SatGuide: A POI-Anchored Vision-Language Learning Framework for Robust Satellite Images-based Urban Socioeconomic Prediction

Tianyang Guo[1], Yuanbo Xu[1,*], Wenbin Liu[1], Jiawei Liu[1],
Lu Jiang[2], Pengyang Wang[3], and Pengfei Wang[4]
[1]*College of Computer Science and Technology, Jilin University, Changchun, China*
[2]*Dalian Maritime University, Dalian, China*
[3]*SKL-IOTSC, University of Macau, Macau, China*
[4]*Computer Network Information Center, Chinese Academy of Sciences, Beijing, China*
guoty24@mails.jlu.edu.cn, yuanbox@jlu.edu.cn, liuwenbin@jlu.edu.cn, jiawei23@mails.jlu.edu.cn,
jiangl761@dlmu.edu.cn, pywang@um.edu.mo, wpf2106@gmail.com

***Abstract*—Predicting socioeconomic indicators from satellite images facilitates formulating urban sustainable development plans. Existing methods impose high dependency on the satellite image quality, exhibiting that performances decreases significantly when the quality is interfered by external factors. In this paper, we propose SatGuide, a robust Point-of-Interest (POI)-anchored vision-language learning framework that reduces sensitivity to satellite imagery quality and provides steady accurate predictions. To achieve this, SatGuide enhances robustness from two perspectives: (1) at the semantic level, we utilize POIs as anchors to assist the LLM generating reliable descriptions, which avoids hallucinations caused by low quality satellite images and improves the consistency between region representations and their practical functions; and (2) at the visual level, we exploit a robustness-centric training strategy that focuses on extracting the inherent structural invariance from augmented and multi-scale image features. Empirical studies on New York and Changchun demonstrate the superiority of SatGuide, which achieves average 12.8% and 12.6% R^2 improvements against state-of-the-art baselines on two benchmarks, respectively. Regarding the robustness validation, SatGuide reduces performance fluctuations averagely by 3.0% under noisy and hazy input conditions, outperforming the strongest baseline by 66.3% Our code is available at https://github.com/Ruiiiy2f/Satguide.**

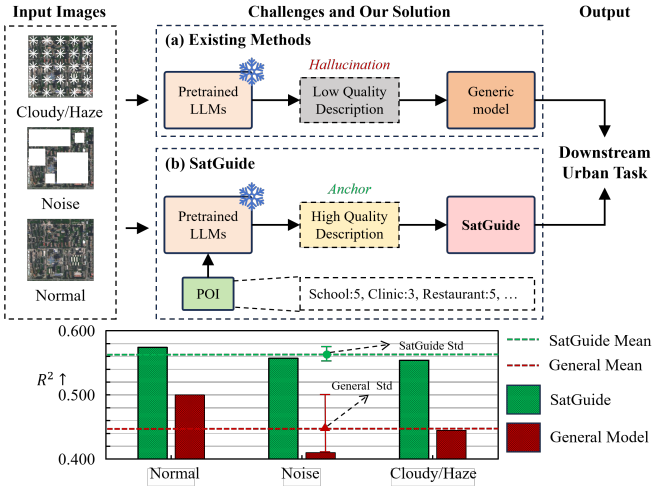

Fig. 1. The illustration of challenges and our solution. **Left** includes real-world satellite images of different qualities. **Middle-upper** explains the reason of robustness issue. **Middle-lower** shows our enhancement towards the robustness. **Histogram** demonstrates that SatGuide alleviate this issue effectively.

I. Introduction

Urban socioeconomic prediction plays a vital role in formulating sustainable urban development plans [1], [2]. Fueled up by deep learning technologies, satellite image-based methods [3]–[5] boost the prediction accuracy, and large vision-language models [6]–[8] further incorporate multi-modal information to realize the comprehensive analysis of urban data.

However, existing methods are commonly challenged by the *robustness* issue that the prediction accuracy exhibits significant fluctuation when processing satellite images of various qualities as shown in the histogram of Figure 1. This limitation bottlenecks the practical deployment of these advanced models for the reason that real-world satellite images can be easily disturbed by weather conditions and sensor operating conditions [9]. To be specific, they are trapped with the following two problems:

(1) Incorrect Semantic Extraction. As shown in the existing methods in Figure 1, due to the external factors (e.g., noises and hazes), LLM might be misled and generate ambiguous descriptions [10], resulting in the incorrect semantic representations of input satellite images. **(2) Redundant Visual Modeling.** Existing methods commonly employ a generic vision model to generate satellite image representations, which inevitably contain redundant details (e.g., lighting intensity or shadow direction) that are irrelevant to the socioeconomic prediction and fail to highlight the most relevant ones.

In this paper, we aim at designing a vision-language learning framework that *possesses robustness to various satellite image qualities and provides steady accurate socioeconomic*

* Corresponding author.

predictions. Specifically, our solutions are listed as follows:
(1) at the semantic level, we utilize POIs contained in satellite images as the explicit guidance for LLMs to generate reliable descriptions. Due to the irrelevance with image qualities, POI information can be served as an anchor during the generation procedure, which guarantees the consistency between representations and semantics, i.e., practical region functions covered in satellite images.
(2) at the visual level, we propose to model the features which are critical for socioeconomic prediction via the contrastive learning approach. Firstly, we design the STA module, an augmentation strategy tailored to satellite images, to simulate the real-world interference scenarios, e.g., noisy and hazy. Secondly, we exploit a feature pyramid network (FPN) to preserve the multi-scale information in satellite images. The ultimate goal of training is to extract the inherent invariance from multi-scale augmented inputs that promotes an accurate, robust and stable socioeconomic prediction model.

To this end, we integrate these enhancements into a two-stage vision-language learning framework, namely SatGuide. During the pre-training stage, SatGuide focuses on improving the robustness to satellite images of different qualities from the semantic level and visual level. During the fine-tuning stage, SatGuide adaptively learns the image representations for specific downstream tasks.

Our primary contributions are summarized as:

- We identify that existing satellite image-based socioeconomic prediction methods are challenged by the robustness issue, and propose SatGuide as a robust alternative that reduces sensitivity to the input quality.
- We introduce quality invariant POIs, as an anchor in LLM-based description generating procedure, to achieve the semantic reliability, and exploit a multi-scale contrast learning strategy coupled with STA and FPN to realize the visual robustness.
- We evaluate the superiority performance of SatGuide on two real-world urban datasets, New York and Changchun, where the average R^2 improvements across multiple prediction tasks against state-of-the-art baselines are 12.8% and 12.6%, respectively. Moreover, our algorithm reduces the performance fluctuations averagely by 3.0% under the noisy and hazy conditions, which outperforms the strongest baseline by 66.3%.

II. Basic Definitions and Problem Statement

Definition 1 (Urban Region). Following prior studies [11], we define a city as a set of urban regions, denoted by $\mathcal{R} = \{r_1, r_2, \ldots, r_N\}$.

Definition 2 (Satellite Image). A satellite image captures detailed ground surface information through remote sensing, providing structural characteristics of an urban region. For an urban region $r_i \in \mathcal{R}$, its associated satellite image is denoted by $I_i \in \mathbb{R}^{H \times W \times 3}$, where H and W represent the height and width of the image, respectively.

Definition 3 (Regional Text Description). A regional text description T_i for an urban region is a concise summary generated using the comprehensive knowledge of well-trained LLMs, encapsulating the region's geographic features.

Problem (Urban Socioeconomic Prediction). Given a set of urban regions with associated satellite images and textual descriptions, the problem is to learn a unified region representation that captures both visual and semantic characteristics. The learned representation can be further leveraged for downstream socioeconomic prediction tasks.

III. Methodology

A. Framework Overview

As illustrated in Figure 2, our framework consists of two primary phases: (1) Pretraining Phase, which aims to learn robust and hierarchical representations from unlabeled satellite imagery; (2) Fine-tuning Phase, where the learned representations are adapted to specific downstream prediction tasks.

B. Semantic Stability: POI-Anchored Text Generation

To ensure that the semantic understanding of a region remains stable despite fluctuations in image quality, we introduce a POI-anchored text generation paradigm. Directly applying LLMs to generate descriptions for satellite imagery often results in factual errors [10] or overly generic text [7], leading to unreliable semantic supervision. We mitigate this by providing the LLM (Gemini) with a multimodal prompt.

For each satellite image I of a region, the corresponding POI data are first extracted, followed by a statistical analysis of the POI categories and their respective counts within the region, formatted as a textual list $\mathcal{P} = \{Category_1 : count_1, Category_2 : count_2, \ldots, Category_N : count_N\}$. This POI list serves as a contextual anchor, providing the LLM with factual, ground-truth information that is decoupled from the visual input. Specifically, by incorporating POI information, we not only leverage the LLM to describe the geographic features of the region but also enable it to infer the region's functionality and potential human activities based on the POI data. The procedure of generating region description can be expressed as:

$$T = \text{LLM}(I, \mathcal{P}). \tag{1}$$

The textual description is first tokenized and processed by a pretrained text encoder (BERT), yielding the final embedding $\tilde{\mathbf{e}}$. To leverage BERT's general language understanding capabilities and mitigate overfitting on limited data, the encoder's parameters are kept frozen and used as a fixed feature extractor. The resulting textual representation $\tilde{\mathbf{e}}$ is subsequently passed through a dedicated projection head $g^{\text{Text}}(\cdot)$. This process can be formally described as:

$$\tilde{\mathbf{e}} = \text{BERT}(T), \tag{2}$$

$$\mathbf{e} = g^{\text{Text}}(\tilde{\mathbf{e}}). \tag{3}$$

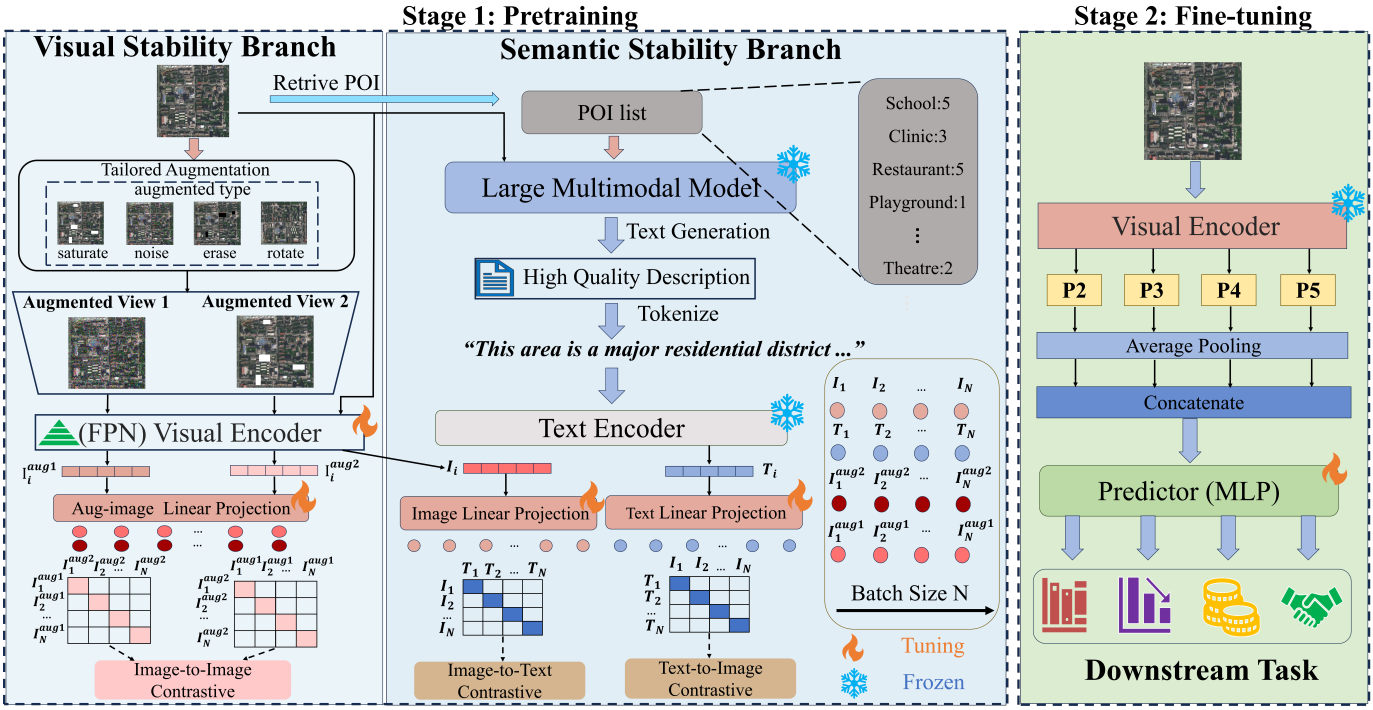

Fig. 2. The main framework of SatGuide model.

C. Visual Feature Invariance Learning

To extract stable visual features that are resilient to environmental noise and variations in spatial details, we propose a dual-robust visual encoding strategy based on the Feature Pyramid Network. Although numerous data augmentation techniques have been developed for natural images, their applicability to satellite imagery is limited due to the inherent domain discrepancies [12].

To address this challenge, we design a **Satellite-Tailored Augmentation** (STA) module that adapts generic data augmentation, which compels the model to learn features that are invariant to superficial changes in satellite imagery. Unlike the conventional augmentations for natural images, our approach focuses on fundamental variations specific to remote sensing data, such as radiometric noise, geometric distortions, and spatial resolution changes.

Specifically, for each original satellite image $I \in \mathcal{I}$, we randomly select and combine two transformations through superposition from a set of carefully designed fundamental operations, including $\{$*random saturation*, *Gaussian noise*, *random erasing*, *random rotation*$\}$, to generate two distinct augmented views, denoted as I^{aug1} and I^{aug2}.

Backbone Network: We employ a ResNet-18 model as the backbone for feature extraction. The backbone processes an input image I and generates a hierarchy of feature maps at different spatial resolutions. We denote the outputs from the four residual blocks (layer 1 to layer 4) as $\{C_2, C_3, C_4, C_5\}$, which form the basis of our bottom-up pathway.

Feature Pyramid Network (FPN): To alleviate the semantic gaps between non-adjacent layers and inherently suppress environmental perturbations without introducing extra auxiliary networks, we adopt a Feature Pyramid Network (FPN) with a residual asymptotic fusion mechanism. Unlike the rigid top-down single-pass aggregation of standard FPN, our FPN progressively fuses features from adjacent layers to non-adjacent layers in an asymptotic manner, while preserving underlying fine-grained spatial structures via residual paths.

First, the channel dimensions of the bottom-up feature maps $\{C_2, C_3, C_4, C_5\}$ are unified to a fixed size d (e.g., 256) via 1×1 convolutional layers, denoted as $X_l = \text{Conv}_{1\times1}(C_l)$ for $l \in \{2, 3, 4, 5\}$. The FPN module then performs progressive multi-stage fusion. To dynamically counteract noise patterns and adjust feature dependencies under varying imagery conditions, we introduce an Adaptive Spatial Fusion (ASF) mechanism within the FPN. At any target pyramid level l, the intermediate fused feature map M_l at spatial coordinate (x, y) is formulated as:

$$M_l(x, y) = \sum_{k=2}^{5} \omega_l^k(x, y) \cdot \psi_{k\to l}(X_k(x, y)), \tag{4}$$

where $\psi_{k\to l}(\cdot)$ denotes the spatial alignment operation (upsampling via bilinear interpolation if $k > l$, or downsampling via strided convolution or pooling if $k < l$) to match the spatial resolution of level l. The term $\omega_l^k(x, y)$ represents the adaptive spatial fusion weight for the feature map from layer k contributing to level l at coordinate (x, y).

Crucially, rather than relying on external quality estimation, the spatial weights are implicitly and natively learned from the multi-scale feature context itself. To ensure the formula strictly complies with the single-column width, the spatial

weight distribution is formalized using a multi-line alignment structure:

$$\begin{aligned}&\left[\omega_l^2(x,y),\omega_l^3(x,y),\omega_l^4(x,y),\omega_l^5(x,y)\right]\\&\quad=\text{Softmax}\left(\text{Conv}(\mathbf{X}_{\text{concat}}(x,y))\right),\end{aligned}\tag{5}$$

where $\mathbf{X}_{\text{concat}}(x,y)$ is the localized concatenation of all aligned feature maps at coordinate (x,y). Facilitated by our robustness-centric training strategy, the internal convolutional layers inherently sense structural distortions or resolution degradation across different scales, adaptively penalizing unreliable feature layers while forming a self-contained robust visual anchor that complements the text-side POI anchor.

To further ensure stable gradient propagation and prevent semantic smoothing during asymptotic fusion, a residual connection is incorporated to generate the final enhanced multi-scale pyramid feature maps $\{P_2,P_3,P_4,P_5\}$:

$$P_l=M_l+X_l,\quad \text{for } l=2,3,4,5.\tag{6}$$

To obtain a single unified vector representation that directly encapsulates these noise-resilient multi-scale characteristics for downstream contrastive alignment, features from each pyramid level are globally average-pooled and concatenated:

$$\tilde{\mathbf{I}}=\text{Concat}\big[\text{AP}(P_2),\text{AP}(P_3),\text{AP}(P_4),\text{AP}(P_5)\big].\tag{7}$$

This FPN encoding process is applied uniformly to the original image I and its augmented views, I^{aug1} and I^{aug2}, sharing identical weight parameters. We wrap the forward passes into two structurally tight rows to comply with the single-column constraint:

$$\begin{aligned}\tilde{\mathbf{I}}&=\text{FPN}(I),\\\tilde{\mathbf{I}}^m&=\text{FPN}(I^m),\quad m\in\{\text{aug}_1,\text{aug}_2\}.\end{aligned}\tag{8}$$

Finally, these robust multi-scale embeddings are mapped into the joint vision-language latent space via two independent projection heads, $g^{\text{Image}}(\cdot)$ and $g^{\text{Aug-Image}}(\cdot)$, allowing the Image-Image Contrastive Loss to directly act upon the FPN-enhanced structural representations:

$$\begin{aligned}\mathbf{I}&=g^{\text{Image}}(\tilde{\mathbf{I}}),\\\mathbf{I}^m&=g^{\text{Aug-Image}}(\tilde{\mathbf{I}}^m),\quad m\in\{\text{aug}_1,\text{aug}_2\}.\end{aligned}\tag{9}$$

D. Joint Optimization via Contrastive Learning

With the visual and textual embeddings prepared, we jointly train the entire pretraining framework by optimizing a composite loss function. This loss function is a weighted sum of two distinct contrastive learning objectives, designed to enforce stability at both the inter-modal (vision-text) and intra-modal (vision-vision) levels. The total loss $\mathcal{L}_{\text{total}}$ is:

$$\mathcal{L}_{\text{total}}=\lambda_1\cdot\mathcal{L}_{\text{text-image}}+\lambda_2\cdot\mathcal{L}_{\text{image-image}},\tag{10}$$

where λ_1 and λ_2 are loss weighting hyperparameters.

Text-Image Contrastive Loss. Inspired by CLIP [13], this loss aims to align the visual and semantic representations in the shared latent space. For a given mini-batch of size N, we compute the cosine similarity between all pairs of L2-normalized visual embeddings $\mathbf{I}$ and text embeddings $\mathbf{e}$. For each sample index $i=1,\ldots,N$, the objective is to maximize the similarity of positive pairs $(\mathbf{I}_i,\mathbf{e}_i)$ while minimizing the similarity of negative pairs $(\mathbf{I}_i,\mathbf{e}_j)$ where $i\neq j$. This is achieved through a symmetric cross-entropy loss:

$$\mathcal{L}_{\text{text-image}}=\frac{1}{2N}\sum_{i=1}^{N}\big(\mathcal{L}_{(\mathbf{I}\to\mathbf{e})}+\mathcal{L}_{(\mathbf{e}\to\mathbf{I})}\big),\tag{11}$$

where the individual loss terms are:

$$\mathcal{L}_{(\mathbf{I}\to\mathbf{e})}=-\log\frac{\exp(\text{sim}(\mathbf{I}_i,\mathbf{e}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{I}_i,\mathbf{e}_j)/\tau)},\tag{12}$$

$$\mathcal{L}_{(\mathbf{e}\to\mathbf{I})}=-\log\frac{\exp(\text{sim}(\mathbf{e}_i,\mathbf{I}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{e}_i,\mathbf{I}_j)/\tau)},\tag{13}$$

where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity and τ is a temperature hyperparameter.

Image-Image Contrastive Loss. Following the SimCLR framework [14], this loss is designed to learn visual representations that are invariant to the augmentations. For each original image, we treat its two augmented views, represented by L2-normalized embeddings $\mathbf{I}^{aug1}$ and $\mathbf{I}^{aug2}$, as a positive pair. A key aspect of our implementation is the use of in-batch negatives from both augmented views to provide a richer set of negative samples. The symmetric loss is calculated as:

$$\mathcal{L}_{\text{image-image}}=\frac{1}{2N}\sum_{i=1}^{N}\big(\mathcal{L}_{(\mathbf{I}^{\text{aug1}}\to\mathbf{I}^{\text{aug2}})}+\mathcal{L}_{(\mathbf{I}^{\text{aug2}}\to\mathbf{I}^{\text{aug1}})}\big).\tag{14}$$

The loss for one direction is defined as:

$$s_{i,j}^{(a,b)}=\frac{\text{sim}\Big(\mathbf{I}_i^{\text{aug}a},\,\mathbf{I}_j^{\text{aug}b}\Big)}{\tau},\tag{15}$$

$$\mathcal{L}_{\mathbf{I}^{\text{aug1}}\to\mathbf{I}^{\text{aug2}}}=-\log\frac{\exp\Big(s_{i,i}^{(1,2)}\Big)}{\sum_{j=1}^{N}\exp\Big(s_{i,j}^{(1,2)}\Big)+\sum_{\substack{k=1\\k\neq i}}^{N}\exp\Big(s_{i,k}^{(1,1)}\Big)}.\tag{16}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature hyperparameter.

The denominator includes the positive pair similarity, similarities to all negative pairs from the other augmented view, and similarities to all other in-batch samples from the same augmented view.

E. Fine-tuning for Downstream Tasks

After pretraining, the model is adapted for specific applications. Crucially, the POI data and regional text descriptions are employed exclusively during the pretraining stage to provide cross-modal semantic guidance; they are completely omitted during fine-tuning. For a given satellite image from a downstream task dataset, it is first passed through this frozen visual encoder to extract its high-level feature representation $\tilde{\mathbf{I}}$ obtained from equation (7). This feature vector is then fed into a lightweight, randomly initialized predictor, which consists of an MLP. Only the parameters of this MLP are tuned on the labeled training data of the specific socioeconomic task.

TABLE I
COMPREHENSIVE PERFORMANCE COMPARISON ON THE NEW YORK DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST ARE UNDERLINED.

City	Dataset	Metric	Methods									
			PCA	Inc-v3	Res-18	ViT	READ	T2V	PG-Sim	U-CLIP	**Ours**	***Impr.***
New York	Edu	R^2 ↑	0.221	0.279	0.384	0.249	0.441	0.471	0.534	0.500	**0.574**	*7.49%*
		RMSE ↓	0.109	0.103	0.096	0.106	0.091	0.089	0.083	0.086	**0.079**	*4.81%*
		MAE ↓	0.082	0.078	0.074	0.080	0.068	0.065	0.063	0.064	**0.059**	*6.34%*
	Income	R^2 ↑	-0.377	-0.014	-0.695	-0.211	-0.118	0.022	0.034	0.388	**0.408**	*5.15%*
		RMSE ↓	0.493	0.423	0.547	0.462	0.443	0.415	0.411	0.328	**0.323**	*1.52%*
		MAE ↓	0.403	0.325	0.411	0.339	0.337	0.307	0.291	0.251	**0.233**	*7.17%*
	Unemploy	R^2 ↑	-0.419	-0.038	-0.119	-0.341	0.113	0.139	0.173	0.193	**0.258**	*33.6%*
		RMSE ↓	0.058	0.046	0.048	0.053	0.044	0.042	0.041	0.041	**0.039**	*4.87%*
		MAE ↓	0.041	0.035	0.036	0.039	0.037	0.033	0.032	0.031	**0.030**	*0.32%*
	Crime	R^2 ↑	0.056	0.036	0.308	0.324	0.344	0.260	0.334	0.420	**0.441**	*5.00%*
		RMSE ↓	0.887	0.914	0.759	0.750	0.741	0.785	0.744	0.695	**0.682**	*1.87%*
		MAE ↓	0.691	0.738	0.585	0.584	0.554	0.601	0.558	0.532	**0.513**	*3.57%*

TABLE II
COMPREHENSIVE PERFORMANCE COMPARISON ON THE CHANGCHUN DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST ARE UNDERLINED.

City	Dataset	Metric	Methods									
			PCA	Inc-v3	Res-18	ViT	READ	T2V	PG-Sim	U-CLIP	**Ours**	***Impr.***
Changchun	Population	R^2 ↑	0.179	0.163	0.324	0.389	0.394	0.472	0.491	0.512	**0.547**	*6.84%*
		RMSE ↓	1.734	1.769	1.556	1.375	1.355	1.189	1.132	1.117	**1.047**	*6.27%*
		MAE ↓	1.374	1.464	1.112	1.106	1.048	1.044	0.998	0.892	**0.831**	*6.84%*
	Carbon	R^2 ↑	0.076	0.045	0.282	0.381	0.305	0.394	0.495	0.598	**0.666**	*11.37%*
		RMSE ↓	1.103	1.229	0.994	0.910	0.953	0.911	0.814	0.793	**0.771**	*2.77%*
		MAE ↓	0.942	0.997	0.741	0.702	0.728	0.655	0.613	0.589	**0.581**	*1.36%*
	Night Light	R^2 ↑	0.156	0.179	0.208	0.323	0.444	0.447	0.648	0.622	**0.775**	*19.60%*
		RMSE ↓	0.997	0.912	0.859	0.832	0.819	0.819	0.695	0.744	**0.610**	*12.23%*
		MAE ↓	0.774	0.738	0.662	0.621	0.600	0.592	0.511	0.532	**0.474**	*7.24%*

IV. EXPERIMENTS

A. Dataset

We conduct experiments on two real-world urban datasets collected from New York City and Changchun City. The New York City dataset follows Census Block Group–level regional divisions and includes satellite imagery paired with multiple socioeconomic indicators, covering crime, education, income, and unemployment. To further evaluate multi-city applicability, we additionally construct a dataset for Changchun City, where each urban region is associated with satellite imagery and indicators reflecting environmental, social, and economic characteristics, including carbon emissions, population, and nighttime light intensity. For both datasets, all target variables are log-transformed, and standard training–validation–test splits are adopted for downstream prediction.

Our model is included in the supplementary material, and the code and data will be open-sourced after publication.

B. Baseline Models

We compare our model with eight established baselines: **PCA** [15] for linear vector transformation; standard ImageNet pre-trained visual encoders including **Inception v3** [16], **Resnet-18** [17], and **ViT** [18]; **READ** [19], a semi-supervised teacher-student distillation approach; **Tile2Vec** [20], which learns visual representations via spatial triplet loss; **PG-SimCLR** [21], a POI-guided self-supervised contrastive learn-

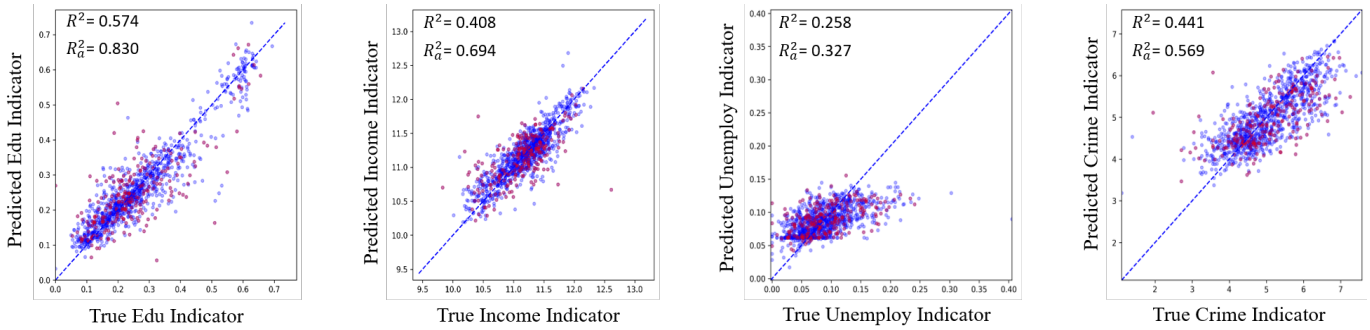

Fig. 3. Predicted values versus true values on four socioeconomic indicator in New York. Blue line is at $45°$, R^2 and R_a^2 correspond to the results of testing regions (red dots) and all regions (red and blue dots), respectively.

ing framework; and **UrbanCLIP** [7], a vision-language contrastive pre-training model for urban region embeddings.

C. Metrics and Implementation

For evaluating model performance, we utilize three widely adopted metrics: coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). Higher R^2 and lower RMSE and MAE indicate superior predictive accuracy. Our model was run on Ubuntu 22.04, with an AMD Ryzen 9 5900X 12-Core Processor, an NVIDIA GeForce RTX 3090 graphics card, and 32 GB of RAM. Model was built by PyTorch 2.7.0. In our experimental setup, the visual encoder FPN with a pretrained ResNet-18 model as its backbone, while the text encoder utilizes a pretrained BERT model. Within the contrastive learning framework, we map the visual features generated by FPN and the text features generated by BERT to a shared 128-dimensional embedding space using a linear transformation layer. To ensure training stability, we set the batch size to 128, the initial learning rate to 3×10^{-4}, and employ the Adam optimizer for parameter optimization.

D. Performance Comparison

Table I reports the performance of the proposed SatGuide framework on the New York dataset, and Table II reports the results on the Changchun dataset. The results validate the effectiveness of our approach, as SatGuide consistently outperforms all baselines. Notably, on the New York tasks, our method achieves R^2 improvements of 5.0%–33.6% over the state-of-the-art baseline. Additionally, Figure 3 visualizes the predicted versus ground-truth values for New York, further confirming the superior accuracy of our framework.

A detailed analysis of the results reveals a clear hierarchy of model capabilities. Unimodal models like ViT and ResNet-18, which rely solely on visual features, exhibit the weakest performance due to their lack of semantic context for interpreting complex socioeconomic patterns. In contrast, self-supervised methods such as Tile2Vec and PG-SimCLR show more competitive results by leveraging implicit similarity signals, but this implicit knowledge lacks the rich, explicit context offered by natural language, thus limiting their effectiveness. UrbanCLIP, which uses LLM-generated text for supervision, emerges as a strong baseline, showcasing the benefit of textual semantics. Yet, its performance is fundamentally constrained by its text generation module's sensitivity to input image quality, which often yields overly generic or factually incorrect descriptions, thereby introducing noise and instability into the training process.

To further investigate the trade-off between the two contrastive objectives, we conduct a sensitivity analysis on the loss weights λ_1 (text-image) and λ_2 (image-image) on the New York dataset, as reported in Table IV. As λ_1 increases from 0.5 to 0.9 (with $\lambda_2 = 1 - \lambda_1$), the performance first improves and then slightly declines. The configuration $\lambda_1 : \lambda_2 = 0.7 : 0.3$ yields the highest R^2 on Income, Unemployment, and Crime, while on Education the best result is achieved at $0.8 : 0.2$ (0.579 vs. 0.574). Considering the overall stability across all four tasks, we select $0.7 : 0.3$ as the default setting, as it provides the best or runner-up performance on every indicator. This balanced weighting confirms that both semantic alignment (text-image) and visual invariance (image-image) are essential for robust urban socioeconomic prediction, and that neither objective dominates the other excessively.

E. Ablation Study

We conducted a series of ablation studies on the New York dataset. We designed five variants of our model by removing or replacing one crucial component at a time: **(1) w/o POI Anchor**, **(2) w/o STA**, **(3) w/o $\mathcal{L}_{\text{image-image}}$**, **(4) w/o $\mathcal{L}_{\text{text-image}}$**, **(5) w/o FPN**. The results on R^2 are depicted in Figure 4.

Effect of POI Anchor. The "w/o POI Anchor" variant, which generates text descriptions solely from satellite imagery without POI guidance, shows a substantial performance degradation across all four tasks. This highlights the critical role of POIs in providing a stable semantic anchor, effectively preventing the LLM from generating ambiguous or inaccurate text when faced with variable visual inputs.

Effect of Contrastive Learning Objectives. Removing either contrastive objective degrades performance. In particular, excluding the text–image loss causes a pronounced drop, most notably on the Unemployment task, indicating that reliable semantic alignment is critical to the framework. Performance

TABLE III
R^2 FLUCTUATION UNDER HETEROGENEOUS SATELLITE IMAGERY IN NEW YORK.

Method	Edu			Income			Unemploy			Crime		
	Normal	Noise	Haze	Normal	Noise	Haze	Normal	Noise	Haze	Normal	Noise	Haze
Inception v3	0.279	0.341	0.190	-0.014	-0.011	-0.014	-0.038	-0.077	-0.075	0.036	-0.001	-0.002
ResNet-18	0.384	0.250	0.215	-0.695	-0.715	-1.124	-0.119	-0.044	-0.165	0.308	0.241	0.258
VIT	0.249	0.307	0.100	-0.211	-0.364	-0.225	-0.341	-0.280	-0.645	0.324	0.297	0.289
READ	0.441	0.417	0.367	-0.118	-0.143	-0.091	0.113	0.086	0.114	0.344	0.323	0.192
Tile2Vec	0.471	0.347	0.340	0.022	-0.194	-0.081	0.139	-0.107	0.043	0.260	0.179	0.232
PG-SimCLR	0.534	0.516	0.472	0.034	-0.079	-0.187	0.173	-0.197	0.135	0.334	0.276	0.258
UrbanCLIP	0.500	0.410	0.441	0.388	0.331	0.301	0.193	0.175	0.163	0.420	0.380	0.331
SatGuide	**0.574**	**0.557**	**0.554**	**0.408**	**0.394**	**0.356**	**0.258**	**0.205**	**0.250**	**0.441**	**0.421**	**0.432**

TABLE IV
SENSITIVITY ANALYSIS OF LOSS WEIGHTS λ_1 (TEXT-IMAGE) AND λ_2 (IMAGE-IMAGE) ON THE NEW YORK DATASET. BEST R^2 FOR EACH TASK IS IN BOLD. BASED ON OVERALL PERFORMANCE, WE SELECT $\lambda_1 : \lambda_2 = 0.7 : 0.3$ AS DEFAULT.

$\lambda_1 : \lambda_2$	Education	Income	Unemployment	Crime
0.5:0.5	0.512	0.345	0.241	0.375
0.6:0.4	0.568	0.402	0.255	0.438
0.7:0.3	0.574	**0.408**	**0.258**	**0.441**
0.8:0.2	**0.579**	0.395	0.250	0.432
0.9:0.1	0.543	0.376	0.235	0.418

also declines without the image–image loss, highlighting the role of augmentation-invariant visual representations in improving robustness.

Effect of Visual Stability Components. Both visual stability components are necessary. Replacing STA with standard augmentations leads to a clear performance drop, indicating the advantage of domain-specific strategies for remote sensing imagery. Removing the FPN further degrades accuracy, particularly on Income, underscoring the importance of multi-scale feature representations for urban interpretation.

F. Visualization of Region Representations

To validate the interpretability of SatGuide, we use PCA to project the learned region representations into 2D space; the result for Changchun is shown in Figure 5. The plot reveals three distinct clusters that exhibit clear semantic alignment with urban development patterns. Specifically, **Cluster 1 (green)** corresponds to dense urban cores, evidenced by the highest peak in the population KDE curve and satellite imagery featuring high-rise buildings. In contrast, **Cluster 2 (orange)** represents suburban or rural areas, characterized by lower population density and visible vegetation. **Cluster 0 (blue)** captures transitional zones between these extremes. These distinct separations demonstrate that SatGuide successfully encodes meaningful socioeconomic structures and discriminative visual features across diverse urban landscapes.

G. Stability Study

This study is motivated by the predictive instability of existing models under heterogeneous satellite image quality. We evaluate the robustness of the pretrained SatGuide model through a stability analysis, with UrbanCLIP as the strongest baseline. Following the fine-tuning protocol, we first use the frozen visual encoders from both models to extract features from the normal (clean) training imagery and train downstream MLP predictors on these features. Subsequently, we freeze both the visual encoders and the trained MLP, and extract features under three distinct conditions: **(1) Normal:** the original, clean test imagery; **(2) Noise:** imagery corrupted with simulated sensor noise; and **(3) Haze:** imagery with simulated atmospheric interference. Predictions are then made using the frozen MLP on these extracted features. Figure 6 plots the average R^2 (solid lines) and standard deviation (shaded regions) across tasks. While UrbanCLIP displays a distinct downward trend on degraded inputs, SatGuide maintains a nearly horizontal trajectory, demonstrating superior robustness. This is quantitatively corroborated by Table III: SatGuide reduces the average R^2 fluctuation from 8.90% (UrbanCLIP) to just 3.00%, representing a 66.3% improvement in stability. These results confirm our framework's resilience to heterogeneous visual quality.

H. Effect of POI-Enhanced Textual Descriptions

To further evaluate the contribution of incorporating POI information into regional descriptions, we construct a variant of the state-of-the-art model UrbanCLIP, denoted as **UrbanCLIP-P**, by replacing the original image-only captions with our POI-enhanced textual descriptions. As illustrated in Figure 7, we prompt the LMM using POI data and satellite imagery to generate high-quality regional textual descriptions. Specifically, for each satellite image, we extract the corresponding POI data and format it as a textual list $\mathcal{P} = \{Category_1 : count_1, Category_2 : count_2, \dots\}$, which serves as a contextual anchor to guide the LLM in generating more reliable and region-specific descriptions.

We evaluate both UrbanCLIP and UrbanCLIP-P on four socioeconomic indicators in the New York dataset. As shown

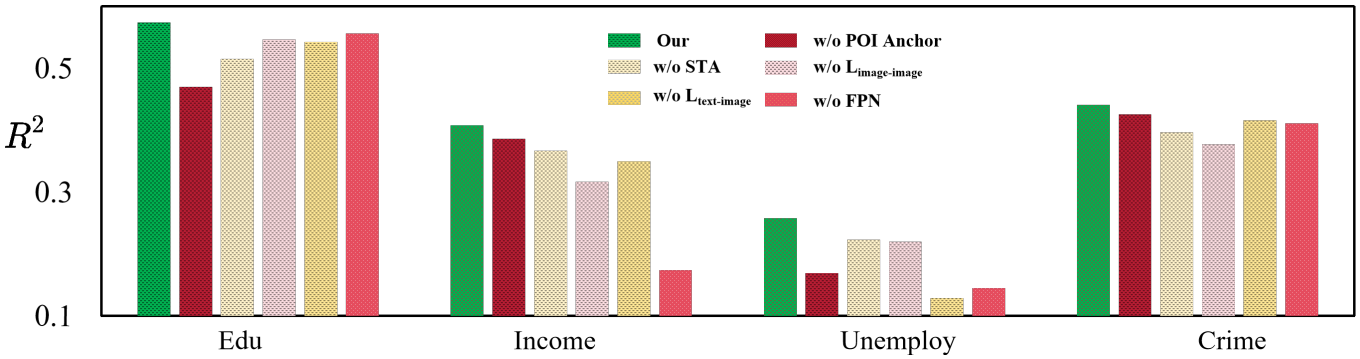

Fig. 4. Results of ablation study in New York.

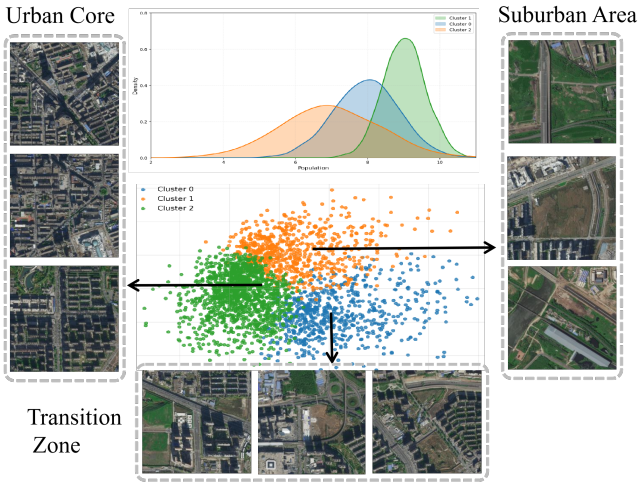

Fig. 5. Representation space visualization in Changchun.

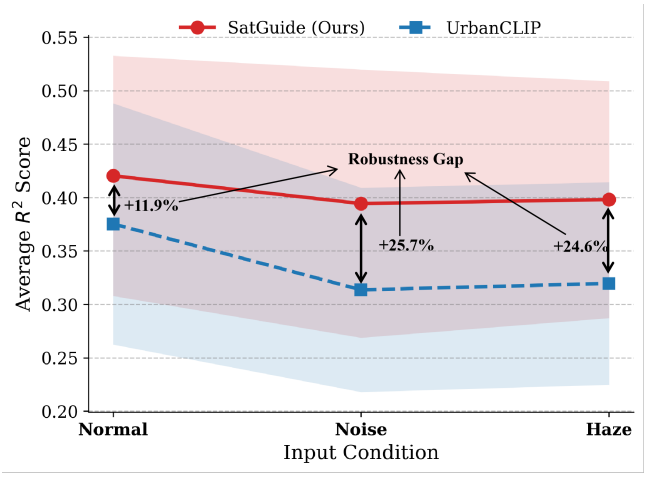

Fig. 6. Stability analysis under heterogeneous imagery. Lines represent the average R^2 across all four socioeconomic tasks in New York, while shaded regions indicate the standard deviation across tasks.

in Figure 8, UrbanCLIP-P consistently outperforms the baseline UrbanCLIP across all four tasks, demonstrating that POI-enhanced textual descriptions provide more informative semantic guidance for urban socioeconomic prediction. The performance improvement is particularly notable on the Income and Unemployment tasks, where the quality of textual descriptions has a substantial impact on the alignment between visual and semantic representations.

I. Vision-Only Branch Analysis for POI-Scarce Regions

Although POI data serve as a powerful semantic anchor in SatGuide, real-world applications in remote or developing areas often suffer from incomplete or missing POI coverage. To evaluate whether our framework remains reliable under such conditions, we design a **vision-only branch** experiment, where the model is trained and evaluated using only the visual branch (including STA and FPN) without any POI-anchored text generation or semantic supervision. This setup directly tests the robustness of the visual branch in scenarios where POI information is unavailable.

We compare the full SatGuide (with POI) against its vision-only variant across seven socioeconomic tasks on both

Prompt based on Satellite Image and POI list:

针对区域ID {region_id}，根据提供的城市区域卫星图像和POI类别及其数量:{餐厅:9, 快餐店:5, 便利店:4, 手机店:17, 超市:13, ... }完成以下内容:
1. 推断该区域的主要功能。2. 描述该区域内可能的居民或人群活动。3. 详细描述图像中的地理特征。请确保描述详细、客观，并充分结合图像和POI信息。

English Translation: For the region ID {region_id}, based on the provided satellite images of the urban area and the POI categories with their quantities: {restaurant: 9, fast_food: 5, convenience: 4, mobile_phone_shop: 17, supermarket: 13, …},complete the following: 1.Infer the main function of the region. 2.Describe the possible residents or crowd activities in the region. 3.Describe in detail the geographical features in the images.
Please ensure the description is detailed, objective, and fully combines the image and POI information.

LMM generation:

该区域是一个典型的商业与居住功能高度融合的混合型社区。既承载着居民的日常生活起居，又汇聚了多样的商业活动，餐饮业态较为发达，覆盖了一日三餐及休闲时段的饮食需求，成为连接商业与居住的重要纽带。图像显示了密集的低矮建筑，包括住宅和商业建筑，道路网络密集，宽度适中，少量绿地分散在建筑物之间。

English Translation: This area is a typical mixed-use community characterized by a high integration of commercial and residential functions. It not only accommodates residents' daily living needs but also aggregates diverse commercial activities. The catering industry here is relatively developed, covering dietary demands for three meals a day as well as leisure hours, thus serving as an important link connecting commerce and residence.Images show dense low-rise buildings, including both residential and commercial structures, with an intensive road network of moderate width. A small amount of green space is scattered among the buildings.

Fig. 7. An example of prompting LMM with POI data and satellite imagery.

New York and Changchun datasets. As shown in Figure 9,

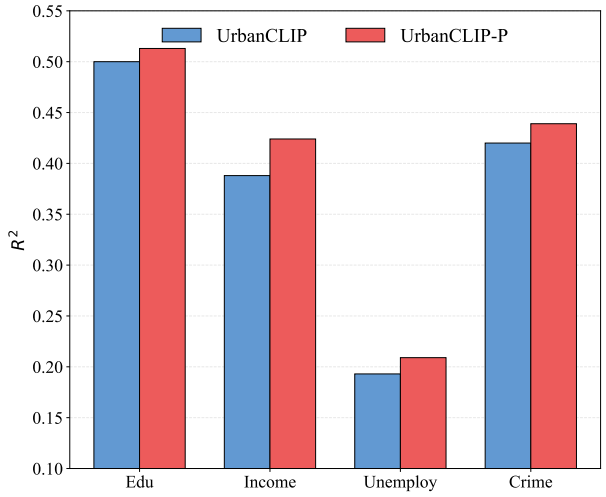

Fig. 8. Comparison of prediction performance (R^2) between the baseline UrbanCLIP and UrbanCLIP-P.

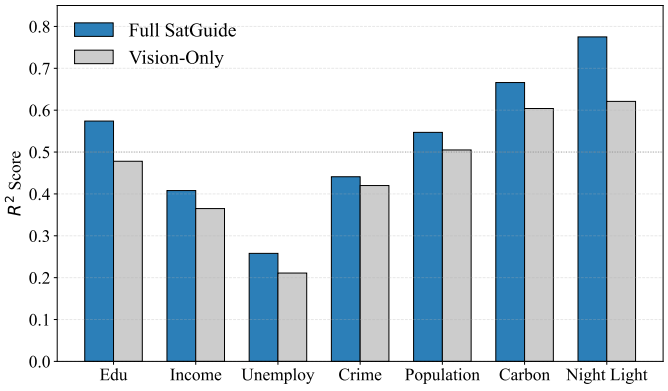

Fig. 9. Performance comparison between full SatGuide (with POI anchor) and its vision-only variant (without POI) across seven prediction tasks. The vision-only branch maintains competitive prediction accuracy even in the absence of POI data, demonstrating the robustness of our visual learning framework for POI-scarce regions.

the vision-only branch still achieves competitive R^2 scores, significantly outperforming naive baselines (e.g., ResNet-18, ViT) reported in Table 1. The performance drop is modest (approximately 5%–15% in R^2), confirming that the visual branch alone can capture meaningful structural cues for socioeconomic prediction. This robustness is attributed to our satellite-tailored augmentation (STA) and multi-scale feature pyramid network (FPN), which extract degradation-invariant representations even without semantic guidance.

More importantly, the gap between the full model and the vision-only variant is significantly smaller than the gap between the vision-only variant and a randomly initialized visual backbone, demonstrating that POI plays a role of *stabilizing* rather than *indispensable* auxiliary. For POI-sparse regions, practitioners can safely deploy the vision-only branch while still obtaining reliable predictions. The full SatGuide remains preferable when both modalities are available.

V. Related Work

Contrastive Learning and Vision-Language Pretraining for Urban Representation. Contrastive learning and vision-language pretraining (VLP) are mainstream for urban and remote sensing representation. Early urban contrastive methods shifted from geographic proximity (, Tile2Vec [20], [22]) to incorporating POI distributions (PG-SimCLR [21]), human mobility (MuseCL [23]), and Urban Knowledge Graphs (KnowCL [24]). Concurrently, foundational contrastive models (, CPC [25], SimCLR [14]) and VLP frameworks (, LXMERT [26], ViLBERT [27], UniCoder-VL [28], ViLT [29], UNITER [30], BLIP/-2 [31], [32]) proved cross-modal transferability. In satellite and urban computing, SatMAE [5] exploits spatial-spectral structures, while CLIP [13], FILIP [33], and UrbanCLIP [7] align imagery with LLM texts. To mitigate LLM hallucination and homogenization, UrbanVLP [6] introduces street views and text calibration, while Health CLIP [34] leverages domain-specific features to enhance socioeconomic indicator prediction.

Multi-Source Fusion and Augmentation for Robust Urban Prediction. To handle environmental variations and spatial complexity, recent literature emphasizes data augmentation, feature aggregation, and multi-source fusion. Satellite data augmentation [12] and multi-scale aggregation via Feature Pyramid Networks (, SegFormer [35]) offer critical strategies for robustness. Meanwhile, multi-source urban representation learning has proven highly effective: Urban2Vec [11] integrates street views and POIs for neighborhood embeddings, while pioneering works combine satellite imagery with auxiliary indicators (, nighttime lights, population) for poverty prediction [36]–[38] or utilize Google Street View [39] for demographic estimation. Together, these multi-source visual and textual cue advancements lay the foundation for robust urban socioeconomic prediction and directly motivate SatGuide.

VI. Conclusion and Future Work

This work tackles predictive instability in satellite-based socioeconomic prediction arising from heterogeneous visual quality. Future extensions of SatGuide may incorporate complementary modalities, including street-view imagery, road networks, and building footprints, to further enrich urban representations.

VII. Acknowledgment

This work is supported by the Natural Science Foundation of China under Grant No. 92567204 and No. 62472196.

References

[1] U. Nations, "Department of economic and social affairs," *Population Division*, 2015.

[2] K. Willis, "The sustainable development goals," in *The Routledge Handbook of Latin American Development*. Routledge, 2018, pp. 121–131.

[3] Y. Xu, E. Wang, Y. Yang, and Y. Chang, "A unified collaborative representation learning for neural-network based recommender systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 11, pp. 5126–5139, 2021.

[4] Y. Jiang, Y. Yang, Y. Xu, and E. Wang, "Spatial-temporal interval aware individual future trajectory prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 10, pp. 5374–5387, 2024.

[5] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon, "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery," Jul 2022.

[6] X. Hao, W. Chen, Y. Yan, S. Zhong, K. Wang, Q. Wen, and Y. Liang, "Urbanvlp: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 27, 2025, pp. 28 061–28 069.

[7] Y. Yan, H. Wen, S. Zhong, W. Chen, H. Chen, Q. Wen, R. Zimmermann, and Y. Liang, "Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web," in *Proceedings of the ACM Web Conference 2024*, 2024, pp. 4006–4017.

[8] M. Chen, Z. Li, W. Huang, Y. Gong, and Y. Yin, "Profiling urban streets: A semi-supervised prediction model based on street view imagery and spatial topology," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 319–328.

[9] C. T. Jerzak, F. Johansson, and A. Daoud, "Image-based treatment effect heterogeneity," *arXiv preprint arXiv:2206.06417*, 2022.

[10] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[11] Z. Wang, H. Li, and R. Rajagopal, "Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding," *Cornell University - arXiv,Cornell University - arXiv*, Jan 2020.

[12] L. M. Hopkins, W.-K. Wong, H. Kerner, F. Li, and R. A. Hutchinson, "Data augmentation approaches for satellite imagery," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 27, 2025, pp. 28 097–28 105.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[15] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, p. 611–622, Sep 1999. [Online]. Available: http://dx.doi.org/10.1111/1467-9868.00196

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2016.308

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2016.90

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] S. Han, D. Ahn, H. Cha, J. Yang, S. Park, and M. Cha, "Lightweight and robust representation of economic scales from satellite imagery," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 428–436.

[20] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, "Tile2vec: Unsupervised representation learning for spatially distributed data," *Proceedings of the AAAI Conference on Artificial Intelligence*, p. 3967–3974, Sep 2019. [Online]. Available: http://dx.doi.org/10.1609/aaai.v33i01.33013967

[21] Y. Xi, T. Li, H. Wang, Y. Li, S. Tarkoma, and P. Hui, "Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests," in *Proceedings of the ACM Web Conference 2022*, Apr 2022. [Online]. Available: http://dx.doi.org/10.1145/3485447.3512149

[22] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, "Tile2vec: Unsupervised representation learning for spatially distributed data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3967–3974.

[23] X. Yong and X. Zhou, "Musecl: Predicting urban socioeconomic indicators via multi-semantic contrastive learning," *arXiv preprint arXiv:2407.09523*, 2024.

[24] Y. Liu, X. Zhang, J. Ding, Y. Xi, and Y. Li, "Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction," Feb 2023.

[25] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[26] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[27] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[28] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 336–11 344.

[29] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International conference on machine learning*. PMLR, 2021, pp. 5583–5594.

[30] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[31] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 800.

[32] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[33] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu, "Filip: Fine-grained interactive language-image pre-training," *arXiv preprint arXiv:2111.07783*, 2021.

[34] T. Ouyang, X. Zhang, Z. Han, Y. Shang, and Y. Li, "Health clip: Depression rate prediction using health related features in satellite and street view images," in *Companion Proceedings of the ACM Web Conference 2024*, 2024, pp. 1142–1145.

[35] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *CoRR*, vol. abs/2105.15203, 2021. [Online]. Available: https://arxiv.org/abs/2105.15203

[36] N. Jean, M. Burke, M. Xie, W. M. Alampay Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016.

[37] A. Kumar, B. Uzkent, M. Burke, D. Lobell, and S. Ermon, "Generating interpretable poverty maps using object detection in satellite images," *Cornell University - arXiv,Cornell University - arXiv*, Feb 2020.

[38] K. Ayush, B. Uzkent, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Efficient poverty mapping from high resolution remote sensing images," *Proceedings of the AAAI Conference on Artificial Intelligence*, p. 12–20, Sep 2022. [Online]. Available: http://dx.doi.org/10.1609/aaai.v35i1.16072

[39] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei, "Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states," *Proceedings of the National Academy of Sciences*, vol. 114, no. 50, pp. 13 108–13 113, 2017.