MuseCL: Predicting Urban Socioeconomic Indicators via Multi-Semantic Contrastive Learning

Xixian Yong, **Xiao Zhou**[*]

Gaoling School of Artificial Intelligence, Renmin University of China

{xixianyong, xiaozhou}@ruc.edu.cn

Abstract

Predicting socioeconomic indicators within urban regions is crucial for fostering inclusivity, resilience, and sustainability in cities and human settlements. While pioneering studies have attempted to leverage multi-modal data for socioeconomic prediction, jointly exploring their underlying semantics remains a significant challenge. To address the gap, this paper introduces a Multi-Semantic Contrastive Learning (MuseCL) framework for fine-grained urban region profiling and socioeconomic prediction. Within this framework, we initiate the process by constructing contrastive sample pairs for street view and remote sensing images, capitalizing on the similarities in human mobility and Point of Interest (POI) distribution to derive semantic features from the visual modality. Additionally, we extract semantic insights from POI texts embedded within these regions, employing a pre-trained text encoder. To merge the acquired visual and textual features, we devise an innovative cross-modality-based attentional fusion module, which leverages a contrastive mechanism for integration. Experimental results across multiple cities and indicators consistently highlight the superiority of MuseCL, demonstrating an average improvement of 10% in R^2 compared to various competitive baseline models. The code of this work is publicly available at https://github.com/XixianYong/MuseCL.

1 Introduction

Urbanization is intricately connected to critical facets of the United Nations Sustainable Development Goals (UNSDGs), affecting energy, environment, economy, climate, etc. [Sachs *et al.*, 2022]. By 2020, over 55% of the global population resided in urban areas, and this trend is projected to persist and intensify in the forthcoming decades [Habitat, 2022]. Embracing urbanization yields numerous advantages, including a thriving cultural milieu, enhanced job prospects, and improved transportation networks, etc. However, it also begets
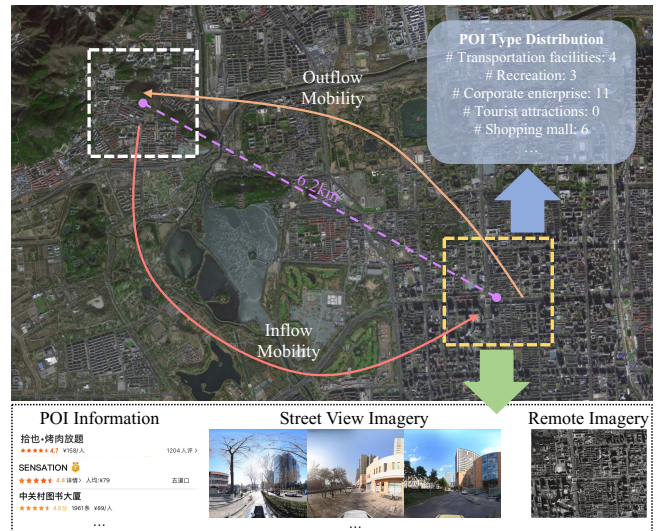
---

[*]Corresponding author.

Figure 1: Multi-modal data representing urban regions. Each region is linked to its remote sensing and street view imagery, POI data, and inter-region connections, encompassing factors like population mobility, fostering comprehensive insights into the urban landscape.

a host of predicaments and hurdles, such as air pollution, traffic congestion, and escalated energy consumption [Zheng *et al.*, 2014]. To address these challenges and achieve SDGs, gaining a comprehensive understanding of the urbanization phenomenon through fine-grained region profiling and accurate socioeconomic indicators becomes crucial.

Traditional approaches have relied on community surveys to gather statistics on metrics like population density and household income, which is both resource-intensive and time-consuming [Custodio *et al.*, 2023]. With the maturation of urban perception technology, diverse forms of data continue to proliferate within cities, which paves the way for fresh opportunities in tracking urban sustainable development indicators. As depicted in Figure 1, these datasets encompass point of interest (POI) information, vehicle movement trajectories, remote sensing data, street view imagery, and social media insights, among others. The utilization of this diverse data pool offers robust support for a myriad of downstream tasks. For instance, social media data is instrumental in predicting crime and unemployment rates [Antenucci *et al.*, 2014;

Aghababaei and Makrehchi, 2016], urban lifestyle mining [Zhou *et al.*, 2018], and significantly contributes to studies on urban sustainability [Ilieva and McPhearson, 2018]. Trajectory data reveals valuable insights into mobility patterns, socioeconomic indicators, and health trends [Cohen *et al.*, 2016; Gao *et al.*, 2017; Zhou *et al.*, 2017; Zhou *et al.*, 2023; Wang *et al.*, 2018b]. POI data aids in discovering new venues to explore [Zhou *et al.*, 2019], deducing regional functions [Yuan *et al.*, 2012], and controlling light pollution [Zhang *et al.*, 2024]. Recent research also delves into the potential of urban imagery, employing remotely sensed images for poverty prediction, land cover classification [Jean *et al.*, 2019; Hong *et al.*, 2020; Burke *et al.*, 2021], and analyzing street view images to estimate pedestrian volume [Chen *et al.*, 2020].

However, utilizing unimodal urban data often yields suboptimal results, prompting a growing inclination towards the integration of multi-modal data. For instance, the simultaneous utilization of streetscape and remote sensing imagery has proven effective in predicting socioeconomic indicators [Wang *et al.*, 2018a; Li *et al.*, 2022]. The combination of urban imagery with POI data has demonstrated its utility in enhancing region representation [Wang *et al.*, 2020; Huang *et al.*, 2021; Liu *et al.*, 2023]. Furthermore, researchers have ventured into the realm of multi-view graphs, leveraging data from diverse sources to comprehensively characterize regions [Qu *et al.*, 2017; Fu *et al.*, 2019]. This shift to multi-modal approaches holds great promise for advancing urban data analysis and interpretation, and helps to better achieve sustainable development goals. Recent efforts [Jean *et al.*, 2019; Wu *et al.*, 2022; Liu *et al.*, 2023] aim to derive latent embeddings for individual regions and employ them in conjunction with regional characteristics to predict a range of socioeconomic indicators, showcasing their notable versatility.

While prior studies have undertaken region profiling and socioeconomic prediction, several challenges persist. Among these, three primary ones emerge: (1) Rapid societal development has reshaped information exchange among regions, prompting a reassessment of the applicability of Tobler's First Law of Geography [Miller, 2004]. Consequently, a more precise method is warranted to assess region similarity. (2) Urban representation predominantly focuses on geography and human activity, necessitating effective modal filtering to meet region representation demands amidst the abundance of urban data. (3) Achieving effective fusion of diverse modal data is crucial yet complex in developing the final region representation, necessitating the advancement of robust multi-modal fusion techniques.

To tackle these challenges, we present a Multi-Semantic Contrastive Learning (MuseCL) framework. The primary contributions of our work can be summarized as follows:

- We pioneer the joint representation of regions using both street view and remote sensing imagery, concurrently integrating POI and mobility flow data to enrich the embedding with multi-dimensional semantic information.

- We enhance the spatial contrastive learning process by factoring in the similarity between regional POI and population mobility, resulting in more effective contrastive learning outcomes.

- We devise a cross-modal fusion model that aligns imagery with textual representation outputs, seamlessly integrating textual semantics into imagery representations.

- We validate the effectiveness of our framework through experiments on socioeconomic indicators in three major metropolises. The results demonstrate the superior performance of our model compared to various competitive state-of-the-art baselines across multiple downstream prediction tasks.

2 Related Work

Urban Representation Learning. With the increasing availability of urban data, representation learning in urban areas has witnessed significant growth in recent years. Numerous studies have capitalized on the proximity of similar regions in the embedding space to address various downstream tasks, such as crime prediction [Wang and Li, 2017; Zhang *et al.*, 2021], land cover classification [Yao *et al.*, 2018; Luo *et al.*, 2022], and socioeconomic feature prediction [Wang *et al.*, 2018a; Li *et al.*, 2022], among others. In this context, various strategies have emerged for urban region representation. For instance, Feng *et al.* [2017] proposed a latent representation model POI2Vec to jointly model the user preference and POI sequential transition influence for predicting potential visitors for a given POI. Wang and Li [2017] introduced a method incorporating temporal dynamics and multi-hop transitions. Zhang *et al.* [2017] presented a novel cross-modal representation learning method, CrossMap, which uncovers urban dynamics with massive geo-tagged social media data. Yao *et al.* [2018] proposed a framework to learn the vector representation of city zones by leveraging large-scale taxi trajectories. In a similar vein, Fu *et al.* [2019] explored multi-view spatial networks, considering geographical distance view and human mobility connectivity view for POIs within each region. Additionally, Wang *et al.* [2020] devised a multi-modal and multi-stage framework integrating image and text data within the neighborhood. These diverse approaches offer unique perspectives and valuable insights for further research in the urban representation learning field.

Socioeconomic Indicators Prediction. Initially, researchers primarily employed supervised and unsupervised learning methods for predicting socioeconomic indicators. For instance, Chakraborty *et al.* [2016] proposed a generative model of real-world events to predict various socioeconomic indicators based on extracted events. Qu *et al.* [2017] introduced a multi-view representation learning approach that fostered collaboration among different views to generate robust representations, subsequently used for socioeconomic indicator prediction. Similarly, He *et al.* [2018] unveiled correlations between visual patterns in satellite images and commercial hotspots. In recent years, self-supervised learning methods, especially contrastive learning, have gained traction for socioeconomic indicator forecasting. Drawing inspiration from Tobler's First Law of Geography [Miller, 2004], Jean *et al.* [2019] employed distance to establish neighborhood similarities in loss functions. Furthermore, Xi *et al.* [2022] incorporated POI similarity into contrastive learning to overcome distance-based limitations.
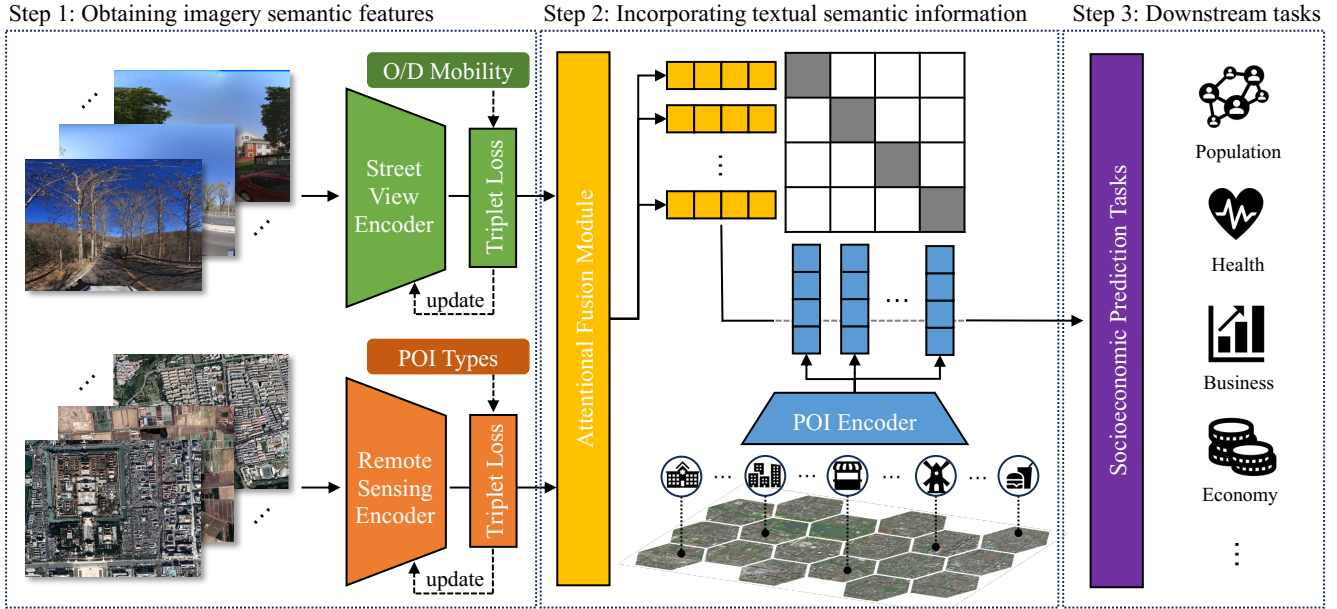
Figure 2: The overall architecture of the proposed MuseCL.

3 Preliminaries & Problem Statement

An urban area typically comprises multiple regions denoted as $\mathcal{R} = \{r_1, r_2, \cdots, r_N\}$. These regions exhibit unique geographic and demographic characteristics, often reflected through various data sources within them. In our study, we focus on analyzing specific attributes of regions $r_i \in \mathcal{R}$ ($i = 1, 2, \cdots, N$), investigating the following aspects:

- **Remote Sensing Imagery** $\mathcal{RV}_i$**.** Remote sensing imagery captures ground surface details, effectively revealing building distribution and thus providing valuable support for region representation.

- **Street View Imagery** $\mathcal{SV}_i = \{s_{i1}, s_{i2}, \cdots, s_{i|\mathcal{SV}_i|}\}$**.** It offers valuable insights into the appearance of streets, buildings, and their immediate surroundings. A region often contains multiple street view images.

- **POI Data** $\mathcal{T}_i = \{T_{i1}, T_{i2}, \cdots, T_{i|\mathcal{T}_i|}\}$**.** We textualize each POI as a bag of words $\{t_1, t_2, \cdots, t_n\}$, where each word is obtained from the POI's categories, ratings, reviews, and other relevant information.

- **Population Mobility** $\mathcal{M}_i = \{m_i^{in}, m_i^{out}\}$**.** m_i^{in} and m_i^{out} refer to the number of people entering and exiting the region r_i over a period of time, respectively. It can reflect the socio-demographic activity of a region.

Given a collection of urban remote sensing images $\mathcal{RV}$, street view images $\mathcal{SV}$, POI data $\mathcal{T}$, and population mobility data $\mathcal{M}$, our primary objective is to derive a low-dimensional representation $\epsilon_i \in \mathbb{R}^d$ for each region $r_i \in R(i = 1, 2, \cdots, N)$, where d signifies the dimension of the representation vectors. By effectively encapsulating the diverse characteristics inherent in each region, our approach aims to generate compact yet informative representations, denoted as $\mathcal{E} = \{\epsilon_1, \epsilon_2, \cdots, \epsilon_N\}$, to enhance various downstream socioeconomic prediction tasks in urban settings.

4 Methodology

4.1 Framework Overview

Figure 2 illustrates our proposed framework for fine-grained urban region profiling to predict socioeconomic indicators. This multi-step contrastive learning model consists of three key components: extracting semantic features from the visual modality, incorporating textual semantic information, and performing downstream tasks.

To begin, we partition the visual semantic learning module into remote sensing imagery representations based on POI similarity and street view imagery representations based on population flow similarity. Contrastive learning sample pairs are curated to acquire imagery features with distinct focal points. Subsequently, we take into account the POI text information associated with each region and leverage a pre-trained encoder-based model to derive the text features for every region. Then, employing a feature-level attentive fusion module, we align the combined remote sensing and street view features with the text representation vectors of each region, thereby imbuing the fused features with both visual and textual semantic insights. Lastly, we evaluate the low-dimensional representations of each region across a range of downstream tasks critical for urban sustainable development.

4.2 Visual Semantic Extraction

Street view and remote sensing imagery often contain information with different emphases. For example, street view imagery can provide characteristics of the social environment and population activity, while remote sensing imagery is more oriented towards geographic attributes and surface features [Liu *et al.*, 2023]. Therefore, we need to get the embedding of both separately and combine them effectively.
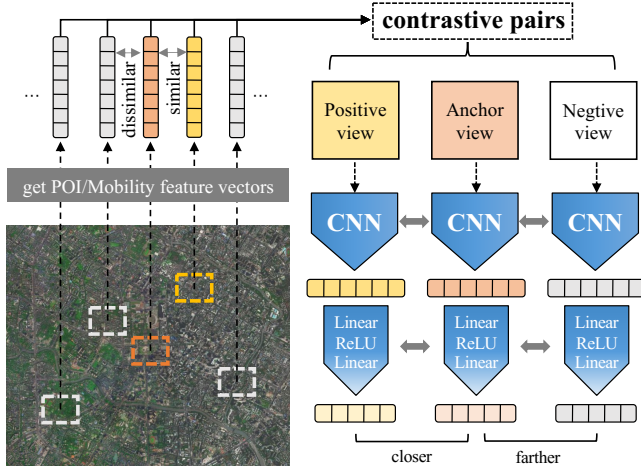
Figure 3: Urban visual representation through contrastive learning based on POI and mobility similarities.

Constructing Contrastive Samples

Recently, Xi *et al.* [2022] highlighted the limitations of Tobler's First Law of Geography [Miller, 2004], noting that relying solely on spatial distance to measure regional similarity is flawed. To address this, we propose a refined approach by constructing contrastive learning pairs for street and remote sensing images based on population flow and POI similarity, respectively. Street images are paired with mobility data because they reflect human movement patterns, while remote sensing images are paired with POI data because they capture the built environment and land use.

Population flow within a region can be gauged by quantifying the influx and efflux of individuals or vehicles over a specified timeframe. If we conceptualize regions as nodes and the movement of individuals or vehicles between regions as edges, the mobility of each region can be captured by tallying the entries and exits at each node. Assuming that the inflow of population to region r_i during a given period is m_i^{in}, and the outflow is m_i^{out}, the population mobility distance between regions r_i and r_j is computed by:

$$dist_{i,j}^{\mathrm{PM}} = \sqrt{\sum_{d \in \{in,out\}} \left(m_i^d - m_j^d\right)^2} \quad (1)$$

Then, the similarity of population mobility between the two regions can be quantified as:

$$\lambda_{i,j}^{\mathrm{PM}} = \frac{1}{dist_{i,j}^{\mathrm{PM}}} \quad (2)$$

We can construct positive samples characterized by higher similarity and negative samples characterized by lower similarity for each street view imagery, based on the parameter $\lambda_{i,j}^{\mathrm{PM}}$. As for remote sensing imagery, assuming that K different POI types are considered, we employ the Euclidean distance to quantify the POI distance between region r_i and r_j as follows:

$$dist_{i,j}^{\mathrm{POI}} = \sqrt{\sum_{k=1}^{K} \left(POI_i^k - POI_j^k\right)^2} \quad (3)$$

Therefore, the POI similarity between region r_i and r_j is:

$$\lambda_{i,j}^{\mathrm{POI}} = \frac{1}{dist_{i,j}^{\mathrm{POI}}} \quad (4)$$

POI / Mobility Triplet Loss

Using the acquired $\lambda_{i,j}^{\mathrm{PM}}$ and $\lambda_{i,j}^{\mathrm{POI}}$, we proceed to form separate pairs of contrastive learning samples for street view and remote sensing imagery. For instance, focusing on street view imagery, we establish each anchor image Anc_i^{SV} along with its corresponding positive sample Pos_i^{SV} and negative sample Neg_i^{SV} based on the population flow similarity $\lambda_{i,j}^{\mathrm{PM}}$. Subsequently, we train a convolutional neural network (CNN) denoted as F_{SV} to map the constructed contrastive learning samples $C^{\mathrm{SV}} = [Anc_i^{\mathrm{SV}}, Pos_i^{\mathrm{SV}}, Neg_i^{\mathrm{SV}}]$ into a low-dimensional vector space: $x_i^{\mathrm{SV}} = F_{\mathrm{SV}}(Anc_i^{\mathrm{SV}})$, $y_i^{\mathrm{SV}} = F_{\mathrm{SV}}(Pos_i^{\mathrm{SV}})$, and $z_i^{\mathrm{SV}} = F_{\mathrm{SV}}(Neg_i^{\mathrm{SV}})$. Similarly, we derive representation vectors for remote sensing imagery denoted as x_i^{RV}, y_i^{RV}, and z_i^{RV}, corresponding to the contrastive learning samples $C^{\mathrm{RV}} = [Anc_i^{\mathrm{RV}}, Pos_i^{\mathrm{RV}}, Neg_i^{\mathrm{RV}}]$.

Loss Optimization

To ensure the minimization of the distance between the anchor image and the positive image, while maximizing the separation from the negative image in the representation space, we employ Triplet Loss [Schroff *et al.*, 2015] as the loss function. The primary objective of this loss function is to bring features with similar labels into close proximity within the representation space, while simultaneously pushing features with dissimilar labels apart. For each pair of samples, we anticipate the fulfillment of the following equations:

$$\mathrm{sim}(x_i^m, y_i^m) + a \leq \mathrm{sim}(x_i^m, z_i^m), m \in \{\mathrm{SV}, \mathrm{RV}\} \quad (5)$$

$$\mathrm{Loss}(C^m) = [a + \mathrm{sim}(x_i^m, y_i^m) - \mathrm{sim}(x_i^m, z_i^m)]_+ \quad (6)$$

where $[\cdot]_+$ is a rectifier function to keep the loss function value non-negative, and $\mathrm{sim}(\cdot)$ denotes the cosine similarity. The value a is used to prevent the features of anchor samples Anc_i^m, positive samples Pos_i^m and negative samples Neg_i^m from aggregating into a small space. The whole training framework is shown in Figure 3.

4.3 Textual Semantic Incorporation

POIs hold significance as data points denoting specific landmarks on a map, often signifying distinct geographic locations such as stores, restaurants, parks, and more in cities. The textual descriptions associated with POIs can effectively capture the geographic attributes of a region. For instance, a clustering of coffee shops within a region could indicate a vibrant locale appealing to young residents, whereas an abundance of parks and green spaces might suggest a neighborhood conducive to family-oriented living.

POI Textual Semantic Extraction

In addition to the imagery features, the textual data associated with POIs plays a crucial role in region profiling. To effectively harness the descriptive potential of POI text for region representation, we employ Gensim in conjunction with Skip-Gram and Huffman Softmax models [Mikolov *et al.*, 2013] for training. The Skip-Gram model, a neural network-based

word vector approach, enables the learning of word vectors by predicting the contextual information of a word. Concurrently, the Huffman Softmax model, which leverages Huffman trees, enhances the neural network's output layer, refining the overall representation process.

Considering the complexity and ambiguity inherent in POI comments, we adopt a two-phase approach to extract the textual semantics. In the training phase, we utilize all POI comments and categories to train the model. However, as we transition to the representation phase, our focus narrows to utilizing solely the categorical information associated with each POI within the target regions. Assuming that for region $r_i \in \mathcal{R}$, its POI data $\mathcal{T}_i = \{T_{i1}, T_{i2}, \cdots, T_{i|\mathcal{T}_i|}\}$ is the category of each POI in the region, and the final mapping of the trained model from words to vectors is W. Then the final POI embedding result for each region is:

$$e_i^{\text{POI}} = \frac{1}{|\mathcal{T}_i|} \sum_{j=1}^{|\mathcal{T}_i|} W(T_{ij}), T_{ij} \in \mathcal{T}_i \tag{7}$$

Attentive Fusion Module

We proceed to integrate the street view features e_i^{SV}, remote sensing features e_i^{RV}, and POI features e_i^{POI}, creating a comprehensive final representation tailored for utilization in various downstream tasks.

Firstly, with the inherent importance of both imagery features e_i^{SV} and e_i^{RV} unknown, we propose the incorporation of an attentive fusion module to derive weights for each of these representations. Considering street view features e_i^{SV} and remote sensing features e_i^{RV} from region r_i, we introduce learnable parameters $\mathbf{c}$, $\mathbf{M}$, and $\mathbf{b}$ to facilitate their fusion:

$$\alpha_i^m = \mathbf{c}^T \cdot \text{ReLU}(\mathbf{M} \cdot e_i^m + \mathbf{b}), m \in \{\text{SV}, \text{RV}\} \tag{8}$$

$$\beta_i^m = \frac{\exp(\alpha_i^m)}{\sum_{m \in \{\text{SV}, \text{RV}\}} \exp(\alpha_i^m)} \tag{9}$$

$$e_i^{\text{Image}} = \sum_{m \in \{\text{SV}, \text{RV}\}} \beta_i^m \cdot e_i^m \tag{10}$$

where e_i^{Image} is the final representation for region's imagery feature and β_i^m ($m \in \{\text{SV}, \text{RV}\}$) are weight coefficients.

Next, in order to incorporate the textual semantic information of POIs, we refer to InfoNCE loss [Oord *et al.*, 2018] to align the features of imagery e_i^{Image} and POI texts e_i^{POI}:

$$\text{Loss}_i = -\log \frac{\exp(\text{sim}(e_i^{\text{Image}}, e_i^{\text{POI}}))}{\sum_{j=1}^n \exp(\text{sim}(e_i^{\text{Image}}, e_j^{\text{POI}}))} \tag{11}$$

where n denotes the mini-batch size. By optimizing the aforementioned loss function, we acquire the region imagery features e_i^{Image} and effectively integrate the semantic information of POIs. Subsequently, these obtained representations for each region can be harnessed to forecast various socioeconomic indicators.

5 Experiments

5.1 Experimental Setups

Datasets

We compile real-world datasets from three major cities: Beijing (BJ), Shanghai (SH), and New York (NY). The city regions are delineated by hexagonal divisions, with a radius of 1 km for Beijing and Shanghai, and 500 meters for New York (New York is much smaller than Beijing and Shanghai). It should be noted that our model is highly adaptable to various division shapes and scales, including road networks and Census Block Groups (CBGs).

For street view imagery, we employ the Baidu Maps API[1] for Beijing and Shanghai, and the Google Maps API[2] for New York. High-resolution (3.6-meter) remote sensing images are acquired through ArcGIS for all three cities. The POI data for Beijing and Shanghai originates from Baidu Maps, while New York's data is sourced from OpenStreetMap[3] (OSM). Socioeconomic indicators, including population density from WorldPop[4], housing data from Lianjia[5], and crime data from NYC Open Data[6], are also integrated.

Baseline Models

We compare our proposed model with various unimodal and multi-modal region representation algorithms, including:

- **Inception v3** proposed in [Szegedy *et al.*, 2016]. It can extract features using convolutional layers with different kernel sizes, max pooling and batch normalization.

- **Resnet-18** proposed in [He *et al.*, 2016]. It uses residual blocks to solve the degeneracy problem of deep networks. We use the Resnet-18 pre-trained in ImageNet.

- **Tile2vec** proposed in [Jean *et al.*, 2019]. It is an unsupervised learning method that uses geographic distance as a criterion for constructing contrastive samples.

- **Urban2vec** proposed in [Wang *et al.*, 2020]. It uses both street view images and POI data to characterize neighborhood features.

- **PG-SimCL** proposed in [Xi *et al.*, 2022]. It uses remote sensing imagery for region representation based on the similarity of geographic distances and POI distributions to perform prediction tasks on socioeconomic indicators.

- **Add-svrv** and **Fusion-svrv**. They represent the summation or attentional fusion of SV and RV embedding.

- **Concat**. We simply concatenate the SV, RV and POI representation results as a variant of our method.

Metrics and Implementation

We adopt rooted mean squared error ($RMSE$) and coefficient of determination (R^2) for evaluation. In our experiments, we use Inception v3 as SV encoder's backbone

---

City	Beijing									
	PD		**HD**		**MC**		**NP**		**NC**	
Methods	$R^2 \uparrow$	$RMSE \downarrow$	$R^2 \uparrow$	$RMSE \downarrow$	$R^2 \uparrow$	$RMSE \downarrow$	$R^2 \uparrow$	$RMSE \downarrow$	$R^2 \uparrow$	$RMSE \downarrow$
Inception v3	0.0023	1.4765	-0.0211	1.0325	0.0051	2.2586	0.0065	1.6991	-0.0077	2.5665
Resnet-18	0.0685	1.4266	-0.0235	1.0337	0.0344	2.2251	0.0356	1.6740	0.0263	2.5229
Tile2vec	0.1102	1.3944	-0.0071	1.0254	0.1104	2.1357	0.1137	1.6048	0.0951	2.4320
Urban2vec	0.4982	1.0471	0.5327	0.6985	<u>0.5943</u>	<u>1.4422</u>	<u>0.7955</u>	<u>0.7708</u>	0.7251	1.3406
PG-SimCL	0.1425	1.3688	0.1119	0.9630	0.1636	2.0708	0.1628	1.5597	0.1332	2.3804
Add-svrv	0.0995	1.4027	0.0984	0.9702	0.0906	2.1593	0.1290	1.5909	0.1402	2.3707
Fusion-svrv	0.1408	1.3701	0.1145	0.9615	0.1716	2.0609	0.1577	1.5645	0.1322	2.3817
Concat	<u>0.4984</u>	<u>1.0469</u>	<u>0.5399</u>	<u>0.6931</u>	0.5738	1.4783	0.7832	0.7937	<u>0.7276</u>	<u>1.3345</u>
Ours	**0.5310**	**1.0123**	**0.5708**	**0.6694**	**0.6229**	**1.3906**	**0.9471**	**0.3921**	**0.8782**	**0.8921**
Impr.	6.54%	3.30%	5.72%	3.42%	4.81%	3.58%	19.06%	49.13%	20.70%	33.15%

Table 1: Prediction results of different socioeconomic indicators for Beijing: Population Density (**PD**), Housing Density (**HD**), Mobility Count (**MC**), Number of POIs (**NP**) and Number of Comments (**NC**). The best results are **in bold** and the second best results are <u>underlined</u>.

(a) Population Density   (b) Housing Density   (c) Mobility Count   (d) Number of POIs   (e) Number of Comments
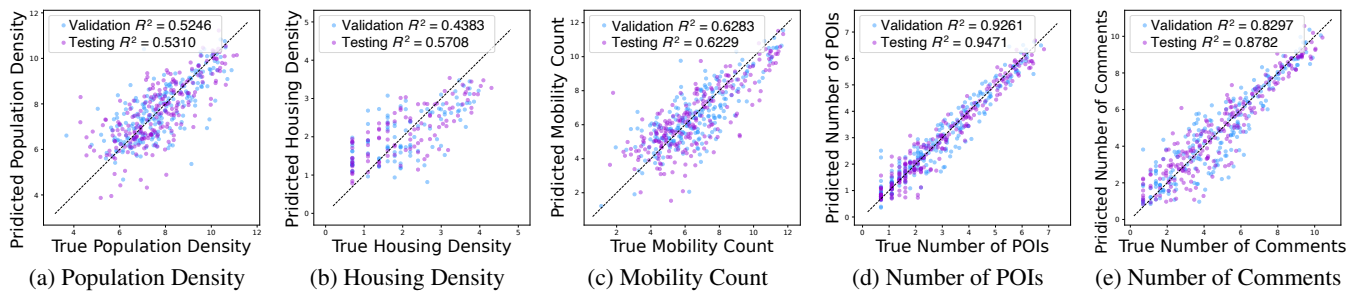
Figure 4: Predicted values versus true values on five socioeconomic datasets in Beijing. The dotted line is 45 degrees. The blue dots represent the regions of the validation set and the purple dots represent the regions of the test set with the respective R^2 illustrated.

and Resnet-18 as RV encoder's backbone with a final linear layer projecting features into the 128-dimensional embedding space. We set the batch size to 32, the learning rate to $5e^{-4}$, and use Adam optimizer. Datasets are split into 60% training, 20% validation, and 20% test sets. All socioeconomic indicators are converted into logarithmic scale.

5.2 Experimental Results

Socioeconomic Indicators Prediction
We utilize the embedding of each region as input and employ a multi-layer perceptron (MLP) to predict the socioeconomic indicators. Table 1 shows the prediction outcomes for Beijing, demonstrating a significant R^2 improvement of 4.81% to 20.70% over the best baselines across the five datasets.

Specifically, the **Inception v3** and **Resnet-18** networks pre-trained on ImageNet achieve average R^2 scores of only -0.0030 and 0.0283 across datasets, failing to adequately capture the relationships between regions. While **Tile2vec** and **PG-SimCL** show some improvements over pre-trained models with an average R^2 of 0.0845 and 0.1428, they both fall short in providing a comprehensive representation solely through remote sensing imagery. Then, **Urban2vec**, a multimodal approach, outperforms other unimodal models with an average R^2 of 0.6292 exhibiting enhanced prediction results by incorporating street view images and POI data. A comparison of **Add-svrv** and **Fusion-svrv** reveals that the use of attentional fusion module is more effective than simple summa-

tion when visual semantic is utilized. Our **MuseCL** notably excels in all Beijing datasets with an average R^2 of 0.7100, which is 13.67% higher than **Concat**, highlighting successful multi-semantic integration across visual and textual modalities. This also demonstrates the effectiveness of MuseCL in representing regional attributes, leading to more precise predictions of socioeconomic indicators. Furthermore, Figure 4 shows the predicted value v.s. true value on socioeconomic indicators for Beijing, indicating that our MuseCL framework shows superior prediction effect on different indicators.

Model Adaptability to Other Cities
We expand our experimental scope to include other welldeveloped cities, thereby testing the adaptability of our model. Shanghai has 746 valid representation regions, while New York has 517. In Shanghai, our experiments cover Population Density (**PD**), Housing Density (**HD**), and Number of POIs (**NP**) indicators. For New York, we incorporate the widely recognized Crime (**CR**) dataset. Consistency in comparisons is maintained by employing the same baseline models as before. The predictive outcomes are presented in Table 2. Notably, our model consistently outperforms across cities with varying sizes and geographical characteristics. Compared to the next-best baselines, we achieve an improvement in R^2 ranging from 3.77% to 19.61%. This robust performance reaffirms the adaptability of our model in addressing the demands of different city types and diverse datasets.

City	Dataset	Metrics	Inception v3	Resnet-18	Tile2vec	Urban2vec	PG-SimCL	Concat	**Ours**	Impr.
Shanghai	**PD**	$R^2 \uparrow$	-0.0384	-0.2813	0.0016	0.3401	0.0261	<u>0.3596</u>	**0.4301**	19.61%
		$RMSE \downarrow$	1.2269	1.3629	1.2031	0.9780	1.1882	<u>0.9635</u>	**0.9089**	5.67%
	HD	$R^2 \uparrow$	-0.0962	-0.0889	0.0424	<u>0.4061</u>	-0.0245	0.3763	**0.4330**	6.62%
		$RMSE \downarrow$	1.0586	1.0551	0.9895	<u>0.7792</u>	1.0234	0.7985	**0.7614**	2.28%
	NP	$R^2 \uparrow$	-0.0069	-0.0540	0.0677	<u>0.8726</u>	0.0902	0.8191	**0.9283**	6.38%
		$RMSE \downarrow$	1.5706	1.6069	1.5113	<u>0.5586</u>	1.4930	0.6657	**0.4191**	24.97%
New York	**PD**	$R^2 \uparrow$	-0.0063	-0.0042	0.0052	<u>0.3551</u>	0.2735	0.3436	**0.4165**	17.29%
		$RMSE \downarrow$	2.1691	2.1669	2.1568	<u>1.7365</u>	1.8431	1.7519	**1.6517**	4.88%
	CR	$R^2 \uparrow$	-0.0001	-0.0146	-0.0442	<u>0.3183</u>	0.1634	0.2979	**0.3303**	3.77%
		$RMSE \downarrow$	1.5713	1.5827	1.6056	<u>1.2973</u>	1.4371	1.3166	**1.2858**	0.89%
	NP	$R^2 \uparrow$	-0.0416	-0.0531	-0.0117	<u>0.2397</u>	0.2102	0.2395	**0.2594**	8.22%
		$RMSE \downarrow$	1.3963	1.4040	1.3761	<u>1.1929</u>	1.2159	1.1931	**1.1774**	1.30%

Table 2: Prediction results of different socioeconomic indicators for Shanghai and New York: Population Density (**PD**), Housing Density (**HD**), Number of POIs (**NP**) and Crime (**CR**). The best results are **in bold** and the second best results are <u>underlined</u>.
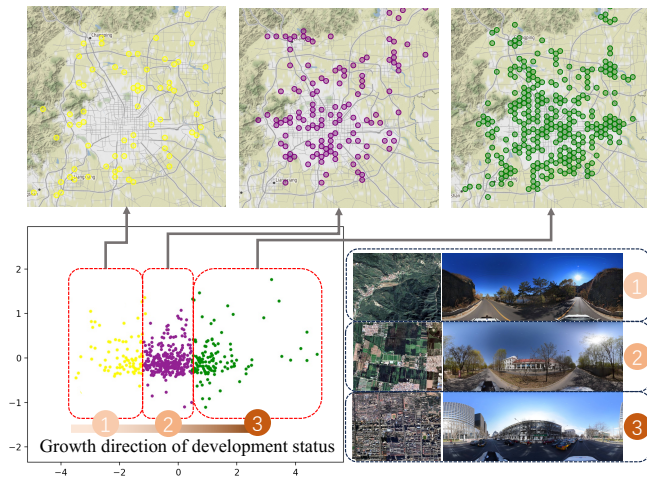

Figure 5: Visualization of the final representation space.

(a) Beijing

(b) New York

Figure 6: Results of ablation study in Beijing and New York.

Visualization of Region Representations

To gain deeper insights into the embedding space, we apply principal components analysis (PCA) [Shlens, 2014] to downscale the representation vectors. We then use the K-means algorithm to classify the regions into three clusters and depict their spatial distribution in Figure 5. This visualization reveals that regions with different levels of development occupy distinct locations in the embedded space.

Specifically, the yellow regions are situated on the outskirts of Beijing, indicating underdevelopment and low socioeconomic attributes. Their remote sensing and street view images depict agricultural landscapes with sparse populations and limited POIs. In addition, purple regions, which are moderately developed, extend across urban and suburban areas, revealing emerging villages in their imagery. These areas maintain non-built spaces but exhibit higher population densities and POI counts compared to the yellow ones. Meanwhile, green regions, in central urban zones, include Beijing's commercial hubs, showing a highly urbanized environment with strong socioeconomic indicators.
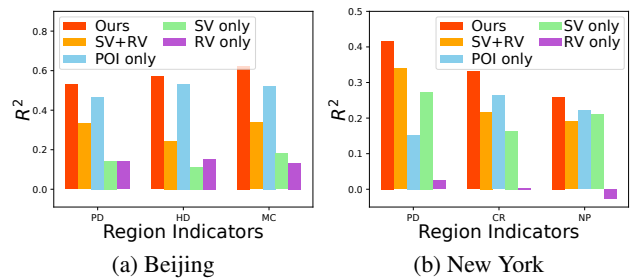
Ablation Study

We conduct ablation experiments using three datasets each from Beijing and New York City. The results, as depicted in Figure 6, indicate that the absence of certain modalities leads to a reduction in the final prediction R^2 value. Notably, relying solely on POI, street view, or remote sensing images yields suboptimal outcomes. When combining street view and remote sensing images without POI information, the performance still falls short of our model's performance, although it fares better than utilizing street view or remote sensing images individually. This reinforces the notion that various modalities contribute distinct insights for predicting downstream tasks and urban region profiling.

6 Conclusion

This paper presents a novel Multi-Semantic Contrastive Learning (MuseCL) framework that skillfully amalgamates semantic insights from visual and textual information to generate embeddings for urban regions. We showcase our model's superiority in socioeconomic indicators prediction across diverse cities and through extended experiments. While our focus is on statically depicting urban regions, it is important to acknowledge their rapid evolution due to development. Therefore, incorporating time into region representation presents an interesting path for future research.

Acknowledgments

References

[Aghababaei and Makrehchi, 2016] Somayyeh Aghababaei and Masoud Makrehchi. Mining social media content for crime prediction. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 526–531. IEEE, 2016.

[Antenucci *et al.*, 2014] Dolan Antenucci, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D Shapiro. Using social media to measure labor market flows. Technical report, National Bureau of Economic Research, 2014.

[Burke *et al.*, 2021] Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021.

[Chakraborty *et al.*, 2016] Sunandan Chakraborty, Ashwin Venkataraman, Srikanth Jagabathula, and Lakshminarayanan Subramanian. Predicting socio-economic indicators using news events. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1455–1464, 2016.

[Chen *et al.*, 2020] Long Chen, Yi Lu, Qiang Sheng, Yu Ye, Ruoyu Wang, and Ye Liu. Estimating pedestrian volume using street view images: A large-scale validation test. *Computers, Environment and Urban Systems*, 81:101481, 2020.

[Cohen *et al.*, 2016] Peter Cohen, Robert Hahn, Jonathan Hall, Steven Levitt, and Robert Metcalfe. Using big data to estimate consumer surplus: The case of uber. Technical report, National Bureau of Economic Research, 2016.

[Custodio *et al.*, 2023] Henry M Custodio, Michalis Hadjikakou, and Brett A Bryan. A review of socioeconomic indicators of sustainability and wellbeing building on the social foundations framework. *Ecological Economics*, 203:107608, 2023.

[Feng *et al.*, 2017] Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee. Poi2vec: Geographical latent representation for predicting future visitors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[Fu *et al.*, 2019] Yanjie Fu, Pengyang Wang, Jiadi Du, Le Wu, and Xiaolin Li. Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 906–913, 2019.

[Gao *et al.*, 2017] Qiang Gao, Fan Zhou, Kunpeng Zhang, Goce Trajcevski, Xucheng Luo, and Fengli Zhang. Identifying human mobility via trajectory embeddings. In *IJCAI*, volume 17, pages 1689–1695, 2017.

[Habitat, 2022] UN Habitat. World cities report 2022: Envisaging the future of cities. *United Nations Human Settlements Programme: Nairobi, Kenya*, pages 41–44, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2018] Zhiyuan He, Su Yang, Weishan Zhang, and Jiulong Zhang. Perceiving commerial activeness over satellite images. In *Companion Proceedings of the The Web Conference 2018*, pages 387–394, 2018.

[Hong *et al.*, 2020] Danfeng Hong, Lianru Gao, Jing Yao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. Graph convolutional networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5966–5978, 2020.

[Huang *et al.*, 2021] Tianyuan Huang, Zhecheng Wang, Hao Sheng, Andrew Y Ng, and Ram Rajagopal. M3g: Learning urban neighborhood representation from multi-modal multi-graph. In *Proceedings of the DeepSpatial 2021: 2nd ACM KDD Workshop on Deep Learning for Spatio-Temporal Data, Applications and Systems*, 2021.

[Ilieva and McPhearson, 2018] Rositsa T Ilieva and Timon McPhearson. Social-media data for urban sustainability. *Nature Sustainability*, 1(10):553–565, 2018.

[Jean *et al.*, 2019] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019.

[Li *et al.*, 2022] Tong Li, Shiduo Xin, Yanxin Xi, Sasu Tarkoma, Pan Hui, and Yong Li. Predicting multi-level socioeconomic indicators from structural urban imagery. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3282–3291, 2022.

[Liu *et al.*, 2023] Yu Liu, Xin Zhang, Jingtao Ding, Yanxin Xi, and Yong Li. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. In *Proceedings of the ACM Web Conference 2023*, pages 4150–4160, 2023.

[Luo *et al.*, 2022] Yan Luo, Fu-lai Chung, and Kai Chen. Urban region profiling via multi-graph representation learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4294–4298, 2022.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of

word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Miller, 2004] Harvey J Miller. Tobler's first law and spatial analysis. *Annals of the association of American geographers*, 94(2):284–289, 2004.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Qu *et al.*, 2017] Meng Qu, Jian Tang, Jingbo Shang, Xiang Ren, Ming Zhang, and Jiawei Han. An attention-based collaboration framework for multi-view network representation learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1767–1776, 2017.

[Sachs *et al.*, 2022] Jeffrey D Sachs, Christian Kroll, Guillame Lafortune, Grayson Fuller, and Finn Woelm. *Sustainable development report 2022*. Cambridge University Press, 2022.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[Shlens, 2014] Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[Wang and Li, 2017] Hongjian Wang and Zhenhui Li. Region representation learning via mobility flow. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 237–246, 2017.

[Wang *et al.*, 2018a] Wenshan Wang, Su Yang, Zhiyuan He, Minjie Wang, Jiulong Zhang, and Weishan Zhang. Urban perception of commercial activeness from satellite images and streetscapes. In *Companion Proceedings of the The Web Conference 2018*, pages 647–654, 2018.

[Wang *et al.*, 2018b] Yingzi Wang, Xiao Zhou, Cecilia Mascolo, Anastasios Noulas, Xing Xie, and Qi Liu. Predicting the spatio-temporal evolution of chronic diseases in population with human mobility data. In *IJCAI*, 2018.

[Wang *et al.*, 2020] Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1013–1020, 2020.

[Wu *et al.*, 2022] Shangbin Wu, Xu Yan, Xiaoliang Fan, Shirui Pan, Shichao Zhu, Chuanpan Zheng, Ming Cheng, and Cheng Wang. Multi-graph fusion networks for urban region embedding. *arXiv preprint arXiv:2201.09760*, 2022.

[Xi *et al.*, 2022] Yanxin Xi, Tong Li, Huandong Wang, Yong Li, Sasu Tarkoma, and Pan Hui. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests. In *Proceedings of the ACM Web Conference 2022*, pages 3308–3316, 2022.

[Yao *et al.*, 2018] Zijun Yao, Yanjie Fu, Bin Liu, Wangsu Hu, and Hui Xiong. Representing urban functions through zone embedding with human mobility patterns. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018.

[Yuan *et al.*, 2012] Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194, 2012.

[Zhang *et al.*, 2017] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *Proceedings of the 26th International Conference on World Wide Web*, pages 361–370, 2017.

[Zhang *et al.*, 2021] Mingyang Zhang, Tong Li, Yong Li, and Pan Hui. Multi-view joint graph representation learning for urban region embedding. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4431–4437, 2021.

[Zhang *et al.*, 2024] Yuyao Zhang, Ke Guo, and Xiao Zhou. Causally aware generative adversarial networks for light pollution control. *arXiv preprint arXiv:2401.06453*, 2024.

[Zheng *et al.*, 2014] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):1–55, 2014.

[Zhou *et al.*, 2017] Xiao Zhou, Desislava Hristova, Anastasios Noulas, Cecilia Mascolo, and Max Sklar. Cultural investment and urban socio-economic development: a geosocial network approach. *Royal Society open science*, 4(9):170413, 2017.

[Zhou *et al.*, 2018] Xiao Zhou, Anastasios Noulas, Cecilia Mascolo, and Zhongxiang Zhao. Discovering latent patterns of urban cultural interactions in wechat for modern city planning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1069–1078, 2018.

[Zhou *et al.*, 2019] Xiao Zhou, Cecilia Mascolo, and Zhongxiang Zhao. Topic-enhanced memory networks for personalised point-of-interest recommendation. In *Proceedings of the 25th ACM SIGKDD International conference on knowledge discovery & data mining*, pages 3018–3028, 2019.

[Zhou *et al.*, 2023] Xiao Zhou, Xiaohu Zhang, Paolo Santi, and Carlo Ratti. Phase-wise evaluation and optimization of non-pharmaceutical interventions to contain the covid-19 pandemic in the us. *Frontiers in Public Health*, 11:1198973, 2023.