Predicting Human Mobility via Self-Supervised Disentanglement Learning

Qiang Gao, Jinyu Hong, Xovee Xu, *Graduate Student Member, IEEE*, Ping Kuang, Fan Zhou, *Member, IEEE*, and Goce Trajcevski, *Member, IEEE*

***Abstract*—Deep neural networks have recently achieved considerable improvements in learning human behavioral patterns and individual preferences from massive spatial-temporal trajectory data. However, most of the existing research concentrates on fusing different semantics underlying sequential trajectories for mobility pattern learning which, in turn, yields a narrow perspective on comprehending human intrinsic motions. In addition, the inherent sparsity and under-explored heterogeneous collaborative items pertaining to human check-ins hinder the potential exploitation of human diverse periodic regularities as well as common interests. Motivated by recent advances in disentanglement learning, we propose a novel disentangled solution called SSDL for tackling the next POI prediction problem. SSDL primarily seeks to disentangle the potential time-invariant and time-varying factors into different latent spaces from massive trajectories, providing an interpretable view to understand the intricate semantics underlying human diverse mobility representations. To address the data sparsity issue, we present two realistic trajectory augmentation approaches to enhance the understanding of both the human intrinsic periodicity/habits and constantly-changing intents. In addition, we devise a POI-centric graph structure to explore heterogeneous collaborative signals underlying historical check-ins. Extensive experiments conducted on four real-world datasets demonstrate that SSDL significantly outperforms the state-of-the-art approaches–for example, it yields up to 8.57% averaged improvement on ACC@1.**

***Index Terms*—Location-based services, human mobility, graph neural network, disentanglement learning, variational Bayes.**

I. Introduction

THE proliferation of geo-tagged social media (GTSM) such as Foursquare and WeChat has enabled numerous users to post interesting places, report daily activities, and make like-minded friends, resulting in the accumulation of massive amounts of contextual data (e.g., check-ins). This, in turn, offers unprecedented opportunities to explore human diverse life experiences (e.g., mobility patterns) and facilitate the development of various user-centric downstream applications such as trajectory identification [1], Point Of Interest (POI) recommendation/prediction [2], itinerary prediction [3] – to name but a few. As a fundamental task in mining check-in data, *predicting human mobility* (often exemplified as next POI prediction) is critical for researchers and practitioners. Exploration and exploitation of the informative semantics and mutual interactions behind human check-ins [4], [5] enables one to precisely ascertain users' future intentions and is fundamental for tasks such as drawing in more potential customers for new ventures [6], [7].

Spatio-temporal check-in sequences (i.e., trajectories) reflect human daily activities upon a set of POIs, which may include certain (periodic) regularities. The majority of the pioneering works in human mobility prediction aimed at modeling human sequential behaviors taking into account spatio-temporal preferences. For instance, in order to predict where a certain user will go in the near future, conventional approaches, such as Markov Chain [8] and Tensor-based Factorization [9] that rely on data-driven paradigms, attempt to incorporate individual visiting preferences and explore *sequential* patterns. However, these approaches depend heavily on hand-crafted characteristics and face the challenge of comprehending the diverse semantics underlying massive volumes of human trajectories. This, in turn, leads to narrow solutions in disclosing human implicit interactive hints/signals regarding historical check-ins.

More recent deep learning techniques such as recurrent neural networks (RNNs) have brought about encouraging achievements of learning informative check-ins (including POIs) from trajectories and become a widespread and popular methodology in tackling miscellaneous mobility learning tasks [4], [10], [11]. For example, Wu et al. [10] present a PLSPL model, which leverages a Long-Short Term Memory (LSTM) neural network to model human short-term sequential preferences while learning contextual features of POIs behind human historical check-ins via attention mechanism. To consider the spatial and temporal influences for next POI recommendation, Kong et al. [12] incorporate the spatial and temporal intervals between two successive check-ins into recurrent hidden states to mitigate the data sparsity of human trajectories. Due to the higher model efficiency and the ability to quantify the contribution of each

Manuscript received 8 December 2022; revised 17 July 2023; accepted 13 September 2023. Date of publication 27 September 2023; date of current version 5 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62102326, in part by the Natural Science Foundation of Sichuan Province under Grant 2023NSFSC1411, in part by the Key Research and Development Project of Sichuan Province under Grant 2022YFG0314, in part by the National Science Foundation SWIFT under Grant 2030249, and in part by Guanghua Talent Project. Recommended for acceptance by B.C.M. Fung. *(Corresponding author: Fan Zhou.)*

Qiang Gao is with the School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China (e-mail: qianggao@swufe.edu.cn).

Jinyu Hong, Xovee Xu, Ping Kuang, and Fan Zhou are with the University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China (e-mail: jinyuhong@std.uestc.edu.cn; xovee.xu@gmail.com; kuangping@uestc.edu.cn; fan.zhou@uestc.edu.cn).

Goce Trajcevski is with the Iowa State University, Ames, IA 50011 USA (e-mail: gocet25@iastate.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TKDE.2023.3317175, provided by the authors.

Digital Object Identifier 10.1109/TKDE.2023.3317175

check-in in a given trajectory, several attention mechanisms like, for example, self-attention and vanilla attention emerged for handling long human historical trajectories [2], [6], [13].

Several state-of-the-art methods have employed graph structure learning to explicitly uncover spatial correlations or collaborative signals to understand individual human interests. More concretely, they attempt to acquire expressive POI representations by considering the rich contexts of highly correlated POIs. For instance, conventional methods such as word2vec-based [14], [15] and deepwalk-based [16], [17] have successfully uncovered the higher-order correlations between consecutive check-ins and offered contextual POI representations. Other schemes using popular graph neural networks (GNNs), such as graph convolutional networks [11] and graph attention networks [18], primarily seek to incorporate the POI-to-POI correlations (e.g., geographical proximity) within massive human trajectories.

Despite the recent achievements in deep human mobility learning, we observe that existing solutions still have three significant drawbacks:

a) *Implicit semantic entanglement:* Although there is a large body of work on human mobility representation learning, the most common scheme is to encode check-in sequences as well as several affiliated contexts such as temporal semantics to obtain each informative mobility representation. Nevertheless, they usually focus on fusing multiple semantics behind sequential trajectories to predict the user's next POI, which could lead to a myopic perspective and produce a non-diverse recommendation result. We call this phenomenon *semantic entanglement*. In practice, human trajectories (e.g., check-in sequences), as typical sequential data, contain rich user mobility patterns that reflect diverse periodic regularities or behavioral habits of humans. More importantly, the intrinsic individual patterns/habits of humans are difficult to change over time, but their near-term intentions/behaviors are prone to be influenced/dictated by certain time instants. Thus, we consider that human mobility patterns can be implicitly disentangled into two aspects: *time-invariant* and *time-varying* behaviors. Existing solutions rely on data-driven models to understand limited mobility patterns, which fail to reveal the nature of human visiting intents. As a result, they only provide a narrow scope to become familiar with human future behaviors, which usually carries the risk of prediction bias due to the limited scale of trajectory data available.

b) *Sparsity in representation learning:* Only when a user decides or is willing to check in via location-based applications can a POI be recorded, which inevitably leads to the sparsity problem when gathering historical human footprints. As a result, the sparsity problem hinders the model from learning a good representation of human mobility. Existing methods either use the next POI as the sole supervisor or complement the representation learning in a semi-supervised manner with unlabeled trajectories. These paradigms mostly follow the merit of text representation in the field of natural language processing (NLP) which, however, easily fails in capturing the innate rules underlying human trajectories such as individual periodic regularity.

c) *Heterogeneous collaborative signals:* Most existing efforts concentrate on learning POI-to-POI relationships (a.k.a. the connectivity of POIs) from a large number of trajectories, such as consecutive correlation and geographic proximity [17], [18], [19]. Despite the successful collaboration of individual human interests via these homogeneous graph structure learning, a notable limitation is that heterogeneous semantics affiliated with the POIs are not investigated well, yielding a limit in the exploration of affluent common preferences behind human diverse trajectories. For example, people may have similar visit time preferences for certain POIs, such as going to a cafe after lunch. In addition, each POI is associated with a textual description (e.g., POI category), reflecting the underlying human activity interest. We conjecture that incorporating the heterogeneous correlations between POI and their category can provide us with a coarse-grained view of the higher-order connective between POIs. For example, people often go to several fashion stores to buy clothes at a time.

To address the aforementioned limitations, we present a novel solution called Self-Supervised Disentanglement Learning (SSDL) framework for understanding human mobility. Rather than previous data-driven representation learning, SSDL performs self-supervision in the latent space and aims at seeking a clean separation of the time-invariant and time-varying vectors for diverse human trajectories, which is inspired by the recent advances of variational inference and contrastive learning. More specifically, SSDL operates the sequential variational autoencoder (VAE) with a mutual information regularization to guide the training of evidence lower bound (ELBO), aiming at promoting the disentanglement of human mobility-related representations. We provide two realistic trajectory augmentation strategies to alleviate the sparsity issue in representation learning, which can further enhance the understanding of human intrinsic periodicity and constantly-changing intents. In addition, we also present a POI-centric graph structure to explore human common interests underlying diverse check-ins, which primarily seeks human consecutive, geospatial, temporal-aspect, and activity-aspect interests. Our contributions can be summarized as follows:

- We introduce a novel disentangled representation learning framework to understand human time-independent and time-dependent behaviors of their individual mobility patterns. To the best of our knowledge, this study is the first work to disentangle human (sequential) mobility and investigate how it can be used for the prediction of the next POI.
- We propose two practical trajectory augmentation methods, guided by the inherent characteristics of individual human mobility patterns, to promote disentanglement.
- To capture heterogeneous collaborative signals behind historical check-ins, we devise a flexible POI-centric graph structure to explore rich human interests in trajectories,

which enhances the performance of downstream next POI prediction task.

- We conduct extensive experiments on four real-world datasets to evaluate the performance of our proposed SSDL. The results demonstrate that our approach outperforms state-of-the-art methods.

II. Related Work

We now provide a review of the related literature, grouped in three categories.

A. Next POI Prediction in Deep Learning

Recent deep learning solutions have stimulated many researchers and practitioners to tackle the problem of detecting human periodic regularities from massive historical check-ins. In particular, deep (recurrent) neural networks such as LSTM [20] and GRU [21] have received widespread interest in the next POI prediction task as they are able to capture the sequential dependencies that can be used for mobility pattern understanding. For instance, [22] extends the vanilla RNN model and integrates the spatial-temporal impacts into each RNN cell, yielding promising results on the next location prediction. A novel ST-LSTM that implements time gates and distance gates into standard LSTM, aiming at capturing the spatio-temporal relation between consecutive check-ins was proposed in [4]. To learn more contextual information, a personalized long- and short-term preference learning scheme to learn the specific user context was proposed in [10], where the different influences of locations and categories of POIs are considered. While most of endeavors focus on pruning or modifying the RNN-based modules [23], [24], [25], researchers also tried to adopt other popular deep neural networks for next POI prediction – e.g., attention-based neural networks [2], [26] and convolutional neural networks [6], [27]. A Transformer architecture as the mobility feature extractor in which it regards the historical trajectory and semantic contexts as the input to handle multiple factors such as temporal and geographic contexts was presented in [2].

B. Mobility Representation Learning

POI embedding and trajectory embedding, as two core components in mobility representation learning, have been investigated in several recent studies.

For POI embeddings, earlier studies such as [22] and [28] set a fixed or learnable matrix as the initial representations of POIs, primarily seeking to alleviate the "Curse of Dimensionality". However, the semantic information between POIs was under-explored in these works. As word embeddings, especially word2vec-based [29], have achieved great performance in NLP, recent studies also proposed various word2vec-based solutions aimed at capturing the proximity semantics of POIs from human check-in sequences (or real-world trajectories). For instance, [30] and [31] regard each POI as a "word" while each human trajectory as a "sequence", and use word2vec to obtain a low-dimensional vector for each POI. POI2Vec is a latent representation model that incorporates geographic influence when using word2vec method for POI embedding [14]. However, training sparse trajectories to obtain POI representations often confronts the problem of poor capability of POI semantics. More recently, the success of graph neural networks (GNNs) has inspired researchers to turn to devising graph-based models to facilitate the learning of human trajectories [11], [18], [32]. For instance, a graph-based model to explore the spatial, temporal, and preference factors behind the POIs was proposed in [18]. However, it only considers homogeneous interactions among the POIs and ignores heterogeneous interactions with other key entities such as activity and check-in time.

Regarding trajectory representation learning, the majority of existing research concentrates on taking the historical trajectory as input and using the next POI as the sole supervision signal [4], [7], [22], [33]. To address the narrow scale of trajectory data, some efforts attempted to employ the unlabeled trajectories as supplements and train them jointly with the labeled trajectories in an unsupervised or self-supervised manner in order to acquire a good representation for each trajectory [6], [34]. In particular, to operate the trajectories in a latent space, recent studies employ generative models such as variational inference or adversarial models to learn the intrinsic distribution underlying massive trajectory data and then turn to fine-tune the model for the next POI prediction tasks. For instance, VANext extended the variational autoencoder (VAE) to consider the uncertainty of user preferences for regularized representation of historical trajectories [6]. A meta-learning technique called METAODE also employed variational Bayes to encode past human movement patterns into latent space [35]. In essence, these approaches principally rely on integrating numerous semantics including sequential information into a unified space while omitting the possibility of disentangling it to expose the characteristics of human mobility patterns.

C. Disentanglement Learning

A distinct aspect of disentanglement learning is that it enables an interpretable perspective to understand the multiple inherent motions/factors behind the intricate data representations, in addition to notable expressiveness. For instance, [36] presents a hypergraph network called DisenHCN to disentangle user representations into three aspects, e.g., location-aware, aiming at exploring the high-order relations among them for activity prediction. Zhang et al. provide us with a factor-controlled learning system that decouples important extraneous factors (e.g., weather) affecting traffic flow to improve the interpretability of future trends [37]. However, the dynamic factors such as time-varying behaviors underlying mobility patterns have not been investigated. Besides, many recent studies developed VAEs to optimize the mutual interaction between different latent factors [38], [39], [40], [41]. For example, β-VAE [38] is a simple but effective variant of the ordinary VAE that severely penalizes the Kullback–Leibler (KL) divergence term for disentanglement learning. A Disentangled Sequential Autoencoder (DSVAE) approach for sequential data (e.g., video), aiming at factorizing the latent variables into static and dynamic parts was presented in [42]. To make the latent variables interpretable and controllable, a latent variable guidance-based generative model called Guided-VAE makes an effort to utilize VAE to learn a transparent representation [43]. A sequential VAE to learn disentangled representations in a self-supervised manner was presented in [44], and an extension of the sequential VAE with a self-supervised learning approach to facilitate the factorization of video representations was presented in [40]. The newly developed self-supervised learning offers a new avenue to drive the acquisition of semantic representations [45]. For example, [46] employs the ideas of latent self-supervision and intention disentanglement to boost the convergence of representation learning and utilize it in sequential recommendation tasks. In sum, the success of these approaches suggests that, in addition to facilitating the understanding of rich semantics underlying data, disentangling the representation into distinct parts can make the representation more transparent and interpretable.

III. Preliminaries

We now formalize the problem and present the background of VAE and contrastive estimation.

A. Problem Definition

Let $l \in \mathcal{L}$ denote a POI tagged by the location-based systems. We assume that each POI, in addition to its corresponding geographic coordinate (e.g., longitude $l.lo$ and latitude $l.la$), also has an associated category $l.ca$ (e.g., restaurant, museum, park, etc.).

Definition 1. (Check-in Sequence): A check-in sequence (or trajectory) $T_u = \{l_1^u, l_2^u, \ldots, l_n^u\}$ left by user u is a sequence of n POIs ordered by visiting time, where l_τ^u means a user u visits POI l at time t_τ ($\tau \in \{1, 2, \ldots, n\}$).

Let $\mathcal{T}_u = \{T_u^1, T_u^2, \ldots, T_u^m\}$ denote a collection of m historical trajectories of the user u, where each trajectory T_u^i contains a sequence of POIs ordered by visiting time, e.g., $T_u^i = \{l_1^{i,u}, l_2^{i,u}, \ldots, l_n^{i,u}\}$.

Formally, given a user u with his/her recently visited check-in sequence $T_u^m = \{l_1^{m,u}, l_2^{m,u}, \ldots, l_n^{m,u}\}$ and entire historical trajectory $\mathcal{T}_u$, our *goal* is to *predict a POI* $l_{n+1}^{m,u}$ for user u to visit next. Notably, we mainly target disentangled representation learning for users' recently visited POI sequences. For simplicity, we will omit user identity (i.e., u) and trajectory index (i.e., m) in the subsequent sections.

B. Variational Bayes

Variational Autoencoder (VAE) [47] containing an encoder and a decoder maps the input data x into a latent space, where the latent variables are denoted by z, and uses the decoder to generate (i.e., reconstruct) data points. The marginal likelihood $\log p(x)$ can be obtained by maximizing the Evidence Lower BOund (ELBO), which is defined as:

$$\log p_\theta(x) \geq \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(x|z)] - KL[q_\phi(z|x)||p(z)]. \quad (1)$$

Herein, $q_\phi(z|x)$ is an approximate posterior distribution, parameterized by ϕ, $p_\theta(x|z)$ with parameters θ is a likelihood function, and $p(z)$ is a prior (e.g., Gaussian prior) over the latent variables.

C. Contrastive Estimation

In recent self-supervised learning paradigms [34], [48], mutual information (MI) is a common measure of the mutual dependence or compatibility between two variables. Specifically, they usually employ the noise contrastive estimation (NCE) [49], [50] to maximize the lower bound on the mutual information, which can be denoted as follows:

$$\mathcal{L}_{NCE} = \mathbb{E}\left[-\log\left(\frac{\exp^{g(x)^\top g(x^+)}}{\exp^{g(x)^\top g(x^+)} + \sum_{j=1}^J \exp^{g(x)^\top g(x_j^-)}}\right)\right], \quad (2)$$

where x, x^+, and x_j^- respectively denote the *anchor*, *positive*, and *negative* instances. The similarity measure (e.g., cosine) between two instances is denoted by $\exp^{g(\cdot)^\top g(\cdot)}$.

IV. Architecture Design and Methodology

In this section, we present the details of our proposed methodology. An overview of the proposed framework SSDL is presented in Fig. 1. As shown, it consists of three main components. First, we build a POI-centric Graph (PGraph) to explore the common interests from the entire user trajectories and make interest aggregation to obtain both homogeneous and heterogeneous semantics underlying each POI. Next, our Self-supervised Disentanglement Learning component attempts to produce the time-invariant and time-varying variables for each trajectory. In particular, SSDL provides two augmentation strategies over the original trajectory to enhance the understanding of human intrinsic periodicity (habit) and changing intentions. Lastly, SSDL uses the disentangled representations as well as the user's long-term preference modeled by an attentive network to predict the next POI. In the sequel, we provide an in-depth discussion of the main components of SSDL.

A. Common Interest Distillation

Exploring multiple common interests from massive check-ins is a prerequisite for understanding human diverse mobility patterns. Prior works usually design an information-free embedding for each interest modeling, aiming at distilling useful information during sequential check-in learning [28], [30]. Also, researchers resort to building a graph structure to correlate user check-in preferences such as GNN-based methods. However, most of them concentrate on exploring or repositioning homogeneous POI-POI relationships into a common latent space by incorporating transitional relations or geographical proximity [11], [18], [51]. Nevertheless, Unifying multiple interests into a single latent space would inevitably lead to the homogenization of heterogeneous semantics and information entanglement. In contrast, we first build a POI-centric graph (PGraph) and then introduce two neural aggregates to refine homogeneous and heterogeneous semantics, respectively.

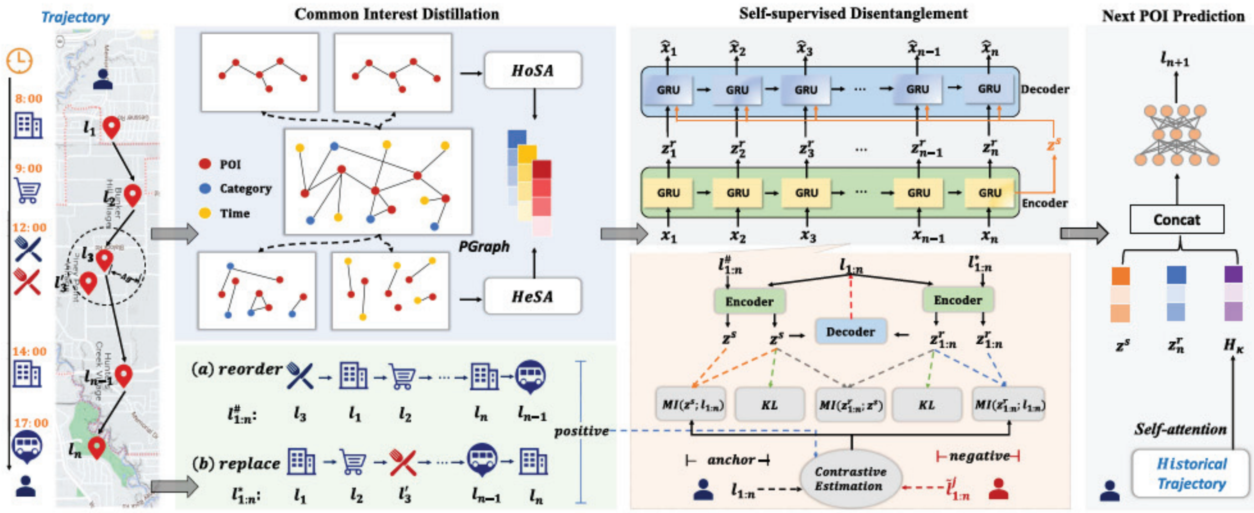


Fig. 1. Pipeline of proposed SSDL.

1) Graph Structure and Building Process: As several elements are recorded by LBSN, e.g., POI identity, geographical coordinate, visiting time, and POI category, we concentrate on exploring four semantic contexts to build our PGraph, including *consecutive*, *geospatial*, *time-aspect*, and *activity-aspect* interests. Let $\mathcal{G} = (\mathcal{V}, E)$ denote the PGraph that models the human common interests, where $\mathcal{V} = (\mathcal{V}_l \cup \mathcal{V}_t \cup \mathcal{V}_a)$ is the set of nodes, and $E = (E_c \cup E_g \cup E_t \cup E_a)$ is the set of edges. Here $\mathcal{V}_l = \mathcal{L}$ represents a collection of different POIs, $\mathcal{V}_t$ is the set of time bins, $\mathcal{V}_a$ denotes the set of POI categories, and E_c, E_g, E_t, E_a respectively indicate the above four semantic contexts. That is to say, $\mathcal{G}$ contains four sub-graphs, each representing an important user interest. The four sub-graphs are described in a greater detail in the following paragraphs.

Consecutive Interest: According to the analysis presented in [31], among millions of POIs in location-based systems: 1) individuals typically visit only a small subset of POIs that appeal to them; and 2) some POIs are visited more frequently than others. This phenomenon demonstrates that human mobility contains some common transitional regularities behind their past check-ins. Therefore, we postulate that it is necessary to capture the consecutive correlations between distinct POIs to reveal human motion-based interests. Correspondingly, we formulate a weighted sub-graph $\mathcal{G}_c = (\mathcal{V}_l, E_c, \boldsymbol{A}_c)$ to describe such diverse correlations, where $\mathcal{V}_l$ is the set of distinct POIs, E_c is the edge set, and $\boldsymbol{A}_c \in \mathbb{R}^{|\mathcal{L}| \times |\mathcal{L}|}$ refers to the adjacency matrix. Given two POIs (e.g., POI l_i and POI l_j) that are visited in succession, we create an edge between them and then calculate the edge weight (i.e., entry $A_c^{ij} \in \boldsymbol{A}_c$) using the corresponding transitional probability. Formally, such an edge weight can be defined as:

$$A_c^{ij} = f_c^{ij} / f_c^i, \tag{3}$$

where f_c^{ij} refers to the frequency of edge $l_i \rightarrow l_j$ appearance in the check-in data, and f_c^i denotes the frequency of POI l_i appearance in the check-in data. As such, we are able to acquire the matrix $\boldsymbol{A}_c$ to preserve the consecutive interests underlying the trajectories.

Geographical Interest: As discussed in [17], people are more likely to visit nearby POIs than distant ones. Motivated by this, we formulate an undirected sub-graph $\mathcal{G}_g = (\mathcal{V}_l, E_g, \boldsymbol{A}_g)$ to describe such interactions, where E_g is the set of edges and $\boldsymbol{A}_g$ $(\in \mathbb{R}^{|\mathcal{L}| \times |\mathcal{L}|})$ denotes the adjacency matrix regarding geographical interests. Given POIs l_i and l_j, the edge weight A_g^{ij} $(\in \boldsymbol{A}_g)$ can be calculated as:

$$A_g^{ij} = \begin{cases} 0, & \mathrm{g}(l_i, l_j) > \Delta g; \\ 1, & \text{otherwise}. \end{cases} \tag{4}$$

Herein, $g(l_i, l_j)$ denotes the orthodromic (i.e., great-circle) distance function and Δg is a predefined threshold to restrict the impact of geographical noise. In this paper, we set $\Delta g = 3$ km. We detail this empirical setting in Appendix 7.1, available online.

Time–Aspect Interest: Each check-in, is associated with a visiting timestamp, reflecting the human temporal semantics. As it is a key factor for understanding any kind of human regularity, we propose to investigate the mutual interactions between POI and visiting time to obtain the time-aspect interest. However, since each visiting timestamp is actually a continuous value, we follow previous studies and aggregate all of the visiting timestamps into time intervals (i.e., time bins) with a duration of an hour (cf. [52], [53]). In addition, since people may respectively show different preferences on weekday and weekend, we thus assign 48 time bins to replace the original visiting timestamps, where the weekday and weekend are specified. With this, we proceed with formulating a weighted sub-graph $\mathcal{G}_t(\mathcal{V}_l \cup \mathcal{V}_t, E_t, \boldsymbol{A}_t)$, where $\boldsymbol{A}_t$ maintains human time-aspect interest. Similar to the above graph $\mathcal{G}_c$, we can also calculate the time-aspect interest between the POI l_i and time bin t_τ by:

$$A_t^{i\tau} = f_t^{i\tau} / f_t^i, \tag{5}$$

where $f_t^{i\tau}$ denotes the frequency of visiting POI l_i at time t_τ, and f_t^i is the total number of visits to the POI l_i has been visited.

Activity-Aspect Interest: A user who wants to post a check-in to LBSNs indicates that he/she is engaged in a specific type of activity that appeals to him/her. In practice, each POI has a contextual description (i.e., POI category) that reflects a real-world property/characteristic, enabling a certain activity. Thus, taking into account such contextual interactions is essential for understanding human preferences. Notably, the number of POI categories is much smaller than the number of POIs. As a result, linking a POI to its category can offer a coarse-grained perspective on the higher-order interactions between various POIs. To this end, we build an undirected graph $\mathcal{G}_a(\mathcal{V}_l \cup \mathcal{V}_a, E_t, \boldsymbol{A}_a)$ to describe the activity-aspect interest. To be more precise, we explicitly build an edge between a POI and the contextual category it belongs to, and then we treat each category as a regular node in $\mathcal{G}_a$.

2) Interest Aggregation: To extract the semantic contexts underlying POIs from the PGraph, we propose to adopt graph neural networks (GNNs) which have been widely applied in numerous graph-based tasks with a remarkable success [54].

Homogeneous Semantic Aggregation (HoSA): According to the structure of the built PGraph, we can find the consecutive interest and geospatial interest that belong to the homogeneous semantics as they only contain the nodes of POI identities. Thus, HoSA attempts to aggregate the underlying information from the nodes of the same type, i.e., POI identity. Since the consecutive correlation matrix $\boldsymbol{A}_c$ reflects human real-world transitional preferences, we can naturally regard each POI's transitional distribution as its prior feature to describe the relationship between a specific POI and its neighbors. To this end, we set each $\boldsymbol{A}_c^i$ as the initial feature of the POI node $\mathcal{V}_l^i$. Besides, the geospatial correlation matrix $\boldsymbol{A}_g$ preserves the geographical closeness between different POIs, providing the weak signal of human potential transitional tendencies. Hence, $\boldsymbol{A}_g$ can be regarded as an augmentation of the consecutive correlation matrix $\boldsymbol{A}_c$. Therefore, we merge these two matrices into a unified matrix $\boldsymbol{A}_h$ to reveal observed and unobserved preferences of transitional dependencies. Specifically, given two distinct POI nodes $\mathcal{V}_l^i$ and $\mathcal{V}_l^j$, the respective correlation score A_h^{ij} is defined as:

$$A_h^{ij} = \begin{cases} A_c^{ij}, \text{ if } A_c^{ij} \neq 0; \\ A_g^{ij}, \text{ others }. \end{cases} \tag{6}$$

For any POI node $\mathcal{V}_l^i$, we embed each POI node to a unified representation as follows:

$$s_i^l = \boldsymbol{A}_h^i \boldsymbol{W}_l + b_l, \tag{7}$$

where $\boldsymbol{W}_l \in \mathbb{R}^{|\mathcal{L}| \times d}$ and $b_l \in \mathbb{R}^d$ are trainable matrices. The dimension of s_i^l is d. Afterwards, each POI has its unique initial representation. To bridge the correlation between POI $\mathcal{V}_l^i$ and each of its neighbor $\mathcal{V}_l^j \in \Omega(\mathcal{V}_l^i)$, we devise a scoring function to evaluate the different contributions of the neighboring nodes. For instance, given a POI node $\mathcal{V}_l^i$ and its neighbor $\mathcal{V}_l^j$, we define the contribution measure as:

$$a(s_i^l, s_j^l) = \boldsymbol{b}_a^T [s_i^l \oplus s_j^l], \tag{8}$$

where $\oplus$ is the concatenation operation and $\boldsymbol{b}_a \in \mathbb{R}^{2d}$ is a learnable vector. Then, we follow the standard GAT [55] and use softmax function to normalize the attention scores across all neighbors of POI $\mathcal{V}_l^i$ (i.e., $\Omega(\mathcal{V}_l^i)$), where each attention score regarding its neighbor $\mathcal{V}_l^j$ can be formulated as:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a(s_i^l, s_j^l)))}{\sum_{k \in \Omega(\mathcal{V}_l^i)} \exp(\text{LeakyReLU}(a(s_i^l, s_k^l)))}. \tag{9}$$

In the end, we obtain the aggregated representation e_i^l of POI node $\mathcal{V}_l^i$ by a sum operation:

$$e_i^l = \sigma\left(\sum_{j \in \Omega(\mathcal{V}_l^i)} \alpha_{ij} s_j^l \boldsymbol{W}_e\right), \tag{10}$$

where σ is the sigmoid activation function and $\boldsymbol{W}_e \in \mathbb{R}^{d \times d}$ is a set of trainable parameters.

Heterogeneous Semantic Aggregation (HeSA): HeSA aims to aggregate the associated information of POIs from the neighboring nodes with different types. In our PGraph, there are two correlations that describe the heterogeneous semantics between different types of nodes, i.e., the time-aspect and activity-aspect interests. In contrast to HoSA, we do not involve the attention mechanism to quantify the different contributions of POI's heterogeneous neighbors. The reason is that the number of them is extremely smaller than that of the POIs, we thus attempt to capture all of the possible heterogeneous neighbors of a given POI directly to enhance the semantic information.

1) For the time–aspect interest, each POI l_i is associated with a probability distribution $\boldsymbol{A}_t^i$ ($\in \boldsymbol{A}_t$) that describes the preference strengths between that POI and time bins. We leverage the message-passing neural network inspired by [11] to incorporate the time-aspect preference of each POI, which can be formulated as:

$$e_i^t = \tanh\left(\boldsymbol{A}_t^i \mathbf{W}_t\right), \tag{11}$$

where $tanh$ is the activation function and $\mathbf{W}_t$ is a trainable matrix. Finally, we can obtain each POI's temporal context.

2) For the activity-aspect interest, we obtain each POI's activity-aware semantic by:

$$e_i^a = \tanh\left(\boldsymbol{A}_a^i \mathbf{W}_a\right), \tag{12}$$

where $\mathbf{W}_a$ is a trainable matrix. Finally, the homogeneous and heterogeneous semantics behind each POI are acquired by HoSA and HeSA, respectively. In the following mobility encoding procedures, we will use these contextual representations as the embeddings of POIs in user trajectories. These embeddings can be jointly optimized during self-supervised learning and task learning.

B. Context-Aware Mobility Encoding

Existing studies often choose recurrent neural networks such as Long-short Term Memory (LSTM) or Gated Recurrent Unit (GRU) to capture human transitional regularities [56]. Since the complex stacked gate operations in LSTM typically struggle with the gradient vanishing problem, we select GRU as the kernel of our mobility encoder. For each l_τ in a given trajectory $T = \{l_1, l_2, \ldots, l_n\}$, we have collected the homogeneous and heterogeneous semantics behind it. In this way, they can be viewed as reflections of different interests in different domains. Therefore, we extend the GRU cell to capture the sequential information as well as the contextual information behind each POI. Correspondingly, the recursive process with GRU can be formulated as follows:

$$c_\tau = [e^l_\tau \oplus e^t_\tau \oplus e^a_\tau]\boldsymbol{W}_f + \boldsymbol{b}_f, \tag{13}$$

$$\boldsymbol{h}_\tau = \text{GRU}(c_\tau, \boldsymbol{h}_{\tau-1}), \tag{14}$$

where $\boldsymbol{h}_\tau$ and $\boldsymbol{h}_{\tau-1}$ are the hidden states of the current POI l_τ and the last visited POI $l_{\tau-1}$, respectively. Herein, c_τ is the contextual embedding of POI l_τ, which is a unified representation that integrates the homogeneous and heterogeneous semantics of POI l_τ (they include e^l_τ, e^t_τ, and e^a_τ). In addition, $\boldsymbol{W}_f$ and $\boldsymbol{b}_f$ are trainable parameters.

C. Self-Supervised Disentanglement Learning

Existing efforts mainly focus on modeling sequential dynamics along with fusing multiple affiliated contexts to formulate a unified representation for each trajectory, inevitably yielding semantic entanglement [2], [12], [22]. To uncover human intrinsic periodicity/behaviors and constantly changing intents, we build self-supervised disentanglement learning in SSDL, aiming at performing self-supervision in the latent space and seeking a clean separation of the time-invariant and time-varying factors.

1) Probabilistic Generative Disentanglement: Given any recent trajectory $T = \{l_1, l_2, \ldots, l_n\}$, we attempt to learn a set of time-varying variables $z^r_{1:n} = \{z^r_1, z^r_2, \ldots, z^r_n\}$ and a time-invariant variable z^s, where $z^r_{1:n}$ aims at exploring the dynamics of human time-dependent interests while z^s undertakes the role of learning human inherent time-independent periodicity (habits). Formally, let z_τ be the entangled latent code of check-in l_τ, and we have $z_\tau = (z^r_\tau, z^s)$. For consistency, let $l_{1:n}$ denote the check-in sequence $\{l_1, l_2, \ldots, l_n\}$. As people's future movements are affected by their previous check-in behaviors, we assume that each z_τ depends on its previous states $z_{<\tau} = \{z_1, z_2, \ldots, z_{\tau-1}\}$. In addition, as user's long-standing interests will not be changed dramatically by recent activities, we assume that $z^r_{1:n}$ and z^s are independent from each other, i.e., $p(z_{1:n}) = p(z^r_{1:n})p(z^s)$. Hence, we formulate the complete probabilistic generative model as follows:

$$\begin{aligned} p(l_{1:n}, z_{1:n}) &= p(z_{1:n})p(l_{1:n}|z_{1:n}) \\ &= [p(z^s)\prod_{\tau=1}^{n} p(z^r_\tau|z^r_{<\tau})] \cdot \prod_{\tau=1}^{n} p(l_\tau|z^r_\tau, z^s). \end{aligned} \tag{15}$$

Prior Settings: We choose the Gaussian distribution as the prior $p(z^s)$, i.e., $p(z^s) \sim \mathcal{N}(0, \boldsymbol{I})$. Following the rule of variational Bayes, we set $\mathcal{N}(\mu(z_{<\tau}), \sigma^2(z_{<\tau}))$ as $p(z^r_\tau|z^r_{<\tau})$, where $\mu(\cdot)$ and $\sigma^2(\cdot)$ can be modeled by popular recursive networks. For instance, we can use GRU to obtain z^r_τ as follows:

$$\begin{aligned} \boldsymbol{h}^r_\tau &= \text{GRU}(\boldsymbol{h}_\tau, z^r_{\tau-1}), \\ z^r_\tau &= \mu(\boldsymbol{h}^r_\tau) + \sigma(\boldsymbol{h}^r_\tau) \odot \epsilon, \end{aligned} \tag{16}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ and $\odot$ is element-wise multiplication.

Posterior Settings: Subsequently, we expect to produce a posterior distribution $q(z_{1:n}|l_{1:n})$ to cater to the learning manner of variational inference. Thus, we can factorize the posterior distribution $q(z_{1:n}|l_{1:n})$ as follows:

$$\begin{aligned} q(z_{1:n}|l_{1:n}) &= q\left(z^s, z^r_{1:n} \mid l_{1:n}\right) \\ &= q(z^r_{1:n}|l_{1:n})q(z^s|l_{1:n}) \\ &= q\left(z^s \mid l_{1:n}\right)\prod_{\tau=1}^{n} q\left(z^r_\tau \mid z^r_{<\tau}, l_{\leq\tau}\right). \end{aligned} \tag{17}$$

In practice, we take our Context-aware Mobility Encoding network as an inference model to generate factorized posterior distribution.

At last, we can formulate the Evidence Lower BOund (ELBO) of our disentanglement learning as follows:

$$\begin{aligned} \textbf{ELBO}: \max_{p,q} \mathbb{E}_{l_{1:n}\sim p_D}\mathbb{E}_{q(z_{1:n}|l_{1:n})}[&\log p\left(l_{1:n} \mid z_{1:n}\right) \\ &- KL\left[q\left(z_{1:n} \mid l_{1:n}\right) \| p(z_{1:n})\right]], \end{aligned} \tag{18}$$

where p_D is the empirical trajectory (mobility) distribution. As $z_{1:n}$ is comprised of mutually independent z^s and $z^r_{1:n}$, the second term of KL-divergence can be disentangled as:

$$\begin{aligned} &KL\left[q\left(z_{1:n} \mid l_{1:n}\right) \| p(z_{1:n})\right] \\ &= KL\left[q\left(z^s \mid l_{1:n}\right) \| p(z^s)\right] + KL\left[q\left(z^r_{1:n} \mid l_{1:n}\right) \| p\left(z^r_{1:n}\right)\right]. \end{aligned} \tag{19}$$

Following the principle of VAE [40], [47], we provide a theoretical derivation to illustrate the above modeling processes:

$$\begin{aligned} &\log p\left(l_{1:n}\right) \\ &\geq -KL\left[q\left(z_{1:n} \mid l_{1:n}\right) \| p\left(z_{1:n} \mid l_{1:n}\right)\right] + \log p\left(l_{1:n}\right) \\ &= -KL\left[q\left(z^s, z^r_{1:n} \mid l_{1:n}\right) \| p\left(z^s, z^r_{1:n} \mid l_{1:n}\right)\right] + \log p\left(l_{1:n}\right) \\ &= \mathbb{E}_{q\left(z^s, z^r_{1:n}|l_{1:n}\right)}[\log p\left(l_{1:n} \mid z^s, z^r_{1:n}\right) \\ &\quad + \log p\left(l_{1:n}\right) - \log q\left(z^s, z^r_{1:n} \mid l_{1:n}\right)] \\ &= \mathbb{E}_{q(z^s, z_{1:n}|l_{1:n})}[\log p\left(l_{1:n} \mid z^s, z^r_{1:n}\right) \\ &\quad - \log q\left(z^s, z^r_{1:n} \mid l_{1:n}\right) + \log p\left(z^s, z^r_{1:n}\right)] \\ &= \mathbb{E}_{q\left(z^s, z^r_{1:n}|l_{1:n}\right)}[\log p\left(l_{1:n} \mid z^s, z^r_{1:n}\right) - \log q\left(z^s \mid l_{1:n}\right) \\ &\quad - \log q\left(z^r_{1:n} \mid l_{1:n}\right) + \log p(z^s) + \log p\left(z^r_{1:n}\right)] \\ &= \mathbb{E}_{q\left(z^s, z^r_{1:n}|l_{1:n}\right)}\left[\log p\left(l_{1:n} \mid z^s, z^r_{1:n}\right)\right] \\ &\quad - KL\left[q\left(z^s \mid l_{1:n}\right) \| p(z^s)\right] \\ &\quad - KL\left[q\left(z^r_{1:n} \mid l_{1:n}\right) \| p\left(z^r_{1:n}\right)\right]. \end{aligned} \tag{20}$$

Recall that the results of the above derivation are similar to the results in standard VAE, which usually confront the agnostic prior distribution that causes posterior collapse problem and leaves the learned latent space still entangled [57]. This phenomenon has been revealed in recent studies, e.g., β-VAE [38] and β-TCVAE [39]. Additionally, [58] provides a clearer perspective that reveals the challenges with disentangled representation in variational inference. Thus, we conjecture that the last two terms regularized by KL-divergence in (20) are hard to close to their corresponding prior, which would make each posterior become non-informative. For the purpose of receiving clean disentanglement of z^s and $z^r_{1:n}$, we are inspired by recent self-supervised learning and enforce disentanglement learning from the perspective of mutual information.

2) Mutual Information Regularization: We now present the details on how to combine contrastive learning with disentangled mobility learning. We first introduce variational mobility learning from the perspective of Mutual Information (MI). The goal of MI is a measure of the mutual dependence between two variables. Since both z^s and $z^r_{1:n}$ are derived from the original trajectory, we thus add three additional MI terms to regularize the latent space of them, which can be defined as follows:

$$\begin{aligned}\mathcal{J}_{self} = \max_{p,q} & \mathbb{E}_{l_{1:n}\sim p_D}\mathbb{E}_{q(z_{1:n}|l_{1:n})}[\log p\left(l_{1:n} \mid z_{1:n}\right) \\ & - \alpha(KL\left[q\left(z^s \mid l_{1:n}\right) \| p(z^s)\right] \\ & + KL\left[q\left(z^r_{1:n} \mid l_{1:n}\right) \| p\left(z^r_{1:n}\right)\right]) \\ & + \beta(MI\left(z^s; l_{1:n}\right) + MI\left(z^r_{1:n}; l_{1:n}\right)) \\ & - \gamma MI\left(z^r_{1:n}; z^s\right),\end{aligned} \tag{21}$$

where α, β and γ are weight coefficients. $MI(\cdot,\cdot)$ refers to MI term. For instance, $MI(z^s; l_{1:n})$ is defined as:

$$\mathbb{E}_{q(z^s,l_{1:n})}\left[\log \frac{q(z^s|l_{1:n})}{q(z^s)}\right]. \tag{22}$$

We note that other MI terms have the similar formulation (see Appendix 7.2, available online for the complete derivation of (21). Now our goal becomes enforcing the posteriors matching with their corresponding priors while ensuring that z^s and $z^r_{1:n}$ are disentangled from each other. To estimate the MI terms, we follow most of recent studies [34], [48], [49] and employ the NCE loss to make contrastive estimation. For instance, a contrastive estimation of $MI(z^s; l_{1:n})$ can be defined as follows:

$$\mathcal{C}_{z^s} \approx \mathbb{E}_{p_D} \log \frac{\psi\left(z^s, l^+_{1:n}\right)}{\psi\left(z^s, l^+_{1:n}\right) + \sum_{j=1}^m \psi\left(z^s, \tilde{l}^j_{1:n}\right)}, \tag{23}$$

where $\psi(\cdot,\cdot) = exp(sim(\cdot,\cdot)/\eta)$, $sim(\cdot,\cdot)$ denotes the cosine similarity function, m is the number of negative trajectories, and $\eta = 0.5$ is a temperature parameter. Notably, we treat $l_{1:n}$ as the *positive* trajectory sequence regarding z^s and specify it using $l^+_{1:n}$. Besides, $\tilde{l}^j_{1:n}$ refers to a *negative* sample (trajectory), which is generated from the other users.

Augmentation for time-invariant factor: Due to the limited scale of *positive* samples, we try to generate more realistic trajectories to augment the original samples. As z^s reveals human intrinsic periodicity and should not be affected by recent moving behaviors (i.e., time-independent), we can thus randomly change the order of a given trajectory $l_{1:n}$ and formulate several augmentation versions w.r.t $l_{1:n}$. We claim that it is a simple but efficient strategy to obtain rich augmented samples. Correspondingly, the contrastive estimation on these samples can be denoted as:

$$\mathcal{C}^{\#}_{z^s} \approx \mathbb{E}_{p_D} \log \frac{\psi\left(z^s, l^{\#}_{1:n}\right)}{\psi\left(z^s, l^{\#}_{1:n}\right) + \sum_{j=1}^m \psi\left(z^s, \tilde{l}^j_{1:n}\right)}, \tag{24}$$

where $l^{\#}_{1:n}$ indicates it is an augmentation version of $l_{1:n}$. For time-invariant factor z^s, we use the collected samples including augmented samples to make the final estimates:

$$MI\left(z^s; l_{1:n}\right) \approx \frac{1}{2}(\mathcal{C}_{z^s} + \mathcal{C}^{\#}_{z^s}). \tag{25}$$

Augmentation for Time-Varying Factor: $z^r_{1:n}$ is a set of latent variables regarding $l_{1:n}$, showing human human time-dependent interests. Similar to (23), we can obtain the contrastive estimation of $MI(z^r_{1:n}; l_{1:n})$ as $\mathcal{C}_{z^r_{1:n}}$. Furthermore, we provide another data augmentation method to enhance the optimization of $MI(z^r_{1:n}; l_{1:n})$. The intuition is that $z^r_{1:n}$ is a set of time-dependent variables. In practice, real-world check-in data could be subject to noise and uncertainty due to the presence of collective POIs [59]. Hence, a user usually posts a fuzzy POI to replace her accurate position, which could weaken human mobility pattern learning and even result in inaccurate predictions. Motivated by [59], [60], it is encouraging that we can use any member of related collective POI to replace the original POI in a trajectory to obtain an augmentation trajectory for time-varying factor training, which will not change any temporal semantics. In addition, another potential benefit of such a practice is to alleviate the uncertainty issue behind diverse human check-in behaviors. In our implementation, we use neighbors of the same category within 300 m of a given POI as members of its collective POI and replace about 30% of the POIs in a given trajectory with their related collective POIs, which does not heavily affect the full semantics behind the original trajectory. As a result, we can obtain a large number of synthetic trajectories that provide multiple views of a given trajectory. Similar to (25), we can get the final estimation regarding $z^r_{1:n}$ as:

$$MI\left(z^r_{1:n}; l_{1:n}\right) \approx \frac{1}{2}(\mathcal{C}_{z^r_{1:n}} + \mathcal{C}^{*}_{z^r_{1:n}}), \tag{26}$$

where $\mathcal{C}^{*}_{z^r_{1:n}}$ is the contrastive estimation of the augmented trajectories regarding time-varying factors. As for the final term $MI(z^r_{1:n}; z^s)$, the variables in it are all in the latent space – thus, we can directly choose the standard mini-batch weighted sampling (MWS) [39] for comparative estimation.

D. Task Learning

So far, we have obtained a set of time-varying variables and a time-invariant variable for each trajectory. We now turn to use our task learning network to predict the next POI. For each user, we actually have his/her entire historical trajectory. Inspired by [6], [61], modeling such a long trajectory would boost the capture of human long-term transitional preferences. It is natural to adopt the RNN to encode the transitional regularity underlying human historical trajectory. However, in practice, there are massive time-ordered POIs in their historical trajectories, which usually result in a serious time-cost problem. Therefore, we employ a self-attention layer with position encoding to capture the taste of the transitional behavior of a user, as well as the long-distance dependencies. Given a user's entire historical trajectory $\mathcal{T}_{1:\mathcal{K}}$ containing K ordered POIs, we first reuse the linear layer (cf. Eq13) to obtain the dense representation of each POI in $\mathcal{T}_{1:\mathcal{K}}$. Correspondingly, we use $\boldsymbol{\mathcal{T}}_{1:\mathcal{K}}$ to denote the trajectory with embedded POIs. To determine the order of POIs in $\boldsymbol{\mathcal{T}}_{1:\mathcal{K}}$, we follow [62] and use the sine/cosine function-based position embedding to formulate the final representation of each POI, which can be denoted as:

$$\boldsymbol{\mathcal{T}}'_i = \boldsymbol{\mathcal{T}}_i + \Phi(\boldsymbol{\mathcal{T}}_i), \tag{27}$$

where $\Phi(\boldsymbol{\mathcal{T}}_i)$ is the position embedding of POI l_i in $\mathcal{T}_{1:\mathcal{K}}$. Then, we employ a one-layer self-attention network to receive a set of hidden states regarding $\boldsymbol{\mathcal{T}}_{1:\mathcal{K}}$, as follows:

$$\boldsymbol{H}_{1:\mathcal{K}} = \text{self-att}(\boldsymbol{\mathcal{T}}'_{1:\mathcal{K}}). \tag{28}$$

In our study, we use the last state $\boldsymbol{H}_{\mathcal{K}}$ to represent $\mathcal{T}_{1:\mathcal{K}}$ and regard it as one of the inputs for task prediction.

Now we take $z^r_{1:n}$, z^s, and $\boldsymbol{H}_{1:\mathcal{K}}$ as the input and employ a one-layer fully-connected network with softmax function to obtain the predicted POI. The process can be expressed as:

$$\hat{l}_{n+1} = \arg\max(\text{softmax}([z^r_n \oplus z^s \oplus \boldsymbol{H}_{\mathcal{K}}]\boldsymbol{W}_t + \boldsymbol{b}_t)). \tag{29}$$

Correspondingly, the loss function for trajectory T can be expressed as follows:

$$\mathcal{L}_T = -l_{n+1} \log \hat{l}_{n+1}. \tag{30}$$

To minimize the above cross-entropy loss, we employ Adam algorithm to optimize the parameters [63]. We outline the complete pipeline of training SSDL in Appendix 7.3, available online.

V. Experiments

We now present our experimental evaluation of the performance of SSDL on real-world datasets, compared to state-of-the-art baselines. We also present an ablation study, along with interpretability and sensitivity analysis. To help other researchers, our source codes are also publicly available at https://github.com/yyyyyhjy/SSDL.

A. Experimental Settings

Datasets: To facilitate reproducible results, we conduct all experiments using data from two publicly available LBS applications: Foursquare [64] and Gowalla [65]. We use two Foursquare datasets, containing check-ins from New York and Tokyo collected from 12 April 2012 to 16 February 2013. Each check-in has a timestamp, GPS coordinates and semantics about it. From Gowalla we also select the data from two cities, i.e., Los Angeles and Houston. Following previous studies [6], [34], we filter out the POIs visited less than eight times. For each user, we concatenate all of his/her chronological check-ins and divide each trajectory into subsequences with the time interval of 24 hours. To specify whether check-ins are collected on weekdays or weekends, we further assign 48 time slots to each check-in time. We take each user's first 80% trajectories as the training set, the remaining 20% as the test set. The statistics of the four datasets are summarized in Table I.

TABLE I
Statistics of the Datasets

City	Users	POIs	Check-ins	Trajectories
Tokyo (TKY)	2102	6789	240056	60365
New York (NYC)	990	4211	79006	23252
Los Angeles (LA)	2346	8676	195231	61542
Houston (HOU)	1351	6994	121502	37514

Baselines: We compare our SSDL with the following representative approaches for next POI prediction task:

- *GRU* [21] is a common approach for sequential data learning as its superiority in incorporating the semantics of long-term dependencies.
- *ST-RNN* [22] is an RNN-based method that incorporates spatio-temporal contexts when predicting the next POI.
- *HST-LSTM* [12] employs the sequence-to-sequence learning scheme to include spatial-temporal influence in LSTM and makes use of contextual information to enhance model performance for sparse data prediction.
- *Flashback* [66] models sparse user mobility footprints by doing flashbacks on hidden states in RNNs. Especially, it explicitly employs the spatio-temporal contexts to search past hidden states with high predictive power. In our experiments, we take the GRU cell as the recurrent component in Flashback for a fair comparison.
- *DeepMove* [61] presents an attention-based RNN to encode human recent trajectories. Furthermore, it employs another RNN to learn user long-term preferences from historical trajectories.
- *VANext* [6] proposes a novel variational attention mechanism to explore human periodic regularities. In addition, it employs a simple convolutional neural network (rather than RNN) to capture human long-term interests.
- *PLSPL* [10] is a unified framework that jointly learns users' long- and short-term interests for next POI prediction.
- *MobTCast* [2] is a Transformer-based approach that considers multiple semantic contexts behind check-ins to enhance the understanding of human mobility. Note that we remove the Social Context Extractor in MobCast as the social relationships are not available in our context.
- *β-VAE* [38] is a widely used representation learning method that is able to separate latent factors into different space by using an adjustable hyperparameter β to the original VAE objective. In this study, we use GRU as the encoder and decoder network structure in β-VAE to model the temporal semantics.
- SML [34] attempts to understand human mobility with trajectory augmentation in a self-supervised learning manner. Especially, it leverages a heuristic strategy to enumerate massive different views of original sparse trajectories for contrastive estimation.

Metrics: To evaluate the performance of SSDL, we follow most of the previous studies [6], [10], [34] and select three commonly used metrics to compare against the baselines. We first use

TABLE II
PERFORMANCE COMPARISONS ON FOUR CITIES

Method	Tokyo					New York				
	ACC@1	ACC@5	ACC@10	AUC	MAP	ACC@1	ACC@5	ACC@10	AUC	MAP
GRU	13.11	27.88	34.28	88.01	7.36	15.37	31.73	36.10	81.40	8.59
ST-RNN	13.38	29.20	36.45	89.82	7.41	13.50	32.86	40.05	81.91	8.20
HST-LSTM	18.70	39.14	46.47	90.66	9.82	17.48	42.77	50.82	86.25	8.53
Flashback	18.23	39.42	46.66	90.40	10.47	22.22	49.52	57.11	87.74	13.59
DeepMove	19.92	40.61	48.25	90.47	12.21	21.56	45.09	52.17	87.30	13.06
VANext	20.21	44.49	52.63	91.30	12.36	22.54	51.26	58.78	89.30	14.02
PLSPL	20.19	43.64	52.45	91.37	12.86	23.02	53.33	63.34	89.21	14.83
MobTCast	19.58	43.41	51.95	89.95	11.67	22.37	54.31	64.18	88.59	14.03
β-VAE	20.10	44.78	53.89	91.35	12.75	22.26	50.71	58.68	89.38	14.07
SML	20.25	44.70	53.58	91.48	12.51	22.62	52.16	60.18	90.17	14.74
SSDL	**22.93**	**46.80**	**55.31**	**92.39**	**14.99**	**25.07**	**56.78**	**65.72**	**90.72**	**16.60**

Method	Los Angeles					Houston				
	ACC@1	ACC@5	ACC@10	AUC	MAP	ACC@1	ACC@5	ACC@10	AUC	MAP
GRU	10.11	19.05	22.67	78.07	4.97	10.74	18.01	21.33	80.47	5.99
ST-RNN	10.01	19.30	23.64	80.48	5.11	11.33	20.21	24.81	82.34	6.88
HST-LSTM	12.01	23.97	29.04	82.57	5.29	13.41	22.86	27.21	82.58	6.58
Flashback	13.81	25.60	30.32	83.74	7.65	14.37	24.52	28.70	84.54	8.79
DeepMove	13.31	25.73	30.35	82.41	7.26	14.13	24.59	29.03	83.29	8.46
VANext	14.36	27.91	33.16	86.22	7.73	14.88	26.78	31.44	86.06	8.31
PLSPL	14.92	28.26	33.86	84.34	7.97	16.06	29.22	34.67	86.11	9.74
MobTCast	14.30	28.62	33.38	83.41	7.65	15.40	28.27	32.87	83.87	8.66
β-VAE	14.39	27.43	32.96	85.95	7.72	15.00	27.09	31.89	86.10	8.23
SML	14.77	28.12	33.35	86.38	7.86	15.15	27.46	32.31	86.48	9.17
SSDL	**15.94**	**31.02**	**36.80**	**86.98**	**8.72**	**16.91**	**30.92**	**36.56**	**87.30**	**10.70**

the ACC@K to evaluate the recommendation performance. In this paper, we report the different testing results of $K = 1, 5, 10$. Additionally, we report area under the ROC curve (AUC) and mean average precision (MAP) metrics that are frequently used in classification tasks.

Implementation Details: We implemented our SSDL and baselines in Python. All methods are based on the Torch library and accelerated by one NVIDIA GTX 1080 GPU. We used Adam [63] to train all deep learning methods. In all baselines except GRU, we also follow their settings for check-in/POI embeddings. For the GRU, we randomly initialized the embeddings for each check-in. In disentanglement learning, the learning rate is initialized as 0.01. We set the coefficient α of KL terms to 1, and β and γ are fixed to be 1 and 0.1, respectively. In task learning, the learning rate is initialized with 5e-4. The dropout rate is set as 0.5, and the batch size is 32. We set other users' trajectories in a mini-batch as the negative samples of a given user. The hidden size of the self-attention network is set to 300. In addition, we set dimension of z^s to 256, while $z^r_{1:n}$ to 32. The dimensions of POI, time, and category are set as 256.

B. *Performance Comparisons*

Table II reports the performance of different approaches on the datasets of four cities, where the best result is represented with **bold** font and the second best is marked with underline. Below, we discuss the main observations.

We observe that ST-RNN does not yield competitive achievements compared to GRU, although it considers the spatial and temporal constraints. A possible reason is that the sparsity issue of check-in data heavily affects the distillation of semantics contexts such as geographical distance. Meanwhile, relying on simple spatio-temporal features and regarding the next POI as the solo supervision usually results in an inference bias or uncertainty problem due to the boundary of available training datasets. To mitigate the data sparsity issue, HST-LSTM which combines spatial and temporal factors with a gate mechanism is able to boost the capture of human mobility patterns by a large margin. Furthermore, HST-LSTM models the periodicity of consecutive check-ins in an end-to-end manner, which brings an encouraging prospect for us to learn the complex distribution behind historical trajectories. Compared with HST-LSTM, which directly adds spatio-temporal factors to hidden states, Flashback achieves competitive performance because it explicitly uses a rich spatio-temporal context to search for past hidden states with high predictive power to predict the next POI.

As for DeepMove and VANext, they both attempt to correlate certain user's recent trajectory and historical trajectory to accurately discover individual periodicity. Our experiments show that they achieve higher gains than models (e.g., ST-RNN) that only consider the past few check-ins. Furthermore, VANext, the first variational inference approach to model human trajectories using a prior assumption, outperforms DeepMove due to the relief of the inherent uncertainty of user mobility. The paradigm of PLSPL is similar to DeepMove, but it operates an attention mechanism to evaluate the importance of each POI in a user's historical check-ins, aiming at exploring the tastes of different users. We found that PLSPL performs better than DeepMove. In addition, MobTCast is a Transformer-based approach that uses self-attention to study the interactive signals between POIs in a given trajectory, as well as multiple semantic contexts, such as category and temporal semantics. We obtain similar performance results compared to PLSPL, indicating that considering multiple semantic contexts indeed helps to discover users' future check-in intentions.

β-VAE is a popular disentanglement learning method that also obtains promising results, which suggests that employing the latent variables produced by variational Bayesian helps in understanding the inherent generative factors underlying human mobility. As for SML, it is the first self-supervised learning solution for the next POI prediction, achieving the best gains on AUC among the baselines. The reason is that it primarily seeks to produce massive synthetic trajectories for data augmentation and leverage contrastive learning to study the diversity of human moving intents behind existing historical check-ins.

In general, our proposed SSDL significantly outperforms the compared approaches by a relatively large margin across all the four datasets. For instance, SSDL respectively yields 8.57% and 11.94% averaged improvement over the best baseline regarding ACC@1 and MAP. Compared to previous deep mobility models, we consider that the superiority of our proposed SSDL is due to two major reasons. First, the PGraph we constructed is able to distill rich meaningful contexts behind human check-ins, which can further facilitate the capture of different interests in human daily movement behaviors. Second, our self-supervised disentanglement learning in SSDL aims to separate time-independent and time-dependent factors behind massive trajectories, allowing us to capture human inherent habits and susceptible near-term intentions. In particular, two realistic trajectory augmentation methods enable the enhancement of trajectory representation learning from the perspective of maximizing the mutual information, primarily seeking to address the inherent sparsity issue of human trajectory data.

C. Ablation Study

To evaluate the contribution of different components of SSDL, we considered several variants from two basic perspectives: POI embedding and trajectory embedding. First, we select 9 popular embedding methods to scrutinize the efficacy of our POI embedding.

- *One-Hot* [61] is the simplest method that maps each POI to a unique vector without any semantic information.
- *Random* [67] uses a dense matrix sampled from a Gaussian distribution to represent the POIs.
- *Word2vec* is a popular embedding technique in NLP, aiming at exploring the surrounding context of a given word. Also, it has successfully applied in POI embedding [14], [31]. We implement the skip-gram model for POI embedding.
- *Causal* [6] is a variant of word2vec that treats the previous footprints of the current POI as its semantic context to incorporate human practical transitional behaviors.
- *Deepwalk* [17] is a data augmentation method that builds a POI graph to integrate users' historical visiting interests and geographical proximity and then leverages the skip-gram technique for POI embedding.
- *GraphAE* [68] is popular method for node embedding. In this paper, we treat each POI as a node and build the same graph as Deepwalk for POI embedding.
- *GraphVAE* [68] is a variant of GraphAE, taking advantage of VAE for POI embedding.
- *HAN* [69] is a recent representative method in heterogeneous information aggregation. We follow it to generate more POI-centric meta-paths and each low-dimensional embedding for each POI.
- *RGCN* [70] is a graph-based method for capturing the heterogeneity. We follow it to use multiple weight matrices to project the POI embeddings into different relation spaces.

Fig. 2 reports the performance of SSDL using different POI embedding methods. In most cases, One-Hot and Random methods perform worse than other methods as they cannot absorb any semantic information. A possible reason for DeepWalk's superiority over other traditional methods is its ability as an augmentation method to enrich the context of adjacent relationships. Besides, the superiority of GraphAE over GraphVAE suggests that GraphAE successfully captures relations between non-adjacent POIs, whereas GraphVAE may confront a severe *Posterior Collapse* problem. In contrast to recent heterogeneous networks (e.g., HAN), we perform interest aggregation in terms of homogeneous and heterogeneous aspects, i.e., HoSA and HeSA, respectively. And this mitigates interference between multiple interests and enhances semantic information capture. We can observe that our embedding method achieves the best results on the vast majority of metrics across the four cities, which indicates its higher effectiveness in capturing multiple human interests behind historical trajectory data.

Next, we turn to investigate the effectiveness of devised components in SSDL. Herein, we conduct experiments with four SSDL variants. The details are shown as follows: **SSDL-Base** is a basic model that removes both graph-based embedding and mutual information regularization of SSDL. Instead, we use the word2vec technique for POI embedding. **SSDL w/o G** only removes the graph-based embedding and uses the word2vec for POI embedding. **SSDL w/o H** only removes the mutual information regularization of SSDL. **SSDL w/o N** only removes heterogeneous nodes and edges of the POI-centric graph.

Fig. 3 illustrates the performance of each variant in four cities. First, we can see that removing any modules would bring significant performance degradation (e.g., SSDL-Base performs the worst), suggesting that both modules in our SSDL benefit to enhance POI prediction. Second, SSDL w/o G performs worse than SSDL across all cities, which demonstrates that considering multiple common interests behind historical check-in data is useful to discover human mobility patterns. Third, SSDL outperforming SSDL w/o H proves that our self-supervised disentanglement learning is an effective module to provide promising representations for task inference. In practice, our mutual information regularization ensures that time-invariant and time-varying factors are disentangled from each other. If such a component is removed, SSDL will degrade to work as β-VAE, which leads to poor performance of disentanglement learning. The results of SSDL w/o N indicate that heterogeneous semantics do help to improve prediction performance since it exposes diverse interests that may affect human future preferences or intents.

D. Disentanglement Interpretability

In this part, we focus on studying the disentangled representations from the interpretability aspect. We first investigate whether z^s and $z^T_{1:n}$ can be well extracted from original

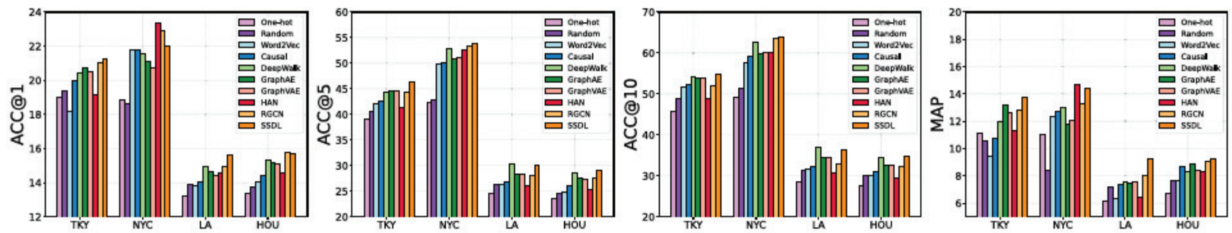


Fig. 2. Effects of different POI embedding methods.

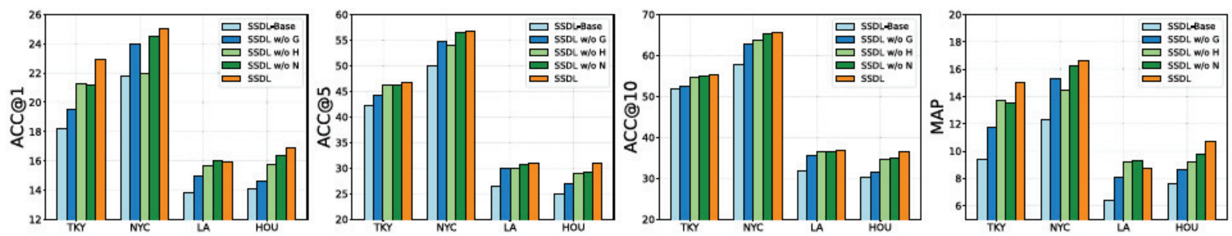
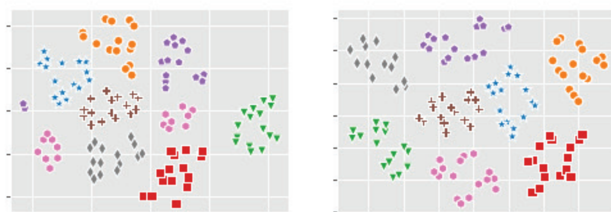


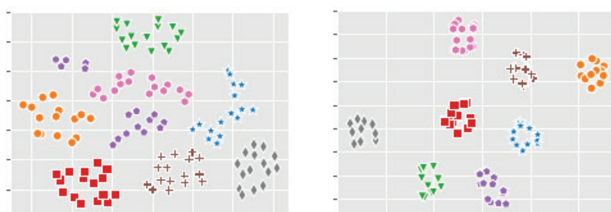
Fig. 3. Effects of different components in SSDL.

(a) z^s in SSDL. (b) z_n^r in SSDL.

Fig. 4. Visualization of latent representations in SSDL.



(a) β-VAE. (b) SSDL w/o Aug.



(c) SSDL w/o Random. (d) SSDL.

Fig. 5. Impact of augmentations on z^s.

trajectories and reflect human time-invariant periodicity/habits and time-varying interests, respectively. To this end, we randomly sample eight different users' trajectories and change their orders to generate several groups of trajectories. Then, we used the TSNE toolkit [71] to visualize the distribution time-invariant representations. As shown in Fig. 4(a), we can find that the representations of z^s produced by SSDL are grouped well, demonstrating that it can successfully separate the time-invariant factors to uncover the inherent preference of users that are not influenced by temporal factors. For $z_{1:n}^r$, as shown in Fig. 4(b), we visualize the distribution of the last states z_n^r for simplicity and we found that they are entangled, indicating that they are really affected by the temporal factors. Therefore, we conclude that z^s and $z_{1:n}^r$ indeed play well the roles of time-invariant and time-varying representations, respectively.

We also study the impact of our data augmentation approaches from a visualization perspective. We visualize the z^s distribution of randomly sampled trajectories of eight different users after task training. As shown in Fig. 5(a), we found that β-VAE can only separate the representations with a small margin. Fig. 5(b) presents the results of SSDL without any data augmentations, and Fig. 5(c) shows the results of SSDL that have no augmentation for time-invariant factors. Compared to Fig. 5(d), we can clearly find that both augmentations used in SSDL can significantly help us distinguish the trajectory representations of different users. This observation further suggests that different user movement patterns can be well refined by our SSDL.

Now we attempt to manipulate z^s or $z_{1:n}^r$ generated by SSDL to estimate whether they indeed learn time-invariant habits (z^s) and time-varying interests ($z_{1:n}^r$), respectively. In practice, the initial goal of SSDL is to encode diverse human trajectories into latent space (i.e., the time-invariant and time-varying vectors). In turn, these latent vectors are able to reconstruct the original trajectory based on variational Bayes [47]. As such, we manipulate the generated z^s and $z_{1:n}^r$ to reconstruct the input trajectories. Note that we follow [72] and use F_1 score

TABLE III
THE PERFORMANCE OF MOBILITY RECONSTRUCTION REGARDING F_1

City	TKY	NYC	LA	HOU
(S1) do nothing	0.8472	0.6715	0.8018	0.8776
(S2) add noise to $z^r_{1:n}$	0.8231	0.6457	0.7794	0.8604
(S3) add noise to z^s	0.6452	0.3733	0.5207	0.6441
(S4) change the order of $z^r_{1:n}$	0.8471	0.6605	0.8080	0.8795
(S5) re-initialize z^s	0.2739	0.0426	0.0626	0.1144

TABLE IV
THE PERFORMANCE OF MOBILITY RECONSTRUCTION REGARDING PAIRS-F_1

City	TKY	NYC	LA	HOU
(S1) do nothing	0.7803	0.6398	0.7633	0.8488
(S2) add noise to $z^r_{1:n}$	0.5801	0.4527	0.5732	0.6855
(S3) add noise to z^s	0.5504	0.3150	0.4594	0.5741
(S4) change the order of $z^r_{1:n}$	0.5182	0.4062	0.5694	0.6044
(S5) re-initialize z^s	0.1547	0.0166	0.0451	0.0705

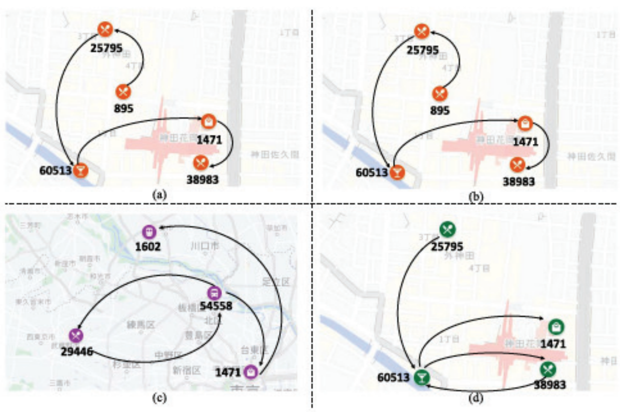


Fig. 6. Case study: (a) Ground truth trajectory; (b) Result based on S1; (c) Result based on S3; and (d) Result based on S4.

and pairs-F_1 score to evaluate the reconstruction performance, where F_1 is the harmonic mean of Precision and Recall of POIs in a trajectory while pairs-F_1 considers both POI correctness and sequential order. Specifically, Tables III and IV report the results with different settings of z^s and $z^r_{1:n}$. We have the following observations: First, **S1** indicates that we use the generated latent vectors to reconstruct the input trajectory directly, and we observe that this setting performs the best. Second, we add a Gaussian noise into each $z^r_{1:n}$ (i.e., **S2**) or change the order of $z^r_{1:n}$ (i.e., **S4**) while doing nothing on z^s. We find the results regarding F_1 are similar to **S1**, which demonstrates that altering $z^r_{1:n}$ while keeping z^s unchanged will not substantially affect the capture of human individual habits. However, manipulating any $z^r_{1:n}$ seriously degrades the performance of pairs-F_1, suggesting that $z^r_{1:n}$ is indeed correlated with time-varying preferences and that any change in $z^r_{1:n}$ affects the sequential order of POIs in the trajectory. Finally, any manipulation on z^s will significantly affect the reconstruction performance regarding both F_1 and pairs-F_1, which indicates that altering z^s will destroy the capture of time-invariant habits.

We provide a case study by visualizing the rebuilt trajectories based on different settings. Fig. 6(a) presents an original trajectory containing five ordered POIs. Fig. 6(b) indicates that using

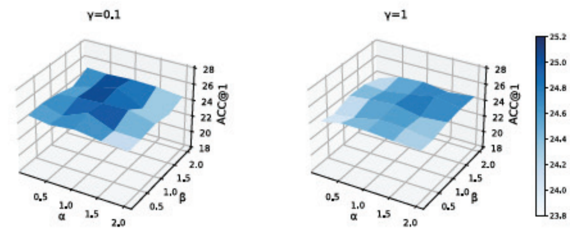


Fig. 7. Influence of weight coefficients in New York dataset.

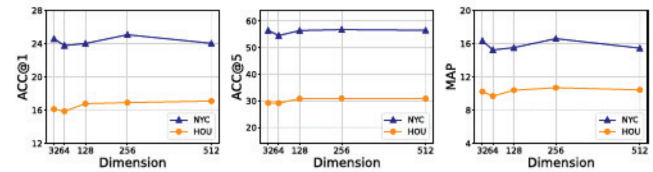


Fig. 8. Influence of the dimension of z^s.

the generated z^s and $z^r_{1:n}$ (cf. **S1**) can successfully rebuild the input trajectory. According to Fig. 6(c), we can find that adding noise to z^s while keeping the $z^r_{1:n}$ unchanged can not reconstruct the original trajectory. According to Fig. 6(d), we change the order of $z^r_{1:n}$ while keeping the z^s unchanged, we observe that the predicted POIs are almost similar to the original trajectory. However, the order is completely different from the original. Hence, we conclude that the time-invariant factors take the role of reacting to which POIs the user visit while time-varying factors take the role of how to schedule these POIs.

E. Sensitivity Analysis

Lastly, we investigate the impact of significant hyperparameters in our SSDL to evaluate the model's robustness.

- *Weight coefficients:* The objective of our representation learning (cf. (21)) contains three coefficients, which would determine the optimization procedure of each relative term. To this end, we generate different combinations of coefficients to investigate their impacts. The results of ACC@1 are shown in Fig. 7. We observe that γ=0.1 obtains better performance than γ=1 in general. We also find that the larger β helps to improve the accuracy of prediction since β represents the importance of mutual information between latent variables and trajectories. Finally, the weight coefficient α cannot be too large, otherwise it would constrain the performance.
- *Dimension of z^s:* Fig. 8 shows the performance variations of SSDL at different sizes of z^s. We find that the larger dimension of z^s does not give us promising results. Hence, for efficiency reasons, we set its dimensionality to 256.
- *Dimension of $z^r_{1:n}$:* Fig. 9 shows how different dimension of $z^r_{1:n}$ would influence the performance of SSDL. The performance decreases when the dimension of $z^r_{1:n}$ in each time step is larger than 32 and stays stable when the dimension increases. To obtain the best performance, we set the dimension of $z^r_{1:n}$ to 32 in our experiments.

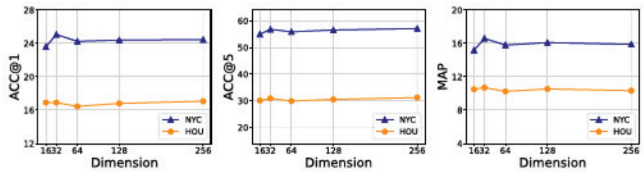


Fig. 9. Influence of the dimension of $z^r_{1:n}$.

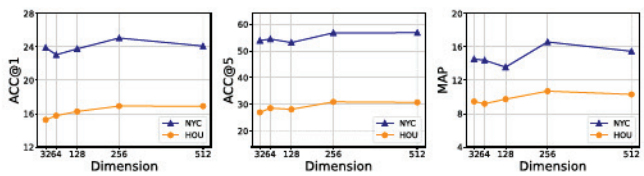


Fig. 10. Influence of embedding size.

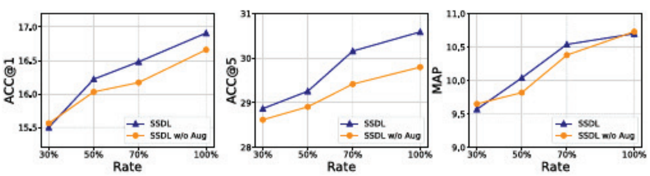


Fig. 11. Sparsity analysis in Houston dataset.

- *Embedding size:* Embedding size is one of the critical factors affecting task prediction performance. Fig. 10 presents the effect of the embedding size. We can observe that the performance of SSDL climbs as the embedding size increases, and degrades or stays stable when the embedding size is larger than 256. In our experiments, we set the embedding size to 256.
- *Sparsity sensitivity:* To show the effectiveness or sensitivity of our augmentation strategies used in the mutual information regularization module, we randomly remove each user's check-in records by setting different sample rates, ranging from 30% to 100%. As Fig. 11 shows, we can find that SSDL outperforms SSDL without augmentation, which suggests that the sparsity problem does affect the model performance, but our augmentation strategies can mitigate this problem well.

VI. Conclusion

We presented SSDL (Self-Supervised Disentanglement Learning) framework to understand human mobility for addressing the next POI prediction problem. In contrast to existing sequential dynamics learning paradigms, SSDL mainly concentrates on disentangling the time-invariant and time-varying factors underlying massive sequential trajectories, which provides an interpretable perspective to become familiar with human complex mobility patterns. We also presented two practical trajectory augmentation strategies to relieve the sparsity issue of check-in data, which enables the disentanglement of latent representations and introduced a flexible graph structure learning method to incorporate multiple heterogeneous collaborative signals from historical check-ins. We believe that several other associated contexts, such as social relations and textual data, can also be easily incorporated into our graph learning. The extensive experiments on four datasets demonstrate the superiority of SSDL compared to state-of-the-art baselines. As our future work, we plan to investigate the more intricate prior assumption during representation learning.

References

[1] J. Feng et al., "User identity linkage via co-attentive neural network from heterogeneous mobility data," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 2, pp. 954–968, Feb. 2022.

[2] H. Xue, F. Salim, Y. Ren, and N. Oliver, "MobTCast: Leveraging auxiliary trajectory forecasting for human mobility prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 30380–30391.

[3] Z. Luo and C. Miao, "RLMob: Deep reinforcement learning for successive mobility prediction," in *Proc. 15th ACM Int. Conf. Web Search Data Mining*, 2022, pp. 648–656.

[4] P. Zhao et al., "Where to go next: A spatio-temporal gated network for next POI recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 5, pp. 2512–2524, May 2022.

[5] S. Wang, J. Cao, and P. Yu, "Deep learning for spatio-temporal data mining: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 8, pp. 3681–3700, Aug. 2022.

[6] Q. Gao, F. Zhou, G. Trajcevski, K. Zhang, T. Zhong, and F. Zhang, "Predicting human mobility via variational attention," in *Proc. World Wide Web Conf.*, 2019, pp. 2750–2756.

[7] H. Zang, D. Han, X. Li, Z. Wan, and M. Wang, "CHA: Categorical hierarchy-based attention for next POI recommendation," *ACM Trans. Inf. Syst.*, vol. 40, no. 1, pp. 1–22, 2021.

[8] W. Mathew, R. Raposo, and B. Martins, "Predicting future locations with hidden Markov models," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 911–918.

[9] D. Massimo and F. Ricci, "Harnessing a generalised user behaviour model for next-POI recommendation," in *Proc. 12th ACM Conf. Recommender Syst.*, 2018, pp. 402–406.

[10] Y. Wu, K. Li, G. Zhao, and Q. Xueming, "Personalized long-and short-term preference learning for next POI recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 1944–1957, Apr. 2022.

[11] X. Rao, L. Chen, Y. Liu, S. Shang, B. Yao, and P. Han, "Graph-flashback network for next location recommendation," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 1463–1471.

[12] D. Kong and F. Wu, "HST-LSTM: A hierarchical spatial-temporal long-short term memory network for location prediction," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 2341–2347.

[13] Y. Luo, Q. Liu, and Z. Liu, "STAN: Spatio-temporal attention network for next location recommendation," in *Proc. Web Conf.*, 2021, pp. 2177–2185.

[14] S. Feng, G. Cong, B. An, and Y. M. Chee, "POI2Vec: Geographical latent representation for predicting future visitors," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 102–108.

[15] S. Zhao, T. Zhao, I. King, and M. R. Lyu, "Geo-Teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation," in *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp. 153–162.

[16] L. Huang, Y. Ma, Y. Liu, and K. He, "DAN-SNR: A deep attentive network for social-aware next point-of-interest recommendation," *ACM Trans. Internet Technol.*, vol. 21, no. 1, pp. 1–27, 2020.

[17] Q. Gao, F. Zhou, T. Zhong, G. Trajcevski, X. Yang, and T. Li, "Contextual spatio-temporal graph representation learning for reinforced human mobility mining," *Inf. Sci.*, vol. 606, pp. 230–249, 2022.

[18] N. Lim et al., "STP-UDGAT: Spatial-temporal-preference user dimensional graph attention network for next POI recommendation," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 845–854.

[19] Y. Li, T. Chen, Y. Luo, H. Yin, and Z. Huang, "Discovering collaborative signals for next POI recommendation with iterative Seq2Graph augmentation," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 1491–1497.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] K. Cho, B. V. M. C. Gulcehre, D. Bahdanau, F. B. H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[22] Q. Liu, S. Wu, L. Wang, and T. Tan, "Predicting the next location: A recurrent model with spatial and temporal contexts," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 194–200.
[23] F. Yu, L. Cui, W. Guo, X. Lu, Q. Li, and H. Lu, "A category-aware deep model for successive POI recommendation on sparse check-in data," in *Proc. Web Conf.*, 2020, pp. 1264–1274.
[24] K. Zhao et al., "Discovering subsequence patterns for next POI recommendation," in *Proc. 29th Int. Conf. Int. Joint Conf. Artif. Intell.*, 2021, pp. 3216–3222.
[25] H. Sun, J. Xu, K. Zheng, P. Zhao, P. Chao, and X. Zhou, "MFNP: A meta-optimized model for few-shot next POI recommendation," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 3017–3023.
[26] M. Zhang, Y. Yang, R. Abbas, K. Deng, J. Li, and B. Zhang, "SNPR: A serendipity-oriented next POI recommendation model," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 2568–2577.
[27] C. Miao, Z. Luo, F. Zeng, and J. Wang, "Predicting human mobility via attentive convolutional network," in *Proc. 13th Int. Conf. Web Search Data Mining*, 2020, pp. 438–446.
[28] Y. Chen, C. Long, G. Cong, and C. Li, "Context-aware deep model for joint mobility and time prediction," in *Proc. 13th Int. Conf. Web Search Data Mining*, 2020, pp. 106–114.
[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[30] X. Liu, Y. Liu, and X. Li, "Exploring the context of locations for personalized location recommendations," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1188–1194.
[31] Q. Gao, F. Zhou, K. Zhang, G. Trajcevski, X. Luo, and F. Zhang, "Identifying human mobility via trajectory embeddings," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 1689–1695.
[32] S. He and K. G. Shin, "Distribution prediction for reconfiguring urban dockless E-scooter sharing systems," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5722–5740, Dec. 2022.
[33] S. Yang, J. Liu, and K. Zhao, "GETNext: Trajectory flow map enhanced transformer for next POI recommendation," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 1144–1153.
[34] F. Zhou, Y. Dai, Q. Gao, P. Wang, and T. Zhong, "Self-supervised human mobility learning for next location prediction and trajectory classification," *Knowl.-Based Syst.*, vol. 228, 2021, Art. no. 107214.
[35] H. Tan, D. Yao, T. Huang, B. Wang, Q. Jing, and J. Bi, "Meta-learning enhanced neural ode for citywide next POI recommendation," in *Proc. IEEE 22nd Int. Conf. Mobile Data Manage.*, 2021, pp. 89–98.
[36] Y. Li, C. Gao, Q. Yao, T. Li, D. Jin, and Y. Li, "DisenHCN: Disentangled hypergraph convolutional networks for spatiotemporal activity prediction," 2022, *arXiv:2208.06794*.
[37] H. Zhang, Y. Wu, H. Tan, H. Dong, F. Ding, and B. Ran, "Understanding and modeling urban mobility dynamics via disentangled representation learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2010–2020, Mar. 2022.
[38] C. P. Burgess et al., "Understanding disentangling in β-VAE," 2018.
[39] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2615–2625.
[40] J. Bai, W. Wang, and C. P. Gomes, "Contrastively disentangled sequential variational autoencoder," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 10 105–10 118, 2021.
[41] S. Zhao, W. Shao, J. Chan, and F. D. Salim, "Spatio-temporal disentangled representation learning for mobility prediction," 2022. [Online]. Available: https://openreview.net/forum?id=2g9m74He1Ky
[42] Y. Li and S. Mandt, "Disentangled sequential autoencoder," 2018, *arXiv: 1803.02991*.
[43] Z. Ding et al., "Guided variational autoencoder for disentanglement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7920–7929.
[44] Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf, "S3VAE: Self-supervised sequential VAE for representation disentanglement and data generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6538–6547.
[45] C. Huang, X. Wang, X. He, and D. Yin, "Self-supervised learning for recommender system," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 3440–3443.
[46] J. Ma, C. Zhou, H. Yang, P. Cui, X. Wang, and W. Zhu, "Disentangled self-supervision in sequential recommenders," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 483–491.
[47] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013, *arXiv:1312.6114*.
[48] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 6827–6839, 2020.
[49] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proc. 30th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 297–304.
[50] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv: 1807.03748*.
[51] Z. Huang, J. Ma, Y. Dong, N. Z. Foutz, and J. Li, "Empowering next POI recommendation with multi-relational modeling," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 2034–2038.
[52] J. Jeon et al., "LightMove: A lightweight next-POI recommendation for taxicab rooftop advertising," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 3857–3866.
[53] Y. Chen, X. Wang, M. Fan, J. Huang, S. Yang, and W. Zhu, "Curriculum meta-learning for next POI recommendation," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2021, pp. 2692–2702.
[54] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.
[55] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–12.
[56] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent advances in recurrent neural networks," 2017, *arXiv: 1801.01078*.
[57] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Information-theoretic regularization for learning global features by sequential VAE," *Mach. Learn.*, vol. 110, no. 8, pp. 2239–2266, 2021.
[58] F. Locatello et al., "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2019, pp. 4114–4124.
[59] Z. Sun, C. Li, Y. Lei, L. Zhang, J. Zhang, and S. Liang, "Point-of-interest recommendation for users-businesses with uncertain check-ins," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5925–5938, Dec. 2022.
[60] L. Zhang et al., "An interactive multi-task learning framework for next POI recommendation with uncertain check-ins," in *Proc. 29th Int. Conf. Int. Joint Conf. Artif. Intell.*, 2021, pp. 3551–3557.
[61] J. Feng et al., "DeepMove: Predicting human mobility with attentional recurrent networks," in *Proc. World Wide Web Conf.*, 2018, pp. 1459–1468.
[62] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[64] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu, "Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 45, no. 1, pp. 129–142, Jan. 2015.
[65] Y. Liu, W. Wei, A. Sun, and C. Miao, "Exploiting geographical neighborhood characteristics for location recommendation," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, 2014, pp. 739–748.
[66] D. Yang, B. Fankhauser, P. Rosso, and P. Cudre-Mauroux, "Location prediction over sparse user mobility traces using RNNs: Flashback in hidden states!," in *Proc. 29th Int. Conf. Int. Joint Conf. Artif. Intell.*, 2021, pp. 2184–2190.
[67] Q. Gao, F. Zhou, K. Zhang, F. Zhang, and G. Trajcevski, "Adversarial human trajectory learning for trip recommendation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 4, pp. 1764–1776, Apr. 2023.
[68] T. N. Kipf and M. Welling, "Variational graph auto-encoders," 2016, *arXiv:1611.07308*.
[69] X. Wang et al., "Heterogeneous graph attention network," in *Proc. World Wide Web Conf.*, 2019, pp. 2022–2032.
[70] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *Proc. 15th Int. Conf. Semantic Web*, Heraklion, Crete, Greece Springer, Jun. 3–7, 2018, pp. 593–607.
[71] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[72] S. M. M. Rashid, M. E. Ali, and M. A. Cheema, "DeepAltTrip: Top-K alternative itineraries for trip recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 9433–9447, Sep. 2023.

Qiang Gao received the PhD degree in software engineering from the University of Electronic Science and Technology of China (UESTC) in 2020. He is currently an Associate Professor with the Southwestern University of Finance and Economics. He was a visiting scholar with Northwestern University, supervised by Dr. Diego Klabjan and Dr. Goce Trajcevski, during 2019-2020. His current research interests include spatio-temporal data mining and deep learning. He currently serves as a PC member or reviewer in several international conferences and journals, e.g., *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Neural Networks and Learning Systems*, KDD, ACM SIGSPATIAL, GeoInformatica, etc.

Jinyu Hong received the BS degree in software engineering from the University of Electronic Science and Technology of China, in 2022. She is currently working toward the MS degree in University of Electronic Science and Technology of China. Her interests include spatio-temporal data mining and deep learning.

Xovee Xu (Graduate Student Member, IEEE) was born in Yulin, Shaanxi, China, in 1996. He received the BS and MS degrees in software engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, Sichuan, China, in 2018 and 2021, respectively. He is currently working toward the PhD degree in computer science with UESTC. His research focuses on understanding spatial-temporal data, information diffusion, user-generated content, and human social behaviors.

Ping Kuang received the BS, MS, and PhD degrees from the University of Electronic Science and Technology of China, in 2000, 2003, and 2006, respectively. He is now a full professor with the University of Electronic Science and Technology of China (UESTC). His research interests include deep learning, computer vision and image processing.

Fan Zhou (Member, IEEE) received the BS degree in computer science from Sichuan University, China, in 2003, and the MS and PhD degrees from the University of Electronic Science and Technology of China, in 2006 and 2011, respectively. He is currently a full professor with the School of Information and Software Engineering, University of Electronic Science and Technology of China. His research interests include machine learning, spatio-temporal data management and social network knowledge discovery.

Goce Trajcevski (Member, IEEE) received the BSc degree from the University of Sts. Kiril and Metodij, and the MS and PhD degrees from the University of Illinois at Chicago. He is currently a Harpole-Pentair Associate Professor with the Department of Electrical and Computer Engineering, Iowa State University. His main research interests are in the areas of spatio-temporal data management, uncertainty and reactive behavior management in different application settings, and incorporating multiple contexts. In addition to a book chapter and three encyclopedia chapters, he has coauthored more than 140 publications in refereed conferences and journals. His research has been funded by the NSF, ONR, BEA, and Northrop Grumman Corp. He was the general co-chair of the IEEE ICDE 2014, ACM SIGSPATIAL 2019, the PC co-chair of the ADBIS 2018 and ACM SIGSPATIAL 2016 and 2017, and has served in various roles in organizing committees in numerous conferences and workshops. He is an associate editor of the ACM TSAS and the *Geoinformatica Journals*.