MMCPP: A MULTI-MODAL CONTRASTIVE PRE-TRAINING MODEL FOR PLACE REPRESENTATION BASED ON THE SPATIO-TEMPORAL FRAMEWORK

Y. Chen [1], X. S. Yu [1], K. Qin [1] *

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan – chenyu8697@whu.edu.cn, yxs1995_s@163.com, qink@whu.edu.cn

KEY WORDS: Place Embedding; Geographic Contexts; Spatial Interaction; Temporal Position Encoding; Self-supervised Learning.

ABSTRACT:

The concept of "place" is crucial for understanding geographical environments from a human perspective. Place representation learning involves converting places into numerical low-dimensional dense vectors and is a fundamental procedure for artificial intelligence in geography (GeoAI). However, most studies ignore multi-level distance constraints and spatial proximity interactions that enable behavioral interactions between places. Furthermore, representing the temporal characteristics of these interactions in trajectory sequences poses a challenge for natural language processing and other field techniques. In addition, most existing methods rely on all modalities from inputs as they use joint training to integrate multiple modalities. To address these issues, we propose a **Multi-Modal Contrastive Pre-training model for Place representation (MMCPP)**. Our model consists of three encoders that capture corresponding place attributes across different modalities, including point of interests (POIs), images, and trajectories. The trajectory encoder, named RodtFormer, takes fine-grained spatio-temporal trajectories as input and leverages self-attention with rotary temporal interval position embedding to simulate dynamic spatial and behavioral proximity interactions between places. By using a coordinated pre-training framework, MMCPP independently encodes place representations across different modalities and improves model reusability. We verify the effectiveness of our model on a taxi trajectory dataset using the location prediction task at next *n* seconds, including 30 seconds(s), 180(s), 300(s). Our results demonstrate that compared to existing embedding methods, our model is capable of learning higher-quality position representations during pre-training, leading to improved performance on downstream tasks.

1. INTRODUCTION

Places offer meaningful insight into the geographical environment from a human perspective, as they contain attributes related to spatial features and human behavior (Liu, Yao, et al. 2020). Place representation learning has emerged as a key component in urban studies and applications, aiming to represent places as numerical low-dimensional vectors. These vectors can be used to reveal the inherent laws of places (Huang et al. 2021) and improve the performance of downstream tasks, such as location prediction (Li et al. 2022), urban function identification (Jenkins et al. 2019; Paul et al. 2021), house price forecasts (Das et al. 2021), etc. Place representation learning drives the development of artificial intelligence in geography (GeoAI) (Mai, Janowicz, et al. 2022; Janowicz et al. 2020).

Representing a place requires describing attributes of the place. These attributes include two aspects: first-order attributes intrinsic to the place itself, and second-order attributes shaped by spatial and behavioral proximity interactions among places. Different modalities of data capture distinct facets of these attributes. For instance, point-of-interests (POIs) (Jenkins et al. 2019) and images (Wang, Wang, et al. 2020) can reveal first-order attributes such as place functions and surface information. Trajectories, can simulate dynamic interactions between places and provide insight into second-order attributes such as spatial autocorrelations and spatial complementarity (Liu et al. 2022).

Several recent studies have demonstrated that combining multi-modal data can yield higher quality place representations compared to using unimodal data (Huang et al. 2021; Zhang et al. 2020). Multi-modal integration technology provides the basis for simultaneous representation of first-order and second-order attributes of a place (Wang, Li, et al. 2020; Zhang et al. 2020).

However, existing research exhibits some deficiencies. Firstly, previous studies have typically focused on the interaction between stay points of trajectories (e.g., pick-up and drop-off points of taxi trajectories) while ignoring the fine-grained trajectory sequences that include multi-level spatial distance constraints and spatial proximity interactions. Secondly, existing representation learning methods in computer science have difficulty capturing the complex geographical features of a place, such as spatial and temporal characteristics. Thirdly, most current research has employed joint representation structures that require all modalities as input, limiting their reusability when only one modality is available or parts of the modalities are missing.

To address these issues, we aim to develop a pre-training model that can simultaneously capture first-order and second-order attributes using multi-modal data, considering not only the attributes of the place itself, but also the interaction of dynamic spatial proximity and behavioral proximity. The model can be used in the tasks involving missing partial modalities while utilizing multi-modal information, ensuring higher reusability.

The main contributions of this paper are summarized as follows:

- We propose MMCPP - a model that uses three types of geographic data: POIs, images, and trajectories. MMCPP

[*] Corresponding author

https://doi.org/10.5194/isprs-annals-X-1-W1-2023-303-2023 | © Author(s) 2023. CC BY 4.0 License.

features three encoders that correspond to the three modalities and can represent places independently, making the model more reusable.

- We propose a trajectory encoder named RodtFormer. This encoder contains a self-attention mechanism with a rotary temporal interval position embedding based on Roformer structure (a variant of Transformer). It allows for simulating dynamic spatial proximity interactions and behavioral proximity interactions among places in fine-grained trajectory sequences.

- A contrastive sample construction method is designed for integrating multi-modal information. By using contrastive learning method, the encoders in MMCPP integrates the information of other modalities, and can represent the place independently when the respective modality data is given, which improves the reusability of the pre-train model.

- We employ MMCPP in the location prediction task at next n seconds, including 30 seconds, 180 seconds, and 300 seconds, and conduct experiments on a taxi trajectory dataset. Our experimental results show significant improvement in prediction performance, demonstrating the superiority of our proposed model.

2. RELATED WORK

Aligning heterogeneous and multi-source geographic data to corresponding places and integrating information of different modalities have become challenges in multimodal place representation modeling.

Some studies extract features of different modalities and concatenate them together as the place representation vector for each geographical unit (Xuan et al. 2016; Li 2018). It can be expressed as $[V_1; V_2; \dots; V_n]$. The place representation vector obtained by these methods have strong interpretability. However, this method has a heavy workload and only considers first-order attributes of the place, ignoring second-order attributes generated by interaction among places.

Alternatively, some studies integrate information of different modalities when constructing the input and learn the place representation containing multi-modal information, which can be expressed as: $\mathrm{Enc}([V_1; V_2; \dots; V_n])$. For example, scholars use the structure of heterogeneous graphs to integrate different modalities' information (Liu et al. 2022; Paul et al. 2021) or compare trajectories to sentences using natural language processing methods to jointly integrate multi-modal information (Wan et al. 2022; Zhao et al. 2017; Zhu et al. 2019).

Moreover, integrating information of different modalities can be achieved by adding a fusion structure in the model, expressed as: $\mathrm{Fusion}(\mathrm{Enc}_1(V_1), \mathrm{Enc}_2(V_2), \dots, \mathrm{Enc}_n(V_n))$. For example, statistical operators such as mean value aggregation (Wang et al. 2021); deep learning modules such as fully connected layers (Jenkins et al. 2019; Zhang et al. 2020; Luo et al. 2022; Wang et al. 2022), attention mechanism variants (Zhang et al. 2020; Luo et al. 2022; Sun et al. 2022) can be utilized. Encoder-decoder and other special training structures also can be used for modality integration (Du et al. 2019; Zhang et al. 2019; Fu et al. 2019).

Some studies design training tasks, i.e., loss functions, to enable the representation model to integrate multi-modal information during the training process. These tasks can be categorized into mask data recovery tasks (Zhang et al. 2017; Lin et al. 2021) and contrastive learning tasks (Huang et al. 2021; Radford et al. 2021).

Most studies using the first three methods above belong to joint representation structures that rely on all modal inputs (Baltrušaitis et al. 2019). The model will not be able to represent when the downstream task lacks any modality, and the model's reusability is low. In contrast, most models using contrastive learning training strategy are coordinated representation structures (Baltrušaitis et al. 2019) applicable to only one modality. However, few studies combine place representation with this method (Huang et al. 2021).

Furthermore, most studies only consider the interaction between stay points in trajectory data, such as pick-up and drop-off points of taxi trajectories. They overlook multi-level distance constraints and spatial proximity interactions in fine-grained trajectory sequences (Yao et al. 2018; Du et al. 2019; Liu, Miranda, et al. 2020).

And it is usually difficult to represente the geographical characteristics of places (e.g. spatio-temporal characteristics) by directly transferred the representation methods from specific fields such as natural language processing. For example, For example, the self-attention mechanism in the Transformer encoder can encode semantics of the sequence context and use absolute sequence position (0, 1, 2, …) to supplement the position information lost in the self-attention mechanism. However, capturing trajectory sequences and temporal characteristics of place interactions is crucial for building place representations. Even though some studies introduce temporal positions (Lin et al. 2021), the representation ability remains insufficient.

3. METHODOLOGY

Figure 1 illustrates the structure of MMCPP, which includes three encoders and two pre-training phases. In phase 1, place attributes described by each modality of data are encoded by three encoders using self-supervised pre-training tasks individually. In phase 2, the encoded place attributes from each modality representation are integrated coordinately using the contrastive learning method. This section explains the structure of MMCPP in greater detail.

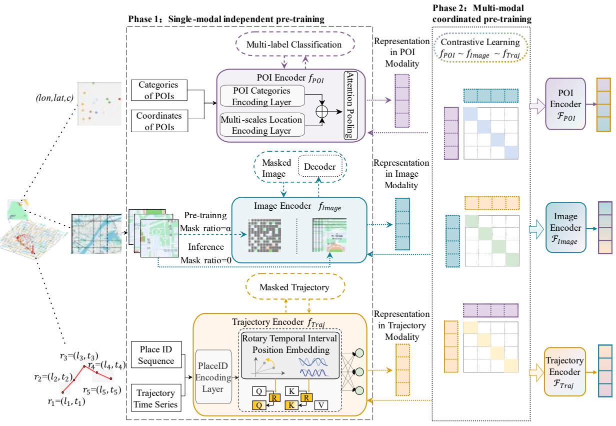

Phase 1：Single-modal independent pre-training
Phase 2：Multi-modal coordinated pre-training
Multi-label Classification
Representation in POI Modality
Contrastive Learning $f_{POI} \sim f_{Image} \sim f_{Traj}$
(lon,lat,c)
Categories of POIs
Coordinates of POIs
POI Encoder f_{POI}
POI Categories Encoding Layer
Multi-scales Location Encoding Layer
Attention Pooling
POI Encoder $\mathcal{F}_{POI}$
Masked Image
Decoder
Representation in Image Modality
Image Encoder f_{Image}
Pre-trainng
Mask ratio=α
Inference
Mask ratio=0
Image Encoder $\mathcal{F}_{Image}$
Masked Trajectory
Representation in Trajectory Modality
Trajectory Encoder f_{Traj}
Place ID Sequence
Trajectory Time Series
PlaceID Encoding Layer
Rotary Temporal Interval Position Embedding
Q
R
K
V
$r_3=(l_3,t_3)$
$r_4=(l_4,t_4)$
$r_2=(l_2,t_2)$
$r_5=(l_5,t_5)$
$r_1=(l_1,t_1)$
Trajectory Encoder $\mathcal{F}_{Traj}$

Figure 1. The structure of multi-modal contrastive pre-training model for place representation (MMCPP)

3.1 POIs Encoder

The POIs encoder extracts functional attributes from categories and geographical location distributions of POI data to encode the first-order attributes of a place. Its structure, shown in Figure 2, includes an input layer and an attention pooling layer for capturing dominant functional properties in the place.

The Input Layer comprises a coding layer of POI categories that converts discrete POI categories into continuous numerical variables. Additionally, a coding layer of POI location uses the sinusoidal multi-scale position encoder (Mai, Xuan, et al. 2022) to encode the two-dimensional coordinates (including latitude and longitude) of POIs within the place. For places without POIs, a special "pseudo-POI" with category name "[NAN]" and coordinates at the geometric center of the place is added to avoid the same representation of places without POIs. After category embedding and coordinate coding, additive operators are used to integrate the POI category and geographical location distribution. Then, the attention pooling layer performs weighted aggregation of different categories of POIs at different locations within the place to capture the main functional attributes of the place. Finally, a feature vector integrated POIs information is output as a representation of the place in this modality.

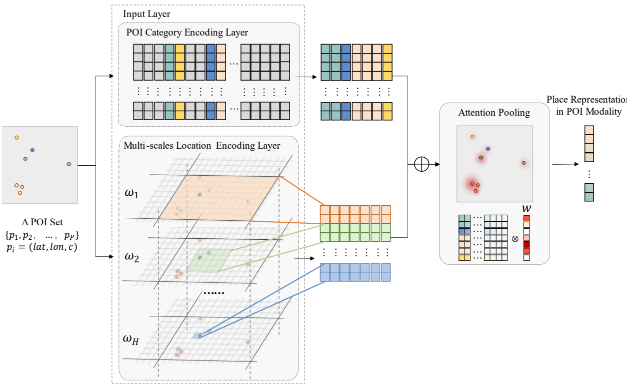

Input Layer
POI Category Encoding Layer
Multi-scales Location Encoding Layer
A POI Set
$\{p_1, p_2, \ldots, p_P\}$
$p_i = (lat, lon, c)$
ω_1
ω_2
......
ω_H
Attention Pooling
W
Place Representation in POI Modality

Figure 2. the structure of POI encoder in MMCPP

3.2 Image Encoder

The image encoder of MMCPP adopts VisionTransformer (ViT) (Dosovitskiy et al. 2021), which is used to extract surface information from images within a place, such as the size, shape, and spatial position distribution of ground objects.

The input to ViT is the images within the place, each containing 200×200 pixels. The size of each image patch is set to 10×10 pixels, resulting in 400 image patchs per image. ViT adds a learning "[CLS]" item before image patch embeddings, which serves as a representation of the entire image corresponding to the place. Therefore, the feature vector of "[CLS]" is used as the representation of the place in the image modality.

3.3 Trajectory Encoder

Fine-grained trajectory sequences describe the dynamic interactions of spatial proximity and behavior proximity among places simultaneously. The representation of a place will spread to the representations of other places along the trajectory sequence containing the movement behavior of the crowd. And the autocorrelation and complementary effects from different places will also spread to the representations of the target places.

This paper proposes a trajectory encoder, named RodtFormer, as shown in Figure 3. It based on the RoFormer (a variant of Transformer) structure, takes fine-grained spatio-temporal trajectories as inputs and uses the self-attention mechanism with rotary temporal interval position embedding to simulate the dynamic interactions among places. It captures spatial proximity interactions, behavioral proximity interactions, and temporal characteristics including absolute chronological order and relative time intervals among places. As a result, the encoder represents second-order properties of places.

RodtFormer begins by converting the coordinate sequence of trajectories into the corresponding place ID sequence. This is followed by the place ID coding layer, which maps each place ID to the corresponding continuous feature vectors.

These vectors are then input into a multi-head self-attention mechanism. However, the self-attention mechanism loses the sequence position information (Vaswani et al. 2017). Transformer and RoFormer use simple sequential orders (0, 1, 2, 3, ...), respectively combined with sinusoidal position encoding and rotary position encoding to complement posotion information. However, for trajectories, the time interval of visiting each place is not uniform and contains important information, such as locations' visited frequencies or stayed durations (Lin et al. 2021).

This paper introduces rotary temporal interval position encoding to capture two aspects of time characteristics in trajectory data: absolute chronological order and relative time intervals. The new encoding method replaces "simple sequential orders" with "first time interval" to introduce time position information. This information is calculated by subtracting the timestamp t_m of each position in the sequence from the timestamp t_1 of the starting point. For instance, given an input trajectory sequence $Traj_{grid} = [g_1, g_2, g_3, \ldots, g_M]$ and its corresponding timestamp sequence $Traj_t = [t_1, t_2, t_3, \ldots, t_M]$, the resulting place ID and time interval sequences are expressed as: $Traj'_{grid} = \left[[BEGIN],\ g_1, g_2, g_3, \ldots, g_M, [END], [PAD], \ldots,\right]$, and $Traj'_{\Delta t} = \left[0, 1, \Delta t_{2,1}, \Delta t_{3,1}, \ldots, \Delta t_{M,1}, \Delta t_{M,1} + 1, 0, \ldots,\right]$. In addition, the factor μ controlled the relative time interval attenuation and the scaling factor s of time interval are introduced to further enhance the ability to express time information. Both of them are 1-dimensional, learnable variables. Moreover, the rotational time interval position encoding matrix $\mathbf{R} = [\mathbf{R}_1, \ldots, \mathbf{R}_m, \ldots, \mathbf{R}_M]$ is constructed, which is used to transform the matrices $\mathbf{Q} = [\mathbf{Q}_1, \ldots, \mathbf{Q}_m, \ldots, \mathbf{Q}_M]$ and $\mathbf{K} = [\mathbf{K}_1, \ldots, \mathbf{K}_n, \ldots, \mathbf{K}_M]$ of the self-attention mechanism. Overall, this enables the modeling of absolute chronological order and relative time intervals of dynamic interactions simultaneously. The calculation equations of the process can be written as follows:

$$\mathbf{Q}_m{}'^{\top}\mathbf{K}_n{}' = (\boldsymbol{R}_m\mathbf{Q}_m)^{\top}(\boldsymbol{R}_n\mathbf{K}_n) \tag{1}$$

$$\boldsymbol{R}_m = \begin{pmatrix} \cos\Delta t_{m,1}\theta_0 & -\sin\Delta t_{m,1}\theta_0 & 0 & 0 & \cdots & 0 & 0 \\ \sin\Delta t_{m,1}\theta_0 & \cos\Delta t_{m,1}\theta_0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos\Delta t_{m,1}\theta_1 & -\sin\Delta t_{m,1}\theta_1 & \cdots & 0 & 0 \\ 0 & 0 & \sin\Delta t_{m,1}\theta_1 & \cos\Delta t_{m,1}\theta_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos\Delta t_{m,1}\theta_{\frac{d}{2}-1} & -\sin\Delta t_{m,1}\theta_{\frac{d}{2}-1} \\ 0 & 0 & 0 & 0 & \cdots & \sin\Delta t_{m,1}\theta_{\frac{d}{2}-1} & \cos\Delta t_{m,1}\theta_{\frac{d}{2}-1} \end{pmatrix} \tag{2}$$

$$\Delta t_{m,1} = s * (t_m - t_1), s \in \mathbb{R} \tag{3}$$

$$\theta_i = \mu^{-\frac{2i}{d}}, \mu \in \mathbb{R} \tag{4}$$

where t_m = timestamp of the m-th position
i = dimension index of the feature vector

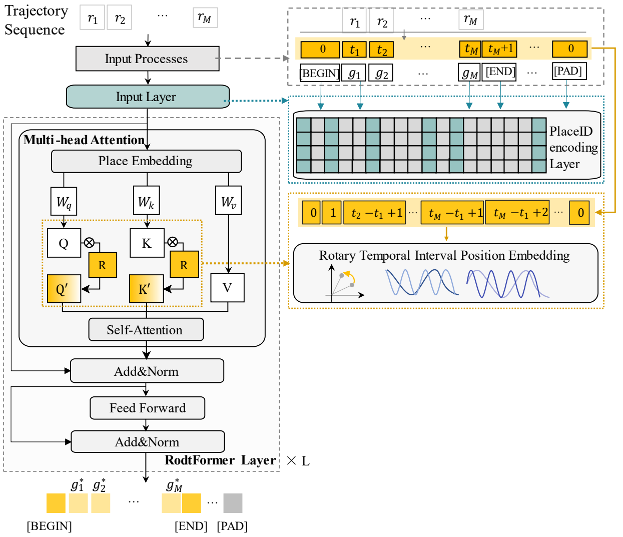


Figure 3. the structure of trajectory encoder in MMCPP

3.4 Model Pre-Training

3.4.1 Phase 1: Single-modal independent pre-training: Three encoders corresponding modalities are pre-trained by using self-supervised learning tasks to capture the place attributes represented in that modality.

The POI encoder is pre-trained using the POI category set (additionally contains “NAN”) for multi-label classification.

The image encoder adopts the Masked Autoencoder (MAE) and uses the mask image restoration task for pre-training (He et al. 2021). Firstly, a scene image is divided into multiple image patches. These patches are then randomly masked and input into a decoder to restore pixel values within the mask regions. A mask rate of 0.85 is used in this study.

The trajectory encoder adopts the pre-training target of Masked Trajectory (MT), which is inspired by the Masked Language Model (MLM) pre-training objective used in BERT(Devlin et al. 2019). MT randomly masks place IDs in trajectories and predicts their original visited place IDs. The mask rate is set to 0.15.

3.4.2 Phase2 Multi-modal coordinated pre-training: Building upon the results of phase 1, we propose a contrastive sample construction method that uses contrastive learning to train MMCPP coordinatedly, allowing the model to integrate first-order and second-order attributes of the place in multi-modal data while maintaining each encoder's independent encoding ability. Figure 4 shows the training process of phase 2.

To begin, a batch of places is randomly selected from the research area, and corresponding POIs and images are collected. Trajectories are then sampled from existing collections of trajectories that include these places as stop points. After the data for each modality is input into its respective encoder, representations of each place in each modality are output.

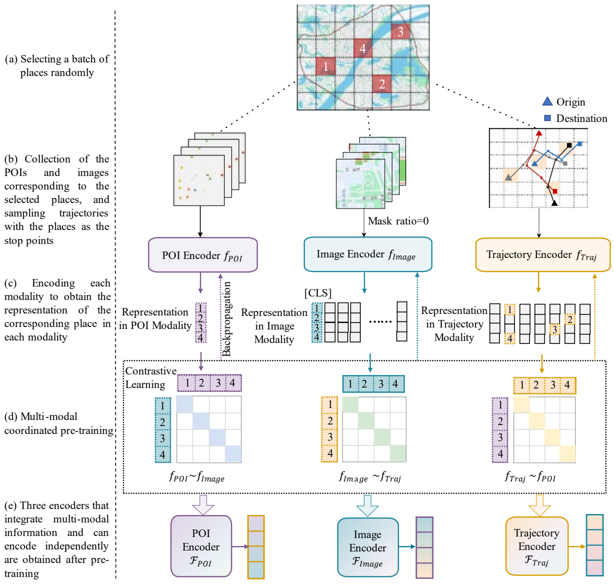

(a) Selecting a batch of places randomly
(b) Collection of the POIs and images corresponding to the selected places, and sampling trajectories with the places as the stop points
(c) Encoding each modality to obtain the representation of the corresponding place in each modality
(d) Multi-modal coordinated pre-training
(e) Three encoders that integrate multi-modal information and can encode independently are obtained after pre-training
1 2 3 4
Origin
Destination
Mask ratio=0
POI Encoder f_{POI}
Image Encoder f_{Image}
Trajectory Encoder f_{Traj}
Representation in POI Modality
Backpropagation
[CLS]
Representation in Image Modality
Representation in Trajectory Modality
Contrastive Learning
f_{POI}~f_{Image}
f_{Image} ~f_{Traj}
f_{Traj} ~f_{POI}
POI Encoder $\mathcal{F}_{POI}$
Image Encoder $\mathcal{F}_{Image}$
Trajectory Encoder $\mathcal{F}_{Traj}$

Figure 4. Phase 2: multi-modal coordinated pre-training

Using the alignment relationship of "information from different modalities describe the same place," representations of the same place in each modality are paired together as positive samples (colored squares on the diagonal in Figure 4(d)). Representations of different places in each modality are paired as negative samples (white squares in Figure 4(d)). The three encoders learn a multi-modal embedding space coordinatedly using contrastive learning by maximizing the cosine similarity of the POIs and image representations, POIs and trajectory representations, as well as the trajectory and image representations of the 3*N positive pairs in the batch while minimizing the cosine similarity of the embeddings of the 3*(N^2 – N) negative pairings. We optimize the sum of the three symmetric cross entropy losses L (Radford et al. 2021) over these similarity scores, which are computed as follows:

$$L = SCE_{POI-image} + SCE_{Image-Traj} + SCE_{Traj-POI} \tag{5}$$

$$SCE = \frac{H(\boldsymbol{q},\boldsymbol{p}) + H(\boldsymbol{q},\boldsymbol{p}^T)}{2} \tag{6}$$

$$H(q,p) = -\mathrm{sum}\big(\boldsymbol{q} \odot \log(\mathrm{softmax}(\boldsymbol{p}))\big) \tag{7}$$

where SCE_{A-B} = symmetric cross entropy losses for A and B modalities
$\boldsymbol{q}$ = a unit matrix
$\boldsymbol{p}$ = a similarity matrix of representations between modalities

For places without POIs or trajectories, we use encoded representations of "pseudo-POIs" and "pseudo-trajectories," respectively. The "pseudo-trajectory" consists of only a corresponding single place ID in the sequence, and its timestamp can be initialized randomly since the "first time interval" of the first point of the trajectory is transferred to 1. This approach enables unification of inputs across all modalities, including cases with partial missing modalities, in multi-modal contrastive learning tasks.

Upon completion of phase 1 and phase 2 pre-training, we obtain three encoder components that integrate different modal information and can still be independently encoded: POI encoder $\mathcal{F}_{POI}$, image encoder $\mathcal{F}_{Image}$, and trajectory encoder $\mathcal{F}_{Traj}$.

4. EXPERIMENTS

To assess the quality of the place representations generated by MMCPP, they were incorporated into a location prediction model and compared with other location embedding methods.

4.1 Study area

The rectangular envelope of the Third Ring Road in Wuhan, China, is taken as the research area. And considering the sampling interval and average speed of most taxi trajectories, the research unit is set as a longitude-latitude geographic grid with a side length of 0.0018° (each grid is about 160 meters×200 meters in the WGS84 coordinate system). The research area is composed of a total of 22,950 geographic grids (170 × 135).

4.2 Datasets

The proposed model uses POIs, maptiles, and taxi trajectories as data sources for the corresponding three modalities.

4.2.1 Point of Interests (POIs)

POIs describe the typical functions and activities of an area and can be used to mine the functional semantics of cities, as shown in previous studies (Jenkins et al. 2019). In this study, POI data was obtained from the Amap platform (https://lbs.amap.com) in 2018, which included 17 categories.

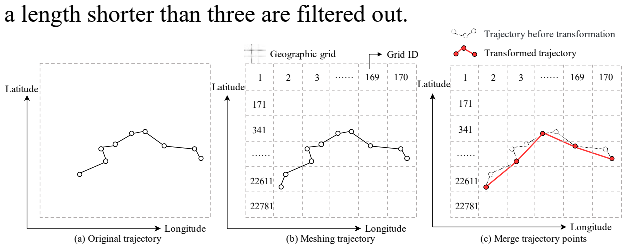
4.2.2 Images

Maptiles offer geospatial information such as the shape, color, and size of ground objects (Wang, Wang, et al. 2020). The maptiles used in this study were obtained from the TencentMap platform, with a zoom level of 17, a resolution of around 1 meter, and three channels. As a data enhancement measure, maptiles of cities that are of the same size as Wuhan were also collected and mixed with Wuhan data for model pre-training. These included parts of four Chinese cities: Hangzhou, Chengdu, Nanjing, and Changsha. All maptiles were resampled to 0.00009°, so that each geographic grid contained 200x200 pixels.

4.2.3 Trajectories

The passenger trajectories are extracted from a taxi trajectory dataset to simulate dynamic interactions between the grids.

"Meshing" is a process that converts trajectory points into corresponding grids based on their latitudes and longitudes, resulting in grid ID sequences. For taxi trajectory data, it aims to model the interactions between grids, therefor trajectory points inside each grid are merged, and only the first trajectory point entering the grid is retained. Specially, in cases where multiple trajectory points fall within the final grid of a sequence, the last trajectory point in the grid is kept to preserve the complete time interval of the trajectory. This transformation process is illustrated in Figure 5. Finally, any transformed trajectories with a length shorter than three are filtered out.



(a) Original trajectory (b) Meshing trajectory (c) Merge trajectory points

Figure 5. The process of meshing and merging trajectories

Finally, the trajectory dataset 2,533,754 passenger trajectories of 4,121 taxis in the first 4 weeks (May 27, 2019-June 23, 2019) and 663,574 passenger trajectories of 4,129 taxis in the last week (June 24, 2019-June 30, 2019). The data cover more than 90% of the grids in the study area.

4.3 Baseline Place Representation Methods

To prove the superiority of MMCPP, we include two classic representation methods, and also two state-of-the-art place embedding methods for comparison.

CBOW(Mikolov et al. 2013): An implementation of Word2Vec, which captures the semantic of sequence points through the correlation between the target place and its context.

Skip-Gram(Mikolov et al. 2013): Another implementation of Word2Vec. It takes target places as input and predicts context places within a certain window range as output, thus capturing the semantic of sequence points.

CTLE(Lin et al. 2021): Context and Time aware Location Embeddingn model uses the Transformer as the backbone with the sinusoidal timestamp position embedding and uses masked trajectory, masked hour, and masked weekday as the pre-training objectives.

RegionEncoder(Jenkins et al. 2019): A deep learning model for learning low-dimensional distributed representations of discrete spatial regions, which utilizes POIs, satellite images and taxi trajectories as data sources. We experimented with satellite data replaced by maptiles in this paper.

4.4 Settings

During the single-modal pre-training in phase 1, we used POI data from Wuhan, maptiles from Wuhan and 4 other cities, as well as trajectory data from the first 4 weeks of Wuhan as the pre-training datasets for each modality. For the POI encoder of MMCPP, we implemented an input layer with a hidden size of 128 and an attention pooling layer with a hidden size of 128. The optimization process for the POI encoder involved using a batch size of 512, a maximum of 150 training epochs, and AdamW with a learning rate of 0.0002. Regarding the MAE structure, the image encoder adopted a 4-layer ViT where the input layer had a hidden size of 128, 8-head self-attention mechanism, and an embedding dimension of the feedforward layer at size 512. The decoder had 2 layers of Transformer with an 8-head self-attention mechanism and a feed-forward layer with a hidden size of 512. The MAE optimization process included using a batch size of 32, a maximum of 50 training epochs, and AdamW with a learning rate of 0.0002. The trajectory encoder used 3 layers of RodtFormer based on an input layer with a hidden size of 128, 8-head self-attention mechanism, and a hidden embedding of feed-forward layer with a hidden size of 512. The optimization for the trajectory encoder entailed utilizing a batch size of 128, a maximum of 20 training epochs, and AdamW with a learning rate of 0.0008. During phase 2 of our multi-modal coordinated pre-training, we utilized the single-modal component that performed the best in phase 1 and the number of sampling trajectories is set to 20. For phase 2 optimization, we used a batch size of 128, a maximum of 20 training epochs, and AdamW with a learning rate of 0.00001. In the two pre-training phases, 95% of these datasets

https://doi.org/10.5194/isprs-annals-X-1-W1-2023-303-2023

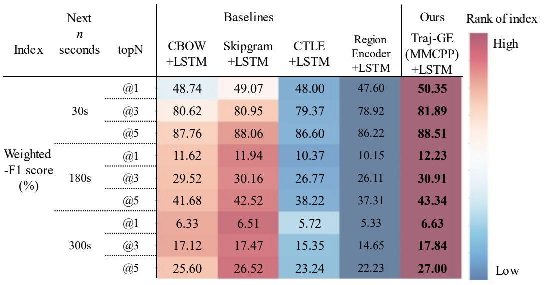
were allocated as the training set, while the remaining 5% of the data was used for validation purposes. And the losses on the validation sets are used to judge whether to stop training early.

To verify performances of models, we used the trajectory dataset from the last week as the downstream task data. Of the total dataset, 94% was allocated to training the models, with 1% set aside for the validation set to select the best downstream task model. The remaining 5% was utilized to evaluate the performance of the downstream models. The location prediction models with different initial grid representation are trained with Cross Entropy loss, and evaluated with weighted-F1 score. All embedding models' dimensions of the vectors of place representaton are set to 128. We implement CTLE based on 3-layer networks with a hidden size of 512. The CTLE was pre-trained on the training sets for 100 epochs and AdamW with learning rate of 0.00008. The same as Lin et al. 2021, for our MMCPP model and CTLE model, we used the placeID encoding layer of the trajectory encoder in MMCPP(marked as Traj-GE(MMCPP)) and in CTLE to obtain the grid representations respectively. The location representations modeled by different baselines and the method in this paper are fixed as non-learnable parameters, and the downstream task models all use a single-layer LSTM with a hidden vector dimension of 128. And all models are trained with the early-stopping mechanism to obtain the best-performing epochs on the evaluation sets. AdamW was finally chosen as optimizer and an initial learning rate of 0.001 for LSTM in the downstream task. We implement all baseline models and our model in PyTorch. All experiments have run on Intel(R) Xeon(R) CPUs, and NVIDIA Tesla T4 GPUs.

4.5 Experimental Results

Figure 6 shows the performance comparison of different models for location prediction at next *n* seconds including 30 seconds(s), 180(s), 300(s). In the chart, a redder color indicates better performance in the downstream task metric, while a bluer color indicates worse performance. Values in bold indicate that the highest number in the row is the top performer.

In the experiment, the model trained with the place representation encoded by MMCPP outperformed the baseline method on the test set, with an average of 1.96% higher in F1 score.



Index	Next *n* seconds	topN	Baselines CBOW +LSTM	 Skipgram +LSTM	 CTLE +LSTM	 Region Encoder +LSTM	Ours Traj-GE (MMCPP) +LSTM
Weighted -F1 score (%)	30s	@1	48.74	49.07	48.00	47.60	**50.35**
		@3	80.62	80.95	79.37	78.92	**81.89**
		@5	87.76	88.06	86.60	86.22	**88.51**
	180s	@1	11.62	11.94	10.37	10.15	**12.23**
		@3	29.52	30.16	26.77	26.11	**30.91**
		@5	41.68	42.52	38.22	37.31	**43.34**
	300s	@1	6.33	6.51	5.72	5.33	**6.63**
		@3	17.12	17.47	15.35	14.65	**17.84**
		@5	25.60	26.52	23.24	22.23	**27.00**

Figure 6. Experimental results of the downstream task using different grid representations encoded by corresponding place representation methods

Both CBOW and Skip-Gram disregard the spatial and temporal information in the trajectory data, which reflect unique attributes of places. As a result, the grid location representations generated by these methods perform slightly worse than MMCPP when applied to downstream models. Given that most taxi trajectory records occur on the same working day or even the same hour, the masked hour and weekday pre-training tasks introduced by the CTLE model may not add value to the actual representation training process. Furthermore, Yan et al. 2019 has found that using the dot product operation in Transformer encoding does not appropriately model time-interval characteristics so that it is less effective than MMCPP in this task. The embedding represented by RegionEncoder performs the worst on the task, possibly due to its weaker convolutional neural network structure than ViT in encoding images. Additionally, the method only considers the interaction between pick-up and drop-off points, ignoring finer details such as visit order and access time intervals contained in the fine-grained trajectories. Lastly, due to the sparseness of taxi flow matrix and POI data, it may be challenging to build sufficient place representations.

MMCPP integrates the information of multiple modalities; and uses the rotary temporal interval position embedding to introduce the time information in the trajectory sequence, and considers the absolute chronological order and relative time intervals among places. These designs enable MMCPP to build higher-quality place representations during pre-training, which can help downstream location prediction models achieve better performance.

4.6 Ablation Study

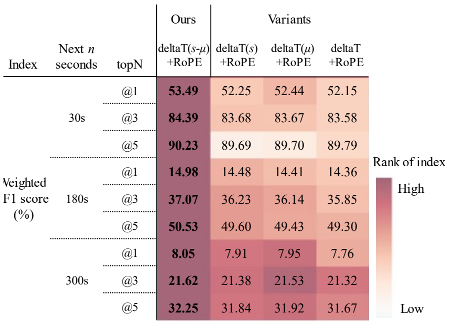
Ablation experiments are used to verify the effectiveness of the rotary temporal interval position embedding and the effectiveness of integrating various multi-modal components.

4.6.1 Verification of the effectiveness of the rotary temporal interval position embedding

This paper compares the position embedding masked as deltaT(*s*-*μ*)+RoPE with three variants shown in Table 1. The results are presented in Figure 7.

	Model	*s*	*μ*
Variants	deltaT(*s*)+RoPE deltaT(*μ*)+RoPE deltaT+RoPE	learnable 1(fixed) 1(fixed)	10000(fixed) learnable 10000(fixed)
Ours	deltaT(*s*-*μ*)+RoPE	learnable	learnable

Table 1. List of models for the ablation experiment with different position embedding methods



Index	Next *n* seconds	topN	Ours deltaT(*s*-*μ*) +RoPE	Variants deltaT(*s*) +RoPE	 deltaT(*μ*) +RoPE	 deltaT +RoPE
Veighted F1 score (%)	30s	@1	**53.49**	52.25	52.44	52.15
		@3	**84.39**	83.68	83.67	83.58
		@5	**90.23**	89.69	89.70	89.79
	180s	@1	**14.98**	14.48	14.41	14.36
		@3	**37.07**	36.23	36.14	35.85
		@5	**50.53**	49.60	49.43	49.30
	300s	@1	**8.05**	7.91	7.95	7.76
		@3	**21.62**	21.38	21.53	21.32
		@5	**32.25**	31.84	31.92	31.67

Figure 7. Experimental results of the downstream task using models with different position embedding methods

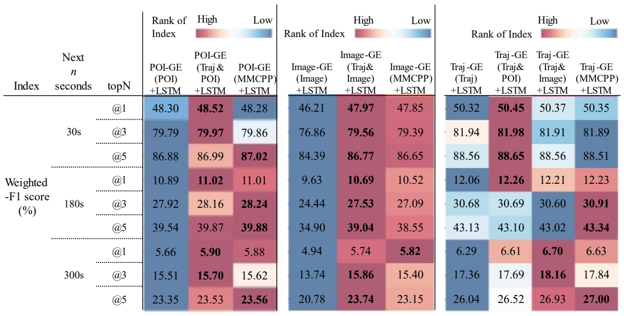
The rotation time difference position encoding proposed in this paper can be improved by approximately 0.17% on average when introducing the scaling factor s of the time interval. And taking the factor *μ* controlled the relative time interval attenuation as a learnable parameter can increase by about 0.16% on average. Combining the two factors can further enhance the performance, resulting in an average increase of approximately 0.64%.

4.6.2 Verification of the effectiveness of the multimodal components

This experiment includes the MMCPP model and 5 variants that employ different modality combinations, involving place representations encoded by 10 encoders. These encoders consist of three different single-modal pre-trained encoders, a POI encoder and a trajectory encoder that integrate POIs and trajectories, an image encoder and a trajectory encoder that integrate images and trajectories, and the integration of all three encoders for the 3-modal data, as shown in Table 2. In particular, "POI-GE" signifies that each grid POI set is encoded by the POI encoder for grid representation. Similarly, "Image-GE" indicates that the mask rate of the image encoder is set to 0, and it encodes the output of each maptile "[CLS]" marked grid representation.

Models	Modality		
	POIs	Image	Trajectory
POI-GE(POI)	√		
POI-GE(Traj&POI)	√		√
POI-GE(MMCPP)	√	√	√
Image-GE(Image)		√	
Image-GE(Traj&Image)		√	√
Image-GE(MMCPP)			
Traj-GE(Traj)			√
Traj-GE(Traj&POI)	√		√
Traj-GE(Traj&Image)		√	√
Traj-GE(MMCPP)	√	√	√

Table 2. List of models for the multi-modal component ablation experiment



Rank of Index
High
Low

Index	Next *n* seconds	topN	POI-GE (POI) +LSTM	POI-GE (Traj & POI) +LSTM	POI-GE (MMCPP) +LSTM	Image-GE (Image) +LSTM	Image-GE (Traj& Image) +LSTM	Image-GE (MMCPP) +LSTM	Traj-GE (Traj) +LSTM	Traj-GE (Traj& POI) +LSTM	Traj-GE (Traj& Image) +LSTM	Traj-GE (MMCPP) +LSTM
Weighted -F1 score (%)	30s	@1	48.30	**48.52**	48.28	46.21	**47.97**	47.85	50.32	**50.45**	50.37	50.35
		@3	79.79	**79.97**	79.86	76.86	**79.56**	79.39	81.94	**81.98**	81.91	81.89
		@5	86.88	86.99	**87.02**	84.39	**86.77**	86.65	88.56	**88.65**	88.56	88.51
	180s	@1	10.89	**11.02**	11.01	9.63	**10.69**	10.52	12.06	**12.26**	12.21	12.23
		@3	27.92	28.16	**28.24**	24.44	**27.53**	27.09	30.68	30.69	30.60	**30.91**
		@5	39.54	39.87	**39.88**	34.90	**39.04**	38.55	43.13	43.10	43.02	**43.34**
	300s	@1	5.66	**5.90**	5.88	4.94	5.74	**5.82**	6.29	6.61	**6.70**	6.63
		@3	15.51	**15.70**	15.62	13.74	**15.86**	15.40	17.36	17.69	**18.16**	17.84
		@5	23.35	23.53	**23.56**	20.78	**23.74**	23.15	26.04	26.52	26.93	**27.00**

Figure 8. Experimental results of the downstream task using models with different combinations of modalities

The experimental results indicate that integrating multi-modal information representation can improve the performance of the downstream task to a certain extent. Integrating POI information enhances the performance of short-term prediction tasks, while integrating maptiles and trajectories improves the performance of long-term prediction tasks. However, in order to simultaneously express the information of all three modalities, MMCPP slightly sacrifices its performance on this task, leading to slightly weaker results on some indicators.

5. CONCLUSION

This paper proposes a novel pre-training model, MMCPP, for representing places with multi-modal data. The model comprises three encoders that capture the attributes and semantics of a place in their corresponding modes, then integrates multi-modal information by using a coordinated pre-training framework. Finally, the three pre-trained encoders in MMCPP can independently encode place representations, which improves the reusability of the model.

We use the taxi trajectory data to verify the effectiveness of the model in the location prediction task at next *n* seconds, including 30 seconds, 180 seconds, and 300 seconds. The results show that in comparison to existing embedding methods, our model is capable of learning higher-quality position representations during pre-training and improves the performance of downstream tasks, benefiting from the integration of multi-modal information and the modeling of dynamic interactions.

REFERENCES

Baltrušaitis T, Ahuja C, Morency L-P., 2019: Multimodal Machine Learning: A Survey and Taxonomy. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. pp.423–443.

Das SSS, Ali ME, Li Y-F, Kang Y-B, Sellis T., 2021: Boosting House Price Predictions Using Geo-spatial Network Embedding. In *Data Mining and Knowledge Discovery*. pp.2221–2250.

Devlin J, Chang M-W, Lee K, Toutanova K., 2019: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S., 2021: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*.

Du J, Zhang Y, Wang P, Leopold J, Fu Y., 2019: Beyond Geo-First Law: Learning Spatial Representations via Integrated Autocorrelations and Complementarity. *2019 IEEE International Conference on Data Mining (ICDM)*, 160–169.

Fu Y, Wang P, Du J, Wu L, Li X., 2019: Efficient Region Embedding with Multi-View Spatial Networks: A Perspective of Locality-Constrained Spatial Autocorrelations. In *Proceedings of the AAAI Conference on Artificial Intelligence*. pp.906–913.

He K, Chen X, Xie S, Li Y, Doll'ar P, Girshick RB., 2021: Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp.15979–15988.

Huang T, Wang Z, Sheng H, Ng AY, Rajagopal R., 2021: Learning Neighborhood Representation from Multi-Modal Multi-Graph: Image, Text, Mobility Graph and Beyond. In *arXiv preprint arXiv:2105.02489*.

Janowicz K, Gao S, McKenzie G, Hu Y, Bhaduri B., 2020: GeoAI: Spatially Explicit Artificial Intelligence Techniques for Geographic Knowledge Discovery and Beyond. In *International Journal of Geographical Information Science*. pp.625–636.

Jenkins P, Farag A, Wang S, Li Z., 2019: Unsupervised Representation Learning of Spatial Data via Multimodal Embedding. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1993–2002.

Li F., 2018: *Evolution and optimization of green space pattern in central Beijing based on multi-source data*. In Beijing, China: Beijing Forestry University.

Li Y, Gao C, Yao Q, Li T, Jin D, Li Y., 2022: DisenHCN: Disentangled Hypergraph Convolutional Networks for Spatiotemporal Activity Prediction. In *arXiv preprint arXiv:2208.06794*.

Lin Y, Wan H, Guo S, Lin Y., 2021: Pre-training Context and Time Aware Location Embeddings from Spatial-Temporal Trajectories for User Next Location Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*. pp.4241–4248.

Liu Y, Ao X, Dong L, Zhang C, Wang J, He Q., 2022: Spatiotemporal Activity Modeling via Hierarchical Cross-Modal Embedding. In *IEEE Transactions on Knowledge and Data Engineering*. pp.462–474.

Liu Y, Yao X, Gong Y, Kang C, Shi X, Wang F, Wang J, Zhang Y, Zhao P, Zhu D, Zhu X., 2020: Analytical methods and applications of spatial interactions in the era of big data. In *Acta Geographica Sinica*. Acta Geographica Sinica, pp.1523.

Liu Z, Miranda F, Xiong W, Yang J, Wang Q, Silva C., 2020: Learning Geo-Contextual Embeddings for Commuting Flow Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*. pp.808–816.

Luo Y, Chung F, Chen K., 2022: Urban Region Profiling via A Multi-Graph Representation Learning Framework. In *arXiv preprint arXiv:2202.02074*.

Mai G, Janowicz K, Hu Y, Gao S, Yan B, Zhu R, Cai L, Lao N., 2022: A Review of Location Encoding for GeoAI: Methods and Applications. In *International Journal of Geographical Information Science*. pp.639–673.

Mai G, Xuan Y, Zuo W, Janowicz K, Lao N., 2022: Sphere2Vec: Multi-Scale Representation Learning over a Spherical Surface for Geospatial Predictions. In *arXiv preprint arXiv:2201.10489*.

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J., 2013: Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, 3111–3119.

Paul D, Li F, Phillips JM., 2021: Semantic Embedding for Regions of Interest. In *The VLDB Journal*. pp.311–331.

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I., 2021: Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning* 139, 8748–8763.

Sun G, Qi H, Shen Y, Yin B., 2022: TCSA-Net: A Temporal-Context-Based Self-Attention Network for Next Location Prediction. In *IEEE Transactions on Intelligent Transportation Systems*. pp.20735–20745.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I., 2017: Attention Is All You Need. *Advances in Neural Information Processing Systems* 30.

Wan H, Lin Y, Guo S, Lin Y., 2022: Pre-training Time-Aware Location Embeddings from Spatial-Temporal Trajectories. In *IEEE Transactions on Knowledge and Data Engineering*. pp.5510–5523.

Wang D, Liu K, Mohaisen D, Wang P, Lu C-T, Fu Y., 2021: Automated Feature-Topic Pairing: Aligning Semantic and Embedding Spaces in Spatial Representation Learning. *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, 450–453.

Wang X, Fukumoto F, Li J, Yu D, Sun X., 2022: STaTRL: Spatial-Temporal and Text Representation Learning for POI Recommendation. In *Applied Intelligence*. pp.1–16.

Wang Y, Wang C, Ling Y, Yokoyama K, Wu H-T, Fang Y., 2020: Leveraging an Efficient and Semantic Location Embedding to Seek New Ports of Bike Share Services. *2020 IEEE International Conference on Big Data (Big Data)*, 1273–1282.

Wang Z, Li H, Rajagopal R., 2020: Urban2Vec: Incorporating Street View Imagery and POIs for Multi-Modal Urban Neighborhood Embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*. pp.1013–1020.

Xuan H, Li Q, Zhang Y., 2016: Spatial Characteristics Analysis of Total Population in Various Cities Based on Geographically Weighted Regression. In *Journal of Biomathematic*. pp.223–228.

Yan H, Deng B, Li X, Qiu X., 2019: TENER: Adapting Transformer Encoder for Named Entity Recognition. In *arXiv preprint arXiv:1911.04474*.

Yao Z, Fu Y, Liu B, Hu W, Xiong H., 2018: Representing Urban Functions through Zone Embedding with Human Mobility Patterns. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 3919–3925.

Zhang C, Zhang K, Yuan Q, Tao F, Zhang L, Hanratty T, Han J., 2017: ReAct: Online Multimodal Embedding for Recency-Aware Spatiotemporal Activity Modeling. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 245–254.

Zhang M, Li T, Li Y, Hui P., 2020: Multi-View Joint Graph Representation Learning for Urban Region Embedding. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 4431–4437.

Zhang Y, Fu Y, Wang P, Li X, Zheng Y., 2019: Unifying Inter-region Autocorrelation and Intra-region Structures for Spatial Embedding via Collective Adversarial Learning. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1700–1708.

Zhao S, Zhao T, King I, Lyu MR., 2017: Geo-Teaser: Geo-Temporal Sequential Embedding Rank for Point-of-interest Recommendation. In *Proceedings of the 26th International Conference on World Wide Web Companion*. pp.153–162.

Zhu M, Chen W, Xia J, Ma Y, Zhang Y, Luo Y, Huang Z, Liu L., 2019: Location2vec: A Situation-Aware Representation for Visual Exploration of Urban Locations. In *IEEE Transactions on Intelligent Transportation Systems*. pp.3981–3990.