Improving Next Location Recommendation Services With Spatial-Temporal Multi-Group Contrastive Learning

Zhixuan Jia, Yushun Fan, Jia Zhang, *Senior Member, IEEE*, Chunyu Wei, Ruyu Yan, and Xing Wu

***Abstract*—Next location recommendation services play a pivotal role in Location-Based Social Networks (LBSNs) due to their ability to provide personalized recommendations of attractive destinations, resulting in substantial benefits for both users and service providers. Recent research indicates that these services are influenced by both sequential and geographical factors. However, we argue that most of these services fail to fully exploit the latent multi-group knowledge of location semantics and user preferences, resulting in suboptimal performance. Therefore, we propose STMGCL, a novel spatial-temporal multi-group contrastive learning-based method to discover intrinsic multi-group information for improving next location recommendation services. Specifically, STMGCL designs Spatial Group Contrastive Learning (SGCL) to extract multiple group knowledge regarding location semantics. Additionally, it develops Temporal Group Contrastive Learning (TGCL) to explore multiple user preference group information through a self-attention based encoder. Finally, we leverage a multi-task learning strategy and a generalized Expectation Maximization (EM) algorithm to ensure that STMGCL is optimized end-to-end with guaranteed convergence. Extensive experiments conducted on four real-world datasets demonstrate the superior performance of STMGCL over baselines.**

***Index Terms*—Next location recommendation, spatial-temporal, contrastive learning, multi-group.**

I. Introduction

LOCATION-BASED Services (LBSs) [1] have experienced significant advancements in recent years, thanks to the prevalence of GPS-enabled mobile devices. Service providers, also known as Location-Based Social Networks (LBSNs) [2], [3], such as Foursquare and Gowalla, offer users the ability to record and share their experiences, tips, and moments at various locations. Due to massive volumes of data being accumulated, location recommendation services lie at the heart of LBSNs, as they have proven successful in alleviating information overload by recommending attractive locations [4], such as restaurants, museums, and shopping malls, to users [5], [6]. These services not only facilitate users in better exploring their surroundings, but also help businesses improve their advertising strategies [7], [8], [9].

Next location recommendation services have emerged as a natural extension of general location recommendation services, gaining increasing momentum in both academia and industry in recent years [2], [10]. These services aim to assist users in discovering the most appropriate next destination by mining their historical check-in sequences and geographical information of locations [11], [12]. They have a wide range of applications, including sightseeing tours, route planning, and location-based advertising [13], [14]. Previous research has studied next location recommendation services using classical methods that explore the impact of users' previous visit behaviors on their subsequent decisions via Tensor Factorization (TF) [15] and Markov Chain (MC) [16]. Nowadays, deep learning-based methods have gained substantial traction in the research community. To model the spatiotemporal transitions in a user's check-in sequence more accurately, Recurrent Neural Networks (RNNs) [17], [18] and their variants, for example, Long Short-Term Memory (LSTM) [5], [19], have become the leading solutions in this field. Cutting-edge techniques leverage attention mechanisms to learn from both successive and non-successive location transitions in a user's check-in sequence [20], [21]. This allows for the capture of long-term dependencies and spatial-temporal correlations among locations. More recently, graph-based methods have been proposed that aim to enhance location representations by modeling the complex transition relationships among locations [22].

Although existing methods have demonstrated promising performance, we argue that the majority of these works overlook the inherent multi-group knowledge of location semantics and user preferences. Location semantics comprise not only geographical attributes but also functionality, which divides locations into several groups. Fig. 1(a) illustrates a variety of locations within a specific geographic grid, such as the basketball court, canteen, and Internet café, that collectively constitute an entertainment area. In addition, locations with similar functions can form group associations. For example, the basketball court, badminton court, table tennis room, and soccer field can be clustered together as sports activity support facilities. Capturing the knowledge of location semantics across multiple groups can

Manuscript received 15 November 2022; revised 14 April 2023; accepted 22 May 2023. Date of publication 29 May 2023; date of current version 8 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62173199. Recommended for acceptance by C.A. Ardagna. *(Corresponding author: Yushun Fan.)*

Zhixuan Jia, Yushun Fan, Chunyu Wei, and Ruyu Yan are with the Beijing National Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, Beijing 100190, China (e-mail: jzx21@mails.tsinghua.edu.cn; fanyus@tsinghua.edu.cn; cy-wei19@mails.tsinghua.edu.cn; yanry18@mails.tsinghua.edu.cn).

Jia Zhang is with the Department of Computer Science, Southern Methodist University, Dallas, TX 75205 USA (e-mail: jiazhang@smu.edu).

Xing Wu is with ByteDance Inc., Shanghai, China (e-mail: wuxing17@mails.tsinghua.edu.cn).

Digital Object Identifier 10.1109/TSC.2023.3280447

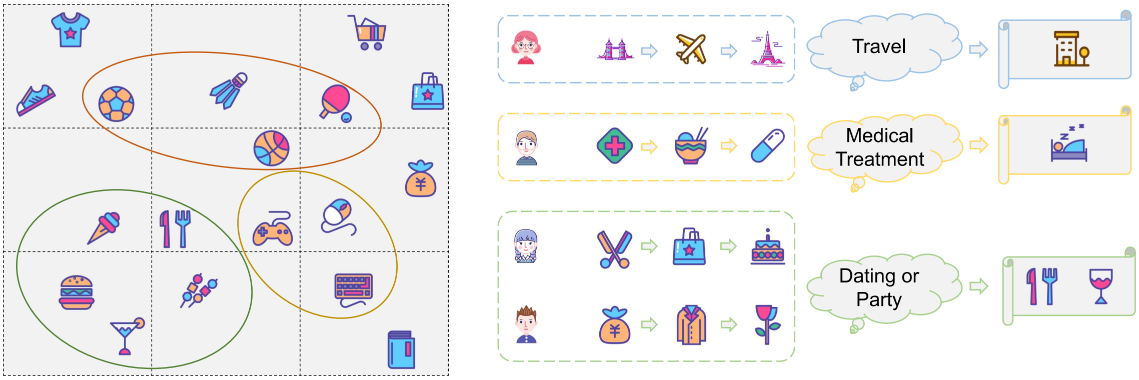


Fig. 1. An example of the multi-group nature of location semantics and user preferences. In (a), the square areas demarcated by dashed lines represent groups organized according to geographical grids, while the oval areas of various colors represent groups defined by different functionalities. In (b), dashed areas of various colors containing user trajectories depict different groups based on user preferences. Cloud-like boxes indicate users' current preferences, while the locations they are likely to visit next are displayed on the right-hand side.

enhance location representation. Similarly, the user preferences inferred from users' historical check-ins over time exhibit a phenomenon of multiple groups. Various user trajectories may reflect different or comparable user preferences. Users in the same preference group typically share common interests or purposes, and therefore may prefer similar locations. As shown in Fig. 1(b), the user in the blue dashed box intends to travel, while the user in the yellow dashed box requires medical treatment. Consequently, distinct recommendations will be generated for each of them. Although the users in the green dashed box have different trajectories, they are likely to share the same intention, which is going out for a date or party. Therefore, recognizing multiple user preference groups can enable the generation of more plausible location recommendations for users.

Unfortunately, it can be a daunting challenge to acquire latent multi-group knowledge and make it conducive to next location recommendation services. First, most scenarios lack multi-group labeled data for location semantics and user preferences. Therefore, there is no panacea for straightforwardly leveraging precise multi-group hints to enhance next location recommendation services. Second, careful consideration must be given to the subtle design of the model training process as an end-to-end optimized process. This is due to the dilemma that the accurate identification of multiple groups is contingent upon well-trained representations of both locations and users, while the high-quality representations must be learned based on precise multi-group information of location semantics and user preferences. Given the aforementioned facts, generating more accurate and reasonable recommendations for a user's next location remains an arduous task.

To fill this gap, we propose a novel method for next location recommendation services based on **S**patial-**T**emporal **M**ulti-**G**roup **C**ontrastive **L**earning (*STMGCL*). It incorporates multi-group information on location semantics and user preferences to enhance the representations of both locations and users. We utilize Geohash-5,[1] which is a simple yet effective method for expressing geographical information in terms of grid regions [23]. The initial location representation is constructed by combining its own representation with that of the region to which it belongs. Since there are no supervision signals for the latent multi-group characteristics, we draw inspiration from the contrastive Self-Supervised Learning (SSL) paradigm and develop Spatial Group Contrastive Learning (SGCL) and Temporal Group Contrastive Learning (TGCL) separately to ensure that the representations capture the inherent multi-group knowledge in spatial-temporal data. To estimate user preferences, we employ three common data augmentations (i.e., cropping, masking, and shuffling) to adequately train the self-attention based encoder. To improve the next location recommendation services, we employ a multi-task learning strategy that combines the contrastive learning tasks related to SGCL and TGCL with the primary task of predicting the next location. To tackle the issue of end-to-end optimization, we adopt a generalized Expectation Maximization (EM) algorithm [24], which alternates between the multi-group inference phase (E-step) and the multi-task learning phase (M-step) until convergence.

The main contributions of our work can be summarized as follows:

- We propose STMGCL, a novel multi-group contrastive learning-based method for improving next location recommendation services. This method harnesses the latent multi-group nature of location semantics and user preferences to benefit representation learning for both locations and users.
- We develop two contrastive learning tasks to progressively capture stable spatial-temporal multi-group knowledge. To benefit the primary task of predicting the next location, we employ a multi-task learning scheme and train them end-to-end jointly with the help of a generalized EM algorithm.
- Extensive experiments show that STMGCL can consistently outperform the baselines and achieve improvements on benchmark datasets.

The remainder of this article is structured as follows. In Section II, we formally define the notations and restate the problem. Section III introduces our proposed STMGCL, while Section IV reports on the details of the experiments and analyzes

[1] https://en.wikipedia.org/wiki/Geohash

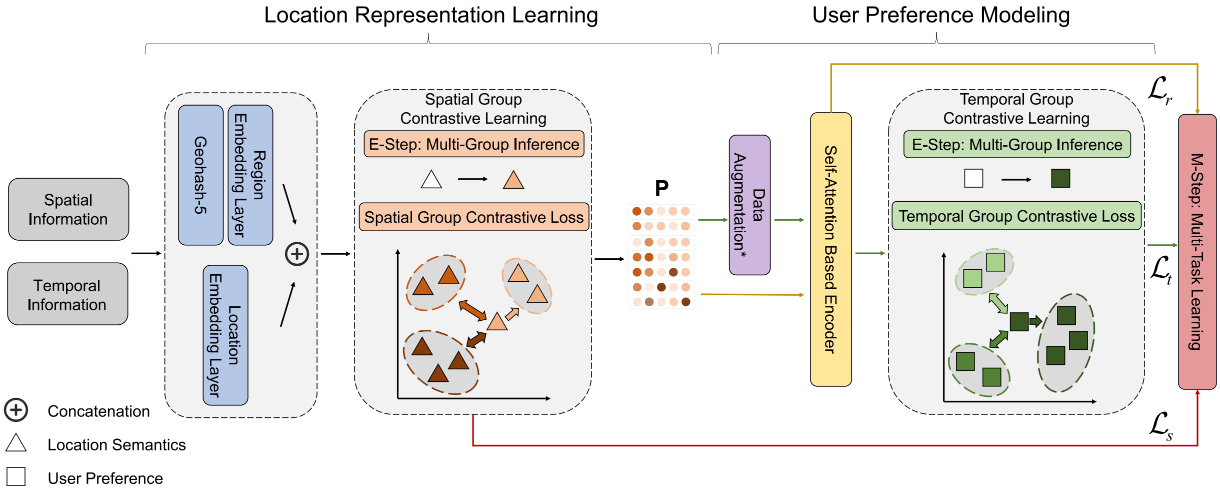

Location Representation Learning
User Preference Modeling
Spatial Information
Temporal Information
Geohash-5
Region Embedding Layer
Location Embedding Layer
Spatial Group Contrastive Learning
E-Step: Multi-Group Inference
Spatial Group Contrastive Loss
P
Data Augmentation*
Self-Attention Based Encoder
Temporal Group Contrastive Learning
E-Step: Multi-Group Inference
Temporal Group Contrastive Loss
M-Step: Multi-Task Learning
$\mathcal{L}_r$
$\mathcal{L}_t$
$\mathcal{L}_s$
Concatenation
Location Semantics
User Preference

Fig. 2. An overview of STMGCL. STMGCL comprises two fundamental components, *Location Representation learning* and *User Preference Modeling*. First, a location is represented by its own feature as well as the feature of the region to which it belongs. The multi-task consists of a SGCL task, a TGCL task, and a primary next location prediction task. These tasks are respectively represented by red, green, and yellow arrows. * means that data augmentations are only utilized in TGCL during the training phase of STMGCL. For the generalized EM algorithm, we define its E-step as multi-group inference and its M-step as multi-task learning. Best viewed in color.

the experimental results. Section V reviews the related work, and finally, Section VI concludes the article.

II. Preliminaries

In this section, we present the definitions of notations, and then formally introduce the problem.

A. Notation Definition

We use $\mathcal{P}$ to denote the set of locations and $|\mathcal{P}|$ to represent the number of locations. We define $p \in \mathcal{P}$ as a location.

Definition 1. Spatial information: The spatial information that is associated with a specific location p can be formally represented as a tuple $Geo_p = (Lat_p, Lon_p)$, where Lat_p and Lon_p respectively represent the latitude and longitude of location p.

Definition 2. Temporal information: We use the term $\mathcal{C}_n$ to refer to a chronologically ordered sequence of check-ins over a certain period, which represents the temporal information associated with user n. Specifically, $\mathcal{C}_n = [p_n^1, p_n^2, \ldots, p_n^t, \ldots, p_n^{|\mathcal{C}_n|}]$, where p_n^t denotes the location visited at time step t and $|\mathcal{C}_n|$ represents the length of $\mathcal{C}_n$.

Definition 3. Spatial group: We define a spatial group as a collection of locations that exhibit similarities in terms of their semantic properties. $\mathcal{S}$ represents the set of spatial groups exist in $\mathcal{P}$, and $|\mathcal{S}|$ is the number of spatial groups.

Definition 4. Temporal group: We refer to a temporal group as a set of similar user preferences extrapolated from temporal information. $\mathcal{T}$ represents the set of temporal groups, and $|\mathcal{T}|$ is the number of temporal groups.

B. Problem Statement

Given both the spatial information $Geo_{\mathcal{P}}$ and the temporal information $\mathcal{C}_n$, the problem is to recommend a location that the user is most likely to visit next, which can be formulated mathematically as follows:

$$\underset{p_n^i \in \mathcal{P}}{\operatorname{argmax}}\, P\left(p_n^{|\mathcal{C}_n|+1} = p_n^i \mid \mathcal{C}_n, Geo_{\mathcal{P}}\right), \tag{1}$$

which can be interpreted as calculating the probability of all candidate locations and selecting the top one for next location recommendation services.

III. Methodology

In this section, we present our proposed STMGCL for next location recommendation services. It comprises two main modules, which are *Location Representation Learning* and *User Preference Modeling*. An overall introduction to STMGCL is shown in Fig. 2.

A. Location Representation Learning

In this section, we provide a detailed account of how to learn location representations that integrate multi-group knowledge. This process can be further subdivided into two key steps: *Location Representation Initialization* and *Spatial Group Contrastive Learning*.

1) Location Representation Initialization: To begin, we utilize Geohash-5 to obtain the region-level IDs of all locations $\mathcal{P}$ based on their spatial information.

$$Reg_{\mathcal{P}} = \text{Geohash}{-}5\left(Geo_{\mathcal{P}}\right). \tag{2}$$

Next, we set up a location embedding layer and a region embedding layer to obtain representations for both the location and its corresponding region. To incorporate geographic awareness into the location representation, we merge these representations by concatenating them together, as follows:

$$\mathbf{P} = \text{Emb}_{loc}\left(\mathcal{P}\right) \,\|\, \text{Emb}_{reg}\left(Reg_{\mathcal{P}}\right), \tag{3}$$

where $\mathrm{Emb}(\cdot)$ represents the embedding layer. $\|$ is the concatenation operation. $\mathbf{P} \in \mathbb{R}^{d\times|\mathcal{P}|}$ and d denotes the dimensionality of the representation.

2) Spatial Group Contrastive Learning: To investigate the multi-group nature of location semantics, we turn to a spatial K-means algorithm, which is a widely-used and effective clustering method for identifying multiple spatial groups related to location semantics.

The spatial multi-group inference is a crucial component of the E-step in the generalized EM algorithm. To enable subsequent calculation of the spatial group contrastive loss, it must be computed initially.

$$\mathbf{G}_s = \mathrm{S} - \mathrm{Kmeans}\left(\mathbf{P}\right), \tag{4}$$

where $\mathrm{S} - \mathrm{Kmeans}(\cdot)$ denotes the spatial K-means operation. $\mathbf{G}_s \in \mathbb{R}^{d\times|\mathcal{S}|}$ represents the centroid representations of multiple spatial groups.

We devise SGCL with a spatial group contrastive loss $\mathcal{L}_s$ as follows. This contrastive loss function is designed to maximize the mutual information between a location and its corresponding spatial group, while distinguishing it from other spatial groups. The calculation of spatial group contrastive loss, which corresponds to the contrastive learning task related to SGCL, constitutes a part of the M-step in the generalized EM algorithm.

$$\mathcal{L}_s = -\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\log\frac{\mathrm{Exp}\left(\mathbf{p}_m{}^{\mathsf{T}}\mathbf{g}_{s_m}/\phi\right)}{\sum_{i\in\mathcal{S}_{\mathcal{M}}}\mathrm{Exp}\left(\mathbf{p}_m{}^{\mathsf{T}}\mathbf{g}_{s_i}/\phi\right)}, \tag{5}$$

where $\mathcal{M}$ denotes a batch of locations with a size of $|\mathcal{M}|$. $\mathrm{Exp}(\cdot)$ represents an exponential function. ϕ denotes the temperature parameter in SGCL and $\mathcal{S}_{\mathcal{M}} \subseteq \mathcal{S}$.

Note that we do not apply data augmentations typically employed in contrastive learning at the level of location representation. We argue that random perturbations at the representation level could significantly distort location semantics, potentially rendering SGCL meaningless.

B. User Preference Modeling

In this section, we focus on estimating users' preferences using temporal information with TGCL to enhance the accuracy of next location prediction. This section consists of two main parts: *User Preference Estimation* and *Temporal Group Contrastive Learning*.

1) User Preference Estimation: First, to capture the position information of locations in temporal information $\mathcal{C}_n$, we add the trainable position representation $\mathbf{P}_{\mathcal{C}_{n_{pos}}} \in \mathbb{R}^{d\times T}$ to the temporal information representation $\mathbf{P}_{\mathcal{C}_n}$. They form a combined input representation $\mathbf{X}$.

$$\mathbf{X} = \mathbf{P}_{\mathcal{C}_n} + \mathbf{P}_{\mathcal{C}_{n_{pos}}}. \tag{6}$$

where T represents the maximum length of $\mathcal{C}_n$ that we set.

Then, the user preference can be estimated based on temporal information $\mathcal{C}_n$. We employ a self-attention based encoder $\mathrm{SAEnc}(\cdot)$ as our primary technique, which operates in the following way.

$$\mathbf{S}_n = \mathrm{SAEnc}\left(\mathbf{X}\right), \tag{7}$$

where $\mathbf{S}_n \in \mathbb{R}^{d\times T}$ consists of a bundle of embedding representations.

Specifically, to better understand complex transitions in temporal information $\mathcal{C}_n$, the self-attention based encoder $\mathrm{SAEnc}(\cdot)$ is made up of multiple stacked self-attention based encoder blocks. A self-attention based encoder block $\mathrm{SAB}(\cdot)$ is shown below.

$$\mathrm{SAB}\left(\mathbf{X}\right) = \mathrm{LayerNorm}\left(\mathbf{E} + \mathrm{Dropout}\left(\mathrm{PFFN}\left(\mathbf{E}\right)\right)\right), \tag{8}$$

$$\mathbf{E} = \mathrm{LayerNorm}\left(\mathbf{X} + \mathrm{Dropout}\left(\mathrm{MHSA}\left(\mathbf{X}\right)\right)\right), \tag{9}$$

where $\mathrm{MHSA}(\cdot)$ denotes the multi-head self-attention module. $\mathrm{PFFN}(\cdot)$ represents the position-wise feed-forward network module. $\mathrm{LayerNorm}(\cdot)$ denotes layer normalization, and $\mathrm{Dropout}(\cdot)$ represents the dropout mechanism [25]. $\mathbf{E} \in \mathbb{R}^{d\times T}$ is a representation matrix used as an intermediate calculation result.

Regarding the multi-head self-attention module $\mathrm{MHSA}(\cdot)$, we provide a brief mathematical description as below.

$$\mathrm{MHSA}\left(\mathbf{X}\right) = \mathbf{W}^Z\left(\mathrm{head}_1 \,\|\, \mathrm{head}_2 \,\|\, \ldots \,\|\, \mathrm{head}_z\right), \tag{10}$$

$$\mathrm{head}_i = \mathrm{Attention}\left(\mathbf{W}_i^Q\mathbf{X}, \mathbf{W}_i^K\mathbf{X}, \mathbf{W}_i^V\mathbf{X}\right), \tag{11}$$

$$\mathrm{Attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \mathbf{V}\mathrm{Softmax}\left(\frac{\mathbf{Q}^{\mathsf{T}}\mathbf{K}}{\sqrt{d/z}}\right), \tag{12}$$

where z represents the number of heads. $\mathbf{W}^Z$, $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$ and $\mathbf{W}_i^V$ are trainable weight matrices. $\mathrm{Softmax}(\cdot)$ denotes the softmax function. The factor $\sqrt{d/z}$ aims to avoid large dot product values for smooth training.

Besides, the position-wise feed-forward network module $\mathrm{PFFN}(\cdot)$ introduces nonlinearity to the self-attention based encoder block $\mathrm{SAB}(\cdot)$ in the following way.

$$\mathrm{PFFN}\left(\mathbf{E}\right) = \mathbf{W}_2\left(\delta\left(\mathbf{W}_1\mathbf{E} + \mathbf{b}_1\right)\right) + \mathbf{b}_2, \tag{13}$$

where $\delta(\cdot)$ is a nonlinear activation function. $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$ and $\mathbf{b}_2$ are trainable parameters.

2) Temporal Group Contrastive Learning: To examine the multi-group case of user preferences, we introduce TGCL as follows. Initially, we utilize the MEAN operation to derive the user preference representation $\mathbf{c}_n$ of temporal information $\mathcal{C}_n$ based on the output $\mathbf{S}_n$ of the self-attention based encoder.

$$\mathbf{c}_n = \mathrm{MEAN}\left(\mathbf{S}_n\right). \tag{14}$$

Analogous to SGCL, we proceed to extract multi-group information regarding user preferences using a temporal K-means algorithm, as outlined below. Temporal multi-group inference constitutes an integral component of the E-step in the generalized EM algorithm. It must be computed initially to facilitate subsequent calculation of temporal group contrastive loss.

$$\mathbf{G}_t = \mathrm{T} - \mathrm{Kmeans}\left(\mathbf{C}\right), \tag{15}$$

where $\mathrm{T} - \mathrm{Kmeans}(\cdot)$ denotes the temporal K-means operation. $\mathbf{G}_t \in \mathbb{R}^{d\times|\mathcal{T}|}$ represents the centroid representations of multiple temporal groups. $\mathbf{C} \in \mathbb{R}^{d\times|\mathcal{U}|}$ and $\mathcal{U}$ indicates the set of all temporal information.

Given a batch of temporal information $\mathcal{N} = \{\mathcal{C}_n\}_{n=1}^{|\mathcal{N}|}$, we obtain a temporal group contrastive loss that facilitates the exploration of multiple temporal groups. The calculation of temporal group contrastive loss, which corresponds to the contrastive learning task related to TGCL, is a part of the M-step in the generalized EM algorithm.

$$\mathcal{L}_t\left(\mathbf{c}_n, \mathbf{g}_{t_n}\right) = -\frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}} \log \frac{\operatorname{Exp}\left({\mathbf{c}_n}^\top \mathbf{g}_{t_n}/\psi\right)}{\sum_{j\in\mathcal{T}_\mathcal{N}} \operatorname{Exp}\left({\mathbf{c}_n}^\top \mathbf{g}_{t_j}/\psi\right)}, \tag{16}$$

where ψ represents the temperature parameter in TGCL and $\mathcal{T}_\mathcal{N} \subseteq \mathcal{T}$.

To fully leverage the available temporal information during the training phase and obtain a well-trained model for subsequent prediction work, we apply three common data augmentations - Cropping, Masking, and Shuffling - to process the temporal information. Cropping is to arbitrarily select a continuous sub-sequence from the original temporal information $\mathcal{C}_n$ with a proportion χ_c. Masking is to randomly mask certain locations in the initial temporal information $\mathcal{C}_n$ with a proportion χ_m. Shuffling means stochastically rearranging a consecutive sub-sequence of temporal information $\mathcal{C}_n$ with a proportion χ_s.

Each temporal information in temporal information batch $\mathcal{N}$ is processed by any two out of three data augmentations, e.g., $\tau^1(\cdot)$ and $\tau^2(\cdot)$, for obtaining $\{\hat{\mathcal{C}}_1^1, \hat{\mathcal{C}}_1^2, \ldots, \hat{\mathcal{C}}_{|\mathcal{N}|}^1, \hat{\mathcal{C}}_{|\mathcal{N}|}^2\}$. Thus, the final temporal group contrastive loss $\mathcal{L}_t$ of TGCL is presented below.

$$\mathcal{L}_t = \frac{1}{2}\cdot\left(\mathcal{L}_t\left(\hat{\mathbf{c}}_n^1, \mathbf{g}_{t_n}\right) + \mathcal{L}_t\left(\hat{\mathbf{c}}_n^2, \mathbf{g}_{t_n}\right)\right), \tag{17}$$

where $\hat{\mathcal{C}}_n^1 \sim \tau_n^1(\mathcal{C}_n)$ and $\hat{\mathcal{C}}_n^2 \sim \tau_n^2(\mathcal{C}_n)$.

C. Multi-Task Learning

Following common practice, we formulate the training objective $\mathcal{L}_r$ for the primary task of predicting the next location using a log-likelihood loss function, which is defined as follows.

$$\mathcal{L}_r = -\frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}}\left[\log\left(\sigma\left(\mathbf{s}_n^{t^\top}\mathbf{p}_n^{t+1}\right)\right) + \log\left(1-\sigma\left(\mathbf{s}_n^{t^\top}\mathbf{p}_-^{t+1}\right)\right)\right], \tag{18}$$

where $\sigma(\cdot)$ is the sigmoid function. s_n^t denotes the location predicted by STMGCL for time step t+1. p_n^{t+1} is the ground truth (i.e., the true location) at time step t+1. p_-^{t+1} represents a randomly sampled negative location at time step t+1, which can be any location not in temporal information $\mathcal{C}_n$.

After establishing SGCL, TGCL, and the primary prediction task, we proceed to deploy a multi-task learning strategy that facilitates predicting the next location of STMGCL by optimizing these tasks jointly. Note that multi-task learning is the M-step in the generalized EM algorithm.

$$\mathcal{L} = \mathcal{L}_r + \alpha\mathcal{L}_s + \beta\mathcal{L}_t, \tag{19}$$

where α and β are different hyperparameters that control the strength of SGCL and TGCL.

Algorithm 1: Training Pipeline of STMGCL.

```
Input: Spatial information ($Geo_{\mathcal{P}}$), temporal information ($\{\mathcal{C}_n\}_{n=1}^{|\mathcal{U}|}$), batch size ($|\mathcal{M}|$ and $|\mathcal{N}|$), the number of spatial groups ($|\mathcal{S}|$), the number of temporal groups ($|\mathcal{T}|$), temperature parameters ($\phi$ and $\psi$), the proportions of data augmentations ($\chi_c$, $\chi_m$ and $\chi_s$), the strength of SGCL ($\alpha$), the strength of TGCL ($\beta$), and trainable parameters $\mathbf{W}$ and $\mathbf{b}$.
Output: Model parameters $\Theta$.
$Reg_{\mathcal{P}} \leftarrow \text{Geohash}-5(Geo_{\mathcal{P}})$
$\mathbf{P} \leftarrow \text{Emb}_{loc}(\mathcal{P}) \parallel \text{Emb}_{reg}(Reg_{\mathcal{P}})$
while Not convergence do
  % E-Step: Multi-group Inference
  $\mathbf{G}_s \leftarrow \mathrm{S}-\mathrm{Kmeans}(\mathbf{P})$ with $|\mathcal{S}|$;
  $\mathbf{G}_t \leftarrow \mathrm{T}-\mathrm{Kmeans}(\mathbf{C})$ with $|\mathcal{T}|$;
  % M-Step: Multi-task Learning
  for $\mathcal{M}$ from $\mathcal{P}$ do
    Calculate $\mathcal{L}_s$ with $\mathbf{G}_s$ and $\phi$;
  end for
  for $\{\mathcal{C}_n\}_{n=1}^{|\mathcal{N}|}$ from $\{\mathcal{C}_n\}_{n=1}^{|\mathcal{U}|}$ do
    Calculate $\mathcal{L}_t$ with $\mathbf{G}_t$, $\psi$, $\chi_c$, $\chi_m$, $\chi_s$, $\mathbf{W}$ and $\mathbf{b}$;
    Calculate $\mathcal{L}_r$ with $\mathbf{W}$ and $\mathbf{b}$;
  end for
  $\mathcal{L} \leftarrow \mathcal{L}_r + \alpha\mathcal{L}_s + \beta\mathcal{L}_t$;
  Update $\Theta$ to minimize $\mathcal{L}$;
end while
Return $\Theta$
```

D. Discussion

To enhance comprehension, we present Algorithm 1 which summarizes the general training pipeline of STMGCL.

1) EM Algorithm: Algorithm 1 can help us provide a detailed analysis regarding the role of the generalized EM algorithm in achieving end-to-end optimization of STMGCL. If there is labeled data on multi-group knowledge, we can use supervised learning to obtain well-trained user and location representations that contain multi-group information intuitively. Then, we can fine-tune them with an encoder to better suit the downstream task of recommending the next location. However, as mentioned in Section I, multi-group knowledge is latent in both location semantics and user preferences in most scenarios. The lack of observable labeled data for multi-group knowledge hinders us from learning high-quality and comprehensive representations straightforwardly. Moreover, accurate multi-group information in location semantics and user preferences cannot be derived without well-trained representations.

Nevertheless, the generalized EM algorithm offers a solution to this predicament, with a guarantee of convergence [26]. STMGCL regards multi-group inference as the E-step and multi-task learning as the M-step. In each iteration of the training phase, STMGCL first executes the E-step to perceive the situation of multiple groups with respect to location semantics and user preferences. Based on the preliminary knowledge acquired, STMGCL calculates the total loss, including multiple tasks, in the M-step. Finally, the trainable parameters of STMGCL are updated in order to minimize the total loss. Therefore, by distinguishing between the E-step and the M-step and alternating between them, STMGCL smoothly completes the end-to-end training process until convergence. With the assistance of the generalized EM algorithm, STMGCL can effectively capture latent multi-group knowledge and apply it to next location prediction.

2) Multi-Task Learning Strategy: As elucidated in Section III-C, we adopt a multi-task learning strategy that integrates the contrastive learning tasks related to SGCL and TGCL with the primary next location prediction task to accomplish the ultimate objective optimization. In contrastive SSL-based methods, it is customary to optimize the contrastive SSL task in conjunction with and for the benefit of the primary prediction task [27], [59]. The utilization of the multi-task learning strategy in STMGCL facilitates the prediction of the next location while progressively identifying multiple groups based on location semantics and user preferences. Note that capturing multi-group knowledge is a gradual process, which is also influenced by the step-by-step optimization of the primary next location prediction task.

3) Computational Complexity: According to Algorithm 1, for the training phase of STMGCL, the computation costs primarily stem from two steps, the E-step (multi-group inference) and the M-step (multi-task learning). It can be roughly estimated as $O(|\mathcal{P}||\mathcal{S}|d + |\mathcal{U}||\mathcal{T}|d + |\mathcal{U}|T^2d + |\mathcal{U}|Td^2)$. In the E-step, the time complexity primarily arises from identifying multiple spatial-temporal groups, which is about $O(|\mathcal{P}||\mathcal{S}|d + |\mathcal{U}||\mathcal{T}|d)$. Regarding the M-step, the main computational cost arises from the operation of the self-attention based encoder SAEnc$(\cdot)$, which is approximately $O(|\mathcal{U}|T^2d + |\mathcal{U}|Td^2)$ due to the multi-head self-attention module and the position-wise feed-forward network module. It is readily apparent that the time complexity of the M-step is dominated by $O(|\mathcal{U}|T^2d)$. Please note that after the training phase, we have already obtained a proficiently trained encoder through SGCL, TGCL and data augmentation. During the test phase, there is no need to use the proposed SGCL, TGCL, and data augmentation. We only utilize the well-trained encoder to generate predictions. Therefore, in practice, the computational cost of STMGCL during the test phase is slightly higher than that of SASRec [28] mainly due to the additional use of Geohash-5.

IV. Experiments

We conduct extensive experiments to evaluate the effectiveness of STMGCL. In this section, we introduce our experimental settings and analyze our experimental results in detail.

A. Datasets and Metrics

We conduct experiments on four real-world datasets collected from two well-known LBSNs, Foursquare[2] and Gowalla,[3] which are abbreviated as US, JP, CA and TX. US and JP include check-in data recorded by Foursquare in the United States of America and Japan. CA and TX contain the check-in data collected by Gowalla within California and Texas regions of the United States of America. Table I summarizes the statistics of the datasets. The data density (i.e., Density) means the ratio of the locations exposed to the users, i.e., the proportion of observed values in the user-location matrix [29]. To avoid the location cold start problem, we remove unpopular locations with less than five check-ins. We partition each user's entire check-in sequence into multiple sub-sequences with a time interval of six hours. In the following experiments, we treat each sub-sequence as an individual check-in sequence.

TABLE I
Statistics of Datasets

Dataset	# User	# Location	# Check-in	Density
US	3,792	35,940	170,159	0.12%
JP	5,736	51,686	559,852	0.19%
CA	11,377	109,343	1,348,361	0.11%
TX	11,651	124,326	1,793,933	0.12%

$$\text{Density} = \frac{\#\text{Check-in}}{\#\text{User} \times \#\text{Location}} \times 100\% \tag{20}$$

We adopt two prevalent top-K metrics, Normalized Discounted Cumulative Gain (NDCG) and Recall rate (Recall), which are widely applied in recommendation systems, to evaluate performance. NDCG is a measure that considers the position of the hit by assigning greater scores to hits that appear at higher ranks within the top-K recommendations. Recall is a measure for computing the fraction of relevant items out of all relevant items. In our experiments, we present the results using these two metrics at K = 10 and 20.

B. Baselines

We compare STMGCL with eight representative baselines, which are described as follows.

- *POP:* It recommends the most popular items based on their frequency of occurrence.
- *SASRec [28]:* It is a sequential recommendation model based on the self-attention mechanism without any recurrent or convolution operations.
- *STGN [18]:* It extends LSTM with time and distance gates to integrate spatial and temporal intervals between successive check-ins.
- *HGN [30]:* It leverages a hierarchical gating network with an item-item product module for sequential recommendations.
- *GeoSAN [20]:* A self-attention based model incorporates geographical information through grid mapping and introduces geographic modeling.
- *SINE [31]:* It activates multiple intentions from a large pool of concepts to generate multiple user interest representations for sequential recommendations.

[2] https://foursquare.com/

[3] https://www.gowalla.com/

- *STAN [21]:* It utilizes self-attention layers to explicitly exploit relative spatial-temporal information of all check-ins along the trajectory.
- *SGRec [22]:* It captures collaborative signals among locations and sets an auxiliary prediction task to enhance recommendation performance.

C. Settings

Taking guidance from [21], [23], [28], [32], [33], [58] and for a fair comparison, we set the maximum length T of temporal information $\mathcal{C}_n$ to 100. This means that if the length of temporal information $\mathcal{C}_n$ is greater than 100, we will intercept the 100 most recent check-ins as the new check-in sequence. If the length of temporal information $\mathcal{C}_n$ is less than 100, we will pad it with zeros on the left side to reach the maximum length T. The dimensionality of the region representation in STMGCL is 8 (included in d). For the self-attention based encoder $\text{SAEnc}(\cdot)$, we set the number of self-attention based encoder blocks to 2 and the number of attention heads z to 2. We choose the temperature parameters ϕ and ψ to be 1. For data augmentations, χ_c, χ_m and χ_s are set to 0.8, 0.2 and 0.2, respectively. For the number of spatial-temporal groups, we investigate $|\mathcal{S}|$ and $|\mathcal{T}|$ within $\{256, 512, 1024, 2048, 4096\}$, respectively. For the strength of STGL and TGCL, α and β are considered from $\{0.01, 0.05, 0.1, 0.2, 0.4, 0.8\}$. We examine the dimensionality of the representation d from $\{16, 32, 64, 128, 256\}$.

Regarding the baselines, we meticulously reproduce them based on their original papers and set their specific hyperparameters according to the reported optimal ones. Subsequently, we fine-tune them to run successfully and achieve optimal performance. We employ a similar approach in [34] to train and test STAN due to its extremely high memory usage. For a fair comparison, we exclude the utilization of location category information in SGRec since other methods do not utilize it [33]. For STMGCL, we optimize all trainable parameters Θ by an AdamW optimizer [35] with a learning rate of 0.001, a batch size ($|\mathcal{M}|$ and $|\mathcal{N}|$) of 512, and a dropout ratio of 0.5. We choose the Gaussian Error Linear Unit (GELU) [36] as the nonlinear activation function $\delta(\cdot)$. For each temporal information $\mathcal{C}_n$ that has $|\mathcal{C}_n|$ check-ins, we divide it into training, validation, and test parts. We utilize a leave-one-out method, meaning we use the first [1, $|\mathcal{C}_n|$ - 2] check-ins as the training part. For the validation part, we use the first [1, $|\mathcal{C}_n|$ - 2] check-ins as the input and the penultimate check-in in $\mathcal{C}_n$ as the label. For the test part, we take the first [1, $|\mathcal{C}_n|$ - 1] check-ins as the input and the last check-in in $\mathcal{C}_n$ as the label. For the evaluation phase, we rank the results on all locations $\mathcal{P}$ without negative sampling, which otherwise leads to biased discoveries [37]. We utilize an early stopping strategy [38], which means we stop the training process if there is no improvement in NDCG@10 on the validation set for 50 consecutive epochs, and we record the results from testing on the test set. For all methods, we execute them seven times with various random seeds and take the average values as the ultimate experimental outcomes.

D. Performance Comparison

In this section, we report the performance of all the methods. The results of the performance comparison are presented in Table II. Based on these results, the following observations can be made.

- Our proposed approach, STMGCL, consistently outperforms all baseline methods, attributed to the developed spatial-temporal multi-group contrastive learning. Specifically, STMGCL demonstrates an improvement over the strongest baseline in terms of NDCG@10 by 9.50%, 8.16%, 7.34%, and 4.61%; Recall@10 by 7.03%, 4.04%, 6.10%, and 6.77% on US, JP, CA, and TX respectively. These results provide empirical evidence for the efficacy of STMGCL in capturing multi-group knowledge related to location semantics and user preferences, thereby enhancing the learning of user and location representations.
- Deep learning-based approaches exhibit superior performance compared to the classical approach, POP. This suggests that deep learning has a significant advantage in accurately learning representations and enhancing downstream tasks.
- In most cases, self-attention based methods outperform other methods. This could be attributed to the self-attention mechanism's ability to capture intra-dependencies in temporal information in a more differentiated manner. The graph-based approach improves performance on CA and TX by modeling complex transition relationships with collaborative signals.
- Despite all of them being based on the self-attention mechanism, GeoSAN and STAN consistently outperform SASRec. This outcome provides clear evidence that the incorporation of spatial information can enhance recommendation performance. This finding aligns well with the fact that a user's mobility is typically confined to a certain geographic area over a period of time, and their subsequent movements are often closely related to that region.

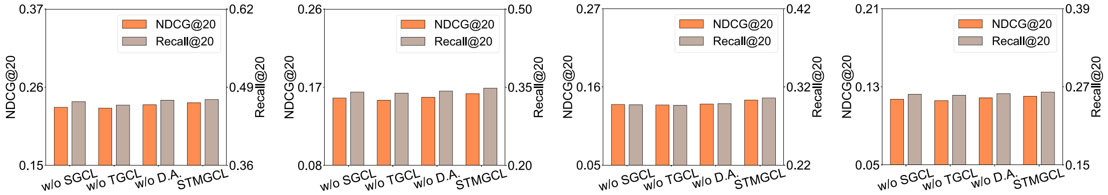
E. Ablation Study

To verify the effectiveness of each main component in STMGCL, including SGCL, TGCL, and data augmentations (D.A.), we conduct an ablation study on four datasets US, JP, CA, and TX. Specifically, we compare the performance of STMGCL with its three variants and present the results for NDCG@20 and Recall@20 in Fig. 3. The three variants are as follows. Here, w/o is an abbreviation for without.

- *w/o SGCL*: This variant removes the spatial group contrastive learning module (SGCL) from STMGCL, which means that knowledge about multiple spatial groups is not considered.
- *w/o TGCL*: This variant removes the temporal group contrastive learning module (TGCL) from STMGCL, resulting in the failure to extract useful signals from multiple temporal groups based on temporal information.
- *w/o D.A.*: This variant removes the data augmentations from the temporal group contrastive learning module

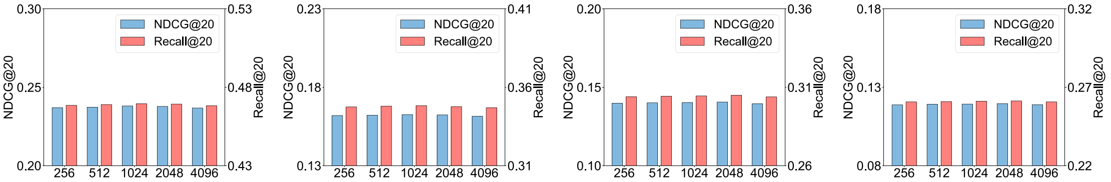
TABLE II
PERFORMANCE COMPARISON OF DIFFERENT METHODS

Dataset	Metric	POP	SASRec	STGN	HGN	GeoSAN	SINE	STAN	SGRec	STMGCL	*Impro.*
US	NDCG@10	0.0007	0.1386	0.0927	0.1499	0.1772	0.1412	0.1957	0.1197	**0.2143**	9.50%
	NDCG@20	0.0011	0.1694	0.1317	0.1695	0.2009	0.1617	0.2185	0.1280	**0.2381**	8.97%
	Recall@10	0.0017	0.2961	0.2016	0.2822	0.3247	0.2800	0.3486	0.2326	**0.3731**	7.03%
	Recall@20	0.0036	0.4179	0.2998	0.3595	0.4227	0.3610	0.4439	0.3172	**0.4695**	5.77%
JP	NDCG@10	0.0181	0.0902	0.0673	0.0944	0.1073	0.0861	0.1286	0.0990	**0.1391**	8.16%
	NDCG@20	0.0211	0.1130	0.0929	0.1105	0.1256	0.1037	0.1533	0.1088	**0.1626**	6.07%
	Recall@10	0.0314	0.1851	0.1384	0.1797	0.2061	0.1701	0.2448	0.1694	**0.2547**	4.04%
	Recall@20	0.0433	0.2756	0.2223	0.2433	0.2912	0.2395	0.3359	0.2468	**0.3483**	3.69%
CA	NDCG@10	0.0127	0.0583	0.0454	0.0683	0.0704	0.0753	0.1076	0.1104	**0.1185**	7.34%
	NDCG@20	0.0164	0.0761	0.0637	0.0820	0.0959	0.0891	0.1298	0.1337	**0.1415**	5.83%
	Recall@10	0.0243	0.1163	0.0895	0.1317	0.1596	0.1457	0.1987	0.2018	**0.2141**	6.10%
	Recall@20	0.0392	0.1876	0.1501	0.1858	0.2515	0.2006	0.2876	0.2951	**0.3059**	3.66%
TX	NDCG@10	0.0094	0.0495	0.0424	0.0688	0.0623	0.0737	0.0942	0.0955	**0.0999**	4.61%
	NDCG@20	0.0112	0.0631	0.0592	0.0807	0.0791	0.0861	0.1151	0.1167	**0.1204**	3.17%
	Recall@10	0.0180	0.0953	0.0746	0.1265	0.1287	0.1403	0.1653	0.1685	**0.1799**	6.77%
	Recall@20	0.0252	0.1499	0.1148	0.1737	0.2198	0.1898	0.2509	0.2541	**0.2619**	3.07%



(a) US (b) JP (c) CA (d) TX

Fig. 3. Ablation study on the main components of STMGCL.



(a) US (b) JP (c) CA (d) TX

Fig. 4. Sensitivity analysis on the number of spatial groups.

(TGCL), indicating that no data augmentation operations are applied to temporal information during training.

By examining Fig. 3, we have made the following findings.

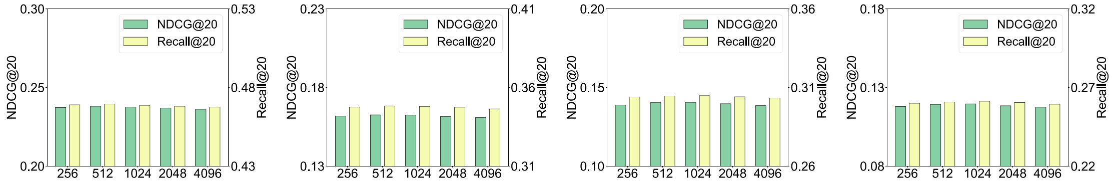
- Removing SGCL results in a decrease in the recommendation performance of STMGCL, which indicates that capturing the knowledge of multiple spatial groups helps refine location representations and improve next location recommendation services.
- The failure of TGCL results in a noticeable degradation of STMGCL performance, and the fine-grained utilization of temporal information is critical for next location recommendation services.
- Eliminating data augmentations leads to a decrease in the performance of STMGCL. Data augmentations help to fully utilize temporal information during training, enabling more effective training of the self-attention based encoder.

F. Sensitivity Analysis

In this section, we present some sensitivity analyses of the main hyperparameters in STMGCL. These include the number of spatial groups ($|\mathcal{S}|$), the number of temporal groups ($|\mathcal{T}|$), the strength of SGCL and TGCL (α and β), and the dimensionality of the representation (d).

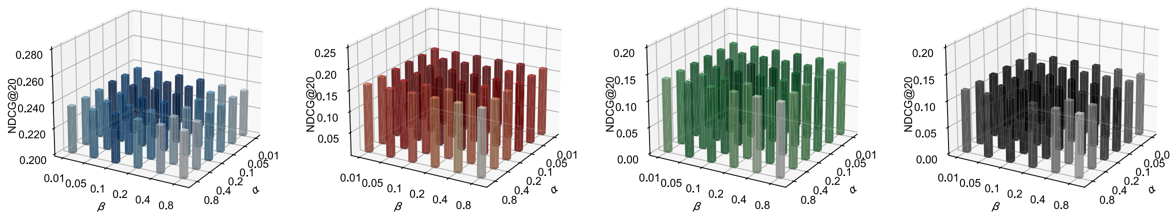
1) The Number of Spatial Groups: As outlined in the settings, we investigate the impact of the number of spatial groups $|\mathcal{S}|$ by varying it from {256, 512, 1024, 2048, 4096} on US, JP, CA, and TX. To provide better visualization, we present the results in terms of NDCG@20 and Recall@20 in Fig. 4.

As shown in Fig. 4, we observe that STMGCL outperforms the baseline methods with different numbers of spatial groups. Furthermore, we find that the best results are obtained when the number of spatial groups $|\mathcal{S}|$ is about 1024 on US and JP, 2048 on CA and TX. This is because a smaller number of



(a) US (b) JP (c) CA (d) TX

Fig. 5. Sensitivity analysis on the number of temporal groups.



(a) US (b) JP (c) CA (d) TX

Fig. 6. Sensitivity analysis on the strength of SGCL and TGCL.

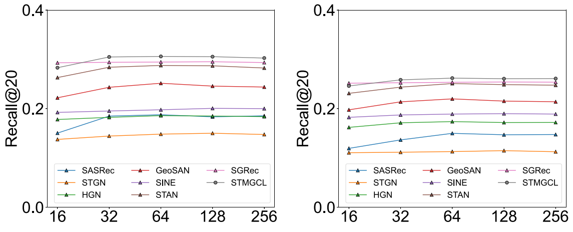
spatial groups $|\mathcal{S}|$ can introduce noise, which causes locations with dissimilar semantics to be placed in the same group and impedes accurate learning of location representations. Yet, if the number of spatial groups $|\mathcal{S}|$ is excessively large, locations with similar semantics may be separated into different spatial groups, resulting in distinct location representations guided by the SGCL task.

2) The Number of Temporal Groups: As mentioned earlier, we adjust the number of temporal groups $|\mathcal{T}|$ from {256, 512, 1024, 2048, 4096} on US, JP, CA, and TX. To improve visualization, we present the results in terms of NDCG@20 and Recall@20 in Fig. 5.

As shown in Fig. 5, STMGCL outperforms the baseline methods with different numbers of temporal groups, which is consistent with the previous section. The most satisfactory results are obtained when the number of temporal groups $|\mathcal{T}|$ is around 512 on US and JP, and 1024 on CA and TX. We again confirm our finding that a smaller number of temporal groups $|\mathcal{T}|$ tends to cluster dissimilar user preferences together, thereby hindering the accurate analysis of user preference types and leading to unsatisfactory location recommendations. Meanwhile, a larger number of temporal groups $|\mathcal{T}|$ results in higher computational costs and a greater likelihood of STMGCL misclassifying similar user preferences as different types, which can negatively impact recommendation performance.

3) The Strength of SGCL and TGCL: We survey the strength of SGCL and TGCL α and β from {0.01, 0.05, 0.1, 0.2, 0.4, 0.8} on the four datasets US, JP, CA, and TX. The results on NDCG@20 are reported in Fig. 6.

From Fig. 6, we conclude that SGCL and TGCL are better suited for overall training objectives when their strengths are set to smaller values, specifically around 0.1. The rationale for this choice stems from the primary goal of STMGCL, which is to provide recommendations for user locations. In this context, the contrastive learning tasks related to SGCL and TGCL serve as auxiliary tasks that enhance the quality of next location recommendation services. It is also consistent with the fundamental principle of the instance discrimination task in contrastive learning. Furthermore, we observe that changes in the strength of TGCL β result in greater performance fluctuations compared to those of SGCL α. This indicates that β has a more significant impact on STMGCL.



(a) CA (b) TX

Fig. 7. Sensitivity analysis on the dimensionality of the representation.

4) The Dimensionality of the Representation: As we stated before, we have fine-tuned the hyperparameters in baselines, including the dimensionality of the representation d. To further explore and clarify the impact of the dimensionality of the representation d, we vary it within {16, 32, 64, 128, 256}. We report the results on CA and TX in terms of Recall@20 in Fig. 7 to facilitate observation. Note that POP does not depend on the dimensionality of the representation, so its corresponding results are not reported here.

As shown in Fig. 7, SASRec, HGN, GeoSAN, STAN, and STMGCL perform optimally on CA and TX when d is equal to 64. STGN, SINE, and SGRec deliver satisfactory results on CA and TX when d is 128. We observe that different values of d have some impact on the performance of these methods across various datasets, highlighting the importance of carefully examining the dimensionality of the representation d.

V. Related Work

Our work mainly touches on three research areas: next location recommendation, sequential recommendation, and contrastive learning for recommendation. In this section, we review notable works in these domains and differentiate our work from them.

A. *Next Location Recommendation*

Next location recommendation has been an important topic in LBSNs, which aims to suggest the next possible location for users based on their historical check-ins and geographical information of locations. For the early classical works, FPMC-LR [16] is proposed to learn a personalized MC for each user by extending FPMC [39]. PRME-G [40] models locations and users in a sequential transition space and a user preference space, respectively. NEXT [41] incorporates a comprehensive framework that encompasses multiple contextual factors, including temporal and geographical influences, sequential relations, and auxiliary information. Currently, the majority of location recommendation services are mainly based on RNNs, self-attention mechanisms, and graphs. STRNN [17] models local spatial-temporal contexts with different transition matrices. ATST-LSTM [5] designs an attention-based spatial-temporal LSTM to better model spatial and temporal contexts. DeepMove [42] combines an attention layer with Gated Recurrent Units (GRU) to learn long-term periodicity and short-term sequential patterns. STGN [18] enhances LSTM by incorporating spatial and temporal gates. LSTPM [43] devises a context-aware non-local network and a geo-dilated LSTM to model both long-term and short-term preferences. GeoSAN [20] incorporates geographical information into a self-attention based method. STAN [21] extracts relative spatiotemporal information between consecutive and non-consecutive locations through the attention layer. GPR [44] considers ingoing and outgoing influences while discovering highly non-linear geographical influences from complex user-location networks. SGRec [22] captures collaborative signals among locations and incorporates location category information into an auxiliary prediction task. Despite notable progress, we assert that current methods are significantly restricted by their limited utilization of the latent multi-group nature of location semantics and user preference. This constraint hinders their ability to learn high-quality user and location representations, ultimately resulting in suboptimal performance.

B. *Sequential Recommendation*

Recent methods for sequential recommendation can be broadly categorized into two schools of models: Markov Chain-based models and deep learning-based models. Markov Chain-based models use transition matrices to predict the probability of future behaviors. For instance, FPMC [39] combines the normal matrix factorization model with a common Markov Chain. Fossil [45] fuses similarity-based models with Markov Chains to personalize sequential behavior. With the emergence of deep learning, an increasing number of sequential recommendation models based on deep learning paradigms have been proposed to overcome the limitations of previous models. Caser [46] proposes to learn sequential patterns by using convolutional filters that can capture essential features. GRU4Rec [47] utilizes a model based on RNNs for sequential recommendation at the session level. HGN [30] develops a hierarchical gating network that uses Bayesian Personalized Ranking to capture both the long-term and short-term interests of users. SASRec [28] is a self-attention based model that can capture long-term semantics and make predictions based on relative actions. BERT4Rec [48] utilizes a sophisticated bi-directional self-attention mechanism to effectively model user behavior sequences. FMLP-Rec [49] is an all multi-layer perception-based model equipped with trainable filters, which offers lower time complexity. In contrast to conventional sequential recommendation approaches that mainly focus on exploiting temporal data, our work sufficiently exploits both spatial data and temporal data by discovering intrinsic multi-group knowledge of location semantics and user preferences. And it achieves satisfactory performance in the field of next location recommendation services.

C. *Contrastive Learning for Recommendation*

Through further research on contrastive learning, methods based on the contrastive learning paradigm have become a key component in SSL-based recommendation systems [50]. Consistent with the original ideology of contrastive learning, these methods aim to bring the views of the same instance closer while pushing those of different instances further apart. In terms of contrastive learning for recommendation systems, CLCRec [51] solves the cold-start problem through contrastive learning, which maximizes the mutual dependencies between item content and collaborative signals. HCCF [52] proposes a hypergraph structure learning module and a cross-view hypergraph contrastive encoding schema. SimGCL [53] introduces directed random noises to the representation and regulates the uniformity of the representation distribution. In the context of time-aware recommendation, S^3-Rec [54] proposes a self-supervised contrastive learning model for sequential recommendation by mining data correlations using the principle of mutual information maximization. CLEA [55] utilizes contrastive learning to extract items relevant to the target item for next basket recommendation. ACVAE [56] integrates contrastive learning into the variational autoencoder for analyzing sequential data. CL4SRec [32] utilizes self-supervised signals obtained solely from raw data to enhance recommendation models. DuoRec [57] incorporates a contrastive regularization technique with both model-level augmentation and supervised positive sampling to construct contrastive samples. Despite the promising results yielded by the contrastive SSL paradigm, its application in the context of next location recommendation services has received relatively little attention. Therefore, our work has the potential to inspire the integration of contrastive SSL techniques into the field of next location recommendation services.

VI. Conclusion

To gain a better understanding of users' dynamic preferences, next location recommendation services make suggestions to users based on their check-in records and geographical information of locations. However, we contend that two crucial multi-group characteristics about location semantics and user preferences, have not been sufficiently leveraged for next location recommendation services. In this article, we present STMGCL, a novel next location recommendation approach based on our crafted spatial group contrastive learning and temporal group contrastive learning. STMGCL mines latent multi-group knowledge of location semantics and user preferences to enhance next location recommendation services, without requiring related or accurate supervision data. The compelling results of extensive experiments unambiguously establish the superiority of STMGCL and verify the benefits of its different components. To advance next location recommendation services, we will strive to incorporate more side information, such as user social information and location category information. Another future endeavor is to enhance STMGCL so that it can generalize to the task of recommending a list of consecutive locations that users are likely to visit in the next period. Additionally, we will consider handling the situation where a location or a user may belong to multiple location semantics groups or multiple user preference groups simultaneously through efficient spatial-temporal clustering methods.

References

[1] V. K. Yadav, N. Andola, S. Verma, and S. Venkatesan, "P2LBS: Privacy provisioning in location-based services," *IEEE Trans. Serv. Comput.*, vol. 16, no. 1, pp. 466–477, Jan./Feb. 2023.

[2] S. Zhao, I. King, and M. R. Lyu, "A survey of point-of-interest recommendation in location-based social networks," 2016, *arXiv:1607.00647*.

[3] Q. Huang, J. Du, G. Yan, Y. Yang, and Q. Wei, "Privacy-preserving spatio-temporal keyword search for outsourced location-based services," *IEEE Trans. Serv. Comput.*, vol. 15, no. 6, pp. 3443–3456, Nov./Dec. 2022.

[4] C. Wang, M. Yuan, R. Zhang, K. Peng, and L. Liu, "Efficient point-of-interest recommendation services with heterogenous hypergraph embedding," *IEEE Trans. Serv. Comput.*, vol. 16, no. 2, pp. 1132–1143, Mar./Apr. 2023.

[5] L. Huang, Y. Ma, S. Wang, and Y. Liu, "An attention-based spatiotemporal LSTM network for next POI recommendation," *IEEE Trans. Serv. Comput.*, vol. 14, no. 6, pp. 1585–1597, Nov./Dec. 2021.

[6] C. Xu, A. S. Ding, and K. Zhao, "A novel POI recommendation method based on trust relationship and spatial–temporal factors," *Electron. Commerce Res. Appl.*, vol. 48, 2021, Art. no. 101060.

[7] L. Liu, F. Lecue, N. Mehandjiev, and L. Xu, "Using context similarity for service recommendation," in *Proc. IEEE 4th Int. Conf. Semantic Comput.*, 2010, pp. 277–284.

[8] G. Cui, Q. He, F. Chen, H. Jin, Y. Xiang, and Y. Yang, "Location privacy protection via delocalization in 5G mobile edge computing environment," *IEEE Trans. Serv. Comput.*, vol. 16, no. 1, pp. 412–423, Jan./Feb. 2023.

[9] H. Mezni, "Temporal knowledge graph embedding for effective service recommendation," *IEEE Trans. Serv. Comput.*, vol. 15, no. 5, pp. 3077–3088, Sep./Oct. 2022.

[10] 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.

[11] H. Gao, J. Tang, and H. Liu, "gSCorr: Modeling geo-social correlations for new check-ins on location-based social networks," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 1582–1586.

[12] S. Hu, Z. Tu, Z. Wang, and X. Xu, "A POI-sensitive knowledge graph based service recommendation method," in *Proc. IEEE Int. Conf. Serv. Comput.*, 2019, pp. 197–201.

[13] Z. Zhang, M. Dong, K. Ota, Y. Zhang, and Y. Kudo, "Context-enhanced probabilistic diffusion for urban point-of-interest recommendation," *IEEE Trans. Serv. Comput.*, vol. 15, no. 6, pp. 3156–3169, Nov./Dec. 2022.

[14] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo, "A survey on deep learning for human mobility," *ACM Comput. Surv.*, vol. 55, no. 1, pp. 1–44, 2021.

[15] X. Li, M. Jiang, H. Hong, and L. Liao, "A time-aware personalized point-of-interest recommendation via high-ordertensor factorization," *ACM Trans. Inf. Syst.*, vol. 35, no. 4, pp. 1–23, 2017.

[16] C. Cheng, H. Yang, M. R. Lyu, and I. King, "Where you like to go next: Successive point-of-interest recommendation," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2605–2611.

[17] Q. Liu, S. Wu, L. Wang, and T. Tan, "Predicting the next location: A recurrent model with spatial and temporal contexts," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 194–200.

[18] P. Zhao et al., "Where to go next: A spatio-temporal gated network for next POI recommendation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, Art. no. 5877.

[19] Z. Wang, Y. Zhu, Q. Zhang, H. Liu, C. Wang, and T. Liu, "Graph-enhanced spatial-temporal network for next POI recommendation," *ACM Trans. Knowl. Discov. Data*, vol. 16, no. 6, pp. 1–21, 2022.

[20] D. Lian, Y. Wu, Y. Ge, X. Xie, and E. Chen, "Geography-aware sequential location recommendation," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 2009–2019.

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

[22] Y. Li, T. Chen, Y. Luo, H. Yin, and Z. Huang, "Discovering collaborative signals for next POI recommendation with iterative Seq2Graph augmentation," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, 2021, pp. 1491–1497, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2021/206

[23] Q. Cui, C. Zhang, Y. Zhang, J. Wang, and M. Cai, "ST-PIL: Spatial-temporal periodic interest learning for next point-of-interest recommendation," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 2960–2964.

[24] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. Hoboken, NJ, USA: Wiley, 2007.

[25] S. Cai, Y. Shu, G. Chen, B. C. Ooi, W. Wang, and M. Zhang, "Effective and efficient dropout for deep convolutional neural networks," 2019, *arXiv:1904.03392*.

[26] N. Sammaknejad, Y. Zhao, and B. Huang, "A review of the expectation maximization algorithm in data-driven process identification," *J. Process Control*, vol. 73, pp. 123–136, 2019.

[27] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.

[28] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *Proc. IEEE Int. Conf. Data Mining*, 2018, pp. 197–206.

[29] C. Gao et al., "KuaiRec: A fully-observed dataset and insights for evaluating recommender systems," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, 2022, pp. 540–550.

[30] C. Ma, P. Kang, and X. Liu, "Hierarchical gating networks for sequential recommendation," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 825–833.

[31] Q. Tan et al., "Sparse-interest network for sequential recommendation," in *Proc. 14th ACM Int. Conf. Web Search Data Mining*, 2021, pp. 598–606.

[32] X. Xie et al., "Contrastive learning for sequential recommendation," in *Proc. IEEE 38th Int. Conf. Data Eng.*, 2022, pp. 1259–1273.

[33] Z. Wang, Y. Zhu, H. Liu, and C. Wang, "Learning graph-based disentangled representations for next POI recommendation," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 1154–1163.

[34] 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.

[35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[36] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[37] W. Krichene and S. Rendle, "On sampled metrics for item recommendation," *Commun. ACM*, vol. 65, no. 7, pp. 75–83, 2022.

[38] Z. Zhong, J. Yan, W. Wu, J. Shao, and C.-L. Liu, "Practical block-wise neural network architecture generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2423–2432.

[39] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, "Factorizing personalized Markov chains for next-basket recommendation," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 811–820.

[40] S. Feng, X. Li, Y. Zeng, G. Cong, Y. M. Chee, and Q. Yuan, "Personalized ranking metric embedding for next new poi recommendation," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 2069–2075.

[41] Z. Zhang, C. Li, Z. Wu, A. Sun, D. Ye, and X. Luo, “NEXT: A neural network framework for next POI recommendation,” *Front. Comput. Sci.*, vol. 14, no. 2, pp. 314–333, 2020.

[42] J. Feng et al., “DeepMove: Predicting human mobility with attentional recurrent networks,” in *Proc. World Wide Web Conf.*, 2018, pp. 1459–1468.

[43] K. Sun, T. Qian, T. Chen, Y. Liang, Q. V. H. Nguyen, and H. Yin, “Where to go next: Modeling long-and short-term user preferences for point-of-interest recommendation,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 214–221.

[44] B. Chang, G. Jang, S. Kim, and J. Kang, “Learning graph-based geographical latent representation for point-of-interest recommendation,” in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 135–144.

[45] R. He and J. McAuley, “Fusing similarity models with Markov chains for sparse sequential recommendation,” in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 191–200.

[46] J. Tang and K. Wang, “Personalized top-N sequential recommendation via convolutional sequence embedding,” in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 565–573.

[47] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, “Session-based recommendations with recurrent neural networks,” 2015, *arXiv:1511.06939*.

[48] F. Sun et al., “BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer,” in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 1441–1450.

[49] K. Zhou, H. Yu, W. X. Zhao, and J.-R. Wen, “Filter-enhanced MLP is all you need for sequential recommendation,” in *Proc. ACM Web Conf.*, 2022, pp. 2388–2399.

[50] J. Yu, H. Yin, X. Xia, T. Chen, J. Li, and Z. Huang, “Self-supervised learning for recommender systems: A survey,” 2022, *arXiv:2203.15876*.

[51] Y. Wei et al., “Contrastive learning for cold-start recommendation,” in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 5382–5390.

[52] L. Xia, C. Huang, Y. Xu, J. Zhao, D. Yin, and J. Huang, “Hypergraph contrastive collaborative filtering,” in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 70–79.

[53] J. Yu, H. Yin, X. Xia, T. Chen, L. Cui, and Q. V. H. Nguyen, “Are graph augmentations necessary? Simple graph contrastive learning for recommendation,” in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 1294–1303.

[54] K. Zhou et al., “S3-Rec: Self-supervised learning for sequential recommendation with mutual information maximization,” in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 1893–1902.

[55] Y. Qin, P. Wang, and C. Li, “The world is binary: Contrastive learning for denoising next basket recommendation,” in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 859–868.

[56] Z. Xie, C. Liu, Y. Zhang, H. Lu, D. Wang, and Y. Ding, “Adversarial and contrastive variational autoencoder for sequential recommendation,” in *Proc. Web Conf.*, 2021, pp. 449–459.

[57] R. Qiu, Z. Huang, H. Yin, and Z. Wang, “Contrastive learning for representation degeneration problem in sequential recommendation,” in *Proc. 15th ACM Int. Conf. Web Search Data Mining*, 2022, pp. 813–823.

[58] Y. Chen et al., “Intent contrastive learning for sequential recommendation,” in *Proc. ACM Web Conf.*, 2022, pp. 2172–2182.

[59] S. Lin et al., “Prototypical graph contrastive learning,” in *Proc. IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

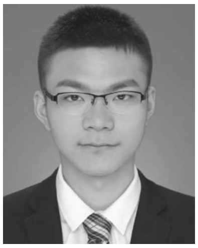
Zhixuan Jia is currently working toward the PhD degree with the Department of Automation, Tsinghua University. His research interests include services computing, service recommendation, and spatial-temporal data mining.

Yushun Fan received the PhD degree in control theory and application from Tsinghua University, China, in 1990. He is currently a tenured professor with the Department of Automation, director of the System Integration Institute, and director of the Networking Manufacturing Laboratory, Tsinghua University. He is the member of IFAC TC 5.1 and TC 5.2, vice director of China Standardization Committee for Automation System and Integration, and an editorial member of the *International Journal of Computer Integrated Manufacturing*. From September 1993 to 1995, he was a visiting scientist, supported by Alexander von Humboldt Stiftung, with the Fraunhofer Institute for Production System and Design Technology (FHG/IPK), Germany. He has authored 10 books in enterprise modeling, workflow technology, intelligent agent, object-oriented complex system analysis, and computer integrated manufacturing. He has published more than 500 research papers in journals and conferences. His research interests include enterprise modeling methods and optimization analysis, business process re-engineering, workflow management, system integration, modern service science and technology, and petri nets modeling and analysis.

Jia Zhang (Senior Member, IEEE) received the BS and MS degrees in computer science from Nanjing University, China, and the PhD degree in computer science from the University of Illinois at Chicago. She is currently the Cruse C. and Marjorie F. Calahan Centennial chair in engineering, professor with the Department of Computer Science, Southern Methodist University. Her research interests emphasize the application of machine learning and information retrieval methods to tackle data science infrastructure problems, with a recent focus on scientific workflows, provenance mining, software discovery, knowledge graph, and interdisciplinary applications of all of these interests in the area of earth science. She has published more than 180 refereed journal articles, book chapters, and conference papers.

Chunyu Wei received the BS degree in control theory and application from Tsinghua University, China, in 2019. He is currently working toward the PhD degree with the Department of Automation, Tsinghua University. His research interests include services computing, service recommendation, and social computing.

Ruyu Yan received the BS degree from Tsinghua University, China, in 2018. She is currently working toward the PhD degree with the Department of Automation, Tsinghua University, China. Her research interests include services computing, recommender systems, and time series prediction.



Xing Wu received the BS and PhD degrees in control theory and application from Tsinghua University, China. He is currently employed with ByteDance Inc., China. His research interests include services computing, Web service recommendation, federated learning, and blockchain.