, Y.; Zheng, Y.; and Gong, J. 2024. More Than Routing: Joint GPS and Route Modeling for Refine Trajectory Representation Learning. In *Proceedings of the ACM on Web Conference 2024*, 3064–3075.

Mao, Z.; Li, Z.; Li, D.; Bai, L.; and Zhao, R. 2022. Jointly contrastive representation learning on road network and trajectory. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 1501–1510.

OpenStreetMap Contributors. 2024. OpenStreetMap: The Free Wiki World Map. https://www.openstreetmap.org. Accessed: 2024-08-15.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 701–710.

Wang, M.-x.; Lee, W.-C.; Fu, T.-y.; and Yu, G. 2019. Learning embeddings of intersections on road networks. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 309–318.

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*, 12(1): 1–27.

Wang, S.; Cao, J.; and Philip, S. Y. 2020. Deep learning for spatio-temporal data mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(8): 3681–3700.

Wu, N.; Zhao, X. W.; Wang, J.; and Pan, D. 2020. Learning effective road network representation with hierarchical graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 6–14.

Xiao, T.; Zhu, H.; Chen, Z.; and Wang, S. 2024. Simple and asymmetric graph contrastive learning without augmentations. *Advances in Neural Information Processing Systems*, 36.

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, P.; Wang, H.; Zhang, Y.; Qin, L.; Zhang, W.; and Lin, X. 2021a. T3S: Effective representation learning for trajectory similarity computation. In *IEEE International Conference on Data Engineering*, 2183–2188.

Yang, S. B.; Guo, C.; Hu, J.; Tang, J.; and Yang, B. 2021b. Unsupervised path representation learning with curriculum negative sampling. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, 3286–3292.

Yao, D.; Cong, G.; Zhang, C.; and Bi, J. 2019. Computing trajectory similarity in linear time: A generic seed-guided neural metric learning approach. In *IEEE International Conference on Data Engineering*, 1358–1369.

Yao, D.; Zhang, C.; Zhu, Z.; Huang, J.; and Bi, J. 2017. Trajectory clustering via deep representation learning. In *IEEE International Joint Conference on Neural Networks*, 3880–3887.

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, J.; Chen, C.; Zhang, W.; Li, R.; Gu, F.; Guo, S.; Luo, J.; and Zheng, Y. 2023. Coupling makes better: an intertwined neural network for taxi and ridesourcing demand coprediction. *IEEE Transactions on Intelligent Transportation Systems*.