Knowledge-infused Contrastive Learning for Urban Imagery-based Socioeconomic Prediction

Yu Liu
Tsinghua University
Beijing, China
liuyu2419@126.com

Xin Zhang
Tsinghua University
Beijing, China
zhangxin4087@163.com

Jingtao Ding*
Tsinghua University
Beijing, China
dingjt15@tsinghua.org.cn

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

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

ABSTRACT

Monitoring sustainable development goals requires accurate and timely socioeconomic statistics, while ubiquitous and frequently-updated urban imagery in web like satellite/street view images has emerged as an important source for socioeconomic prediction. Especially, recent studies turn to self-supervised contrastive learning with manually designed similarity metrics for urban imagery representation learning and further socioeconomic prediction, which however suffers from effectiveness and robustness issues. To address such issues, in this paper, we propose a Knowledge-infused Contrastive Learning (KnowCL) model for urban imagery-based socioeconomic prediction. Specifically, we firstly introduce knowledge graph (KG) to effectively model the urban knowledge in spatiality, mobility, etc., and then build neural network based encoders to learn representations of an urban image in associated semantic and visual spaces, respectively. Finally, we design a cross-modality based contrastive learning framework with a novel image-KG contrastive loss, which maximizes the mutual information between semantic and visual representations for knowledge infusion. Extensive experiments of applying the learnt visual representations for socioeconomic prediction on three datasets demonstrate the superior performance of KnowCL with over 30% improvements on R^2 compared with baselines. Especially, our proposed KnowCL model can apply to both satellite and street imagery with both effectiveness and transferability achieved, which provides insights into urban imagery-based socioeconomic prediction.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; **Computer vision representations**; • **Applied computing** → *Law, social and behavioral sciences.*

*Corresponding author.

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583876

KEYWORDS

Urban knowledge graph, socioeconomic prediction, urban imagery, contrastive learning

ACM Reference Format:
Yu Liu, Xin Zhang, Jingtao Ding, Yanxin Xi, and Yong Li. 2023. Knowledge-infused Contrastive Learning for Urban Imagery-based Socioeconomic Prediction. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543507.3583876

1 INTRODUCTION

Driven by the rapid urbanization, more than half of the world population-4.4 billion inhabitants-live in cities and contribute over 80% of global GDP today [4], which makes cities an increasingly important role in achieving United Nations Sustainable Development Goals (SDGs) on economy, education, environment, health, etc. [32, 33]. Especially, socioeconomic indicators like population, educational background and household income are good proxies for SDG monitoring [11]. The traditional door-to-door surveys for such statistics however are costly, labor-intensive, time-consuming and further affected by recent COVID-19 pandemic [33]. In contrast, the inclusive, ubiquitous and frequently-updated web applications paves the way for high-quality, economical and timely SDG monitoring. Recently, researchers predict socioeconomic indicators with the enormous amount of web data especially the urban imagery [6, 25, 43], i.e., the satellite imagery and the street view imagery provided in web map services like Google Map and web platforms like Instagram.

Built upon the great success of deep learning in computer vision [9, 21], most studies adopt convolutional neural networks (CNNs) to learn visual representations of urban imagery for socioeconomic prediction. Specifically, earlier studies follow the task-specific supervised learning for visual representations with neighborhood demographics [1, 11] and country poverty [17, 48] as supervision signals, which require massive labeled data for training and suffer from generalization issues [37]. To overcome such issues, recent studies turn to self-supervised learning with contrastive objectives, a.k.a, contrastive learning [27], and learn a single representation vector for one image to generalize across diverse prediction tasks [2, 44]. Based on manually designed similarity metrics, these studies learn visual representations of urban imagery by maximizing the agreement between similar images in latent space under an image

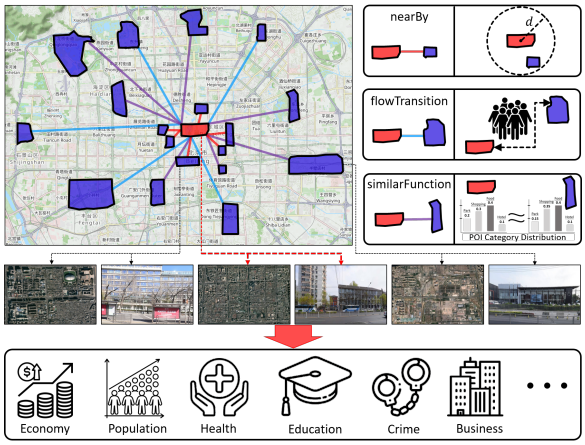


Figure 1: Illustration of infusing various types of knowledge for urban imagery-based socioeconomic prediction, e.g., "nearBy", "flowTransition" and "similarFunction" relational links describe urban knowledge in spatiality, mobility and function (POI category distribution), respectively.

view-based contrastive learning framework [50]. For example, a commonly used metric on spatiality knowledge is from Tobler's First Law of Geography [30] that spatially near images should have similar semantics and thus closer representations [5, 18, 20, 45]. Moreover, a recent study [47] applies the typical contrastive learning framework, SimCLR [7], for satellite imagery-based socioeconomic prediction, assuming that images with similar point of interest (POI) features should have closer representations [1]. Owing to the task-agnostic representation learning from unlabeled data, contrastive learning becomes a promising avenue for urban imagery-based socioeconomic prediction.

Despite this, existing contrastive learning based methods heavily rely on manually designed similarity metrics for urban imagery representation learning, which only capture one or two types of semantic knowledge in urban environment and thus affect the practical performance in socioeconomic prediction. According to recent works of leveraging multi-source urban data for socioeconomic prediction [46, 49], there are various types of semantic knowledge available for urban imagery-based socioeconomic prediction, e.g., the spatiality knowledge of spatial neighborhood, the mobility knowledge of significant flow transitions and the function knowledge of similar POI category distributions, as shown in Figure 1. Thus, how to infuse comprehensive knowledge into contrastive learning for urban imagery-based socioeconomic prediction becomes an important research problem, which however is challenging in:

- **Effective structure for knowledge identification.** Unlike the well-known domain knowledge like Tobler's First Law of Geography, other types of aforementioned knowledge lack explicit domain definition. Moreover, further knowledge infusion requires an effective structure to store and represent the knowledge, increasing the difficulty of knowledge identification for urban imagery-based socioeconomic prediction.
- **Contrastive learning for knowledge infusion.** Existing studies adopt the image view-based contrastive learning framework where the similarity metric is manually designed with a single type of knowledge. Therefore, they fail to infuse various types of knowledge for urban imagery-based socioeconomic prediction.

To address such challenges, in this paper, we propose a Knowledge-infused Contrastive Learning model for urban imagery-based socioeconomic prediction, termed as KnowCL. Firstly, motivated by the recent success of the structured knowledge graph (KG) for urban knowledge modeling [26, 29, 40, 41, 53], we introduce urban knowledge graph (UrbanKG) to identify comprehensive knowledge in multi-source urban data. In the UrbanKG, entity nodes characterize urban elements like regions, POIs and business centers, while relation edges describe semantic connections between them, i.e., the knowledge in spatiality, mobility and function. Moreover, we present cross-modality based contrastive learning for knowledge infusion by exploiting the naturally associated pairing of urban imagery and regions in UrbanKG[2]. To be specific, for an urban image, KnowCL develops a visual encoder to extract its visual representation, and a semantic encoder to extract KG embedding of its associated region entity in UrbanKG [42], which are further optimized for maximum agreement with contrastive loss on image-KG pairs. The learnt visual representations of urban imagery are further fed into traditional regression models for diverse socioeconomic prediction tasks. Therefore, KnowCL firstly leverages UrbanKG for knowledge identification, then represents comprehensive knowledge with KG embedding, and combines with cross-modality based contrastive learning to achieve knowledge infusion for urban imagery-based socioeconomic prediction. The main contributions of this paper are summarized as follows:

- To the best of our knowledge, we are the first to investigate KG for urban imagery-based socioeconomic prediction, which provides an effective structure to comprehensively identify the semantic knowledge in spatiality, mobility, function, etc.
- We propose a cross-modality based contrastive learning framework, which infuses semantic knowledge into visual representations of urban imagery via the novel contrastive objective between the image and KG modalities. The proposed framework might shed light on urban imagery representation learning.
- We conduct extensive experiments on three cities of Beijing, Shanghai and New York across six socioeconomic indicators. The results on both satellite and street view imagery demonstrate that our proposed framework achieves significant performance improvement compared with state-of-the-art models. Further ablation studies and analysis confirm the effectiveness and transferability of knowledge infusion for urban imagery-based socioeconomic prediction.

2 RELATED WORK

As described before, the urban imagery-based socioeconomic prediction studies focus on urban imagery representation learning, which extracts visual representations for downstream socioeconomic prediction tasks. Based on whether the urban imagery representation learning process needs supervision signals from downstream tasks,

[1]Here POI features correspond to POI category distributions of regions identified in urban imagery.

[2]The knowledge graph is identified as another kind of modality data versus urban imagery data in visual modality.

related studies can be classified into supervised learning, unsupervised learning and self-supervised learning[3].

Supervised Urban Imagery Representation Learning for Socioeconomic Prediction. We first discuss about the input source of satellite imagery. With the CNN model pre-trained on ImageNet [9] and light intensity as supervision signal, both Jean *et al.* [17] and Yeh *et al.* [48] extract satellite imagery representations for assets prediction in Africa. Similar frameworks are proposed in [15, 35] for economic indicator prediction. Han *et al.* [12] train a teacher-student network with limited labels to predict demographics like household and income. As for street imagery case, Gerbu *et al.* [11] train a CNN model to identify the types and number of cars in street view imagery, which are further used to estimate socioeconomic indicators like race and education. Lee *et al.* [23] leverage semantic segmentation and graph convolution network (GCN) to predict livelihood indicators of wealth index and BMI. Moreover, Law *et al.* [22] extract features from both satellite and street view imagery to estimate the house prices. However, above studies learn urban imagery representations supervised by a specific downstream task, i.e., the learnt representations cannot generalize to various socioeconomic prediction tasks.

Unsupervised Urban Imagery Representation Learning for Socioeconomic Prediction. Han *et al.* [13] adopt clustering algorithm and partial order graph to distinguish economic development of satellite imagery, which are further used to train a scoring model for urbanization prediction. Suel *et al.* [38] apply the pre-trained CNN model to extract street view imagery representations for inequality measurement in urban environment. Besides, He *et al.* [15] extract traditional image features like histogram of oriented gradients from both satellite and street view imagery to predict commercial activeness. However, such unsupervised learning methods only capture shallow features of urban imagery, which lead to inferior performance.

Self-supervised Urban Imagery Representation Learning for Socioeconomic Prediction. Motivated by the milestones of self-supervised learning achieved in computer vision [7, 19, 44], researchers also leverage self-supervised learning especially contrastive learning for urban imagery-based socioeconomic prediction, and focus on similarity metric design to distill expressive representations of urban imagery. Especially, most studies follow the Tobler's First Law of Geography [30] that "everything is related to everything else, but near things are more related than distant things", and design corresponding similarity metrics or loss forms. For example, Jean *et al.* [18] employs the triplet loss to minimize the distance between representations of spatially near satellite images but maximize the distance between those of spatially distant pairs, while Wang *et al.* [45] employs the similar loss for street view imagery case. Moreover, recent studies [5, 20, 47] mainly adopt the InfoNCE loss [34] with SimCLR framework [7] to encode such spatiality knowledge into visual representations. Xi *et al.* [47] further incorporate a POI-based similarity metric such that images corresponding to similar POI category distributions should have closer visual representations. Furthermore, Li *et al.* [25] consider the spatiality based similarity metrics for both satellite imagery and street imagery. According to the discussion above, most existing studies adopt the image view-based contrastive learning framework, where a pair of images are compared to capture spatiality knowledge, failing to infuse various types of knowledge together. In comparison, our proposed KnowCL model captures comprehensive knowledge via UrbanKG and achieves effective knowledge infusion with cross-modality based contrastive learning framework.

[3]The self-supervised learning is separated from the unsupervised one, which emphasizes using supervision signals generated from data itself.

3 PRELIMINARIES & PROBLEM STATEMENT

As stated in *The sustainable Development Goals Report* [33], the SDGs determine the survival of humanity, facing the current confluence of crises. Especially, various socioeconomic indicators are characterized for SDG monitoring [8]:

Definition 1 (Socioeconomic Indicator). Socioeconomic indicators measure the status of the region/nation on the socioeconomic scale, determined by a combination of social and economic factors such as population, amount and kind of education, household consumption, crime rate, etc.

Moreover, the urban region becomes an important subject for socioeconomic indicator investigation, which are defined as:

Definition 2 (Urban Region). A city can be partitioned into a set of urban regions $\mathcal{A}$, following certain partition criteria like road network division and administrative division [12, 45].

The urban imagery includes the satellite imagery and the street view imagery, which are visual appearances of the city from overhead view and ground-level view, respectively [25, 43]. Figure 1 provides some examples of the urban imagery. Specifically, the satellite images are collected by satellites, which capture the structure of regions in the city. The street view images are taken by automobiles or citizens along the street, which capture the internal environment of regions in the city. Moreover, an urban region usually associates with one satellite image but multiple street view images taken at different locations therein, which are defined as:

Definition 3 (Urban Imagery). Given a city, the urban imagery set is denoted by $\mathcal{I} = \mathcal{I}^{\mathrm{SI}}/\mathcal{I}^{\mathrm{SV}}$ with the satellite imagery set $\mathcal{I}^{\mathrm{SI}}$ and the street view imagery set $\mathcal{I}^{\mathrm{SV}}$. For $\forall a \in \mathcal{A}$, its associated satellite image is denoted by $I_a^{\mathrm{SI}} \in \mathcal{I}^{\mathrm{SI}}$, and its associated n street view images are denoted by $I_a^{\mathrm{SV}} = \{I_{a,1}^{\mathrm{SV}}, \cdots, I_{a,n}^{\mathrm{SV}}\}$ with $I_{a,1}^{\mathrm{SV}}, \cdots, I_{a,n}^{\mathrm{SV}} \in \mathcal{I}^{\mathrm{SV}}$.

The UrbanKG generalizes the commonly used KG concept [16, 42] to urban domain for urban knowledge modeling [26, 29, 40, 53], which is defined as:

Definition 4 (Urban Knowledge Graph). An UrbanKG is defined as a multi-relational graph $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})$, where $\mathcal{E}$, $\mathcal{R}$ and $\mathcal{F}$ are the sets of entities, relations and facts, respectively, with $\mathcal{F} = \{(e_h, r, e_t)|e_h, e_t \in \mathcal{E}, r \in \mathcal{R}\}$ hold. Especially, the entity set $\mathcal{E}$ includes urban elements like regions, POIs, business centers and categories, while the relation set $\mathcal{R}$ describes their semantic connections on spatiality, mobility, function and business. The set of region entities in $\mathcal{G}$ corresponds to above defined region set $\mathcal{A}$.

Details of the UrbanKG will be presented in detail in Section 4.3. To capture the semantic knowledge for downstream applications, recent studies learn to embed entities and relations of a KG into low-dimensional vector space, a.k.a., KG embeddings [16, 39, 42].

Based on the preliminaries above, we then formally define the urban imagery-based socioeconomic prediction problem as follows.

Problem 1 (Urban Imagery-based Socioeconomic Prediction). Given the region set $\mathcal{A}$ with its associated urban imagery set $\mathcal{I}$, for $\forall a \in \mathcal{A}$, the main goal is to learn the visual representation $\boldsymbol{I}_a$ and well estimate the socioeconomic indicator y_a. The ground truth values of y_a is assumed to be unknown in urban imagery representation learning.

4 METHODOLOGY

4.1 Framework Overview

Figure 2 presents the main framework of our proposed KnowCL model for urban imagery-based socioeconomic prediction problem. Since we consider the socioeconomic indicators on region level, we solve the challenges of knowledge identification and knowledge infusion with focus on regions in the city.

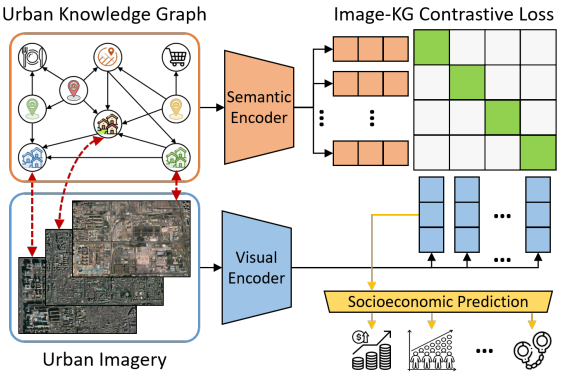


Figure 2: The main framework of KnowCL model, where the urban imagery input can be either satellite imagery or street view imagery. The projector heads after encoders are omitted for simplicity in the illustration.

To identify the comprehensive knowledge in urban environment, we firstly introduce the recently proposed UrbanKG structure to store and represent urban knowledge related with regions, which is then fed into a GCN-based semantic encoder to learn KG embeddings for region entities therein. As for satellite/street view images associated with regions, a CNN-based visual encoder is adopted for visual representations. Furthermore, we propose cross-modality based contrastive learning framework to achieve knowledge infusion. Especially, the designed image-KG contrastive loss encourages one region's KG embedding and its associated urban imagery representation to exhibit high mutual information, through which the semantic knowledge preserved in KG embedding is successfully infused into urban imagery representation. Finally, the knowledge-infused urban imagery representations are leveraged for diverse socioeconomic prediction tasks.

4.2 Urban Knowledge Identification

As defined in Section 3, we introduce UrbanKG to identify the urban knowledge for socioeconomic prediction. Specifically, the entities in UrbanKG include regions partitioned by road network, business centers of commercial and consumption activities, POIs of infrastructures like restaurants, markets and schools, as well as categories of POI attributes, e.g., food, shopping, education, etc. Thus, the entity types in UrbanKG are Region, Business Center (BC), POI and Category.

Table 1: The captured knowledge and corresponding relational structures in UrbanKG.

Knowledge	Relation	Head Entity	Tail Entity
Spatiality	*borderBy*	Region	Region
	nearBy	Region	Region
	locateAt	POI	Region
Mobility	*flowTransition*	Region	Region
Function	*similarFunction*	Region	Region
	coCheckin	POI	POI
	cateOf	POI	Category
Business	*provideService*	BC	Region
	belongTo	POI	BC
	competitive	POI	POI

Moreover, we model the comprehensive knowledge in multi-source urban data as semantic relations, which are summarized in Table 1. Various types of knowledge are represented in triple form with relation, head entity and tail entity.

- **Spatiality.** We determine *borderBy* and *nearBy* relational links by spatial distance between regions, and use *locateAt* to identify POIs' spatially located regions.
- **Mobility.** we aggregate individual mobility trajectories to induce the significant flow transition between regions, which are linked by *flowTransition*.
- **Function.** We consider widely used features in urban computing tasks [51] as function knowledge. For example, *coCheckin* connects highly correlated POIs in terms of concurrence in check-in data, which implies two POIs are consecutively visited by several people, e.g., a cinema and a neighboring restaurant [28]. Since the region function is usually featured by POI category distribution therein [47], we connect regions with similar POI category distribution via *similarFunction*. Besides, *cateOf* describes the category attribute of POIs.
- **Business.** We connect regions with their neighboring business centers via *provideService*, to capture the economic status of regions. Similarly, POIs are connected with neighboring business centers via *belongTo*. To further identify the competitiveness between POIs, neighboring POIs with the same category are connected via *competitive* [24].

Besides, reverse edges are added to model inverse relations, e.g., (POI, *locateAt*, Region) and (Region, ~*locateAt*, POI). Following the structure above, we construct the UrbanKG with urban knowledge identified for socioeconomic prediction. The construction details can be referred to Section A.

4.3 Encoder Design

4.3.1 Semantic Encoder Design. The semantic encoder aims to learn region embeddings with urban knowledge represented. To fully exploit both semantic information of various relations and structural information of graph topology in UrbanKG, we adopt the GCN-based encoder for region embeddings [28, 39].

Given the UrbanKG $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})$, for $\forall v \in \mathcal{E}, r \in \mathcal{R}$, their d-dimensional embeddings after l layer are denoted by $\boldsymbol{e}_v^{l+1}$ and

$\boldsymbol{r}^{l+1}$, respectively. The neighborhood of v is denoted by $\mathcal{N}_v = \{(u,r)|(u,r,v) \in \mathcal{F}\}$, and $\boldsymbol{e}_v^{l+1} \in \mathbb{R}^d$ can be calculated as:

$$\boldsymbol{e}_v^{l+1} = \sigma\left(\sum_{(u,r)\in\mathcal{N}_v} \boldsymbol{W}_{\text{dir}(r)}^l \phi(\boldsymbol{e}_u^l, \boldsymbol{r}^l) + \boldsymbol{W}_{\text{self}}^l \boldsymbol{e}_v^l\right), \quad (1)$$

where $\boldsymbol{W}_{\text{dir(r)}}^l$ and $\boldsymbol{W}_{\text{self}}^l$ are direction-specific projection matrices for incoming/outgoing relations and self loop relation, respectively. $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is the composition function for message calculation in GCN, e.g., element-wise summation and element-wise product [39]. $\sigma(\cdot)$ is an activation function. Besides, we use pre-trained embeddings from TuckER [3] for initialized embeddings.

Let $f^{\text{KG}}(\cdot)$ denote the semantic encoder with L layers of GCN following above design, and the region embedding for $a \in \mathcal{A}$ can be calculated as $\boldsymbol{e}_a = f^{\text{KG}}(\mathcal{G}, a)$, i.e., $\boldsymbol{e}_a = \boldsymbol{e}_a^L$.

4.3.2 Visual Encoder Design. Our proposed KnowCL model allows various choices of network architectures for visual encoder design. For simplicity, we adopt the commonly used ResNet [14] to obtain visual representations of urban imagery.

For $a \in \mathcal{A}$ with satellite imagery I_a^{SI}, the visual representation can be calculated as $\boldsymbol{I}_a^{\text{SI}} = \text{ResNet}(I_a^{\text{SI}})$. As for street view imagery $I_a^{\text{SV}} = \{I_{a,1}^{\text{SV}}, \cdots, I_{a,n}^{\text{SV}}\}$, the visual representation is calculated by average pooling on street view images therein:

$$\boldsymbol{I}_a^{\text{SV}} = \frac{1}{n}\sum_{i=1}^{n} \text{ResNet}(I_{a,i}^{\text{SV}}). \quad (2)$$

Thus, let $f^{\text{Image}}(\cdot)$ denote the visual encoder designed above, and the urban imagery representation can be obtained by $\boldsymbol{I}_a = f^{\text{Image}}(I_a)$ with $I_a = I_a^{\text{SI}}/I_a^{\text{SV}}$ and $\boldsymbol{I}_a = \boldsymbol{I}_a^{\text{SI}}/\boldsymbol{I}_a^{\text{SV}}$.

4.4 Contrastive Loss Design & Optimization

Motivated by cross-modality based contrastive learning between image and text modalities [36, 50], we design a novel image-KG contrastive loss for knowledge infusion. The core insight here is that both semantic representation (KG embedding) and visual representation (urban imagery representation) of a region should be close to each other.

First, for better representation quality, we introduce two independent projection heads $g^{\text{KG}}(\cdot)$ and $g^{\text{Image}}(\cdot)$ after semantic encoder and visual encoder, respectively, as validated in empirical studies [5, 7]. The corresponding outputs of $\tilde{\boldsymbol{e}}_a$ and $\tilde{\boldsymbol{I}}_a$ for $a \in \mathcal{A}$ can be expressed as:

$$\tilde{\boldsymbol{e}}_a = g^{\text{KG}}(\boldsymbol{e}_a) = \boldsymbol{W}_2^{\text{KG}}\text{ReLU}(\boldsymbol{W}_1^{\text{KG}}\boldsymbol{e}_a) \quad (3)$$

$$\tilde{\boldsymbol{I}}_a = g^{\text{Image}}(\boldsymbol{I}_a) = \boldsymbol{W}_2^{\text{Image}}\text{ReLU}(\boldsymbol{W}_1^{\text{Image}}\boldsymbol{I}_a), \quad (4)$$

where four projection matrices are used to project representations for both modalities from their encoder space to the same space for contrastive learning.

Moreover, following the core insight above, we extend the traditional InfoNCE loss [34] to image-KG contrastive loss, and the loss function for $a \in \mathcal{A}$ is expressed as:

$$\begin{aligned}\mathcal{L}_a &= \mathcal{L}_a^{\text{Image}\to\text{KG}} + \mathcal{L}_a^{\text{KG}\to\text{Image}} \\ &= -\log\frac{\exp(\text{sim}(\tilde{\boldsymbol{I}}_a, \tilde{\boldsymbol{e}}_a))}{\sum_{i=1}^m \exp(\text{sim}(\tilde{\boldsymbol{I}}_a, \tilde{\boldsymbol{e}}_i))} - \log\frac{\exp(\text{sim}(\tilde{\boldsymbol{e}}_a, \tilde{\boldsymbol{I}}_a))}{\sum_{i=1}^m \exp(\text{sim}(\tilde{\boldsymbol{e}}_a, \tilde{\boldsymbol{I}}_i))}, \quad (5)\end{aligned}$$

where the loss is computed in a minibatch of m samples, and $\text{sim}(\cdot)$ represents the inner product. $\mathcal{L}_a^{\text{Image}\to\text{KG}}$ and $\mathcal{L}_a^{\text{KG}\to\text{Image}}$ are image-to-KG and KG-to-image contrastive losses, respectively, which maximally preserve the mutual information between image-KG pairs. Unlike existing urban imagery-based socioeconomic prediction studies using image view-based contrastive loss in the same modality [25, 47], our proposed image-KG contrastive loss is based on cross modalities of inputs, which successfully infuses the comprehensive knowledge captured in region embeddings into urban imagery representations.

By optimizing the image-KG contrastive loss on the whole data, we obtain knowledge-infused urban imagery representations $\boldsymbol{I}_{a\in\mathcal{A}}$ from the visual encoder, which are further fed into the regression module of multi-layer perceptron (MLP) for socioeconomic indicator training and prediction.

5 EXPERIMENTS AND RESULTS

5.1 Experimental Setup

5.1.1 Datasets. We collect three datasets with urban imagery and socioeconomic indicator data for evaluation: Beijing (BJ), Shanghai (SH) and New York (NY). Regions in Beijing and Shanghai are partitioned by road network, while regions in New York are Census Block Groups (CBGs) used by US Census Bureau. The 256×256-pixel satellite images with about 4.7 m-resolution are obtained from ArcGIS, which are further merged along irregular region boundaries for input satellite images. The 1024×512-pixel street view images in Beijing and Shanghai as well as the 512×512-pixel street view images in New York are obtained from Baidu Map API and Google Street API, respectively.

As for socioeconomic indicator data, Beijing dataset includes **(1) Pop.**: population data from WorldPop, **(2) Econ.**: economic activity data from [10], **(3) Rest.**: restaurant business (takeaway order data) and **(4) Consp.**: consumption data from a life service platform, while Shanghai dataset includes **(1) Pop.**: population and **(2) Econ.**: economic activity data from same sources. New York dataset includes **(1) Pop.**: population and **(2) Edu.**: education data from SafeGraph, and **(3) Crime**: crime data from NYC Open Data. The UrbanKG data for three datasets are from [28, 41], and business knowledge related relations are omitted in New York dataset due to a lack of source data. All socioeconomic indicators are converted into logarithmic scale, i.e. $y = \ln(y_{\text{raw}} + 1)$. Besides, regions in datasets with over 40 street view images are selected and randomly split into train/valid/test sets by a proportion of 6:2:2 in the socioeconomic prediction step. Table 2 summarizes dataset statistics and details are provided in Section A.

Table 2: Dataset Statistics.

Dataset	#Region	#SV	#SI	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{R}\rvert$	$\lvert\mathcal{F}\rvert$
Beijing	789	31,560	18,289	36,752	10	188,985
Shanghai	1,553	62,120	5,904	58,145	10	363,159
New York	1,142	45,680	1,560	87,020	6	357,464

5.1.2 Baselines. We compare our model with several baselines in urban imagery-based socioeconomic prediction studies. The satellite imagery-based baselines include:

- **Tile2Vec** [18]. Tile2Vec uses the triplet loss to minimize visual representations of spatially near satellite images and maximize those of distant pairs.

Table 3: Satellite imagery-based socioeconomic prediction results on three datasets. Best results are in bold and the best results are underlined. The last row shows relative improvement in percentage.

Dataset	Beijing								Shanghai				New York					
Model	Pop.		Econ.		Rest.		Consp.		Pop.		Econ.		Pop.		Edu.		Crime	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
ResNet-18	0.277	0.887	0.168	1.465	0.146	2.569	0.125	3.435	0.007	1.027	0.082	1.674	-0.404	0.788	0.518	0.085	0.324	0.751
Tile2Vec	0.274	0.888	0.092	1.531	0.108	2.626	0.074	3.534	0.125	1.041	0.065	1.689	0.143	0.615	0.533	0.084	0.382	0.718
READ	0.300	0.872	0.173	1.461	0.222	2.451	0.213	3.258	0.154	0.949	0.097	1.660	-0.034	0.676	0.534	0.084	0.413	0.700
PG-SimCLR	0.356	0.837	0.361	1.285	0.275	2.368	0.269	3.140	0.307	0.858	0.166	1.596	-0.223	0.735	0.622	0.075	0.434	0.687
KnowCL	**0.479**	**0.752**	**0.532**	**1.100**	**0.493**	**1.979**	**0.443**	**2.741**	**0.424**	**0.783**	**0.325**	**1.436**	**0.153**	**0.612**	**0.658**	**0.042**	**0.536**	**0.622**
Improv.	34.5%	10.2%	47.3%	14.4%	79.3%	16.4%	64.7%	12.7%	38.1%	8.7%	95.8%	10.0%	7.0%	0.5%	5.8%	44.0%	23.5%	9.5%

Table 4: Street view imagery-based socioeconomic prediction results on three datasets. Best results are in bold and the best results are underlined. The last row shows relative improvement in percentage.

Dataset	Beijing								Shanghai				New York					
Model	Pop.		Econ.		Rest.		Consp.		Pop.		Econ.		Pop.		Edu.		Crime	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
ResNet-18	0.085	0.997	0.215	1.423	0.262	2.388	0.257	3.166	0.046	1.007	0.033	1.718	0.151	0.612	0.402	0.095	0.340	0.742
Urban2Vec	0.026	1.029	0.059	1.559	0.094	2.646	0.103	3.478	0.012	1.025	0.007	1.741	0.046	0.649	0.232	0.107	0.020	0.904
SceneParse	0.073	1.004	0.157	1.476	0.183	2.512	0.193	3.299	0.058	1.001	0.049	1.704	0.154	0.611	0.426	0.093	0.222	0.806
PG-SimCLR	0.237	0.911	0.288	1.356	0.409	2.136	0.439	2.750	0.015	1.023	0.112	1.646	0.283	0.563	0.569	0.080	0.482	0.657
KnowCL	**0.416**	**0.796**	**0.557**	**1.069**	**0.470**	**2.024**	**0.449**	**2.725**	**0.359**	**0.826**	**0.281**	**1.482**	**0.377**	**0.524**	**0.586**	**0.079**	**0.552**	**0.612**
Improv.	75.6%	12.6%	44.3%	21.2%	14.9%	5.2%	2.3%	1.0%	519.0%	17.5%	150.9%	10.0%	33.2%	6.9%	3.0%	1.3%	14.5%	6.8%

- **READ** [12]. READ uses limited label data to train a teacher-student network with satellite imagery. The pre-trained model in original paper is used for comparison.

The street view imagery-based baselines includes:

- **Urban2Vec** [45]. Urban2Vec follows the similar design with Tile2Vec but focuses on street view imagery.
- **SceneParse** [52]. We use this scene parsing model to extract street view imagery representations following [23, 25].

We also apply two baselines for both satellite and street view imagery-based socioeconomic prediction:

- **ResNet-18** [14]. ResNet-18 is pre-trained on ImageNet [9], which is a backbone adopted in several related studies, and thus selected for comparison in both satellite and street view imagery.
- **PG-SimCLR** [47]. PG-SimCLR originally employs SimCLR [7] for satellite imagery-based socioeconomic prediction, with spatiality and POI category distribution considered in similarity metric design. We select it for both satellite and street view imagery cases considering its competitive performance.

We implement the baselines following reported settings or using pre-trained models in their original papers, and the obtained urban imagery representations are fed into the MLP-based socioeconomic indicator regression module for training and prediction.

5.1.3 Metrics & Implementation. We adopt the widely used rooted mean squared error (RMSE) and coefficient of determination (R^2) [12, 17, 47] for evaluation metrics. For the implementation, ResNet-18 [14] and CompGCN [39] are adopted for visual and semantic encoders, respectively. We select Adam optimizer for parameter learning. In the contrastive learning step, we set the KG embedding dimension as 64 while the number of GCN layers is selected from $\{1, 2, 3, 4\}$. The learning rate is set as 0.0003. In the socioeconomic prediction step, for each region, the learnt single urban imagery representation vector is used to predict various socioeconomic indicators with the learning rate and dropout searched from $\{0.0005, 0.001, 0.005\}$ and $\{0.1, 0.3, 0.5\}$. Besides, we randomly select 10 street view images for each region in the main experiment. The implementation codes are available at the link[4].

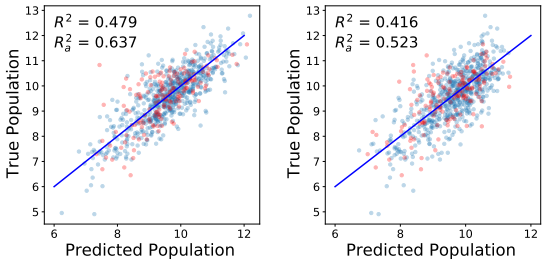
5.2 Performance Comparison

We evaluate the satellite imagery-based socioeconomic prediction performance in Table 3. Results on three datasets across six types of socioeconomic indicators demonstrate the superiority of our proposed KnowCL model, which improves the best baseline (PG-SimCLR) by 7%-79% on R^2 in all cases. Especially, KnowCL achieves the significant performance improvements owing to the comprehensive knowledge infused by the cross-modality based contrastive learning. As for the performance comparison with contrastive learning based models like Tile2Vec and PG-SimCLR, the results show that introducing more knowledge can bring better performance. For example, Tile2Vec only considers spatiality knowledge in similarity metric design while PG-SimCLR further considers function knowledge, leading to over 15% improvements on R^2 on average. Besides, traditional models focus on satellite images in grid shape, failing to extract high-quality visual representations for the more practical case of irregular shape partitioned by road network.

As for the street view imagery-based socioeconomic prediction performance in Table 4, KnowCL also achieves state-of-the-art results, which further validate the effectiveness and robustness. Limited by the repetitive street view images collected in Shanghai dataset, all baselines perform poorly across population and economic activeness prediction tasks therein, while KnowCL leverages UrbanKGs for informative representation learning from urban imagery with competitive performance achieved.

[4] https://github.com/tsinghua-fib-lab/UrbanKG-KnowCL

According to the absolute performance in Table 3 and Table 4, the socioeconomic indicators of different cities show diverse preference to urban imagery, e.g., KnowCL with satellite imagery obtains the best absolute performance for population prediction in Beijing and Shanghai, while the street view imagery becomes the better choice for population and crime prediction in New York. This phenomenon is mainly determined by city structures and socioeconomic indicator characteristics. Different from complex city structures in Beijing and Shanghai, New York follows the grid layout with block regions in similar shapes, which provides limited information for population estimation. Besides, the street view imagery can provide an internal view for urban environment safety perception, as validated in [31]. Such results further indicate that both satellite and street view imagery can provide complementary information to each other, and our proposed KnowCL model can fully exploit the value of urban imagery, which is quite essential for urban environment perception and SDG monitoring.



(a) Satellite Imagery (b) Street View Imagery

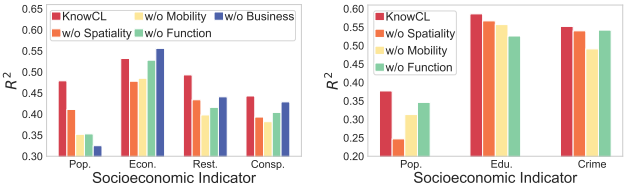
Figure 3: Predicted population versus true population across all regions on Beijing dataset. Blue line is at 45°. R^2 and R^2_a correspond to the results of testing regions (red dots) and all regions (red and blue dots), respectively

To further analyze the predictive power of our proposed KnowCL model, in Figure 3, we compare the predicted and true population for all regions in Beijing dataset, and results for other datasets are provided in Section B. The results show that KnowCL can well replicate the population of most regions (see the dots along 45° line) via urban imagery, explaining 52%-63% of the variation in population on two datasets.

5.3 Ablation Study

5.3.1 Effectiveness of Knowledge Identification. To validate the effectiveness of identified knowledge in UrbanKG, Figure 4 presents the performance comparison of UrbanKG without certain type of knowledge. We select Beijing and New York datasets for evaluation, on which satellite and street view imagery inputs achieve the best absolute performance, respectively. The business knowledge on New York dataset is not provided in UrbanKG and thus not reported.

All four types of semantic knowledge identified by UrbanKG are essential for socioeconomic prediction, according to the findings in Figure 4. Particularly, the knowledge that is hardly captured by urban imagery is more important, e.g., the mobility knowledge of crowd flow transitions between regions brings 5%-30% gains for predicting all socioeconomic indicators, because both satellite and street view imagery cannot capture such dynamic information



(a) Satellite Imagery (b) Street View Imagery

Figure 4: Performance comparison of different identified knowledge on Beijing and New York datasets with satellite and street view imagery, respectively.

without additional knowledge infused. Additionally, the impacts of various types of semantic knowledge vary to socioeconomic indicators. For example, in New York dataset, the education indicator is highly correlated with function knowledge while the population indicator prefers to spatiality knowledge, which enlightens us to identify comprehensive knowledge for a broader urban imagery-based socioeconomic prediction with more indicators considered.

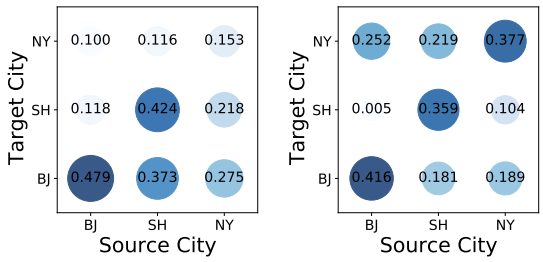
5.3.2 Effectiveness of Knowledge Infusion. A major novelty of this paper is introducing the cross-modality based contrastive learning with the image-KG contrastive loss for knowledge infusion, which is different from the single-modality based ones in existing studies [18, 47]. To validate the effectiveness, we develop a direct image-image contrastive loss for knowledge infusion, which is similar to PG-SimCLR [47], termed as KG-SimCLR. Specifically, KG-SimCLR calculates KG embedding similarity for positive region pairs, and requires their associated urban images to be closer in visual representation space.

Table 5: Performance comparison R^2 of different knowledge infusion ways on Beijing and New York datasets.

		Beijing				New York		
Model		Pop.	Econ.	Rest.	Consp.	Pop.	Edu.	Crime
SI	KG-SimCLR	0.272	0.197	0.192	0.175	-0.302	0.555	0.341
	KnowCL	0.479	0.532	0.493	0.443	0.153	0.658	0.536
SV	KG-SimCLR	0.209	0.229	0.299	0.329	0.053	0.382	0.294
	KnowCL	0.416	0.557	0.470	0.449	0.377	0.586	0.552

(a) Satellite Imagery Matching (BJ→ SH) (b) Street View Imagery Matching (BJ→ SH)

Figure 5: Most similar urban imagery matching between Beijing and Shanghai datasets via learnt urban imagery representations by KnowCL. The population indicator and cosine similarity are presented below the images. Satellite images might be in irregular shape due to the shape of associated regions.



(a) Satellite Imagery (b) Street View Imagery

Figure 6: The R^2 for the transferability test on satellite and street view imagery-based population prediction.

Table 5 presents the performance comparison between KG-SimCLR and KnowCL with different urban imagery inputs. The significant performance gaps between two models on both satellite and street view imagery-based socioeconomic prediction indicate that simply using existing image view-based contrastive loss cannot achieve effective knowledge infusion. Especially, directly modeling the comprehensive knowledge in KG as a similarity metric provides a quite weak self-supervision signal for visual representation learning, while our proposed cross-modality based contrastive learning framework infuses such knowledge via similarity matching in representation space. Overall, the ablation studies demonstrate the effectiveness of our proposed knowledge infusion design and can potentially apply in various urban imagery-based research.

5.4 Transferability Study

5.4.1 Prediction Performance Across Cities/Countries. The experiment results above validate the effectiveness of UrbanKG, which however might be not available in underdeveloped and developing cities/countries due to data deficiency. Thus, here we investigate the practical case of socioeconomic prediction in transfer setting [35]: Given the visual encoder of a KnowCL model trained on a source city with urban imagery and UrbanKG data, we apply it for socioeconomic prediction in target cities where only urban imagery data are available. The transferability task checks whether KnowCL infuses shared knowledge across cities into the visual encoder for urban imagery-based socioeconomic prediction.

We vary the source-target city pairs and report the population prediction performance in Figure 6, where both satellite imagery and street view imagery are evaluated. Here 40 street view images for each region in off-diagonal transfer experiments are used for robust performance. According to the results, the diagonal line shows the highest correlation due to the same city transferred from the source to the target. Moreover, compared with baselines trained and evaluated on the same dataset in Table 3 and Table 4, KnowCL achieves competitive transfer performance for both satellite and street view imagery-based population prediction in Bejing and New York datasets, as validated by similar scatter sizes in each row. For example, SH→BJ transfer experiment achieves a R^2 of 0.373 compared with 0.356 of the best baseline (PG-SimCLR) achieved in non-transfer setting. Such results validate the transferability of our proposed KnowCL model for socioeconomic prediction across cities/countries, which mainly owes to the shared knowledge identified in UrbanKG and infused in visual encoder by cross-modality based contrastive learning. Hence, pre-trained KnowCL model can be leveraged for socioeconomic prediction in cities without UrbanKG. Besides, we also investigate the transferability of the best baseline PG-SimCLR in Section B, which is less competitive due to limited knowledge considered.

5.4.2 Visual Analogies Across Cities. We also investigate the visual similarity between urban imagery across cities. Specifically, given an urban image in source city, we compute the cosine similarity between its visual representation and all visual representations in another city, and select the most similar ones for comparison [45], as shown in Figure 5. As for the satellite imagery matching in Figure 5(a), similar regions share the similar distribution of buildings as well as populations. On the other hand, the street view imagery matching in Figure 5(b) successfully identifies regions with similar physical appearance and populations. Thus, the knowledge-infused urban imagery representations capture not only visual features but also socioeconomic information associated with regions.

6 CONCLUSION

In this paper, we present a novel approach to make predictions on population, economic activity, consumption, education, and public safety indicators from web-collected urban imagery covering both satellite and street view imagery. Our proposed knowledge-infused contrastive learning model KnowCL is built upon the comprehensive knowledge identified by urban knowledge graph, and further designs an image-KG contrastive loss for effective knowledge infusion into urban imagery representations. KnowCL is the first solution that introduces the knowledge graph and the cross-modality based contrastive learning framework for urban imagery-based socioeconomic prediction. Extensive experiments validate the model effectiveness and transferability in different cities across several socioeconomic indicators.

We demonstrated model's potential in socioeconomic prediction with ubiquitous urban imagery, which is of great importance to sustainable development in data-poor regions and countries. Although our model outperforms baselines, the results may be less interpretable, so in future work we will consider exploring in-depth the semantics of UrbanKG for interpretability.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under 2020AAA0106000, the National Nature Science Foundation of China under 61972223, 61971267, U1936217.

REFERENCES

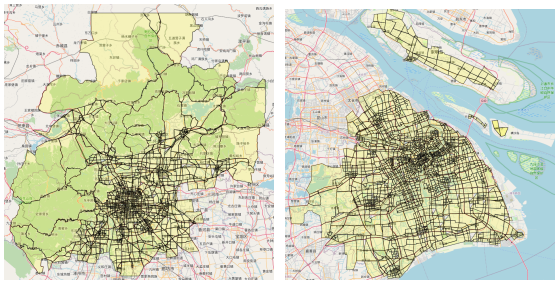
[1] Jacob Levy Abitbol and Marton Karsai. 2020. Interpretable Socioeconomic Status Inference from Aerial Imagery through Urban Patterns. *Nature Machine Intelligence* 2, 11 (2020), 684–692.
[2] Donghyun Ahn, Meeyoung Cha, Sungwon Han, Jihee Kim, Susang Lee, Sangyoon Park, Sungwon Park, Hyunjoo Yang, and Jeasurk Yang. 2020. Teaching Machines to Measure Economic Activities from Satellite Images: Challenges and Solutions. (2020).
[3] Ivana Balažević, Carl Allen, and Timothy M Hospedales. 2019. TuckER: Tensor Factorization for Knowledge Graph Completion. In *EMNLP*. 5185–5194.
[4] The World Bank. 2022. Urban Development. https://www.worldbank.org/en/topic/urbandevelopment/overview. Accessed: 2022-11-01.
[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. In *AAAI*, Vol. 35. 14711–14720.
[6] Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. 2021. Using Satellite Imagery to Understand and Promote Sustainable Development. *Science* 371, 6535 (2021), eabe8628.
[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*. 1597–1607.
[8] Daniel J Corsi, Melissa Neuman, Jocelyn E Finlay, and SV Subramanian. 2012. Demographic and Health Surveys: A Profile. *International Journal of Epidemiology* 41, 6 (2012), 1602–1613.
[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A Large-scale Hierarchical Image Database. In *CVPR*. 248–255.
[10] Lei Dong, Xiaohui Yuan, Meng Li, Carlo Ratti, and Yu Liu. 2021. A Gridded Establishment Dataset as A Proxy for Economic Activity in China. *Scientific Data* 8, 1 (2021), 1–9.
[11] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. Using Deep Learning and Google Street View to Estimate the Demographic Makeup of Neighborhoods across the United States. *PNAS* 114, 50 (2017), 13108–13113.
[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 *AAAI*, 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 *KDD*. 2970–2979.
[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.
[15] Zhiyuan He, Su Yang, Weishan Zhang, and Jiulong Zhang. 2018. Perceiving Commerial Activeness Over Satellite Images. In *WWW*. 387–394.
[16] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge Graphs. *ACM Comput. Surv.* 54, 4 (2021), 1–37.
[17] 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.
[18] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. 2019. Tile2vec: Unsupervised Representation Learning for Spatially Distributed Data. In *AAAI*, Vol. 33. 3967–3974.
[19] Longlong Jing and Yingli Tian. 2020. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE TPAMI* 43, 11 (2020), 4037–4058.
[20] 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.
[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet Classification with Deep Convolutional Neural Networks. *Commun. ACM* 60, 6 (2017), 84–90.
[22] Stephen Law, Brooks Paige, and Chris Russell. 2019. Take a Look Around: Using Street View and Satellite Images to Estimate House Prices. *ACM TIST* 10, 5 (2019), 1–19.
[23] Jihyeon Lee, Dylan Grosz, Burak Uzkent, Sicheng Zeng, Marshall Burke, David Lobell, and Stefano Ermon. 2021. Predicting Livelihood Indicators from Community-Generated Street-Level Imagery. In *AAAI*, Vol. 35. 268–276.
[24] Shuangli Li, Jingbo Zhou, Tong Xu, Hao Liu, Xinjiang Lu, and Hui Xiong. 2020. Competitive Analysis for Points of Interest. In *KDD*. 1265–1274.
[25] Tong Li, Shiduo Xin, Yanxin Xi, Sasu Tarkoma, Pan Hui, and Yong Li. 2022. Predicting Multi-level Socioeconomic Indicators from Structural Urban Imagery. In *CIKM*. 3282–3291.
[26] Jia Liu, Tianrui Li, Shenggong Ji, Peng Xie, Shengdong Du, Fei Teng, and Junbo Zhang. 2021. Urban Flow Pattern Mining based on Multi-source Heterogeneous Data Fusion and Knowledge Graph Embedding. *IEEE TKDE* (2021).
[27] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised Learning: Generative or Contrastive. *IEEE TKDE* (2021).
[28] Yu Liu, Jingtao Ding, and Yong Li. 2021. Knowledge-driven Site Selection via Urban Knowledge Graph. *arXiv preprint arXiv:2111.00787* (2021).
[29] Yu Liu, Jingtao Ding, and Yong Li. 2022. Developing Knowledge Graph based System for Urban Computing. In *ACM SIGSPATIAL Geospatial Knowledge Graphs Workshop*. 3–7.
[30] Harvey J Miller. 2004. Tobler's First Law and Spatial Analysis. *Annals of the Association of American Geographers* 94, 2 (2004), 284–289.
[31] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. 2014. Streetscore-predicting the Perceived Safety of One Million Streetscapes. In *CVPR*. 779–785.
[32] UN Department of Economic and Social Affairs. 2018. World Urbanization Prospects: The 2018 Revision.
[33] UN Department of Economic and Social Affairs. 2022. The Sustainable Development Goals Report 2022.
[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748* (2018).
[35] Sungwon Park, Sungwon Han, Donghyun Ahn, Jaeyeon Kim, Jeasurk Yang, Susang Lee, Seunghoon Hong, Jihee Kim, Sangyoon Park, Hyunjoo Yang, et al. 2022. Learning Economic Indicators by Aggregating Multi-level Geospatial Information. In *AAAI*.
[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*. 8748–8763.
[37] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. 2021. A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery. *Nature Communications* 12, 1 (2021), 1–11.
[38] Esra Suel, John W Polak, James E Bennett, and Majid Ezzati. 2019. Measuring Social, Environmental and Health Inequalities Using Deep Learning and Street Imagery. *Scientific Reports* 9, 1 (2019), 1–10.
[39] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2019. Composition-based Multi-relational Graph Convolutional Networks. In *ICLR*.
[40] Huandong Wang, Qiaohong Yu, Yu Liu, Depeng Jin, and Yong Li. 2021. Spatio-Temporal Urban Knowledge Graph Enabled Mobility Prediction. *ACM UbiComp* 5, 4 (2021), 1–24.
[41] Pengyang Wang, Kunpeng Liu, Lu Jiang, Xiaolin Li, and Yanjie Fu. 2020. Incremental Mobile User Profiling: Reinforcement Learning with Spatial Knowledge Graph for Modeling Event Streams. In *KDD*. 853–861.
[42] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE TKDE* 29, 12 (2017), 2724–2743.
[43] 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 *WWW*. 647–654.
[44] Yi Wang, Conrad Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiaoxiang Zhu. 2022. Self-Supervised Learning in Remote Sensing: A Review. *IEEE Geoscience and Remote Sensing Magazine* (2022).
[45] Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. 2020. Urban2vec: Incorporating Street View Imagery and POIs for Multi-modal Urban Neighborhood Embedding. In *AAAI*, Vol. 34. 1013–1020.
[46] Shangbin Wu, Xu Yan, Xiaoliang Fan, Shirui Pan, Shichao Zhu, Chuanpan Zheng, Ming Cheng, and Cheng Wang. 2022. Multi-Graph Fusion Networks for Urban Region Embedding. *arXiv preprint arXiv:2201.09760* (2022).
[47] 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 *WWW*. 3308–3316.
[48] 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.
[49] Mingyang Zhang, Tong Li, Yong Li, and Pan Hui. 2021. Multi-view Joint Graph Representation Learning for Urban Region Embedding. In *IJCAI*. 4431–4437.
[50] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020. Contrastive Learning of Medical Visual Representations from Paired Images and Text. *arXiv preprint arXiv:2010.00747* (2020).
[51] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban Computing: Concepts, Methodologies, and Applications. *ACM TIST* 5, 3 (2014), 1–55.
[52] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing through ADE20K Dataset. In *CVPR*. 633–641.
[53] Chenyi Zhuang, Nicholas Jing Yuan, Ruihua Song, Xing Xie, and Qiang Ma. 2017. Understanding People Lifestyles: Construction of Urban Movement Knowledge Graph from GPS Trajectory. In *IJCAI*. 3616–3623.

A DETAILS OF DATASET

A.1 UrbanKG Construction

Here we introduce the details of urbanKG construction. For region entities in UrbanKG for Beijing and Shanghai datasets, we partition the city into multiple regions by the road network, which are shown in Figure 7. The region entities in New York dataset follows the CBG division by US Census Bureau, whose visualization can be referred to the official link[5]. Each region entity is provided with a sequence of longitude-latitude pairs $L_a = \{(lng_a^1, lat_a^1), \cdots, (lng_a^k, lat_a^k)\}$ as region boundary. POI entities and business center entities are provided with location information of longitude-latitude pairs like $l_i = (lng_i, lat_i)$. The category entities are POI properties identified by experts, e.g., food, shopping, accommodation, business, residence, education, etc.



(a) Beijing (b) Shanghai

Figure 7: Visualization of region entities in UrbanKG for Beijing and Shanghai datasets.

Based on the entities above, the relational links defined in Table 1 can be extracted as follows.

- *borderBy.* Given two regions a, b, they are connected by *borderBy* if $|L_a \cap L_b| > 0$, i.e., sharing the same boundary points.
- *nearBy.* Given two regions a, b, they are connected by *nearBy* if $\|\bar{L}_a - \bar{L}_b\| \leq 1km$, where $\bar{L}_a, \bar{L}_b$ are center location of regions.
- *locateAt.* Given a POI p and a region a, they are connected by *locateAt* if l_p is in the closure by region boundary L_a.
- *flowTransition.* Given two regions a, b, they are connected by *flowTransition* if the aggregated flow transition between two regions exceeds the threshold.
- *similarFunction.* Given two regions a, b and the category distribution vectors of POIs therein z_a, z_b, they are connected by *similarrFunction* if $cos(z_a, z_b) \geq 0.95$ with cosine similarity.
- *coCheckin.* Given two POIs p_1, p_2, they are connected by *coCheckin* if the number of records that consecutively visit p_1 and p_2 exceeds the threshold.
- *cateOf.* A POI is connected to its associated category by *cateOf.*
- *provideService.* Given a region a and a business center bc, they are connected by *provideService* if $\|\bar{L}_a - l_{bc}\| \leq 3km$.
- *belongTo.* Given a POI p and a business center bc, they are connected by *belongTo* if $\|l_p - l_{bc}\| \leq 3km$.
- *competitive.* Given two POIs p_1, p_2, they are connected by *competitive* if $\|l_{p_1} - l_{p_2}\| \leq 500m$ and they are in the same category.

[5]https://data.cityofnewyork.us/City-Government/2010-Census-Blocks/v2h8-6mxf

A.2 Data Sources & Preprocessing

The data sources in our experiments and their links are provided as follows.

- **Satellite imagery data.** ArGIS, https://geoenrich.arcgis.com/.
- **Street view imagery data in Beijing & Shanghai.** Baidu Map API, http://api.map.baidu.com.
- **Street view imagery data in New York.** Google Street API, https://maps.googleapis.com/.
- **Population data in Beijing & Shanghai.** WorldPop. https://www.worldpop.org/.
- **Population & education data in New York.** SafeGraph, https://www.safegraph.com/.
- **Crime data in New York.** NYC Open Data, https://opendata.cityofnewyork.us/.

Since we focus on the more practical case of irregular region boundaries partitioned by road network, in the data preprocessing step, we merge multiple grid-based satellite images to match the region boundary. Thus, the satellite images for regions might be in irregular shape, as shown in Figure 5(a).

B MODEL DETAILS & EXPERIMENT RESULTS

B.1 Training Algorithm

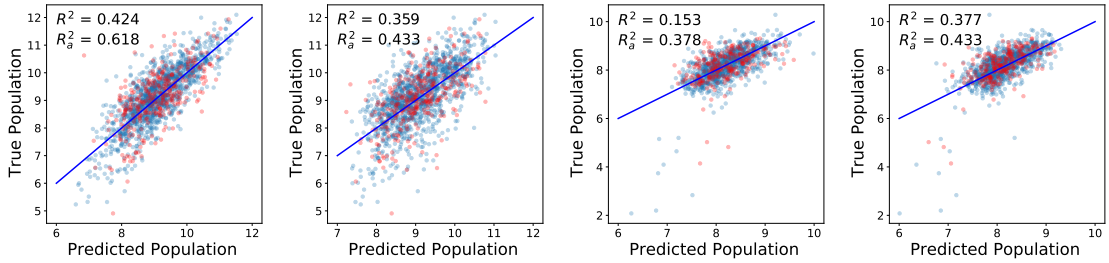
Algorithm 1 summarizes the learning procedure of our proposed KnowCL model for urban imagery-based socioeconomic prediction. The overall framework is divided into two steps of knowledge-infused contrastive learning and socioeconomic prediction. In the first step, lines 4-5 build semantic encoder and visual encoder for different modalities of inputs, and lines 6-11 execute the cross-modality based contrastive learning in a minibatch way with model parameter updated. As for the second step in lines 12-14, only the regression module is trained with observed socioeconomic indicator data, which is then used for socioeconomic prediction.

Algorithm 1 Learning procedure of KnowCL model.

1: **Input**: UrbanKG $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})$, urban imagery data $\mathcal{I}$, region set $\mathcal{A}$, socioeconomic indicator data $\mathcal{D} = \{(a, y_a) | a \in \mathcal{A}\}$.
2: **Output**: The socioeconomic indicator $y_{a'}$ for a region a' without observed label.
3: **Step** 1: Knowledge-infused Contrastive Learning
4: Initialize semantic encoder $f^{\text{KG}}(\cdot)$;
5: Initialize visual encoder $f^{\text{Image}}(\cdot)$;
6: **for** $i = 1, 2, \cdots, n_{\text{iter}}$ **do**
7: Sample a minibatch $\mathcal{A}_{\text{batch}} \in \mathcal{A}$ of size m;
8: $\boldsymbol{e}_a = f^{\text{KG}}(\mathcal{G}, a), \boldsymbol{I}_a = f^{\text{Image}}(I_a), \forall a \in \mathcal{A}_{\text{batch}}$;
9: $\tilde{\boldsymbol{e}}_a = g^{\text{KG}}(\boldsymbol{e}_a), \tilde{\boldsymbol{I}}_a = g^{\text{Image}}(\boldsymbol{I}_a), \forall a \in \mathcal{A}_{\text{batch}}$;
10: Compute the image-KG contrastive loss $\mathcal{L}_a$ using (5);
11: Update encoder parameters w.r.t. the gradients, $\nabla \mathcal{L}_a$.
12: **Step** 2: Socioeconomic Prediction
13: Train regression module MLP$(\cdot)$ on $\mathcal{D}$;
14: Predict socioeconomic indicator $y_{a'} = \text{MLP}(f^{\text{Image}}(I_{a'}))$.

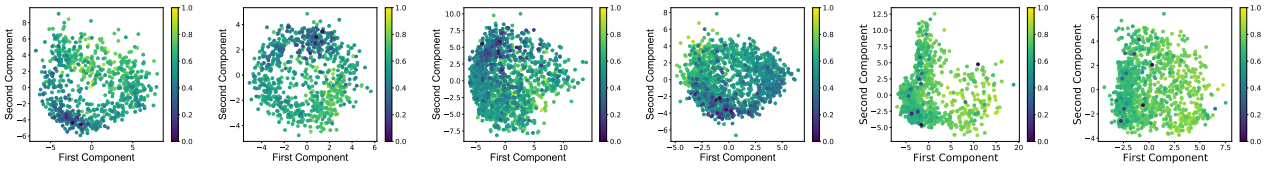
B.2 Predicted v.s. True Indicators

Similar to the setting in Figure 3, we present the comparison of predicted and true population results on Shanghai and New York in



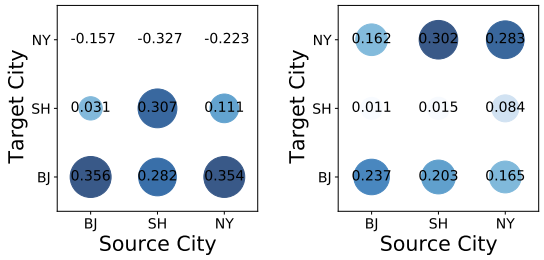
(a) Satellite Imagery (SH) (b) Street View Imagery (SH) (c) Satellite Imagery (NY) (d) Street View Imagery (NY)

Figure 8: Predicted population versus true population across all regions on Shanghai and New York datasets. Blue line is at 45°. R^2 and R^2_a correspond the results of testing regions (red dots) and all regions (red and blue dots), respectively



(a) Satellite (BJ) (b) Street View (BJ) (c) Satellite (SH) (d) Street View (SH) (e) Satellite (NY) (f) Street View (NY)

Figure 9: Visualization of urban imagery based on PCA algorithm, where dot color represents the value of corresponding population to the urban imagery.



(a) Satellite Imagery (b) Street View Imagery

Figure 10: The R^2 for the transferability test of PG-SimCLR on satellite and street view imagery-based population prediction. The visual encoder is trained on one source city and then evaluated on other target cities.

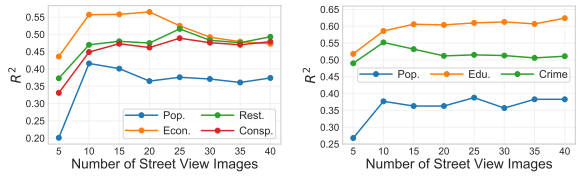
Figure 8. It can be observed that most predicted samples are located along the line at 45°.

B.3 Transferability Study

To validate the transferability of our proposed KnowCL model, we also investigate the transferability of the best baseline PG-SimCLR in Figure 10. Compared with results of KnowCL in Figure 6, PG-SimCLR is less competitive in transfer setting.

B.4 Parameter Study

Figure 11 further investigates the influence of street view images on Beijing and New York datasets. Specifically, in the knowledge-infused contrastive learning step, we keep the number of street view images per region to 10, and tune the number of street view images used in socioeconomic prediction step. According to the results, with the increasing of used street view images, the prediction performance across most of socioeconomic indicators first



(a) Beijing (b) New York

Figure 11: R^2 versus the number of street view images per region used for socioeconomic prediction on two datasets.

increases and then converges. This phenomenon may partly owe to the image quality as well as the limited information captured in street view imagery. Moreover, such results also imply that introducing more stree view images for socioeconomic prediction may not bring performance improvement, which is heavily affected by the noise therein.

B.5 Component Analyses

To analyze the information captured in urban imagery representations, we employ principal component analysis (PCA) algorithm on learnt visual representations by KnowCL for dimension reduction, which are presented in Figure 9. We use the dot color to indicate the population at corresponding regions. Especially, the clustering phenomenon in respective of populations can be observed in Figure 9(a)-(d) on Beijing and Shanghai datasets, which validate the effectiveness of cross-modality based contrastive learning even without population supervision signals. As for the results in New York dataset, such phenomenon is not that obvious because the population in New York is uniformly distributed in block based regions, as we can see that most dots in Figure 9(e)-(f) are in similar colors. All the experiment results validate the effectiveness and robustness of our proposed knowledge-infused contrastive learning model for urban imagery-based socioeconomic prediction.