Beyond the First Law of Geography: Learning Representations of Satellite Imagery by Leveraging Point-of-Interests

Yanxin Xi
University of Helsinki
Finland
yanxin.xi@helsinki.fi

Tong Li§
Tsinghua University, China
University of Helsinki, Finland
tongli@mail.tsinghua.edu.cn

Huandong Wang
Tsinghua University
China
wanghuandong@tsinghua.edu.cn

Yong Li
Tsinghua University
China
liyong07@tsinghua.edu.cn

Sasu Tarkoma
University of Helsinki
Finland
sasu.tarkoma@cs.helsinki.fi

Pan Hui
Hong Kong University of Science and Technology, Hong Kong
panhui@cse.ust.hk

ABSTRACT

Satellite imagery depicts the earth's surface remotely and provides comprehensive information for many applications, such as land use monitoring and urban planning. Existing studies on unsupervised representation learning for satellite images only take into account the images' geographic information, ignoring human activity factors. To bridge this gap, we propose using Point-of-Interest (POI) data to capture human factors and design a contrastive learning-based framework to consolidate the representation of satellite imagery with POI information. Also, we design an attention model that merges the representations from the geographic and POI perspectives adaptively. On the basis of real-world datasets collected from Beijing, we evaluate our method for predicting socioeconomic indicators. The results show that the representation containing POI information outperforms the geographic representation in estimating commercial activity-related indicators. Our proposed framework can estimate the socioeconomic indicators with an R^2 of 0.874 and outperforms the baseline methods.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**.

KEYWORDS

Representation learning, socioeconomic indicator prediction, satellite imagery, data mining

ACM Reference Format:
Yanxin Xi, Tong Li§, Huandong Wang, Yong Li, Sasu Tarkoma, and Pan Hui. 2022. Beyond the First Law of Geography: Learning Representations of Satellite Imagery by Leveraging Point-of-Interests. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3485447.3512149

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512149

1 INTRODUCTION

Satellite images acquired through ubiquitous remote sensing technology depict the Earth's surface from a bird's eye perspective, providing comprehensive data for a variety of applications, ranging from land cover monitoring [2, 20] to socioeconomic status inference [1, 7]. In comparison to traditional data sources, such as field surveys, satellite imagery is collected in a more time and cost-efficient manner. Thus providing enormous potential for timely monitoring of land cover and human activities on a large scale.

Combined with current advances in computer vision and deep learning, studies have successfully leveraged satellite imagery to classify land cover [17], predict commercial activeness [15, 31], estimate population [7], and infer economic status [34]. These achievements were made by task-specific supervised learning, which requires massive labeled data for training. However, in most remote sensing applications, obtaining a large volume of high-quality annotated data is extremely difficult [20, 32]. In general, remote sensing delivers large amounts of data, like satellite imagery; but a lack of labels makes many downstream applications difficult to implement.

To reduce the need for labeled data, the research community has turned to unsupervised representation learning for satellite images [12, 13, 20]. The task of such representation learning is to find a low-dimensional representation of a satellite image while persevering associations between objects. Such learning does require labeled data. The learned representations are multipurpose and can be used for different downstream tasks [4]. Similar to word embeddings in natural language processing (NLP) [25, 26], the critical issue in learning satellite imagery representations is to define a similarity/association metric between satellite images and encode them into quantitative representations. The majority of existing studies [20, 22, 32] built their similarity metrics by following Tobler's First Law of Geography [27], which states that 'everything is related to everything else, but near things are more related than distant things.' In other words, geographically adjacent satellite images are more likely to have similar meanings and thus representations. In uninhabited regions, this law works well. However, the law does have deficiencies in human-inhabited areas that have been substantially altered by human activity. Two human-inhabited regions might have different land uses and configurations, even though they are geographic neighbors. To bridge this gap, Han *et*

§Corresponding Author.

al. [12] incorporated human efforts into their representation learning framework by having experts annotate a small set of satellite images and then distilling knowledge from this annotated dataset. Notably, such a method is still time-consuming and, to some extent, limits the performance.

In this paper, we aim to learn compressed yet informative representations for unlabeled satellite imagery. Specifically, there are three challenges to achieving this objective. 1). As previously discussed, relying solely on the geographic features of satellite imagery has some limitations. Human activity is more important in describing the surrounding areas in urban regions. As a result, the first challenge is how to depict and capture information about human activity factors. 2). After capturing human factors, the second challenge is how to develop a framework for introducing human factor features into the satellite imagery representation learning process. 3). Apart from human factors, geographic information contained in satellite images is also useful for learning representations. As a result, the third challenge is to figure out how to combine the features of these two aspects in an auto-adaptive manner to produce better final representations.

To address these issues, we propose a representation learning framework for unlabeled satellite imagery based on the three key designs. 1). We use the POI data to capture the nature of human activity factors. POIs, which show the locations related to human activity in populated areas, can reflect the characteristics of human activity in a given region. For example, if a region has a lot of restaurant POIs, this suggests that there are many people dining there, and if a region has education POIs, this suggests that people come to the region primarily to study. 2). A contrastive learning module is designed to take advantage of POIs that support learning representations of unlabeled satellite images. In particular, we measure the similarity across all satellite images in the POI domain and learn common features between similar instances and distinguishing features between dissimilar instances. By doing so, we can extract more representative features from satellite images that reflect human activity. 3). We design an attentional fusion model that can adaptively fuse representations from two different perspectives: human and geographic factors. Thus the importance of different perspectives in predicting socioeconomic indicators can be automatically learned.

The contribution of our work is summarized as follows.

- We introduce POI data to capture the characteristics of human activity. Such data is an important supplement to existing assumptions (e.g., First Law of Geography) that enable unsupervised representation learning for satellite imagery.
- We propose a framework based on contrastive learning. Our framework learns informative representations of satellite imagery containing human activity factors by maximizing the similarity of the representations for satellite images with similar POI features.
- We develop an attentional fusion model to adaptively fuse representations related to human activity and geographic factors, enhancing the adaptability of learned representations across multiple tasks such as population prediction and commercial activeness prediction.
- We evaluate our proposed framework on real-world large-scale datasets. In the task of predicting socioeconomic indicators, our method outperforms the baselines by 4.3% in terms of R^2. We also demonstrate the transferability of our model across different cities.

We envision new Web-based services enabled by machine learning and data fusion to provide insights regarding socioeconomic indicators and human behavior. To this end, our work advances Web technologies by contributing a new machine learning method for combining Web-accessible data, such as POIs and satellite imagery.

2 PRELIMINARIES AND FRAMEWORK OVERVIEW

2.1 Data Overview

Satellite Imagery. Satellite imagery captures images of the earth's surface by using space satellites. It can reflect near real-time information about the ground surface. In practice, Environmental Systems Research Institute (Esri)[1] provides visible-band spectral satellite imagery of various resolutions.

Point-of-Interests. Point of interest (POI) data represents various venues in human-inhabited areas, like shopping malls and theaters. Existing studies [11, 35] have found that human activity has a strong link with POIs. Inspired by these studies, we incorporate POI information into the representation learning for satellite imagery in our study. In practice, we crawl 1,481,100 POIs from Tencent Map Service covering the entire area of Beijing with 14 POI categories.

Socioeconomic Indicators. The socio-economic development status of human-inhabited areas is an important characteristic that can be measured using socioeconomic indicators. Previous research [1, 13, 34] has shown that satellite images can be used to estimate a region's socio-economic development. Thus, we use learned satellite imagery representations to estimate multiple socioeconomic indicators to assess performance. Specifically, we use the number of takeaway orders and online comments on commercial entities as ground-truth indicators of commercial activeness. On the other hand, we use population and population density as ground-truth social indicators.

• *Number of takeaway orders.* The dataset covers the takeaway order records of over 25,000 restaurants in Beijing. The data was gathered from Meituan, China's most popular online shopping platform for local consumer goods and retail services.

• *Number of comments.* The online comments were gathered from Dianping, a popular Chinese platform for restaurant reviews. The collected dataset contains the comments of around 140,000 commercial entities in Beijing.

• *Population.* The WorldPop organization[2] provided us with Beijing population statistics for 2020.

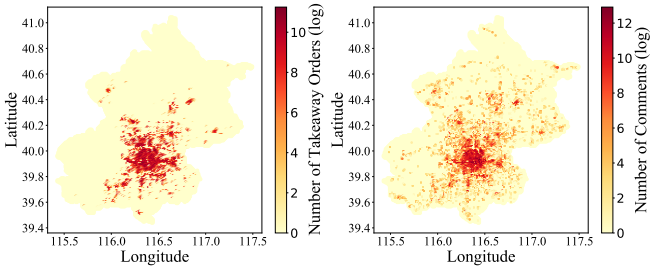
• *Population Density.* The density of people in a given area is referred to as population density. In our case, the population density data was collected in 2020 by the WorldPop organization.

Figure 1 shows the geographic distribution of the aforementioned indicators. We can see from the visualizations that the geographic

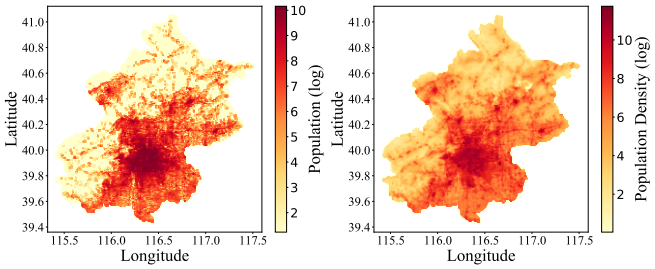
[1] https://www.esri.com/en-us/home.
[2] https://www.worldpop.org/.

Table 1: Correlation between POI categories and socioeconomic indicators.

POI category	Population	Ranking	Population density	Ranking	Number of comments	Ranking	Number of takeaway orders	Ranking
Transportation facility	0.4654	7	0.3828	7	0.2353	10	0.2571	12
Leisure & Sport	0.5324	3	0.4572	4	0.5307	1	0.4521	4
Residence	0.4929	5	0.4221	6	0.1440	11	0.3751	6
Company	0.4358	8	0.3682	8	0.4055	3	0.3532	7
Medical service	0.3743	10	0.3368	9	0.2392	9	0.3000	9
Factory & Agriculture	0.2758	13	0.1910	13	0.0365	14	0.1009	13
Government & Organization	0.5497	2	0.5253	1	0.2946	7	0.4301	5
Education	0.4828	6	0.4740	3	0.3402	6	0.4684	3
Scenic spot	0.0951	14	0.0871	14	0.0453	13	0.0368	14
Automobile service	0.3976	9	0.3056	12	0.1420	12	0.2608	11
Life service	0.5907	1	0.5029	2	0.4040	4	0.5048	1
Shopping mall	0.3481	12	0.3085	10	0.3675	5	0.2677	10
Hotel	0.3532	11	0.3058	11	0.2873	8	0.3070	8
Restaurant	0.5223	4	0.4315	5	0.5282	2	0.4753	2



(a) Number of Takeaway Orders (log). (b) Number of Comments (log).



(c) Population (log). (d) Population Density (log).

Figure 1: Geographic distribution of socioeconomic indicators in Beijing.

distributions of socioeconomic indicators do not follow Tobler's First Law of Geography very well. The distributions of takeaway orders and online comments, in particular, do not appear to be geographically consistent. As a result, relying solely on geographic autocorrelation to model socioeconomic indicators is insufficient.

2.2 Motivation

Learning representations for satellite imagery based solely on geographic information may have the drawback of ignoring human factors, resulting in poor performance in predicting socioeconomic indicators related to human activity. This motivates us to leverage POI data to represent the human factor for satellite imagery

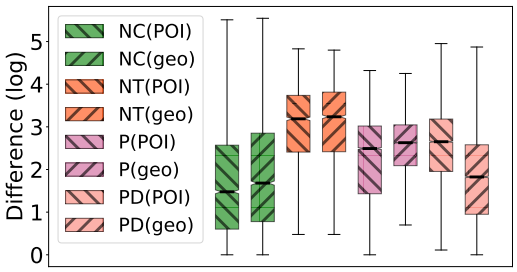


Figure 2: Differences in socioeconomic indicators between two satellite images with similar POI distributions or that are geographically adjacent. ('geo': 'Geographic', 'NC': Number of Comments, 'NT': Number of Takeaway Orders, 'P': Population Count, 'PD': Population Density)

representation learning. We begin by conducting a preliminary investigation into the correlations between the number of POI categories and the socioeconomic data in each satellite image. Table 1 displays the results, with the top three POI categories for each indicator colored in blue from darkest to lightest. We can observe that the POI data, particularly some POI categories such as life service, shopping, and entertainment, is highly correlated with socioeconomic indicators.

We also explore the difference of a single socioeconomic indicator in two geographically adjacent satellite images and in two satellite images with the most similar POI distribution. The distributions of differences are shown in Figure 2 with box plots. We can deduce that the differences in the number of takeaway orders, number of comments, and population count between two satellite images with similar POI distribution are smaller than the differences between two geographically adjacent images. This reinforces the importance of incorporating POI data into the representation of satellite imagery to reflect human activity factors.

2.3 Problem Statement and Framework Overview

We next formally define the representation learning problem for the unlabeled satellite imagery. Given a set of unlabeled satellite

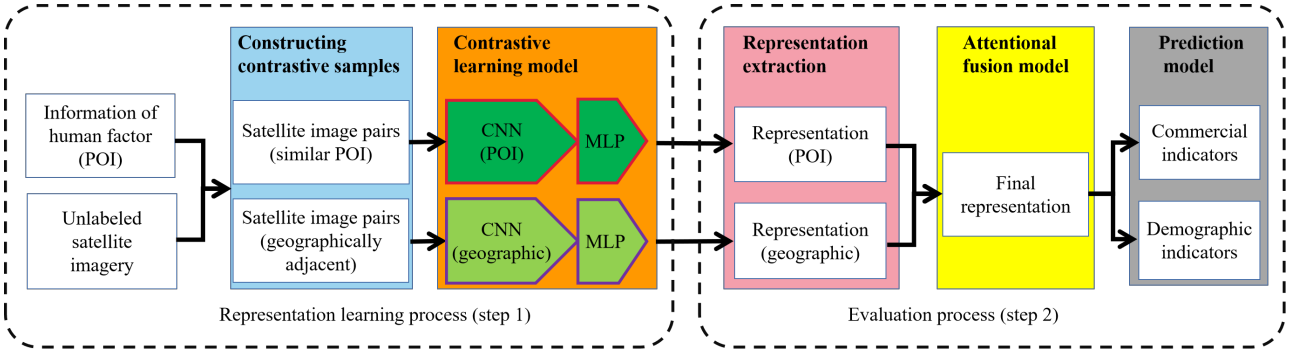


Figure 3: Framework Overview.

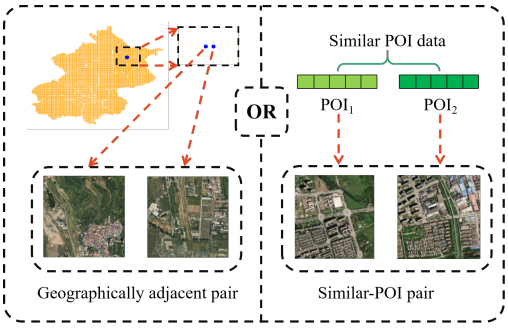


Figure 4: Architecture of constructing contrastive samples.

images **I** and corresponding POIs that fall within the scope of the satellite images, our objective is to learn a representation $\mathbf{r}_i$ for each satellite image I_i through an unsupervised model F, where $\mathbf{r}_i = F(I_i)$.

We present an overview of our proposed framework in Figure 3. There are two principal steps in our framework: the representation learning process and the evaluation process. To be specific, according to the POI information or the coordinates of the satellite images, we construct contrastive samples for satellite images with similar POI distribution and for geographically adjacent images, respectively. Based on the samples, we train two different contrastive learning models for imagery representations with POI information and with geographic information. After that, we design an attentional fusion model to adaptively combine the representations from two different perspectives. With the socioeconomic indicators, we evaluate the effectiveness of our proposed method.

3 METHOD

We formally present the design of our proposed framework in this section, as well as key modules including constructing contrastive samples, contrastive learning model, and attentional fusion model.

3.1 Constructing Contrastive Samples

Contrastive samples are the image pairs that are used to train the contrastive learning model. In this phase, we find a contrastive partner (thus creating a pair) for every satellite image in the POI view by selecting the satellite image with the most similar POI distribution. Assuming there are overall K POI categories in the dataset, for each satellite image I_i, we can define a K-dimensional POI vector $POI_i = [POI_{i1}, POI_{i2}, ..., POI_{iK}]$, with each dimension $POI_{ik} (k = 1, 2, \cdots, K)$ representing the number of k-th POI category. Then, to determine the satellite image with the most similar POI distribution, we use the Euclidean distance between POI vectors of two satellite images I_i and I_j as follows:

$$dist_{i,j}^{POI} = \sqrt{\left(\sum_{k=1}^{K} \left(POI_{ik} - POI_{jk}\right)^2\right)}. \quad (1)$$

Those images with the smallest distance are considered to be the most similar satellite image partners in POI view. If there are multiple satellite images that have the same POI distance to a certain image, we select one of them at random as the most similar.

We also select for each satellite image the geographically most adjacent image, which is the satellite image with the minimum geographic distance calculated from the geographic coordinates. After this step, a satellite image has a geographically adjacent pair and a most similar pair in POI view.

3.2 Contrastive Learning Model

A contrastive learning model learns from the contrastive samples to make a compressed representation of unlabeled satellite images. The model is separated into POI-view and geographic-view contrastive learning models. The former assumes that two satellite images with similar POI distribution should possess similar human activity information, while the latter assumes that geographically adjacent satellite images should be more semantically similar than the geographically distant images. In practice, the contrastive learning model maximizes the representation similarity of the satellite images from the same pair, i.e., satellite images with similar POIs or geographically adjacent images, and enlarges the representation dissimilarity of the satellite images from different pairs, i.e., satellite images with different POIs or geographically distant satellite images.

From POI-view contrastive pairs, we consider an unlabeled satellite imagery I_i and its POI-similar pair C_i. Then, we train a CNN (convolutional neural network) F_P to map the satellite images I_i and C_i into a low-dimensional representation space:

$$\mathbf{r}_i^P = F_P(I_i), \mathbf{v}_i^P = F_P(C_i). \quad (2)$$

The similarity between $\mathbf{r}_i^P$ and $\mathbf{v}_i^P$ should reflect that these two satellite images have similar human activities, i.e., they have similar POI distributions. Inspired by [30, 33], we adopt the Normalized Temperature-scaled Cross Entropy loss, called NT_Xent loss in

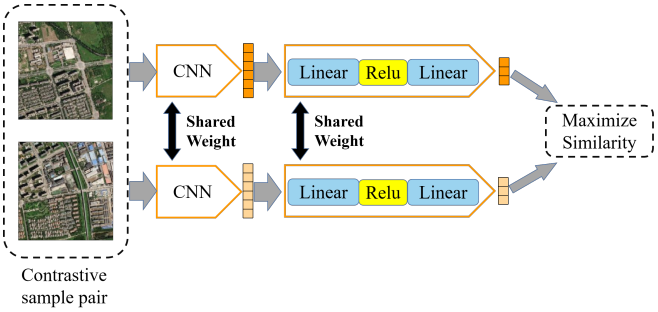


Figure 5: Architecture of the contrastive learning model.

[8]. For N randomly selected satellite images I_i $(i = 1, 2, \cdots, N)$ in a minibatch, we get N corresponding images C_i $(i = 1, 2, \cdots, N)$ that have similar POI distribution. By applying the CNN F_P, there are $2N$ representations for the N satellite image pairs: $\mathbf{r}_i^P$ and $\mathbf{v}_i^P$ $(i = 1, 2, \cdots, N)$. When computing the loss, we treat the satellite imagery I_i and its matching pair C_i as positive samples and the other $2(N-1)$ satellite images in this minibatch as negative samples. The loss for image pair (I_i, C_i) is as follows:

$$\text{loss}(I_i, C_i) = -\log \frac{\exp\left(2 \cdot \text{sim}\left(\mathbf{r}_i^P, \mathbf{v}_i^P\right)\right)}{d1 + d2}, \tag{3}$$

where $\text{sim}(\cdot)$ denotes the cosine similarity, $d1$ and $d2$ are calculated using the negative samples as:

$$d1 = \sum_{k=1}^{N} \mathbf{1}_{[I_i \neq I_k]} \exp\left(2 \cdot \text{sim}\left(\mathbf{r}_i^{\text{P}}, \mathbf{r}_k^{\text{P}}\right)\right), \tag{4}$$

$$d2 = \sum_{k=1}^{N} \exp\left(2 \cdot \text{sim}\left(\mathbf{r}_i^{\text{P}}, \mathbf{v}_k^{\text{P}}\right)\right), \tag{5}$$

where $\mathbf{1}_{[I_i \neq I_k]}$ is an indicator function: $\mathbf{1}_{[I_i \neq I_k]} = 1$ if $I_i \neq I_k$ and $\mathbf{1}_{[I_i \neq I_k]} = 0$ in other situations. The loss is computed across all contrastive samples (I_i, C_i) and (C_i, I_i) in the minibatch. Previous studies suggest that the quality of the learned representations will increase if we, in the training step, add a multi-layer perception (MLP) on top of the CNN for calculating the contrastive loss and use the representations of the final layer of CNN for the downstream tasks[8]. Therefore, we use an MLP with two linear layers and a ReLU activation function as the projection module for the CNN. To be specific, we define the output of projection module H as

$$\mathbf{z}_i^P = H\left(\mathbf{r}_i^P\right) = \mathbf{W}^{(2)}\text{ReLU}\left(\mathbf{W}^{(1)}\mathbf{r}_{\text{i}}^{\text{P}}\right), \tag{6}$$

where $\mathbf{r}_i^P$ is the output representation of F_P for satellite image I_i, and $\mathbf{W}^{(2)}$ and $\mathbf{W}^{(1)}$ are the parameters in the projection module. In the optimization process, we use the output $\mathbf{z}_i^P$ of the projection module to calculate the loss in (3), rather than the direct output $\mathbf{r}_i^P$ of the POI-view model F_P.

Similarly, we denote the geographic-view model as F_G, and use a CNN and projection module of the same architecture as in the POI-view contrastive learning model. We then have

$$\mathbf{r}_i^G = F_G\left(\hat{I}_i\right), \tag{7}$$

$$\mathbf{z}_i^G = H\left(\mathbf{r}_i^G\right) = \mathbf{W}^{(2)}\text{ReLU}\left(\mathbf{W}^{(1)}\mathbf{r}_{\text{i}}^{\text{G}}\right), \tag{8}$$

where $\mathbf{r}_i^G$ is the representation of satellite image $\hat{I}_i$ through F_G, and $\mathbf{z}_i^G$ is the output vector of the projection module.

3.3 Attentional Fusion Model

Next we consider merging the two representations to construct a more informative final representation for use in the downstream tasks. As discussed previously, the two representations are from different modalities: the representation from the POI-view model F_P emphasizes information about POI data (human factors), while the representation from the geographic-view model F_G emphasizes spatial location (the First Law of Geography). In addition, the importance of the different representations in estimating different socioeconomic indicators is still unknown, so we add an attentional fusion model before the final prediction process to automatically determine the weights for each kind of representation.

For $\mathbf{r}_i^P$ (POI-view representation) and $\mathbf{r}_i^G$ (geographic-view representation) from one satellite image, we define learnable parameters $\mathbf{c}$, $\mathbf{V}$, and $\mathbf{b}$ to adaptively fuse them. We have

$$\alpha_i^m = \mathbf{c}^T \cdot \text{Tanh}\left(\mathbf{V} \cdot \mathbf{r}_i^m + \mathbf{b}\right), \quad m \in \{P, G\}, \tag{9}$$

$$\beta_i^m = \frac{\exp\left(\alpha_{\text{i}}^m\right)}{\sum_{m \in \{P,G\}} \exp\left(\alpha_i^m\right)}, \tag{10}$$

$$\mathbf{r}_i^{final} = \sum_{m \in \{P,G\}} \beta_i^m \cdot \mathbf{r}_i^m, \tag{11}$$

where $\mathbf{r}_i^{final}$ is the final representation for satellite image I_i and β_i^m $(m \in \{P, G\})$ are weight coefficients. We then use an MLP with the ReLU activation function to predict the socioeconomic indicator y_i from $\mathbf{r}_i^{final}$ as follows,

$$y_i = \text{MLP}(\mathbf{r}_i^{final}). \tag{12}$$

4 EVALUATION

In this section, we conduct extensive experiments on real-world datasets to evaluate the effectiveness of our method and discuss the case studies of the learned representations.

4.1 Datasets

The datasets in the experiments include satellite imagery, POI data, and four socioeconomic indicators collected from Beijing. 1). The satellite images are of a fixed size 256*256 and spatial resolution $\approx$ 4.7m. For Beijing, the number of total satellite images is 18,289. 2). The POI data is collected from November 2018 to January 2020. There are 1, 481, 100 POI in Beijing, which are divided into 14 categories. 3). Number of takeaway orders. The takeaway order records are collected from July 2020 to December 2020. 4). Number of comments. The number of comments data is collected from 2017 to 2018 [10]. The total number of restaurants is 139, 131. 5). Population[3]. The dataset is of a resolution of approximately 100m. The units are the number of people per grid cell in 2020. 6). Population density[4]. The dataset is the population density in 2020 per grid cell and is of a resolution of approximately $1km$.

[3]https://www.worldpop.org/geodata/summary?id=49919.
[4]https://www.worldpop.org/geodata/summary?id=44834.

Table 2: Socioeconomic indicators prediction results in Beijing.

	Number of Takeaway Orders			Number of Comments			Population			Population Density		
Methods	*RMSE*	R^2	*MAPE*	*RMSE*	R^2	*MAPE*	*RMSE*	R^2	*MAPE*	*RMSE*	R^2	*MAPE*
Autoencoder	2.2116	0.1307	0.7077	3.0886	-0.0258	0.8807	1.5851	0.3755	0.2611	1.5451	0.4647	0.2434
PCA	2.2057	0.1354	0.7186	2.9138	0.0870	0.8720	1.5280	0.4196	0.2603	1.5961	0.4288	0.3398
ICA	2.2610	0.0916	0.7946	2.8825	0.1065	0.8834	1.5093	0.4338	0.3103	1.7815	0.2884	0.2749
Resnet-18	1.3599	0.6552	0.4371	2.2614	0.4501	0.6546	1.0118	0.7746	0.1633	0.9939	0.7806	0.1668
Tile2vec	1.4199	0.6241	0.4767	2.3324	0.4150	0.6771	1.0959	0.7014	0.1825	0.9253	0.8098	0.1441
READ	1.3359	0.6673	0.4186	2.1122	0.5202	0.5755	0.9582	0.7718	0.1530	0.9409	0.8034	0.1595
Geographic	1.3894	0.6401	0.4654	2.1600	0.4983	0.6560	1.0283	0.7372	0.1709	0.8084	0.8549	0.1300
POI	1.2445	0.7113	0.4118	1.9924	0.5731	0.5313	0.9431	0.7789	0.1526	0.9714	0.7904	0.1671
Concat	1.1997	0.7240	0.4035	1.9038	0.6424	0.4997	0.9421	0.7794	0.1501	0.8771	0.8292	0.1405
Ours	**1.1771**	**0.7486**	**0.3947**	**1.8962**	**0.6453**	**0.4903**	**0.8551**	**0.8183**	**0.1371**	**0.7523**	**0.8743**	**0.1197**

4.2 Experiment and Baseline Settings

4.2.1 Experiment settings. In our experiments, we use Resnet-18 [14] as a backbone. With a batch size of 128 and a learning rate of $3e^{-4}$, the Adam optimizer [23] is used to minimize the training loss. After training 100 epochs, we extract a low-dimensional vector representation for each satellite imagery. In the prediction tasks, for each indicator, we randomly split the dataset into 60% training, 20% validation, and 20% test sets.

4.2.2 Baselines. To evaluate the effectiveness of our methods in estimating multiple socioeconomic indicators, we compare our methods with various baselines as introduced below:

Autoencoder [24]. An autoencoder is a neural network that learns representations for unlabeled data. In our case the autoencoder is trained by minimizing the reconstruction error of unlabeled input satellite images.

PCA [29]/ICA [18]. Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are dimension reduction methods in signal processing. We unravel the original satellite imagery into a long vector and apply PCA/ICA to compute the first 10 principal/independent components of each satellite image.

Resnet-18 [14]. Resnet-18 is a deep learning model trained on Imagenet. We use it as the feature extractor to show the limitations of directly applying a model trained on natural images to satellite images.

Tile2vec [20]. Tile2vec is an unsupervised method, which uses geographic distance as a form of weak supervision. For each satellite imagery, Tile2vec finds a geographic neighbor image as a positive sample and a distant image as a negative sample. Tile2vec then minimizes the representation distance between the positive samples and maximizes the distance between negative samples.

READ [12]. Representation Extraction over an Arbitrary District (READ) is a semi-supervised model. It is trained with a small subset of human-labeled satellite images and a large number of unlabeled images. We directly use the embedding model of the original paper to extract representations for our satellite imagery.

POI/Geographic. These are the representations learned by utilizing only the POI-view model or geographic-view model.

Concat. This is a variant of our method, which concatenates representations obtained from the POI-view and geographic-view models. This is to demonstrate the superiority of the attentional model.

4.3 Performance Analysis

We present the RMSE (Root Mean Square Error), R^2, and MAPE (Mean Absolute Percentage Error) as the evaluation metrics of prediction results of the socioeconomic indicators in Table 2, where the best performance is in boldface. Our method with the attentional fusion model outperforms all the baseline representations when using MLP to predict the socioeconomic indicators. The autoencoder performs the worst, possibly because its compressed representation cannot adequately capture human-related factors. PCA and ICA use dimension reduction to compute principal/independent components, which may not be relevant to socioeconomic indicators. Also, a model trained on natural images has limitations in capturing socioeconomic status from satellite images, as shown by Resnet-18's performance. Tile2vec and READ show comparable performance with our methods, because Tile2vec uses geographic distance as a side information while READ uses human knowledge about development status in satellite imagery. However, they yet to consider human-factor information. Notably, the POI-view model has better performance regarding takeaway orders, comments, and population count, while the geographic-view model performs better on population density. The results are consistent with our preliminary studies (see Figure 2). The concatenation of the representations shows higher performance than most of the baselines, but is still lower than our model with an attentional module, which demonstrates the superiority of our proposed attentional fusion model.

4.4 Case Studies

4.4.1 Visualization of the POI-view Representations. In this phase, we map the POI-view representations into a 2-dimensional space through PCA, and visualize how the satellite images of different socioeconomic values are located in the representation space in Figure 6. We select six anchor points as shown in Figure 6 and display the corresponding satellite images. In general, the development status increases from anchor point 1 to 6. At anchor point 1, the satellite images mostly contain farmland areas where few people live. Anchor point 2 images mostly contain small villages, and at anchor point 3, the satellite images contain more residential areas. We can see the images at anchor point 4 are composed of a large number of buildings, but there are still some non-building areas. Finally, at anchor points 5 and 6, the satellite images mainly contain

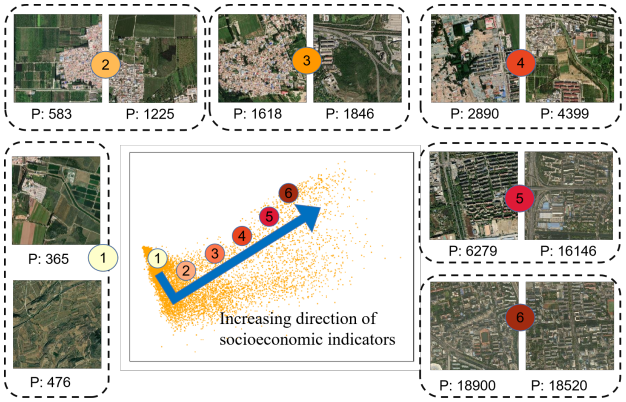


Figure 6: Visualization of the representation space. We mark the urbanization trend (the increasing direction of socioeconomic values) with an arrow in blue. 'P' denotes the population in the satellite image.

urban areas and display highly urbanized cityscapes. We use the population data of satellite images to verify our analysis. We see the population increase from anchor point 1 to 6, and the increasing trend is marked by an arrow in blue. Therefore, we can conclude that our proposed POI-view model can learn highly informative representations containing human factors.

We next select the points at the top-right corner of the representation space in Figure 6 and plot their actual geographic locations in Figure 7. The selected points are blue, and the others are orange. After querying the specific locations, we find the 'blue points' are in commercial centers in urban and suburban areas. Location 1 is near the government building of the Yanqing district (a district in Beijing). Similarly, the locations 2 to 5 are near the government buildings of Miyun, Huairou, Changping, and Pinggu districts (administrative districts in Beijing), respectively. Most government buildings are located in relatively developed areas where the surroundings typically have high commercial activity. Locations 6 and 7 are close to Shimen and Beijing Jiaotong University subway stations. Many commerce POIs are found in the surrounding areas. From the above analysis, we can conclude that, through the learned POI-view representations, regions with high socioeconomic values can be distinguished from regions with low socioeconomic indicators without the help of actual ground surveys.

4.4.2 Model Transferability to Other Cities. To test the generalizability of our proposed model, we conduct experiments with another well developed city (i.e., Shanghai) and an underdeveloped city (Shenyang) in China. The number of satellite images is 5,904 for Shanghai and 15,095 for Shenyang and population data is also collected from WorldPop. Additionally, we collect online comment datasets for both cities from Meituan. We begin by applying the trained Beijing's model to satellite images of Shanghai and Shenyang. The corresponding socioeconomic indicators for Shanghai and Shenyang are then predicted. Even in cities with different levels of development, our contrastive learning method with an attentional model outperforms the baselines in predicting socioeconomic indicators. A potential reason for the differing performance of our method in different cities is the diversity of human activity patterns across these cities.

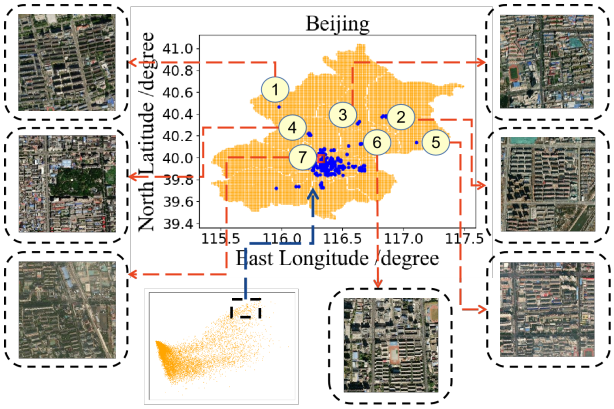


Figure 7: Geographic locations of the points at the top-right corner in the representation space in Figure 6 and corresponding satellite images. The points selected are marked in a rectangle and their geographic locations are blue.

4.4.3 Region Similarity Analysis. Representations of satellite imagery can also be used to depict the similarity between two regions. Given satellite images of Beijing, we examine whether we can find satellite images of Shanghai with similar levels of a given socioeconomic indicator. We randomly select three satellite images of Beijing with varying populations and compute the cosine similarity between the POI-view representations of the selected images and all the images of Shanghai. Then we show the satellite images of Shanghai with high cosine similarity in Figure 8. Although the images of Beijing have varying populations, similar images of Shanghai can be found through the representations. Therefore, our learned representation can find similar regions across different cities with high performance.

5 RELATED WORK AND DISCUSSION

5.1 Related Work

5.1.1 Representation Learning for Satellite Imagery. Representation learning for satellite imagery transforms satellite images into compressed yet informative vectors for various downstream tasks. The representation learning methods can be classified into supervised and unsupervised methods. For the supervised case, Ayush *et al.* [3] trained an object detection network for satellite imagery to generate an interpretable representation. Jean *et al.* [19] used the nighttime light intensity as the label for satellite imagery to extract representations related to poverty. He *et al.* [15] used OpenStreetMap tags as labels for representation learning. However, labeled satellite images are hard to obtain in most cases, therefore many works focus on unsupervised methods. Han *et al.* [12] designed a semi-supervised method by labeling a small set of satellite images manually and used knowledge distillation to train the deep learning model. Jean *et al.* [20] proposed Tile2vec that uses the geographic distance of the satellite imagery to extract the representations from unlabeled satellite imagery. Bjorck *et al.* [5] applied the geographic information in representation learning of satellite images for the management of invasive species. Wang *et al.* [31] utilized traditional feature extraction methods, such as HOG and GIST, and a 'Bag Of Features' model to generate representations for satellite images. Unlike the previous

Table 3: Socioeconomic indicators prediction results in multiple cities.

City	Indicator	Metrics	Autoencoder	PCA	ICA	Resnet-18	Tile2vec	READ	Geographic	POI	**Ours**
Shanghai	Population	*RMSE*	1.4462	1.3993	1.5102	0.9902	0.9744	0.9384	0.9887	0.9518	**0.8743**
		R^2	0.2768	0.3227	0.2115	0.6683	0.6597	0.6727	0.6780	0.7166	**0.7339**
		MAPE	0.1739	0.1631	0.1866	0.1309	0.1338	0.1164	0.1282	0.1204	**0.1152**
Shanghai	Number of Comments	*RMSE*	3.1245	2.5403	2.5295	1.8587	2.0656	1.8281	1.9485	1.8845	**1.8245**
		R^2	-0.5315	-0.0123	-0.0037	0.4221	0.2865	0.4411	0.3650	0.4061	**0.4433**
		MAPE	0.9916	0.8811	0.8802	0.5705	0.6591	0.5533	0.5903	0.5640	**0.5443**
Shenyang	Population	*RMSE*	1.6645	1.4209	1.7966	1.1668	1.2779	1.1554	1.1895	1.1821	**1.1334**
		R^2	0.1384	0.3721	-0.0038	0.5658	0.4792	0.5743	0.5488	0.5544	**0.5904**
		MAPE	0.3809	0.3304	0.40418	0.2588	0.2916	0.2561	0.2666	0.2631	**0.2526**
Shenyang	Number of Comments	*RMSE*	4.0102	3.3529	3.1894	1.8392	1.9408	1.7147	1.8326	1.7764	**1.6084**
		R^2	-0.9572	-0.4155	-0.2809	0.6202	0.5771	0.6699	0.6300	0.6457	**0.7095**
		MAPE	1.3817	1.2626	0.9525	0.6068	0.6177	0.5795	0.5776	0.5666	**0.5496**

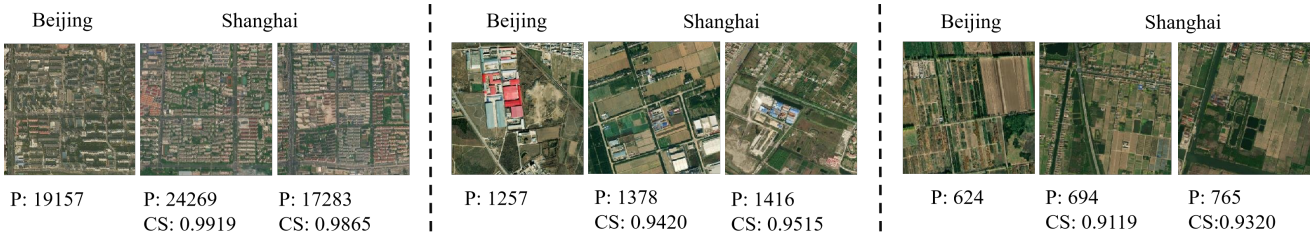


Figure 8: Region Similarity Analysis. ('P' is the abbreviation for population and 'CS' is the abbreviation for cosine similarity.)

methods, we use POI data to capture human factors in satellite imagery and a contrastive learning model to extract human-activity related representations from unlabeled satellite imagery.

5.1.2 Socioeconomic Indicators Estimation from Satellite Imagery. The advancement of remote sensing technology has made it possible to perform socioeconomic estimations using satellite imagery, which previously required costly field surveys. There are two types of indicators in previous studies: economic and social indicators. The economic indicators mostly deal with wealth and commercial activeness. Abitbol *et al.* [1] trained a deep learning model to predict the income status of Paris from satellite images. Yeh *et al.* [34] predicted the asset wealth in African villages from satellite imagery. Mirza *et al.* [28] utilized the nighttime light data to study the inequality problem globally. Han *et al.* [13] incorporated human intelligence and machine intelligence to design a scoring model for the development status based on satellite imagery. Wang *et al.* [31] proposed to predict the commercial activeness of urban commercial districts with satellite imagery and street view imagery. He *et al.* [15] predicted commercial activeness from satellite imagery and OpenStreetMap tags. Jean *et al.* [19] used nighttime imagery as a proxy to predict poverty in five African countries. Head *et al.* [16] explored the potential to measure human development indicators using satellite imagery. Chen *et al.* [6] analyzed regional economic development based on land cover and land use data. In the case of social indicators, Han *et al.* [12] used a semi-supervised model to predict population density, age, and household data from satellite images. There are also studies that combine satellite images and social sensing data to map populations (Cheng *et al.* [9], Jing *et al.* [21]). The advantage of our work is the simultaneous consideration of the social and economic indicators and the incorporation of human factors (POI information) in the satellite images for prediction.

5.2 Discussion

The accurate and timely measurement of socioeconomic indicators is important for urban planning. Relatedly, the increasing availability of detailed satellite images can help enable the measurement of such socioeconomic indicators. However, the scarcity of labeled satellite images forces researchers to turn to unlabeled images. To help leverage this unlabeled data, this work presents the first use of POI data for capturing the human factors in unlabeled satellite images. We show how our contrastive learning model takes advantage of POI information to incorporate human-related factors into the representations of satellite images. We then demonstrate the POI-view representations to verify they are informative about the socioeconomic indicators. Our research shed light on incorporating POI data for representation learning of unlabeled satellite imagery. In summary, our research establishes new performance benchmarks for tasks of representation learning and predicting socioeconomic indicators from satellite imagery.

We believe there are significant research directions for future work based on further study and understanding of representations. Specifically, we plan to analyze the situations where POI data can lead to better representations than the geo data and how different categories of POI contribute to the POI-view representation. Additionally studying how representations contribute to predicting different kinds of socioeconomic indicators will be important for broadening the application areas.

6 CONCLUSION

In this paper, we introduce an unsupervised learning method to learn representations for unlabeled satellite imagery. Apart from a representation learning method that utilizes the First Law of Geography, we propose to use the POI data to capture the human activity factor in representations and design a contrastive learning-based framework to combine the POI data and satellite imagery in representation learning. On top of the representations using spatial information and POI information, we construct an attentional fusion model to fuse the representations from the two modalities automatically. Experiments for predicting various socioeconomic indicators demonstrate that our proposed model can learn more effective representations. Overall, our research takes a fresh look at how to learn representations for unlabeled satellite imagery.

ACKNOWLEDGMENTS

This research has been supported in part by the National Key Research and Development Program of China under Grant 2020YFB210400 and Grant 2020AAA0106000; in part by the National Natural Science Foundation of China under Grant 61972223, Grant U1936217, Grant U20B2060, and Grant 61971267; in part by the International Postdoctoral Exchange Fellowship Program (Talent-Introduction Program) under YJ20210274; and in part by the Academy of Finland under Project 319669, Project 319670, Project 325570, Project 326305, Project 325774, and Project 335934.

REFERENCES

[1] Jacob Levy Abitbol and Marton Karsai. 2020. Interpretable socioeconomic status inference from aerial imagery through urban patterns. *Nature Machine Intelligence* 2, 11 (2020), 684–692.

[2] Adrian Albert, Jasleen Kaur, and Marta C Gonzalez. 2017. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1357–1366.

[3] Kumar Ayush, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. 2020. Generating interpretable poverty maps using object detection in satellite images. *arXiv preprint arXiv:2002.01612* (2020).

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[5] Johan Bjorck, Brendan H. Rappazzo, Qinru Shi, Carrie Brown-Lima, Jennifer Dean, Angela Fuller, and Carla Gomes. 2021. Accelerating Ecological Sciences from Above: Spatial Contrastive Learning for Remote Sensing. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 14711–14720. https://ojs.aaai.org/index.php/AAAI/article/view/17728

[6] Chao Chen, Xinyue He, Zhisong Liu, Weiwei Sun, Heng Dong, and Yanli Chu. 2020. Analysis of regional economic development based on land use and land cover change information derived from Landsat imagery. *Scientific Reports* 10, 1 (2020), 1–16.

[7] Longbiao Chen, Chenhui Lu, Fangxu Yuan, Zhihan Jiang, Leye Wang, Daqing Zhang, Ruixiang Luo, Xiaoliang Fan, and Cheng Wang. 2021. UVLens: Urban Village Boundary Identification and Population Estimation Leveraging Open Government Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–26.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[9] Luxiao Cheng, Lizhe Wang, Ruyi Feng, and Jining Yan. 2021. Remote Sensing and Social Sensing Data Fusion for Fine-Resolution Population Mapping With a Multimodel Neural Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), 5973–5987. https://doi.org/10.1109/JSTARS.2021.3086139

[10] Lei Dong, Carlo Ratti, and Siqi Zheng. 2019. Predicting neighborhoods' socioeconomic attributes using restaurant data. *Proceedings of the National Academy of Sciences* 116, 31 (2019), 15447–15452.

[11] Zhihan Fang, Fan Zhang, Ling Yin, and Desheng Zhang. 2018. MultiCell: Urban population modeling based on multiple cellphone networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–25.

[12] Sungwon Han, Donghyun Ahn, Hyunji Cha, Jeasurk Yang, Sungwon Park, and Meeyoung Cha. 2020. Lightweight and robust representation of economic scales from satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 428–436.

[13] Sungwon Han, Donghyun Ahn, Sungwon Park, Jeasurk Yang, Susang Lee, Jihee Kim, Hyunjoo Yang, Sangyoon Park, and Meeyoung Cha. 2020. Learning to score economic development from satellite imagery. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2970–2979.

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

[15] Zhiyuan He, Su Yang, Weishan Zhang, and Jiulong Zhang. 2018. Perceiving Commerial Activeness Over Satellite Images. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 387–394. https://doi.org/10.1145/3184558.3186353

[16] Andrew Head, Mélanie Manguin, Nhat Tran, and Joshua E Blumenstock. 2017. Can human development be measured with satellite imagery?. In *Ictd*. 8–1.

[17] Danfeng Hong, Lianru Gao, Jing Yao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. 2021. Graph Convolutional Networks for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* 59, 7 (2021), 5966–5978. https://doi.org/10.1109/TGRS.2020.3015157

[18] Aapo Hyvärinen and Erkki Oja. 2000. Independent component analysis: algorithms and applications. *Neural networks* 13, 4-5 (2000), 411–430.

[19] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794.

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

[21] Chuanbao Jing, Weiqi Zhou, Yuguo Qian, and Jingli Yan. 2020. Mapping the Urban Population in Residential Neighborhoods by Integrating Remote Sensing and Crowdsourcing Data. *Remote Sensing* 12, 19 (2020). https://www.mdpi.com/2072-4292/12/19/3235

[22] Jian Kang, Ruben Fernandez-Beltran, Puhong Duan, Sicong Liu, and Antonio J Plaza. 2020. Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast. *IEEE Transactions on Geoscience and Remote Sensing* 59, 3 (2020), 2598–2610.

[23] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[24] Mark A Kramer. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal* 37, 2 (1991), 233–243.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

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

[28] M Usman Mirza, Chi Xu, Bas van Bavel, Egbert H van Nes, and Marten Scheffer. 2021. Global inequality remotely sensed. *Proceedings of the National Academy of Sciences* 118, 18 (2021).

[29] Michael E Tipping and Christopher M Bishop. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61, 3 (1999), 611–622.

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748 [cs.LG]

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

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

[33] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. 2018. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. arXiv:1805.01978 [cs.CV]

[34] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature communications* 11, 1 (2020), 1–11.

[35] Mingyang Zhang, Tong Li, Yong Li, and Pan Hui. 2020. Multi-View Joint Graph Representation Learning for Urban Region Embedding.. In *IJCAI*. 4431–4437.