

Trajectory-User Linking via Hierarchical Spatio-Temporal Attention Networks

WEI CHEN, Ocean University of China, China

CHAO HUANG, The University of Hong Kong, China

YANWEI YU*, Ocean University of China, China

YONGGUO JIANG, Ocean University of China, China

JUNYU DONG, Ocean University of China, China

Trajectory-User Linking (TUL) is crucial for human mobility modeling by linking different trajectories to users with the exploration of complex mobility patterns. Existing works mainly rely on the recurrent neural framework to encode the temporal dependencies in trajectories, have fall short in capturing spatial-temporal global context for TUL prediction. To fill this gap, this work presents a new hierarchical spatio-temporal attention neural network, called *AttnTUL*, to jointly encode the local trajectory transitional patterns and global spatial dependencies for TUL. Specifically, our first model component is built over the graph neural architecture to preserve the local and global context and enhance the representation paradigm of geographical regions and user trajectories. Additionally, a hierarchically structured attention network is designed to simultaneously encode the intra-trajectory and inter-trajectory dependencies, with the integration of the temporal attention mechanism and global elastic attentional encoder. Extensive experiments demonstrate the superiority of our *AttnTUL* method as compared to state-of-the-art baselines on various trajectory datasets. The source code of our model is available at https://anonymous.4open.science/r/Attn_TUL.

CCS Concepts: • **Information systems** → **Spatial-temporal systems**; • **Computing methodologies** → **Learning latent representations**; *Neural networks*.

Additional Key Words and Phrases: Trajectory-user linking, attention neural networks, trajectory representation learning, spatio-temporal data

ACM Reference Format:

Wei Chen, Chao Huang, Yanwei Yu, Yongguo Jiang, and Junyu Dong. 2022. Trajectory-User Linking via Hierarchical Spatio-Temporal Attention Networks. *ACM Trans. Knowl. Discov. Data.* xx, xx, Article xxx (December 2022), 21 pages. <https://doi.org/10.1145/nnnnnnnn>.

1 INTRODUCTION

The development of mobile computing techniques enables the collection of massive mobility data from various sources, including location-based social networks, geo-tagged social media and GPS enabled mobile applications [30, 39]. Among various mobility modeling applications, a recently introduced *Trajectory-User Linking* (TUL) task, which aims to link

*Corresponding author.

Authors' addresses: Wei Chen, weichen@stu.ouc.edu.cn, Ocean University of China, Songling RD 238, Qingdao, Shandong, China, 266100; Chao Huang, chaohuang75@gmail.com, The University of Hong Kong, Pokfulam, Hong Kong, China; Yanwei Yu, yuyanwei@ouc.edu.cn, Ocean University of China, Songling RD 238, Qingdao, Shandong, China, 266100; Yongguo Jiang, jiangyg@ouc.edu.cn, Ocean University of China, Songling RD 238, Qingdao, Shandong, China, 266100; Junyu Dong, dongjunyu@ouc.edu.cn, Ocean University of China, Songling RD 238, Qingdao, Shandong, China, 266100.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

Manuscript submitted to ACM

Manuscript submitted to ACM

anonymous trajectories to users who generate them, is crucial and beneficial for a broad range of spatial-temporal mining applications, ranging from location-based recommendations [19] to urban anomaly detection [10, 38]. For example, vehicle-mounted terminals and ride-sharing services usually anonymize user identities in collecting mobility data due to the privacy issue. Effective linking trajectories with their corresponding users can not only enable accurate location-based recommendations, but also identify abnormal events from users' GPS traces. Therefore, the linking results between users and their anonymous trajectories are important for advancing the business intelligence and smart city applications [3, 5].

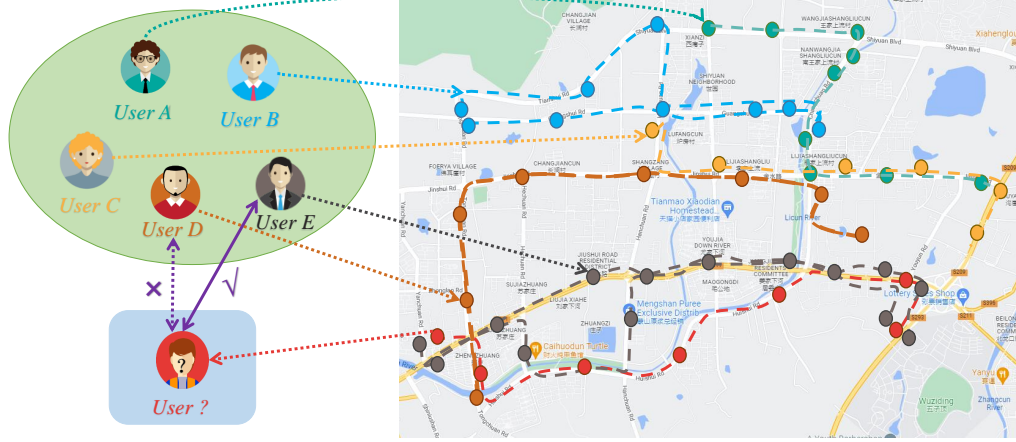


Fig. 1. An illustrated example of linking anonymous mobility trajectories with corresponding users. Based on the mobility trajectories of five users, we find that the anonymous trajectory (red) is much closer to the trajectory of user E (dark gray). Thus, we can link the anonymous trajectory with user E.

In this paper, we tackle the trajectory-user linking challenge via exploring the spatio-temporal movement patterns from the mobility data. Figure 1 presents an intuitive example of linking anonymous trajectories with the corresponding users. Our goal in this work is to link anonymous mobility trajectories to the users who are most likely to produce them based on the historical trajectories. Due to the strong representation ability of deep learning techniques, existing studies attempt to solve the TUL problem based on various trajectory representation learning models [8, 21, 43]. For example, TULVAE [43] incorporates VAE model into the TUL problem to learn hierarchical semantics of check-in trajectories with RNN to improve the prediction accuracy. DeepTUL [21] proposes to integrate recurrent networks with attention mechanism to model higher-order and multi-periodic mobility patterns by learning from historical traces. AdattTUL [6] and TGAN [44] utilize Generation Adversarial Network (GAN) to study the TUL problem with adversarial mobility modeling. Recently, SML-TUL [42] uses contrastive learning to encode the predictive representations from the user mobility itself constrained by spatio-temporal factors.

Despite their effectiveness, four key limitations exist in these methods. *First*, all existing methods still suffer from *data sparsity*, and cannot learn quality representations over low-sampling trajectories (e.g., check-ins of inactive users). *Second*, existing methods generally rely on the recurrent neural frameworks to model trajectory sequence, and thus can hardly capture the *long-term dependencies* among the long sequence mobility traces [16]. Previous approaches show sub-optimal performance when dealing with long sequence data in our evaluation. *Third*, most of existing works are limited to inject

the global spatial and temporal context into the intra- and inter-trajectory pattern modeling. In practical scenarios, human trajectory are often exhibited with hierarchical mobility patterns across different time granularities [43]. *Lastly*, few of previous approaches incorporate the *rich external contextual features* into the representation of human mobility data.

To address the aforementioned challenges, we propose AttnTUL, a hierarchical spatio-temporal attention network, to realize high predictability for TUL problem. Particularly, AttnTUL is built on a graph neural architecture to encode both local and global trajectory transitional patterns. Our graph-based message passing paradigm is able to alleviate the data sparsity issue by effectively performing knowledge transfer among geographical regions and trajectories based on our generated hierarchical spatial graphs. To capture the intra-trajectory transitional regularities, we develop a temporal self-attention mechanism to encode the long-term mobility dependencies. To model the inter-trajectory relationships, AttnTUL enhances the trajectory representation paradigm by preserving the spatial context across the entire urban space, with the introduced global elastic attentive encoder. The integrated local and global embeddings will be fed into the prediction layer for linking trajectories to their corresponding users. With the design of local and global hierarchical modeling, our proposed AttnTUL method is general and robust for both dense GPS trajectory data and sparse check-in trajectory data

In our evaluation, experimental results on three types of real-life mobility datasets show that our model significantly outperforms several strong baselines (21.75% ACC@1 gain and 24.34% Macro-F1 gain on average) in TUL task on both sparse and dense trajectory data. Additionally, our perform ablation study to justify the model design rationale with component-wise effect investigation.

Our contributions can be summarized as follows:

- We propose AttnTUL, a hierarchical spatio-temporal attention network model to solve the TUL problem. Our AttnTUL simultaneously models local and global spatial and temporal characteristics of users' mobility trajectories over a graph neural architecture.
- We design a hierarchical spatio-temporal attention network which contains a temporal self-attention encoder to learn the local sequential dependencies within trajectory, and a global elastic attention encoder to capture the complex inter-trajectory dependencies.
- We conduct extensive experiments on three types of real-life mobility datasets. Results show that our model significantly outperforms state-of-the-art baselines by 12.73%~46.85% and 18.90%~50.97% improvements in terms of ACC@1 and Macro-F1.

2 RELATED WORK

In literature, measuring the similarity or distance between trajectories which is essential to trajectory pattern mining. Similarity measures, *e.g.*, Dynamic Time Warping (DTW) [15], Longest Common Sub-Sequence (LCSS) [35], Trajectory-Hausdorff Distance [1], Spatio-Temporal Linear Combine distance [24], and Spatiotemporal Signature [13], are often used to discover the user similarity from their trajectories. *However, such approaches are artificially designed, and thus only suitable for specific scenario.* Recently, deep representation learning has been used for trajectory similarity computation [18, 32–34, 37]. *However, these methods focus more on improving the efficiency of trajectory similarity computation.* Trajectory classification is another way to understand mobility patterns. Existing trajectory classification works focus on labeling trajectories as different motion patterns, such as Driving, Biking and Walking in transportation

classification [40] and Occupied, Non-occupied and Parked in taxi status inference [46]. These approaches mainly rely on extraction of spatio-temporal characteristics of trajectories.

TUL problem was recently introduced in [8], which links trajectories to their generating-users, and gradually becomes a hot topic in spatio-temporal data mining by classifying trajectories by users. Due to its broad range of applications in personalized recommendation systems, location-based services and urban planning, it gradually becomes a hot topic in spatio-temporal data mining. Several methods have been proposed to solve the TUL problem [8, 21, 25, 36, 43]. TULER [8] utilizes RNN based models to learn sequential transition patterns from trajectories, and links them to users. It first uses word embedding to learn the representations for locations in the trajectories and feed them into RNN model to capture mobility patterns for TUL. However, the standard RNN based models suffer from data sparsity problem and lacking of understanding hierarchical semantics of human mobility. In their follow-up work [43], TULVAE is proposed to improve the prediction accuracy by incorporating VAE into TUL task to learn hierarchical semantics of check-in sequences. However, it fails to utilize existing abundant features and also do not consider multi-periodic mobility regularities. DeepTUL [21] proposes using the attentive recurrent network in TUL to alleviate the data sparsity problem by leveraging historical data, and capture multi-periodic regularities of human mobility to improve prediction accuracy. AdattTUL [6] and TGAN [44] introduce Generation Adversarial Network (GAN) to deal with the TUL problem. Recently, SML-TUL [42] uses contrastive learning to learn the predictive representations from the user mobility itself constrained by the spatio-temporal factors. *Nevertheless, these methods use RNNs for modeling or prediction, which cannot effectively model long-term dependencies of trajectories, and all above methods ignore the contribution of global spatial modeling to TUL prediction.*

In addition, a few recent works are proposed to identify users across different mobility datasets [4, 5]. DPLink [5] is designed to model the correlations between trajectories to measure the trajectory similarity from different data source. Recently, attention mechanism [2] has also been studied in spatio-temporal data mining. Researchers combine attention with RNN for mobility modeling, such as mobility prediction [7] and mobility inference [45], and personalized route recommendation [28]. Thanks to the characteristics of attention mechanism, it can make up for the deficiency of RNN in capturing long-term dependencies to some extent. *Different from the previous studies, to our best knowledge, we are the first to abandon RNN and adopt the fully attention neural network to solve TUL problem.*

3 PRELIMINARIES

In this section, we first introduce some preliminary concepts and then formally define the problem of TUL.

Let $\mathcal{U} = \{u_1, u_2, \dots, u_i\}$ denote a set of moving users.

DEFINITION 1 (SPATIO-TEMPORAL POINT). *A spatio-temporal point is a uniquely entity in the form of $\langle t, \ell \rangle$, where t is the visited timestamp, and ℓ represents the geographical coordinates of the point (i.e., longitude and latitude).*

DEFINITION 2 (TRAJECTORY). *A trajectory is a sequence of spatio-temporal points $(\langle t_1, \ell_1 \rangle, \langle t_2, \ell_2 \rangle, \dots, \langle t_m, \ell_m \rangle)$ generated by user u_i in chronological order during a certain time interval τ , which is denoted by $Tr_{u_i}^\tau$.*

The time interval can be one hour, one day, one week or even one month. A trajectory is called *unlinked*, if we do not know the user who generated it. Let T denote the whole time interval. The known trajectories generated in previous time intervals by all users, i.e., linked trajectories, are denoted by $Tr_u = \{Tr_{u_i}^\tau | u_i \in \mathcal{U} \wedge \tau \in T\}$.

Based on the above definitions, we now state our studied problem as below:

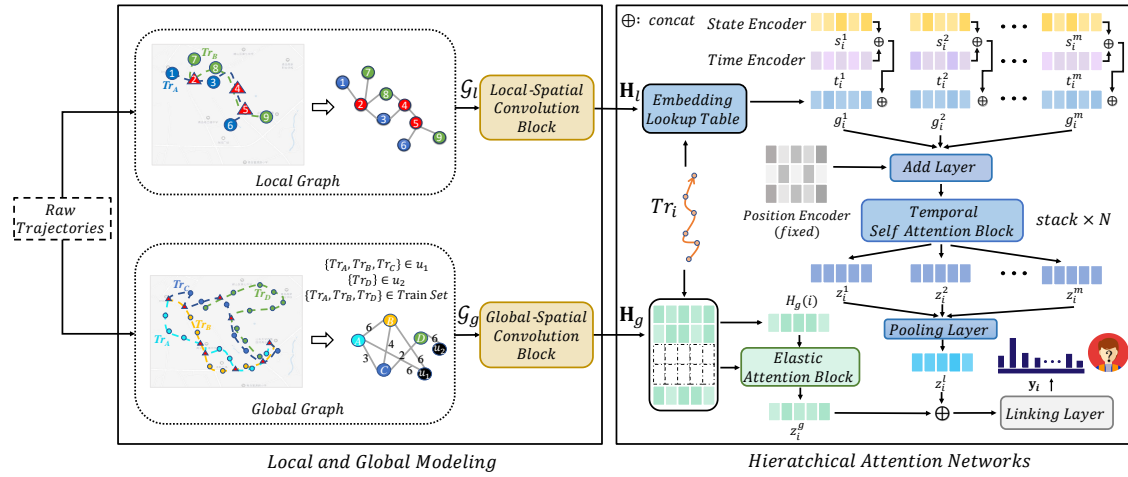


Fig. 2. The overview of the proposed framework

PROBLEM (TRAJECTORY-USER LINKING). Given a set of unlinked trajectories \overline{Tr} generated by users \mathcal{U} and corresponding linked trajectories Tr_u , our goal is to learn a mapping function $f : \overline{Tr} \rightarrow \mathcal{U}$ that links anonymous trajectories to users.

Key notations used in the paper are summarized in Table 1.

Table 1. Main notations and their definitions.

Notation	Definition
\mathcal{U}	the set of moving users
$\langle t, \ell \rangle$	a spatio-temporal point
Tr_i	a trajectory
m	the number of spatio-temporal points in Tr_i
τ	the time interval
$Tr_{u_i}^\tau$	a trajectory generated by u_i
T	the whole time interval
Tr_u	the set of linked trajectories
\overline{Tr}	the set of unlinked trajectories
$\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l)$	the local spatial graph
A_l	the adjacency matrix of local spatial graph
X_l	the feature matrix of local spatial graph
$\mathcal{G}_g = (\mathcal{V}_g, \mathcal{E}_g)$	the global spatial graph
A_g	the adjacency matrix of global spatial graph
X_g	the feature matrix of global spatial graph
H_l	the embeddings of all grids
H_g	the global embeddings of all trajectories
z_i^l	the local representation for Tr_i
z_i^g	the global representation for Tr_i

4 METHODOLOGY

In this section, we present the details of our neural network model AttnTUL (as shown in Figure 2), consisting of four key components: (1) *local and global graph modeling*, (2) *spatial convolutional networks*, (3) *hierarchical spatio-temporal attention networks*, and (4) *linking layer*. *First*, we construct a local graph and a global graph to model micro and macro spatial relationships for all trajectories, respectively. *Second*, we present the spatial convolutional networks on the two graphs to learn initial embedding for each divided grid and each trajectory. *Third*, we employ a multi-head temporal self-attention network to capture the temporal dependency in local embeddings of each trajectory and fuse the local embeddings. We also design an global elastic attention encoder to obtain global representation. *Finally*, we use a linking layer to classify trajectories by users.

4.1 Local and Global Graph Modeling

In this section, we introduce how to construct local and global spatial graphs.

4.1.1 Preprocessing. To facilitate the construction of local spatial graph, we employ a gridding technique to divide the geographic space of all trajectories into discrete grids. More specifically, given grid size s , the whole geographic space is divided into n grids. Consequently, we construct a grid mapping function $f_g : \ell_i \rightarrow g_i$, so that each trajectory $Tr = (\langle t_1, \ell_1 \rangle, \langle t_2, \ell_2 \rangle, \dots, \langle t_m, \ell_m \rangle)$ can be mapped to a grid sequence $Tr' = (\langle t_1, g_1 \rangle, \langle t_2, g_2 \rangle, \dots, \langle t_m, g_m \rangle)$.

4.1.2 Local Spatial Graph Construction. To capture the correlations among spatial locations in all trajectories, we first construct a local spatial graph $\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l)$, where each grid is a node in \mathcal{G}_l , and edges indicate the connectivity between grids. That is, an edge is created between grid g_i and grid g_j if a trajectory contains a consecutive snippet from $g_i \setminus g_j$ to $g_j \setminus g_i$. In high-sampling trajectories, some consecutive spatio-temporal points may be mapped to same grid. For this case, we remove the self-loop edges in \mathcal{G}_l . The weight on edge $e_{i,j} \in \mathcal{E}_l$ is defined as the number of trajectories that contains the consecutive snippet $(g_i \setminus g_j \rightarrow g_j \setminus g_i)$.

Furthermore, the semantic context of nodes can be expressed by a feature matrix. In this paper, we adopt a one-hot vector (the length is the number of grids) for each node g_i to express the uniqueness of its location in the feature matrix. We use \mathbf{A}_l and \mathbf{X}_l to denote the adjacency matrix and feature matrix of local spatial graph, respectively.

4.1.3 Global Spatial Graph Construction. To model spatial correlations between trajectories and known linkages between users and trajectories, we next construct a global spatial graph $\mathcal{G}_g = (\mathcal{V}_g, \mathcal{E}_g)$, which is a heterogeneous graph that contains the above two relationships.

Specifically, \mathcal{G}_g includes two types of nodes – trajectory nodes and user nodes, *i.e.*, each trajectory is treated as a node and each user is also regarded as a node. If two trajectories share common grids, then an edge is created between these two trajectory nodes. The weight of the edge between two trajectory nodes is defined as the number of shared grids in two trajectories. In addition, the labeled trajectories are connected to their users, and the weights on the edges between user node and trajectory node are uniformly defined as the maximum weight between trajectory nodes, which indicates that the relationship strength between labeled trajectories and their users is maximized, and further the relationship between the trajectories of the same user is also strengthened.

For each trajectory node Tr_i and user node u_j , we encode its spatial features (passed grids) into a multi-hot vector (the length is the number of grids) to represent their spatial characteristics. We use \mathbf{A}_g and \mathbf{X}_g to denote the adjacency matrix and feature matrix of global spatial graph, respectively.

Notably, when the number of trajectories is very large, if the traversal method is used to search for pairs of interactive trajectories (*i.e.*, share common grids), the construction of the global spatial graph is actually very time-consuming. To improve the construction efficiency and reduce the time cost, we propose a fast implementation trick. We first use grid coding to represent each trajectory, that is, use one hot coding, for each grid, if the trajectory contains the grid, then the code is 1, otherwise 0. Thus we can obtain the grid coding matrix of all trajectories, denoted by C_g . Therefore, the adjacency matrix A_g can be calculated directly by: $A_g = C_g \cdot C_g^T$.

4.2 Spatial Graph Convolutional Networks

4.2.1 Local Graph Convolution. GCNs have been widely used in graph representation learning to capture graph topological structure by aggregating neighbor features and have achieved great success [9, 31]. We employ GCN on the local spatial graph and global spatial graph constructed by different granularities to learn the hierarchical embeddings of trajectories.

To jointly capture the topological structures and spatial location information among grids, we first perform convolution operation on local graph G_l . Following [17], multi-layer spatial convolution network performs following layer-wise propagation rule:

$$\mathbf{H}_l^{(i+1)} = \text{ReLU} \left(\tilde{\mathbf{D}}_l^{-\frac{1}{2}} \tilde{\mathbf{A}}_l \tilde{\mathbf{D}}_l^{-\frac{1}{2}} \mathbf{H}_l^{(i)} \mathbf{W}_l^{(i)} \right), \quad (1)$$

where $\mathbf{W}_l^{(i)}$ is a layer-specific trainable weight matrix, $\tilde{\mathbf{A}}_l = \mathbf{A}_l + \mathbf{I}$, and $\tilde{\mathbf{D}}_{l_{ii}} = \sum_j \tilde{\mathbf{A}}_{l_{ij}}$. $\mathbf{H}_l^{(0)} = \mathbf{X}_l$, where \mathbf{X}_l is the feature matrix of \mathcal{G}_l . $\mathbf{H}_l^{(i)} \in \mathbb{R}^{n \times d}$ is the output of i -th layer where d is the embedding dimension, denoting the initial embeddings of all grids.

4.2.2 Global Graph Convolution. Although local graph convolution can learn the embeddings for all grids, it may not be able to capture the correlations between trajectories and between users and their produced trajectories from a global perspective. Therefore, to learn the embeddings of trajectories and users, we next perform convolution operation on global spatial graph:

$$\mathbf{H}_g^{(i+1)} = \text{ReLU} \left(\tilde{\mathbf{D}}_g^{-\frac{1}{2}} \tilde{\mathbf{A}}_g \tilde{\mathbf{D}}_g^{-\frac{1}{2}} \mathbf{H}_g^{(i)} \mathbf{W}_g^{(i)} \right), \quad (2)$$

where $\mathbf{W}_g^{(i)}$ is a layer-specific trainable weight matrix, and $\mathbf{H}_g^{(0)} = \mathbf{X}_g$. We can learn the embedding for each trajectory and each user via global graph convolutional network, which captures the key spatial characteristics of trajectories from the whole.

4.3 Hierarchical Spatio-Temporal Attention Networks

4.3.1 Semantic Location Encoder. In addition to spatial information, contextual features in mobility data, *e.g.*, motion state and time feature, can be considered into location embedding for trajectory sequence modeling. Therefore, location encoder is a multi-modal embedding module. Following [29], we divide the motion state into (i) speed-related operations (*i.e.*, acceleration, deceleration, and constant speed) and (ii) direction-related operations (*i.e.*, turning left, turning right, and moving straight), and combine them into nine motion states. For time feature, we divide the whole time into time windows according to a certain time granularity (*e.g.*, 10 minutes). Figure 3 shows an example of state encoder and time encoder in semantic location encoder.

Specifically, we design two sparse linear embedding layers to encode motion state s_i and time window t_i (*e.g.*, one-hot), and then concatenate the corresponding grid embedding to obtain an ensemble vector x_i . The formulation of

location encoder is as follows:

$$x_i = \text{Tanh} (FC([\mathbf{W}_t t_i + b_t; \mathbf{W}_s s_i + b_s; \mathbf{H}_l(g_i)])) , \quad (3)$$

where \mathbf{W}_t , \mathbf{W}_s , b_t and b_s are learnable parameters of embedding layers, $\mathbf{H}_l(g_i)$ denotes the local embedding of grid g_i , and $[\cdot; \cdot]$ denotes the concatenate function. $\text{Tanh}(\cdot)$ is the non-linear activation function, and $FC(\cdot)$ is a fully connected layer.

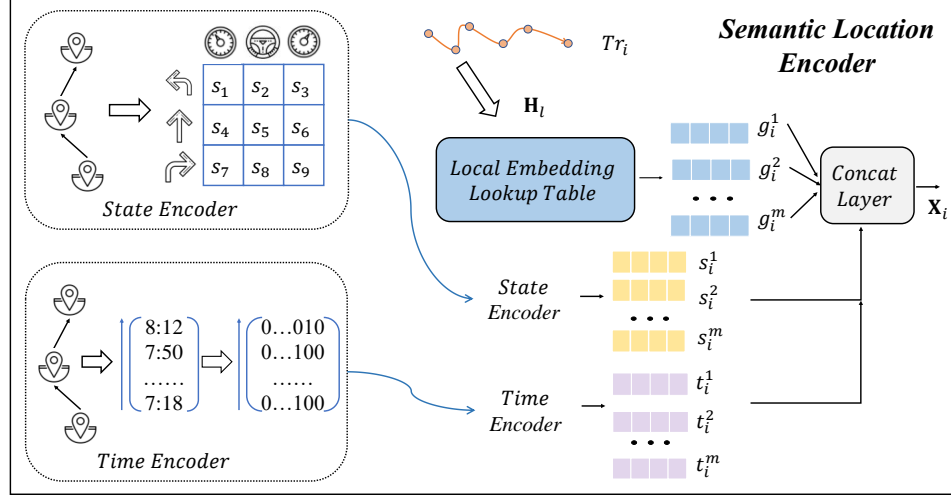


Fig. 3. State/time encoder in semantic location encoder

4.3.2 Temporal Self-Attention Encoder. To capture the long-term temporal dependencies in trajectory sequence which is difficult to model and learn by RNN-based models, we employ a multi-head temporal self-attention mechanism to learn the intra-trajectory correlations.

To supplement sequence information, the input first adds the position encoding:

$$\mathbf{M}_i = \mathbf{X}_i + \mathbf{P}, \quad (4)$$

where $\mathbf{X}_i = \{x_1, x_2, \dots, x_m\}$ are location embeddings for all grids in trajectory Tr_i , and \mathbf{P} denotes the position encoding, which aims to distinguish the sequence position. Notice that \mathbf{P} is fixed and defined as in [27].

Then, \mathbf{M}_i is then sent to the self-attention module to integrate the sequence information:

$$\begin{aligned} (Q_i, K_i, V_i)^T &= \mathbf{M}_i (\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V)^T, \\ Z_i &= \text{Softmax} \left(\frac{Q_i K_i^T}{\sqrt{d}} \right) V_i, \end{aligned} \quad (5)$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times d}$ are learnable projection matrices. The spatiotemporal correlation is calculated by dot product. $\text{Softmax}(\cdot)$ is used to normalize the spatiotemporal dependencies, and the scale \sqrt{d} prevents the saturation

led by Softmax function. Besides, we also employ a residual connection to solve the problem of network degradation, followed by layer normalization.

In addition, multiple dependency patterns from location embeddings in a single trajectory can be learned through a multi-head attention mechanism:

$$\mathbf{Z}_i = FC(\text{concat}(\mathbf{Z}_i^{(1)}, \mathbf{Z}_i^{(2)}, \dots, \mathbf{Z}_i^{(\#head)})), \quad (6)$$

By stacking multiple temporal self-attention layers, we can effectively capture the complex temporal dependencies in trajectory sequence.

Furthermore, the updated location embeddings $\mathbf{Z}_i = \{z_i^1, z_i^2, \dots, z_i^m\}$ of trajectory Tr_i are sent to a pooling layer to extract important location information in the trajectory:

$$z_i^l = \text{Pooling}(z_i^1, z_i^2, \dots, z_i^m), \quad (7)$$

For Pooling, we select the Max-Pooling to detect the significant information. Consequently, the local representation $z_i^l \in \mathbb{R}^d$ for trajectory Tr_i is obtained by fusing the local location information.

4.3.3 Global Elastic Attention Encoder. Inspired by [20, 22], we design an elastic attention module to select the most relevant global information. Specifically, we first find the global embedding $\mathbf{H}_g(i)$ for given trajectory Tr_i , and the attention score w.r.t. trajectory Tr_j is defined as:

$$\begin{aligned} a_{i,j}^s &= \frac{\mathbf{H}_g(i)\mathbf{H}_g(j)^T}{\|\mathbf{H}_g(i)\| \|\mathbf{H}_g(j)\|^T}, \\ \mathbf{A}_i^s &= \{a_{i,1}^s, a_{i,2}^s, \dots, a_{i,\|Tr\|}^s\}, \\ \mathbf{W}_i^g &= \text{Sparsemax}(\mathbf{A}_i^s), \end{aligned} \quad (8)$$

where $\|Tr\|$ is the number of trajectories, and Sparsemax is an alternative to Softmax which tends to yield sparse probability distributions:

$$\text{Sparsemax}(\mathbf{x}) = \underset{\mathbf{p} \in \Delta^d}{\text{argmin}} \|\mathbf{p} - \mathbf{x}\|^2, \quad (9)$$

where $\Delta^d = \{\mathbf{p} \in \mathbb{R}^d : \mathbf{p} \geq 0, \|\mathbf{p}\|_1 = 1\}$. The predictive distribution $\mathbf{p}^* = \text{Sparsemax}(\mathbf{x})$ is likely to assign exactly zero probability to low-scoring choices, which retains the most important factors.

Finally, the output of global elastic attention encoder is the weighted sum of the global embedding vectors based on attention weights \mathbf{W}_i^g :

$$z_i^g = \mathbf{W}_i^g \mathbf{H}_g, \quad (10)$$

where $z_i^g \in \mathbb{R}^d$ is the global representation of trajectory Tr_i .

4.4 Linking Layer

The linking layer aims to obtain the probability distribution of users (labels) from the representations learned by hierarchical spatio-temporal attention networks. The linking layer first concatenates z_i^l and z_i^g for each trajectory to obtain a higher-level representation. Then, a fully connected layer is used to project the high-level representation into a

vector with $|\mathcal{U}|$ dimension. The formulation is as follows:

$$\mathbf{y}_i = \left(W_c \begin{bmatrix} z_i^l \\ z_i^g \end{bmatrix} + b_c \right) \quad (11)$$

where $\mathbf{W}_c \in \mathbb{R}^{|\mathcal{U}| \times 2d}$ and $b_c \in \mathbb{R}^{|\mathcal{U}|}$ are learnable weight matrix and bias, and the vector \mathbf{y}_i is the estimated probability of users for trajectory Tr_i .

To train our model, we apply cross-entropy as loss function and use back propagation algorithm to optimize our model. We define our cross-entropy-based loss function as follows:

$$\mathcal{L}(\Theta) = -\frac{1}{\zeta} \sum_{i=1}^{\zeta} c_i \log(\sigma(\mathbf{y}_i)) + \frac{\lambda}{2} \|\Theta\|^2 \quad (12)$$

where c_i is the one-hot ground truth label of trajectory Tr_i , $\sigma(\cdot)$ is the softmax function, ζ is the number of training trajectories, and Θ is the set of all trainable parameters. λ is an L2 regularization hyperparameter to alleviate overfitting.

4.5 Algorithm Pseudo-Code

Algorithm 1 shows the pseudo-code of our proposed AttnTUL framework guided by the above objective function (*i.e.*, Eq. (12)).

Algorithm 1 The Learning Process of AttnTUL

Input: Input local spatial graph \mathcal{G}_l , global spatial graph \mathcal{G}_g , feature matrix \mathbf{X}_l and \mathbf{X}_g , embedding dimension d , the number of convolution layers l_n

Output: Model parameters Θ

```

1: while ! convergence do
2:   Perform graph convolutional networks
3:   for  $i = 1$  to  $l_n$  do
4:      $\mathbf{H}_l^{(i)} \leftarrow \text{ReLU} \left( \tilde{\mathbf{D}}_l^{-\frac{1}{2}} \tilde{\mathbf{A}}_l \tilde{\mathbf{D}}_l^{-\frac{1}{2}} \mathbf{H}_l^{(i-1)} \mathbf{W}_l^{(i-1)} \right)$ ;
5:      $\mathbf{H}_g^{(i)} \leftarrow \text{ReLU} \left( \tilde{\mathbf{D}}_g^{-\frac{1}{2}} \tilde{\mathbf{A}}_g \tilde{\mathbf{D}}_g^{-\frac{1}{2}} \mathbf{H}_g^{(i-1)} \mathbf{W}_g^{(i-1)} \right)$ ;
6:   end for
7:    $\mathbf{X} \leftarrow \text{Tanh} (FC([\mathbf{W}_t t + b_t; \mathbf{W}_s s + b_s; \mathbf{H}_l]))$ ;
8:   Get  $z_i^l$  using Eqs. (4)-(7);
9:   Get  $z_i^g$  using Eqs. (9)-(10);
10:   $\mathbf{y}_i \leftarrow \left( W_c \begin{bmatrix} z_i^l \\ z_i^g \end{bmatrix} + b_c \right)$ 
11:  Calculate  $\mathcal{L}$  using Eq. (12);
12:  Back propagation and update parameters in AttnTUL;
13: end while
14: Return  $\Theta$ ;
```

We now analyze the time complexity of our model. The time consumption of our model mainly lies in spatial GCNs, temporal self-attention encoder, and global elastic attention encoder. First, the time complexity of spatial GCN block is $O(|\mathcal{E}|d)$ and $O(|\mathcal{V}|^2 d)$ for the GCN operation and the spatial attention operation, respectively, for every iteration. Second, for each trajectory, the time complexity of temporal self-attention encoder is $O(m^2 d)$ for the multi-head self-attention operation for each layer. It is clear that the time complexity of the global elastic attention encoder is much smaller than that of the temporal self-attention encoder. Therefore, the time complexity of our model is mainly

Table 2. Statistics of the datasets. \overline{len} : average length of trajectories, and H: hour(s).

Datasets	#users	#trajectories	#points	#POIs	τ	\overline{len}
Gowalla	547	38,567	15,967	15,892	6H	2
	222	18,808	11,979	11,960	6H	3
PrivateCar	71	4,178	135,085	10,493	1H	32
	42	2,797	87,719	7,514	1H	31
GeoLife	90	6,035	945,971	23,369	3H	157
	56	4,064	768,533	19,384	3H	189

composed of the spatial GCNs and the temporal self-attention encoder, and also depends on the number of layers of networks, the number of trajectories, and the average length of trajectories.

5 EXPERIMENTS

In this section, we evaluate our proposed model on three types of real-world mobility datasets. The following research questions (RQs) are used to guide our experiments:

- **RQ1.** How does our AttnTUL perform in TUL problem on real-world datasets compared to existing methods?
- **RQ2.** How does each component contribute to the performance of the proposed AttnTUL?
- **RQ3.** How does AttnTUL perform with different parameter settings (*e.g.*, grid size s , dimension d , and time window size)?
- **RQ4.** Whether the trajectory representation learned by our AttnTUL is better than DNN-based baselines?
- **RQ5.** How does our AttnTUL perform efficiently compared to existing methods?

5.1 Datasets

We conduct extensive experiments on three publicly available real-world datasets: Gowalla¹, PrivateCar², and GeoLife³.

- **Gowalla** [8] is a user check-in dataset from a location based social network service. For each user, we concatenate all check-in locations to form a trajectory which will be further divided into sub-trajectories based on given time interval.
- **PrivateCar** [11] is a GPS trajectory dataset of private cars collected in Shenzhen, China from January 1 to January 15 in 2016. The sampling rate of the private car trajectory dataset is relatively high, 1~60 seconds (average: 15.3 seconds and standard deviation: 7.8 seconds).
- **GeoLife** [41] is a GPS trajectory dataset collected in Beijing from April 2007 to August 2012. This dataset includes a broad range of users' outdoor movements, including sightseeing, hiking, cycling and so on. 91% of the trajectories are logged densely, *e.g.*, every 1~5 seconds or every 5~10 meters per point.

Notice that our model does not requires POI data. Some baselines (*e.g.*, TULER and TULVAE) are designed to work on sparse check-in data. To ensure fair comparison, we map GPS trajectories to POI check-in data to make these baselines can handle GPS trajectory data. To fairly reproduce the results of baselines, we crawl POIs in Shenzhen and Beijing from BaiduMap⁴ as the additional geographical context for PrivateCar and GeoLife datasets.

¹<http://snap.stanford.edu/data/loc-gowalla.html>

²<https://github.com/HunanUniversityZhuXiao/PrivateCarTrajectoryData>

³<https://www.microsoft.com/en-us/research/project/geolife-building-social-networks-using-human-location-history/>

⁴<https://map.baidu.com/>

In order to check the robustness of our model, we select two different number of users from each dataset. The statistics of these datasets are summarized in Table 2. Following [8], τ is set as a reasonable time interval on different datasets (*i.e.*, 6 hours, 1 hour and 3 hours for Gowalla, PrivateCar and GeoLife datasets, respectively).

5.2 Baselines

In our evaluation, we compare our AttnTUL against the following two categories of baselines.

5.2.1 Classic models.

- **LCSS** [35] – It adopts LCSS to compute trajectory similarity, and searches the most similar trajectory in training set to find the corresponding user.
- **LDA** [23] – We apply Linear Discriminant Analysis (LDA) to TUL task by embedding trajectories into one-hot vectors and using SVD to decompose the within-class scatter matrix.
- **DT** [12] – Decision Tree (DT) is a classic classification method for trajectory data. We use entropy as the criterion in TUL problem, which shows better performance than Gini index.
- **SR** [13, 14] – Signature Representation (SR) is a state-of-the-art trajectory similarity measure for moving object linking.

5.2.2 Deep neural network models.

- **TULER** [8] – This is the original RNN model for solving TUL task. There are three variants: RNN with Gated Recurrent Unit (**TULER-G**), Long Short-Term Memory (**TULER-L**) and bidirectional LSTM (**Bi-TULER**).
- **TULVAE** [43] – It utilizes VAE to learn the hierarchical semantics of trajectory with stochastic latent variables that span hidden states in RNN.
- **DeepTUL** [21] – This is a recurrent network with attention mechanism to solve TUL problem, which is a state-of-the-art method. It learns from labeled historical trajectory to capture multi-periodic nature of user mobility and alleviate the data sparsity problem.
- **DPLink** [4, 5] – DPLink is a state-of-the-art method to link user accounts from heterogeneous mobility data.
- **T3S** [32] – T3S is a state-of-the-art trajectory representation learning method for trajectory similarity computation.

Notice that we extend T3S with multi-class classification supervision to support TUL prediction. DPLink also cannot be directly applied to TUL problem, thus we extend it to perform user identification in a single mobility dataset in our experiments. The source code of our model is available at https://anonymous.4open.science/r/Attn_TUL.

5.3 Evaluation Metrics

We use the widely used $ACC@k$, Macro-P, Macro-R and Macro-F1 to evaluate the performance, which are common metrics in multi-classification task. Specifically, $ACC@K$ is used to evaluate the accuracy of TUL prediction as:

$$ACC@K = \frac{|\{Tr_i \in \overline{Tr} : u^*(Tr_i) \in \mathcal{U}_K(Tr_i)\}|}{|\overline{Tr}|}, \quad (13)$$

where $u^*(Tr_i)$ is the ground truth user, and $\mathcal{U}_K(Tr_i)$ is the predicted top- K user set. It is considered correct if the ground truth user $u^*(Tr_i)$ lies within the predicted top- k user set $\mathcal{U}_K(Tr_i)$.

Macro-F1 is regarded as an overall performance indicator, taking into account the precision and recall across all classes in multi-classification task, which is defined as:

$$\begin{aligned} Macro-P &= \frac{1}{|c|} \sum_{i=1}^{|c|} P_i, \\ Macro-R &= \frac{1}{|c|} \sum_{i=1}^{|c|} R_i, \\ Macro-F1 &= \frac{1}{|c|} \sum_{i=1}^{|c|} \frac{2 \times P_i \times R_i}{P_i + R_i} \end{aligned} \quad (14)$$

where $|c|$ is number of classes, and P_i and R_i are *precision* and *recall* of each class (user in TUL) respectively.

5.4 Experimental Settings

In our experiments, we use the first 60% of each user's trajectories as the training set on all datasets, the following 20% as the validation set, and the remaining 20% as test set.

We use the source code released by authors for baselines, and use the parameter setting recommended in the original paper and fine-tune them on each dataset to be optimal. In the experiment, we set embedding dimension d to 128, grid size s to 40 meters on Gowalla and 120 meters on the other two datasets, the length of time window to 2 hours, 10 minutes and 20 minutes on Gowalla, PrivateCar and GeoLife, the number of GCN layers to 2, the number of self-attention layers to 3, $\#head$ to 4, and λ to $5e-4$ for our AttnTUL. For fair comparison, for all learning methods, we set epoch to 80, batch size to 16, dropout to 0.5, tune learning rate from 0.0001 to 0.01, use early stopping mechanism, and set patience to 10 to avoid over fitting. In each experiment, we repeat 10 runs and report the average.

All experiments for model efficiency evaluation are conducted on a machine with Intel Xeon Gold 6126@2.60GHz 12 cores CPU, 192GB memory, and NVIDIA Tesla V100-SXM2 (16GB) GPU.

Notice that all the user personal information in our datasets are anonymized to protect user's privacy.

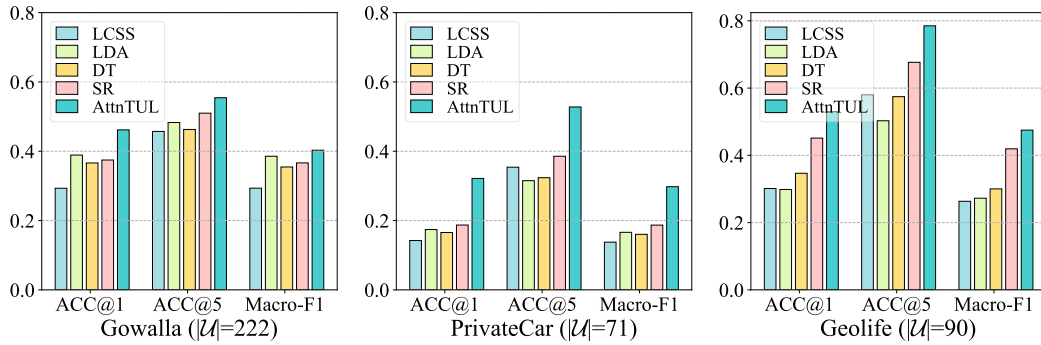


Fig. 4. Performance comparison with classic models.

5.5 Experiment Results

5.5.1 Overall Performance (RQ1). We first evaluate the overall performance of the proposed model on three categories of trajectory datasets in comparison with state-of-the-art baselines. The overall performance is reported in Table 3.

Table 3. Performance comparison with deep neural network models on three real-world datasets.

Data	Methods	ACC@1	ACC@5	Macro-P	Macro-R	Macro-F1	ACC@1	ACC@5	Macro-P	Macro-R	Macro-F1
		$ \mathcal{U} = 222$					$ \mathcal{U} = 547$				
Gowalla	TULER-L	41.67%	51.23%	43.45%	34.11%	36.03%	37.61%	46.97%	40.96%	32.47%	34.24%
	TULER-G	41.56%	50.96%	41.06%	33.08%	34.64%	38.88%	48.61%	42.20%	33.69%	35.15%
	Bi-TULER	41.19%	50.36%	41.88%	33.92%	35.83%	36.88%	46.85%	41.09%	32.20%	33.80%
	TULVAE	40.13%	49.72%	39.71%	32.01%	33.68%	37.18%	46.44%	40.39%	32.30%	33.67%
	DeepTUL	<u>42.36%</u>	51.87%	43.81%	35.32%	<u>37.22%</u>	37.99%	48.16%	41.15%	32.96%	34.48%
	DPLink	41.51%	52.94%	41.32%	35.36%	36.15%	37.19%	49.72%	36.11%	33.18%	32.65%
	T3S	40.80%	50.04%	44.52%	33.68%	35.77%	34.53%	45.97%	35.58%	30.76%	30.50%
	AttnTUL	46.17%	55.44%	45.92%	39.54%	40.03%	44.90%	54.72%	45.40%	38.44%	39.63%
PrivateCar		$ \mathcal{U} = 42$					$ \mathcal{U} = 71$				
	TULER-L	21.92%	44.90%	19.41%	21.05%	18.42%	15.43%	32.12%	15.20%	12.47%	12.16%
	TULER-G	21.92%	45.94%	19.35%	19.61%	18.01%	16.13%	32.77%	15.96%	14.37%	13.29%
	Bi-TULER	22.46%	47.79%	21.52%	21.54%	19.76%	16.57%	36.62%	15.76%	15.14%	13.80%
	TULVAE	23.98%	<u>50.44%</u>	25.34%	20.34%	20.45%	16.89%	31.81%	16.92%	15.38%	14.59%
	DeepTUL	22.99%	49.71%	22.95%	22.53%	20.16%	17.65%	33.09%	15.68%	16.97%	14.24%
	DPLink	23.47%	47.12%	25.36%	23.32%	22.23%	20.71%	<u>41.79%</u>	19.14%	16.83%	15.27%
	T3S	26.53%	47.90%	<u>31.30%</u>	25.07%	25.59%	21.28%	39.45%	24.13%	20.05%	20.01%
	AttnTUL	35.11%	60.40%	33.24%	32.80%	31.49%	31.25%	54.74%	32.25%	32.24%	30.21%
GeoLife		$ \mathcal{U} = 56$					$ \mathcal{U} = 90$				
	TULER-L	41.79%	71.81%	33.78%	34.94%	31.70%	36.84%	60.28%	32.91%	30.41%	29.32%
	TULER-G	43.93%	70.08%	37.09%	36.50%	33.33%	35.51%	61.24%	33.88%	31.73%	30.25%
	Bi-TULER	44.50%	74.09%	38.14%	35.82%	33.76%	37.99%	61.85%	35.82%	33.16%	32.12%
	TULVAE	46.04%	70.99%	42.32%	39.73%	36.87%	39.27%	64.30%	36.23%	33.83%	32.31%
	DeepTUL	51.23%	79.19%	45.84%	41.82%	39.36%	44.92%	69.93%	38.19%	38.35%	35.99%
	DPLink	<u>53.80%</u>	<u>80.03%</u>	<u>48.23%</u>	<u>46.63%</u>	<u>45.03%</u>	<u>47.83%</u>	<u>73.95%</u>	<u>43.93%</u>	<u>40.80%</u>	<u>39.63%</u>
	T3S	51.10%	77.40%	45.47%	43.84%	43.13%	44.33%	70.65%	40.18%	39.52%	38.18%
	AttnTUL	61.37%	85.87%	59.20%	59.59%	58.53%	53.92%	79.47%	49.94%	50.29%	48.24%

where the best is shown in **bold** and the second best is shown as underlined. We also show the comparison with classic models in Figure 4.

Notice that SML-TUL [42] and TGAN [44] are not compared in our experiments because there are no publicly available source codes for them. However, according to the results reported in [42], our AttnTUL significantly outperforms SML-TUL and TGAN in terms of Acc@ k and Macro-F1 on Gowalla, even though our AttnTUL links more users on the same Gowalla dataset. The specific results are as follows: AttnTUL vs. SML-TUL vs. TGAN: 46.17% vs. 45.71% vs. 43.79% in ACC@1, and 40.03% vs 36.15% vs. 33.06% in Macro-F1 on Gowalla.

From the results in Figure 4, our AttnTUL significantly outperforms all classic baseline methods in terms of all evaluation metrics, achieving 23.22%, 67.02%, 17.44% improvements in terms of ACC@1 over SR method on Gowalla, PrivateCar, and Geolife datasets, respectively. SR performs better than other classic baselines in terms of ACC@1, ACC@5, and Macro-F1 on three datasets as it considers the similarity of two trajectories from four representation strategies (*i.e.*, sequential, temporal, spatial and spatiotemporal). However, SR is still based on the defined distance calculation, and cannot automatically learn the similarity or dependency between trajectories based on the relationship among trajectories on different mobility datasets.

From the results in Table 3, we can see that AttnTUL achieves the best performance in terms of all metrics on three different categories of trajectory datasets. This is because our designed local and global graph model effectively captures both micro- and macro-spatial features for spatio-temporal trajectories. In addition, our hierarchical spatio-temporal attention network can learn the long-term dependencies in time dimension and adaptively fuse the local and global representations. This is main reason that our AttnTUL performs the best on both sparse check-in dataset (*i.e.*, Gowalla) and dense GPS trajectory datasets (*i.e.*, PrivateCar and GeoLife). Results in Table 3 show that our model significantly outperforms state-of-the-art baselines by 12.73%~46.85% and 18.90%~50.97% improvements in terms of ACC@1 and Macro-F1. Specifically, AttnTUL achieves average gains of 21.75% ACC@1 and 24.34% Macro-F1 score in comparison to the best performed baseline across all datasets. Considering that the performance gain in TUL prediction reported in recent works [21, 43] is usually around 2.74-9.26% ACC@1 and 1.11-11.00% Macro-F1, this performance improvement achieved by our AttnTUL is significant. Among various baselines, DPLink performs the best on most of metrics, because they extract comprehensive trajectory features. However, our AttnTUL consistently outperforms DPLink by 11.23-50.89% ACC@1 and 10.73-97.84% Macro-F1 score across all datasets, which suggests the rationality of our designed hierarchical spatio-temporal attention network architecture. Although T3S and DeepTUL use the attentive recurrent neural framework, they are still limited by the only consideration of the information within the trajectory, and do not consider the complex cross-trajectory correlations.

We also observe that model performance on data with fewer users is better than that on data with more users. This is intuitive as the more users the more difficult the classification becomes. Among all baselines, DeepTUL and T3S perform best in terms of ACC@1 and Macro-F1 on both Gowalla and PrivateCar datasets, and DPLink performs the best in terms of ACC@1 and Macro-F1 on GeoLife dataset. However, our AttnTUL achieves more improvement over the best performed baseline in terms of ACC@1 and Macro-F1 on data with more users, *e.g.*, on Gowalla, 15.48% improvement in ACC@1 for $|\mathcal{U}| = 547$ vs. 8.99% improvement for $|\mathcal{U}| = 222$, and on PrivateCar, 46.85% improvement in ACC@1 for $|\mathcal{U}| = 71$ vs. 32.34% improvement for $|\mathcal{U}| = 42$.

5.5.2 Ablation Study (RQ2). To validate the effectiveness of each component in AttnTUL, we further conduct the ablation study. We compare our AttnTUL with five carefully designed variants. Despite the changed part(s), all variants have the same framework structure and parameter settings.

- **TUL-L** – It removes the whole local spatial module to demonstrate the importance of local modeling.
- **TUL-G** – It removes the whole global spatial module to demonstrate the importance of global modeling.
- **TUL-SA** – It removes the self attention module and directly sends the location embeddings to pooling layer to fuse local information.
- **TUL-EA** – It uses Softmax to replace our proposed elastic attention in global module.
- **TUL-TS** – In this variant, we remove the time and state encoders to verify the importance of them.

The results of ablation study on three datasets are shown in Table 4. As we can see, AttnTUL outperforms all variants on three datasets, which demonstrates that our key components all contribute to performance improvement of AttnTUL. A noteworthy phenomenon is that TUL-TS performs the second best on three datasets. A potential reason is that time information is not as important as spatial information in TUL problem. Similar conclusion has also been verified in [13, 14].

The comparison between TUL-L and AttnTUL highlights the effectiveness of the proposed local modeling for capturing intra-trajectory dependencies in TUL problem. The comparison between TUL-G and AttnTUL reflects the effectiveness of the proposed global modeling for extracting inter-trajectory corrections in TUL problem. Compared to

Table 4. Results of ablation study. Mac: Macro.

Datasets	Methods	ACC@1	ACC@5	Macro-P	Macro-R	Macro-F1
Gowalla ($ \mathcal{U} =222$)	TUL-L	43.66%	53.28%	45.14%	37.95%	38.56%
	TUL-G	41.87%	52.61%	37.03%	33.31%	33.29%
	TUL-SA	43.20%	52.70%	45.82%	36.58%	37.99%
	TUL-EA	39.99%	51.58%	37.12%	31.73%	32.08%
	TUL-TS	44.01%	54.13%	45.47%	38.72%	39.92%
	AttnTUL	46.17%	55.44%	45.92%	39.54%	40.03%
PrivateCar ($ \mathcal{U} =71$)	TUL-L	19.83%	44.74%	18.26%	20.86%	16.01%
	TUL-G	30.09%	53.14%	27.50%	31.05%	27.16%
	TUL-SA	28.31%	53.92%	27.40%	27.67%	24.97%
	TUL-EA	28.49%	52.12%	28.36%	24.54%	23.64%
	TUL-TS	30.58%	53.20%	26.96%	31.52%	27.41%
	AttnTUL	31.25%	54.74%	32.25%	32.24%	30.21%
GeoLife ($ \mathcal{U} =90$)	TUL-L	49.51%	78.43%	45.49%	45.48%	43.04%
	TUL-G	48.97%	75.46%	43.82%	43.61%	42.03%
	TUL-SA	51.06%	78.65%	47.01%	47.15%	44.01%
	TUL-EA	46.38%	74.35%	41.98%	40.56%	38.59%
	TUL-TS	52.07%	79.25%	47.58%	47.41%	45.02%
	AttnTUL	53.92%	79.47%	49.94%	50.29%	48.24%

AttnTUL, TUL-L and TUL-G perform much worse demonstrating the importance of both local and global modelling in our model.

In addition, we can see that TUL-SA performs much worse than AttnTUL in terms of ACC@1 on PrivateCar and GeoLife, indicating that our designed temporal self-attention network could better capture long-term dependencies for long sequence trajectories. Nevertheless, TUL-SA also performs worse than AttnTUL on Gowalla datasets, which demonstrates that the self-attention mechanism can also capture temporal dependencies for sparse check-in trajectories. Furthermore, we also find an important observation that TUL-EA produces worse results than TUL-G in terms of most metrics on three datasets. This adequately demonstrates the crucial role of our designed elastic attention of global module in extracting relevant global representation for TUL prediction.

5.5.3 Robustness Analysis (RQ3). We also evaluate the robustness of AttnTUL w.r.t. different settings of grid size s , embedding dimension d , the number of heads $\#head$ and the length of time window T_W . The results in terms of ACC@1, Macro-P and Macro-F1 on both sparse Gowalla ($|\mathcal{U}| = 222$) and dense PrivateCar ($|\mathcal{U}| = 71$) are depicted in Figure 5.

The effect of the grid size. From two left sub-figures in Figure 5, we find that the performance of AttnTUL decreases as s increases on Gowalla, while first increases and then decreases with the growth of s from 40 to 200 meters. This phenomenon is easy to understand. Recall that Gowalla is a sparse check-in data, while PrivateCar is a dense GPS trajectory data. If grid is too large, the spatial information would be lost, which is reflected more obvious on sparse Gowalla than on dense PrivateCar. If grid is too small, some spatial interactions among trajectories would not be reflected, which affects the prediction performance.

The effect of the embedding dimension. As shown in the two sub-figures of the second group, the performance of AttnTUL first increases and then decreases on both Gowalla and PrivateCar as d increases, reaching the best at $d = 64$ and $d = 128$ on Gowalla and PrivateCar, respectively. This is because Gowalla is very sparse, and thus few grids are

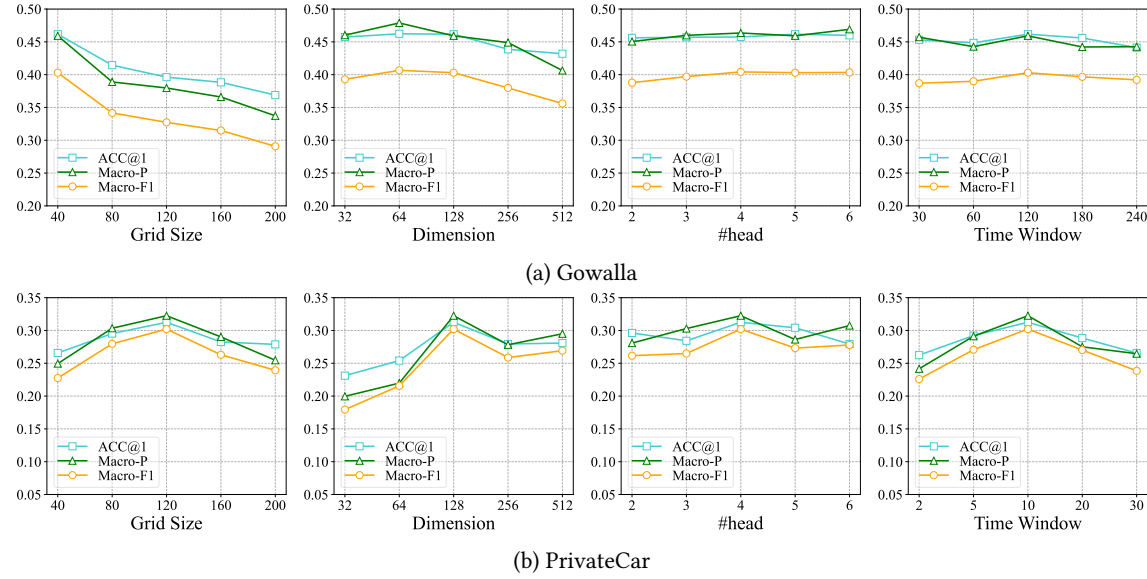


Fig. 5. Parameter sensitivity w.r.t. grid size s , embedding dimensionality d , the number of attentive representation heads $\#head$ and time window length T_W on Gowalla and PrivateCar data. Performances are measured by ACC@1, Macro-P, Macro-F1.

involved. The lower embedding dimensionality has little effect on it, while PrivateCar is denser. When the dimensionality is very low, it has a greater impact on the prediction performance. However, larger embedding dimensionality may contain superfluous information which hurts the TUL prediction performance.

The effect of the number of heads. To investigate the effect of the number of heads in temporal self-attention encoder, we show the evaluation results with the settings of different head numbers in the third group of sub-figures in Figure 5. As we can see, the performance of AttnTUL has little effect w.r.t. the number of heads $\#head$ on Gowalla, first increases and then decreases on PrivateCar as $\#head$ increases, and the model performance achieves the best when $\#head = 4$ on PrivateCar. The possible reason is that sub-trajectories in Gowalla are too sparse, so the intra-trajectory dependencies are not complicated, and thus the effect of multi-head attention mechanism is not obvious. While sub-trajectories in PrivateCar are relatively dense, and the use of multi-head attention mechanism effectively captures the complex temporal dependencies within the trajectory.

The effect of the length of time window. From two right sub-figures in Figure 5, we can see that the model performance also first increases and then decreases as the length of time window increases on both Gowalla and PrivateCar datasets. Considering the sparsity of Gowalla, denseness of PrivateCar, and the length of trajectories in these two datasets, we vary the length of time window from 30 to 240 minutes on Gowalla and from 2 to 30 minutes on PrivateCar. Specifically, the model performance achieves the best when T_W is set to 2 hours on Gowalla and 10 minutes on PrivateCar. That is, on the sparse check-in mobility data Gowalla, when two-hour time window is set to the same encoding, the best results are obtained, indicating that check-ins within such a length of time can be considered to be indistinguishable in time. Similarly, on the dense GPS trajectory data PrivateCar, the best performance is achieved with the same encoding for the spatial-temporal points within 10 minutes, which also shows that there is no temporal difference between the different sub-trajectories in 10-minute time window.

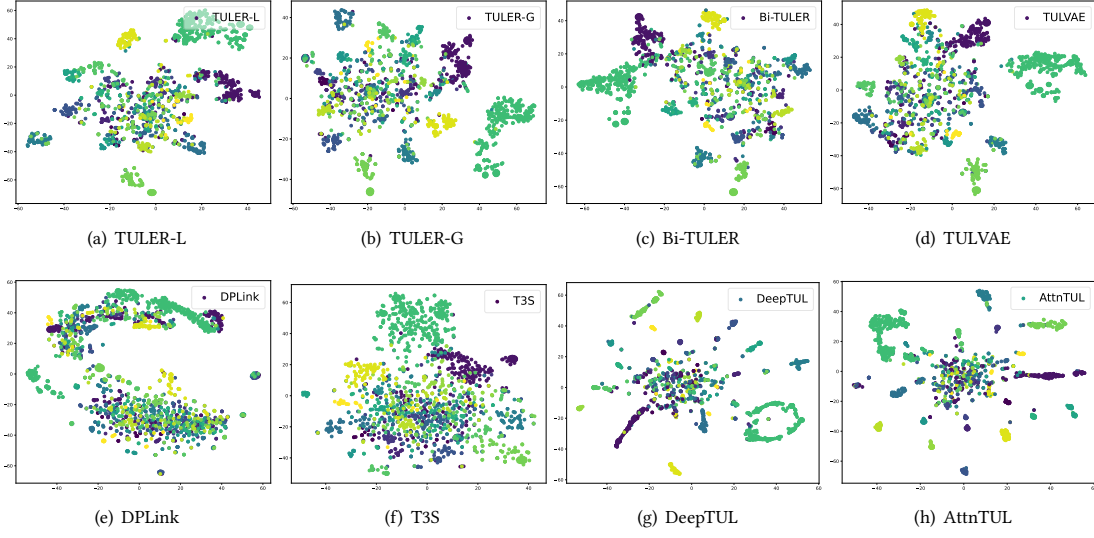


Fig. 6. Spatial-temporal representation visualization of trajectories learned by AttnTUL and other baselines on Gowalla.

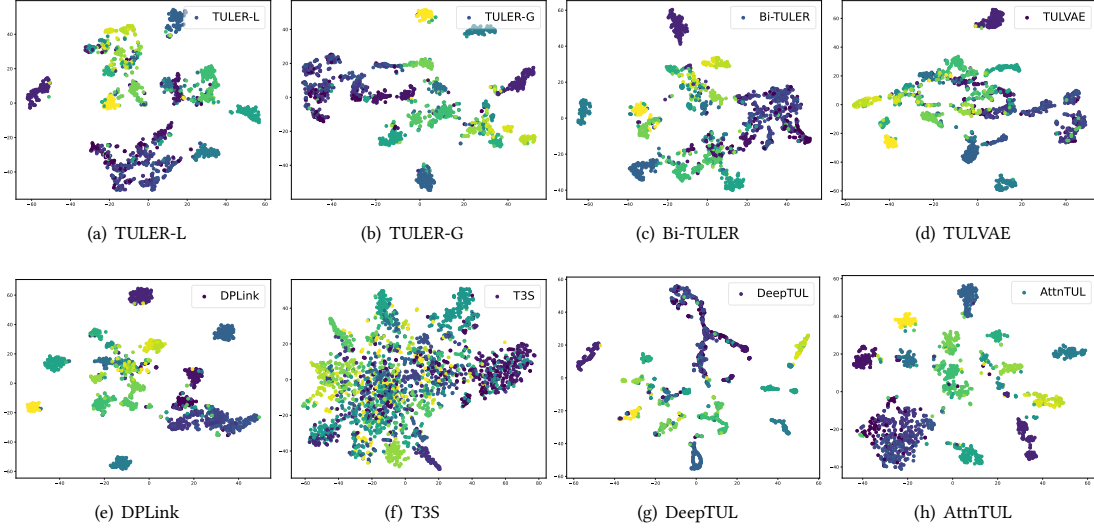


Fig. 7. Spatial-temporal representation visualization of trajectories learned by AttnTUL and other baselines on GeoLife.

5.5.4 Visualization (RQ4). To further verify the effectiveness of our model in learning spatiotemporal representation of trajectories, we adopt a visualization way to compare the representation vectors of trajectories learned by different models.

To this end, we use t-SNE [26] to plot the latent space of trajectories learned by DNN-based models. Specifically, we randomly select 20 users and their corresponding trajectories from Gowalla and GeoLife. The learned representation of each trajectory is projected to the 2D space. The visualization results of learned trajectory reorientation on Gowalla

($|\mathcal{U}| = 222$) and GeoLife ($|\mathcal{U}| = 90$) are shown in Figure 6 and Figure 7, where points with the same color represent trajectories from the same user.

From Figures 6 and 7, we can observe that the trajectory representations generated by our AttnTUL show an apparent clustering effect, and the clustering is tighter than other methods (*i.e.*, Figure 6(h) and Figure 7(h)). This indicates that AttnTUL is able to effectively distinguish trajectories generated by different users, which is crucial for TUL prediction task. Although DeepTUL and DPLink can also distinguish some clusters (users), there are obviously plenty of trajectories from different users intertwined together, and thus it is difficult to identify the correct users for these trajectories.

Furthermore, we also observe that all baselines perform poorly on sparse check-in dataset (*i.e.*, Gowalla) compared to the easily distinguishable dense GeoLife dataset, while our model effectively distinguishes the trajectories produced by most users (see Figure 6(h)).

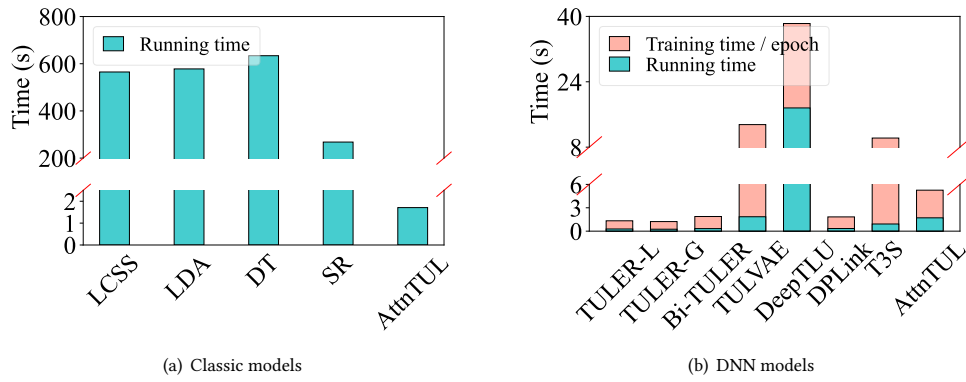


Fig. 8. Model Efficiency Evaluation

5.5.5 Model Efficiency Analysis (RQ5). We also compare the efficiency of our AttnTUL with all baselines on Gowalla ($|\mathcal{U}| = 222$) data. We report the experiment results *w.r.t.* classic methods and DNN-based methods in Figure 8.

Since classic methods have no training time, we only compare the running time. As can be seen from Figure 8(a), the prediction time of our AttnTUL is much less than the running time of classic methods. This is because the classic methods need to calculate a large number of distances between trajectories, so the time cost is relatively high. For example, LCSS, LDA, and DT all consume more than 300 times the prediction time of our model.

From the results in Figure 8(b), we can see that the running time (*i.e.*, the testing time) of DNN-based models is much less than that of classic methods. Although the testing time of our model is not the best, it is faster than state-of-the-art DeepTUL, and the testing time is also less than 2 seconds. Since we adopt the early stopping mechanism for all DNN-based models, the number of rounds of early stopping is different for different models. For a fair comparison, we report the average training time per each epoch. In terms of model training time, the efficiency of our AttnTUL is in the middle level, better than state-of-the-art models (*i.e.*, TULVAE, DeepTUL, and T3S), and worse than the simple TUL models (*i.e.*, TULER and its variants) and DPLink. It takes around 5.6 seconds for each epoch on Gowalla using a Tesla V100 GPU card. Specifically, our AttnTUL is 2.4 times and 7.3 times faster than TULVAE and DeepTUL per epoch, respectively. The potential reason is that DeepTUL needs to use historical trajectory data, so a large number of

trajectory distances need to be calculated, while TULVAE needs to use variational inference, which requires expensive time cost.

6 CONCLUSION

In this paper, we present a novel hierarchical spatio-temporal attention network, called AttnTUL, for TUL problem. AttnTUL effectively learns the local and global representations for each trajectory by the designed hierarchical spatio-temporal attention network to classify trajectories by users. AttnTUL first employs GCN on the constructed local and global graphs to learn embeddings for grids and trajectories. Then, it uses a hierarchical spatio-temporal attention network to obtain the local and global representations for each trajectory, respectively. Eventually, a linking layer is designed to fuse the two representation to classify trajectories by users. Experiments on three real-world mobility datasets demonstrate that our model significantly outperforms state-of-the-art baselines in terms of all metrics for TUL problem. For future work, we plan to enhance our proposed AttnTUL model with the ability of handling streaming mobility trace data for real-time trajectory-user linking scenario.

REFERENCES

- [1] Stefan Atev, Grant Miller, and Nikolaos P. Papanikolopoulos. 2010. Clustering of Vehicle Trajectories. *IEEE Transactions on Intelligent Transportation Systems* 11, 3 (2010), 647–657. <https://doi.org/10.1109/TITS.2010.2048101>
- [2] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.
- [3] Yudong Chen, Xin Wang, Miao Fan, Jizhou Huang, Shengwen Yang, and Wenwu Zhu. 2021. Curriculum Meta-Learning for Next POI Recommendation. In *KDD*. 2692–2702.
- [4] Jie Feng, Yong Li, Mingyang Zhang, Zeyu Yang, Huandong Wang, Han Cao, and Depeng Jin. 2020. User Identity Linkage via Co-Attentional Neural Network From Heterogeneous Mobility Data. *TKDE* (2020), 1–1. <https://doi.org/10.1109/TKDE.2020.2989732>
- [5] Jie Feng, Mingyang Zhang, Huandong Wang, Zeyu Yang, Chao Zhang, Yong Li, and Depeng Jin. 2019. Dplink: User identity linkage via deep neural network from heterogeneous mobility data. In *WWW*. 459–469.
- [6] Qiang Gao, Fengli Zhang, Fuming Yao, Ailing Li, Lin Mei, and Fan Zhou. 2020. Adversarial mobility learning for human trajectory classification. *IEEE Access* 8 (2020), 20563–20576.
- [7] Qiang Gao, Fan Zhou, Goce Trajcevski, Kunpeng Zhang, Ting Zhong, and Fengli Zhang. 2019. Predicting human mobility via variational attention. In *WWW*. 2750–2756.
- [8] Qiang Gao, Fan Zhou, Kunpeng Zhang, Goce Trajcevski, Xucheng Luo, and Fengli Zhang. 2017. Identifying Human Mobility via Trajectory Embeddings. In *IJCAI*, Vol. 17. 1689–1695.
- [9] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NeurIPS*. 1025–1035.
- [10] Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V Chawla. 2018. DeepCrime: Attentive hierarchical recurrent networks for crime prediction. In *CIKM*. 1423–1432.
- [11] Yourong Huang, Zhu Xiao, Xiaoyou Yu, Dong Wang, Vincent Havyarimana, and Jing Bai. 2019. Road network construction with complex intersections based on sparsely sampled private car trajectory data. *TKDD* 13, 3 (2019), 1–28.
- [12] Zhe Jiang. 2018. A survey on spatial prediction methods. *TKDE* 31, 9 (2018), 1645–1664.
- [13] Fengmei Jin, Wen Hua, Jiajie Xu, and Xiaofang Zhou. 2019. Moving object linking based on historical trace. In *ICDE*. IEEE, 1058–1069.
- [14] Fengmei Jin, Wen Hua, Thomas Zhou, Jiajie Xu, Matteo Francia, Maria Orowska, and Xiaofang Zhou. 2020. Trajectory-Based Spatiotemporal Entity Linking. *TKDE* (2020).
- [15] Eamonn J Keogh and Michael J Pazzani. 2000. Scaling up dynamic time warping for datamining applications. In *KDD*. 285–289.
- [16] Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context. In *ACL*. 284–294.
- [17] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.
- [18] Xiucheng Li, Kaiqi Zhao, Gao Cong, Christian S Jensen, and Wei Wei. 2018. Deep representation learning for trajectory similarity computation. In *ICDE*. IEEE, 617–628.
- [19] Wei Liu, Zhi-Jie Wang, Bin Yao, and Jian Yin. 2019. Geo-ALM: POI Recommendation by Fusing Geographical Information and Adversarial Learning Mechanism. In *IJCAI*, Vol. 7. 1807–1813.
- [20] Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *ICML*. PMLR, 1614–1623.

- [21] Congcong Miao, Jilong Wang, Heng Yu, Weichen Zhang, and Yinyao Qi. 2020. Trajectory-user linking with attentive recurrent network. In *AAMAS*. 878–886.
- [22] Ben Peters, Vlad Niculae, and André FT Martins. 2019. Sparse Sequence-to-Sequence Models. In *ACL*. 1504–1519.
- [23] Hamid Reza Shahdoosti and Fardin Mirzapour. 2017. Spectral-spatial feature extraction using orthogonal linear discriminant analysis for classification of hyperspectral data. *European Journal of Remote Sensing* 50, 1 (2017), 111–124.
- [24] Shuo Shang, Lisi Chen, Zhewei Wei, Christian Søndergaard Jensen, Kai Zheng, and Panos Kalnis. 2017. Trajectory similarity join in spatial networks. *Vldb* 10, 11 (2017).
- [25] Tao Sun, Yongjun Xu, Fei Wang, Lin Wu, Tangwen Qian, and Zezhi Shao. 2021. Trajectory-User Link with Attention Recurrent Networks. In *ICPR*. IEEE, 4589–4596.
- [26] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*. 5998–6008.
- [28] Jingyuan Wang, Ning Wu, Wayne Xin Zhao, Fanzhang Peng, and Xin Lin. 2019. Empowering A* search algorithms with neural networks for personalized route recommendation. In *KDD*. 539–547.
- [29] Pengyang Wang, Xiaolin Li, Yu Zheng, Charu Aggarwal, and Yanjie Fu. 2019. Spatiotemporal representation learning for driving behavior analysis: A joint perspective of peer and temporal dependencies. *TKDE* (2019).
- [30] Senzhang Wang, Jiannong Cao, and Philip Yu. 2020. Deep learning for spatio-temporal data mining: A survey. *TKDE* (2020).
- [31] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *TNNLS* 32, 1 (2020), 4–24.
- [32] Peilun Yang, Hanchen Wang, Ying Zhang, Lu Qin, Wenjie Zhang, and Xuemin Lin. 2021. T3S: Effective Representation Learning for Trajectory Similarity Computation. In *ICDE*. IEEE, 2183–2188.
- [33] Di Yao, Gao Cong, Chao Zhang, and Jingping Bi. 2019. Computing trajectory similarity in linear time: A generic seed-guided neural metric learning approach. In *ICDE*. IEEE, 1358–1369.
- [34] Di Yao, Gao Cong, Chao Zhang, Xuying Meng, Rongchang Duan, and Jingping Bi. 2020. A Linear Time Approach to Computing Time Series Similarity based on Deep Metric Learning. *TKDE* (2020).
- [35] Josh Jia-Ching Ying, Eric Hsueh-Chan Lu, Wang-Chien Lee, Tz-Chiao Weng, and Vincent S Tseng. 2010. Mining user similarity from semantic trajectories. In *SIGSPATIAL Workshop on LBSNs*. 19–26.
- [36] Yong Yu, Haina Tang, Fei Wang, Lin Wu, Tangwen Qian, Tao Sun, and Yongjun Xu. 2020. TULSN: Siamese Network for Trajectory-user Linking. In *IJCNN*. IEEE, 1–8.
- [37] Hanyuan Zhang, Xinyu Zhang, Qize Jiang, Baihua Zheng, Zhenbang Sun, Weiwei Sun, and Changhu Wang. 2020. Trajectory Similarity Learning with Auxiliary Supervision and Optimal Matching. In *IJCAI*. AAAI, 3209–3215.
- [38] Mingyang Zhang, Tong Li, Hongzhi Shi, Yong Li, Pan Hui, et al. 2019. A decomposition approach for urban anomaly detection across spatiotemporal data. In *IJCAI*.
- [39] Yu Zheng. 2015. Trajectory data mining: an overview. *TIST* 6, 3 (2015), 1–41.
- [40] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. 2008. Understanding mobility based on GPS data. In *UbiComp*. 312–321.
- [41] Yu Zheng, Xing Xie, Wei-Ying Ma, et al. 2010. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.* 33, 2 (2010), 32–39.
- [42] Fan Zhou, Yurou Dai, Qiang Gao, Pengyu Wang, and Ting Zhong. 2021. Self-supervised human mobility learning for next location prediction and trajectory classification. *Knowledge-Based Systems* (2021), 107214.
- [43] Fan Zhou, Qiang Gao, Goce Trajcevski, Kunpeng Zhang, Ting Zhong, and Fengli Zhang. 2018. Trajectory-User Linking via Variational AutoEncoder. In *IJCAI*. 3212–3218.
- [44] Fan Zhou, Ruiyang Yin, Goce Trajcevski, Kunpeng Zhang, Jin Wu, and Ashfaq Khokhar. 2021. Improving human mobility identification with trajectory augmentation. *GeoInformatica* 25, 3 (2021), 453–483.
- [45] Fan Zhou, Xiaoli Yue, Goce Trajcevski, Ting Zhong, and Kunpeng Zhang. 2019. Context-aware variational trajectory encoding and human mobility inference. In *WWW*. 3469–3475.
- [46] Yin Zhu, Yu Zheng, Lihang Zhang, Darshan Santani, Xing Xie, and Qiang Yang. 2012. Inferring taxi status using gps trajectories. *arXiv preprint arXiv:1205.4378* (2012).