UrbanVLP: Multi-Granularity Vision-Language Pretraining for Urban Socioeconomic Indicator Prediction

Xixuan Hao[1*], Wei Chen[1*], Yibo Yan[1], Siru Zhong[1], Kun Wang[2], Qingsong Wen[3], Yuxuan Liang[1†]

[1]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[2]National University of Singapore, Singapore
[3]Squirrel AI, Bellevue, USA
{xhao390, wchen110, szhong691}@connect.hkust-gz.edu.cn;
{yanyibo70, qingsonedu}@gmail.com; wk520529@mail.ustc.edu.cn; yuxliang@outlook.com

Abstract

Urban socioeconomic indicator prediction aims to infer various metrics related to sustainable development in diverse urban landscapes using data-driven methods. However, prevalent pretrained models, particularly those reliant on satellite imagery, face dual challenges. Firstly, concentrating solely on macro-level patterns from satellite data may introduce bias, lacking nuanced details at micro levels, such as architectural details at a place. Secondly, the text generated by the precursor work UrbanCLIP, which fully utilizes the extensive knowledge of LLMs, frequently exhibits issues such as hallucination and homogenization, resulting in a lack of reliable quality. In response to these issues, we devise a novel framework entitled UrbanVLP based on Vision-Language Pretraining. Our UrbanVLP seamlessly integrates multi-granularity information from both macro (satellite) and micro (street-view) levels, overcoming the limitations of prior pretrained models. Moreover, it introduces automatic text generation and calibration, providing a robust guarantee for producing high-quality text descriptions of urban imagery. Rigorous experiments conducted across six socioeconomic indicator prediction tasks underscore its superior performance.

Code — https://github.com/skyerhxx/UrbanVLP

Introduction

Urban Socioeconomic Indicator (USI) Prediction, a scholarly pursuit in the realm of urban computing (Zou et al. 2024), deploys data-driven methodologies to forecast socioeconomic metrics (e.g., GDP, population, carbon emission). Rooted in the escalating global prominence of urban environments and the imperative for judicious urban planning, this academic discipline finds motivation in enhancing the efficacy of policy-making, optimizing resource allocation, and mitigating challenges endemic to sustainable development (Xi et al. 2022; Yan et al. 2024).

Given the widespread accessibility of satellite imagery on platforms such as Google Maps (Li et al. 2022; Liu et al. 2023; Xi et al. 2022), coupled with its wealth of information on regional features (*e.g.*, road networks, building density, and vegetation coverage), a predominant majority of initiatives *harness satellite imagery as the foundational modality* for learning urban region representations and making predictions (Zhang et al. 2024). Through a comprehensive review of the current literature, we summarize mainstream methods for USI prediction into two categories:

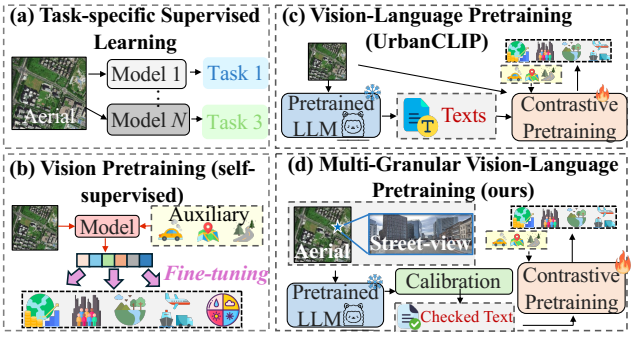

Figure 1: USI prediction frameworks. Compared to existing arts, we present the first attempt to introduce multi-granular visual information and high-quality calibrated texts.

- **Task-specific supervised learning** commonly realizes USI prediction in a fully supervised task, e.g., identifying poverty levels (Han et al. 2020; Ayush et al. 2020, 2021), crop yields (Rußwurm and Körner 2020; Martinez et al. 2021; M Rustowicz et al. 2019; Yeh et al. 2021), and commercial activeness (He et al. 2018; Liu et al. 2023). However, its inherent task specificity (*i.e.,* reliant on ample labeled data), may impede the model's capacity for broad generalization to other downstream tasks.
- **Vision pretraining**, also known as a type of Self-Supervised Learning (Jaiswal et al. 2020; Jiang et al. 2024), aims to learn general visual features from urban satellite imagery, which are subsequently fine-tuned on a specific task for enhancing performance (Bai et al. 2023; Xi et al. 2022; Liu et al. 2023). The inclusion of more auxiliary data, *e.g.*, Points of Interests (POIs) (Zhang et al. 2021; Li et al. 2024b) and road networks (Liang et al. 2018, 2021; Chen et al. 2024), leads to richer, more accurate, and more useful representations of urban areas. For clarity, Figure 1 (a-b) depict a sketch of these two streams

*These authors contributed equally.
†Corresponding author.

of approaches.

Despite the success of utilizing satellite imagery for USI prediction, urban environments exhibit a spatial hierarchy in reality, from a macro *region* level to a micro *location* level (*e.g.*, architectural details, street furniture). Previous frameworks in Figure 1 (a-b) mainly focus on single granularity, neglecting finer-grained visual clues. Nonetheless, as the saying goes, *"the devil is in the details."*. By zooming into micro levels using corresponding street-view images, a more nuanced and fine-grained understanding emerges, as shown in the Figure 1(d) Left. Therefore, the integration and alignment of multi-granularity information remain ongoing challenges in USI prediction that require further exploration.

In recent years, by leveraging the extensive knowledge embedded in LLMs (Manvi et al. 2023) and the inherent interpretability of text modality, a wide range of tasks across various domains are empowered. Textual data can serve as an effective auxiliary tool in multimodal learning across a wide range of scenarios, such as geo-localization (Li et al. 2024a) and recommender systems (Gao et al. 2024).

While in USI prediciton, the efficacy of leveraging textual modality for semantic enrichment is still less explored. UrbanCLIP (Yan et al. 2024) stands out as the innovative solution, which generates text description by a pretrained LLM for satellite imagery and achieves learning urban region representations through natural language supervision.

Nevertheless, the text generation process of UrbanCLIP raises several critical concerns: *i) Hallucination*: The generation of textual content sometimes deviates from or introduces information not present in the input satellite imagery. *ii) Homogenization*: The generated descriptions appear overly simplified and general, potentially leading to homogenization and inadequate differentiation. To better align with the intricacies of satellite (or street-view) imagery, *we necessitate a more powerful approach that goes beyond the intuitive text generation method in UrbanCLIP, ensuring a more faithful and explainable representation of urban regions*.

In this paper, we present a Vision-Language Pretraining framework (i.e., **UrbanVLP**) for urban socioeconomic indicator prediction. As depicted in Figure 1(d), our model elaborately integrates multi-granularity information from both satellite (macro-level) and street-view (micro-level) imagery to produce comprehensive urban region representations, while simultaneously harnessing the interpretability inherent in high-quality textual descriptions. Targeting the first challenge, we introduce a novel *Multi-Granularity Cross-Modal Alignment* module, which utilizes dual-branch contrastive learning to establish alignment between information derived from two semantic granularities. To address the second issue (*i.e.*, hallucination and homogenization in LLM generated texts), we devise an *Automatic Text Generation* together with a *Calibration* mechanism to uphold text quality standards. To guarantee the quality of LLM-generated descriptions, we propose a reference-free metric called `PerceptionScore`, motivated by the human evaluation system.

In summary, our contributions lie in the following aspects:

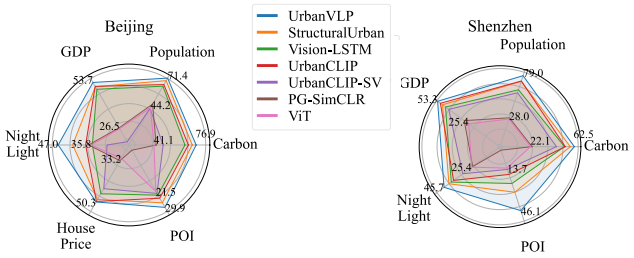

Figure 2: R^2 results in Beijing and Shenzhen.

- *Multi-Granularity Cross-Modal Alignment*. We explore the role of two distinct visual data modalities at various semantic granularities – satellite imagery and street-view imagery. We inject fine-grained semantic information by integrating street-view data through token-level contrastive learning.
- *Automatic Text Generation and Calibration*. Powered by image-to-text LLMs, we generate text descriptions and implement a robust evaluation mechanism based on a new reference-free metric, ensuring the fidelity between the generated texts and the corresponding image content.
- *A New Benchmark & Empirical Evidence*. We plan to open-source the first vision-text and multi-granularity urban dataset upon paper notification, consisting of six downstream tasks across the socio-economy. Extensive experiments demonstrate that our UrbanVLP outperforms existing approaches by an average improvement of 3.95% on the R^2 metric, as illustrated in Figure 2.

Related Work

We formally define the problem of USI prediction. Given a satellite image I_g^{sa}, a set of street-view images $\mathcal{I}_g^{sv}$ belonging to the target coverage, and their corresponding latitude and longitude pair $\mathcal{L}_g$. The objective is to employ a learning function $\mathcal{F}$ to map them to representation vectors $\mathbf{e}_g = \mathcal{F}(I_g^{sa}, \mathcal{I}_g^{sv}, \mathcal{L}_g)$. The representation can then be further utilized in downstream tasks, such as inferring socioeconomic indicators $\mathbf{Y}_g$ for the given region g. The detailed formulation and definitions related to this task are provided in Appendix A.

Urban Socioeconomic Indicator Prediction. In literature, many studies have focused on learning task-specific region representations from various urban data, especially urban imagery due to its consistent updates and easy accessibility (Liu et al. 2023; Xi et al. 2022; Chen et al.). For example, Urban2Vec (Wang, Li, and Rajagopal 2020) integrated *street-view imagery* and POI data to learn neighborhood embeddings. Some contrastive learning approaches like PG-SimCLR (Xi et al. 2022) and UrbanCLIP (Yan et al. 2024) have shown success in representing urban regions through *satellite images*. However, the aforementioned works exclusively considered a single type of urban imagery, *overlooking the potential synergy between satellite imagery and street-view imagery, which can complement each other.*

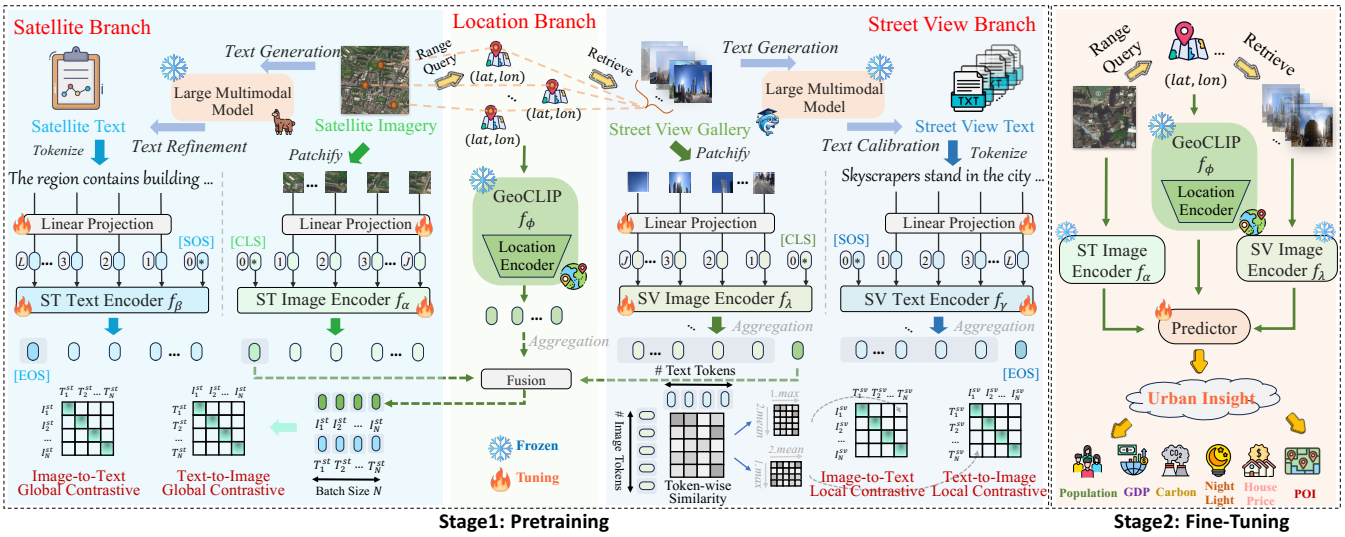

Figure 3: Overall framework of our proposed UrbanVLP.

Vision-Language Pretraining (VLP). VLP aims to jointly encode vision and language in a fusion model. A milestone work CLIP (Radford et al. 2021) and its variants (Li et al. 2021; Yao et al. 2022) highlight the efficacy of contrastive learning in cross-modal downstream tasks, such as zero-shot learning and cross-modal retrieval. Recent works (Tsimpoukelli et al. 2021; Alayrac et al. 2022) shift towards leveraging LLMs knowledge for VLP. *In USI prediction, the potential of VLP paradigm remains untapped, with limited exploration of the benefits of textual information.*

Methodology

Figure 3 shows our framework with two stages. In the pretraining stage, we first devise an automatic text generation and calibration module using ShareGPT4V (Chen et al. 2023) to generate textual descriptions for street-view images with geographical and visual prompts. To guarantee the quality of the generated text, we introduced a novel metric called `PerceptionScore`. Then a multi-granularity cross-modal alignment framework is designed to utilize multi-level contrastive learning and fine-grained information injection. During the fine-tuning stage, we employ frozen encoders from Stage 1 to extract features and fine-tune a lightweight MLP for accurate predictions.

Automatic Text Generation and Calibration

Text Generation. For each street-view image, we use advanced Large Multimodal Models (LMMs) due to their robust cross-modal knowledge retention abilities, to provide comprehensive textual descriptions. However, this confronts two primary challenges. Firstly, due to *substantial financial costs* associated with API usage, employing closed source LMMs like GPT4V (Yang et al. 2023) or Gemini (Fu et al. 2023) for generating tens of thousands of text descriptions is impractical. Secondly, street-view images often encompass diverse details, *necessitating a well-designed prompt template* for obtaining high-quality textual descriptions.

To address these issues, we first employ an open-source model closely related to GPT-4V, known as ShareGPT4V (Chen et al. 2023), to generate the street-view image descriptions. This model is a powerful LLM with vision capabilities trained on 100k GPT4V-generated captions, revealing captioning ability comparable to GPT4V. Moreover, in the process of designing text prompt templates, we believe that the proportion of each element in street-view images reflects their relative abundance, thereby facilitating LMMs in evaluating their significance. Therefore, we employ a pretrained segmentation model (CSAILVision 2024) to decouple visual elements and calculate the proportion of segmentation for each element. As evidenced by existing research (Fan et al. 2023b), the accuracy of current segmentation methods is already adequate to recognize various visual components. Simultaneously, geospatial coordinates of street-view images are incorporated as prompts to aid in the generation of more precise textual descriptions. Template and segmentation ratio details can be found in Appendix H.

Text Calibration. Existing works (Fan et al. 2023a; Yan et al. 2024) have already demonstrated that the quality of the textural descriptions is crucial for enhancing model performance. However, previous research (Yan et al. 2024) primarily employs simplistic rules and manual refinement processes for refining text description of satellite imagery, leading to potential unresolved hallucination problems (Rawte, Sheth, and Das 2023). Besides, the rule-based methods are overly general and fail to address specific issues in each image description as they are tailored to particular cases. On the other hand, manual rewriting is highly labor-intensive, which limits its scalability.

Therefore, an effective method for assessing the quality of generated text and calibrating the alignment between the text and image is essential to ensure a high level of LMM-inherent knowledge integration. Compared to classical methods that assess LMM-generated descriptions by

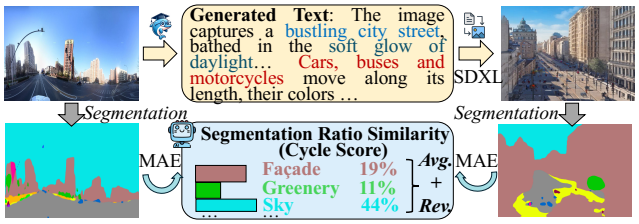

Figure 4: The procedure of `CycleScore` calculation.

relying on manually-annotated references to measure content overlap (Zhang et al. 2020; Lee et al. 2020), we devise an automatic mechanism to simulate human evaluation systems. This mechanism places greater emphasis on semantic consistency rather than merely assessing typical content overlap, by introducing a novel quality metric called `PerceptionScore`. Concretely, it comprises two parts: text semantic quality and visual recall quality. The former one directly adopts the efficient implementation of `CLIPScore` (Hessel et al. 2021), which is based on CLIP's strong zero-shot capabilities and demonstrates high correlation with human semantic judgment. However, `CLIPScore` solely measures textual quality without consideration of the omission of specific visual elements in the text, thereby *lacking an assessment of visual recall* (Hu et al. 2023).

To address this problem, we propose `CycleScore`, a metric designed to align fine-grained elemental information within the image I. As illustrated in Figure 4, given each description T generated by LMMs, we leverage the state-of-the-art open-sourced text-to-image model $\mathrm{SDXL}(\cdot)$ (Podell et al. 2023) to generate a consistent image I', reflecting various information contained in the text intuitively. Subsequently, we utilize a pretrained segmentation model $\mathrm{Seg}(\cdot)$ (CSAILVision 2024) to decouple consistent semantic elements (Fan et al. 2023b) in both the input and output images and calculate the Mean Absolute Error, *i.e.* $\mathrm{MAE}(\cdot)$ score, enforcing visual-semantic consistency. We formalize the process as follows:

$$I' = \mathrm{SDXL}(I), \tag{1}$$

$$\texttt{CycleScore}(I, I') = 1 - \mathrm{MAE}(\mathrm{Seg}(I), \mathrm{Seg}(I')), \tag{2}$$

$$\texttt{PerceptionScore}(I, T) = (\texttt{CLIPScore}(I, T) + \texttt{CycleScore}(I, I'))/2. \tag{3}$$

Multi-Granularity Cross-modal Alignment

Modality Representation. Utilizing automated text generation methods, we obtain a dataset of high-quality image-text pairs $D = (I, T)$. Here, I represents satellite images I^{st} or street-view images I^{sv}, while T indicates satellite text descriptions T^{st} or street-view text descriptions T^{sv}.

We first deploy Vision Transformer (ViT) (Dosovitskiy et al. 2021) as the image encoder to process images derived from satellite and street-view sources. Then we use the output $\mathbf{z}_l^0$ corresponding to the global tokens I_{cls} inserted at the beginning as the final output of the image encoder, rewrite as $\mathbf{z}_I$ for image representation. Concurrently, we leverage the basic Transformer-Encoder (Vaswani et al. 2017) as the text encoder to comprehend and encode the semantics of textual descriptions in parallel. Furthermore, considering that the geospatial location information associated with street-view images can assist in pinpointing the precise geospatial coordinates of the area, we utilize the open-sourced GeoCLIP's (Vivanco, Nayak, and Shah 2023) location encoder to capture longitude and latitude $\mathcal{L}_g$, generating semantically informative geospatial features, denoted as $\mathbf{z}_L$. Additional details can be found in Appendix G.

Modality Alignment. To capture the comprehensive visual information, beyond deriving global visual features for the region from satellite images, the aggregation of location-level clues within the region can enhance the injection of details. Therefore, we further incorporate fine-grained information from street-view images, along with the corresponding geographical features. The formalized process is as follows:

$$\mathbf{z}_{g_{st}} = f(\mathbf{z}_{I_{st}}, \mathrm{Aggr}(\mathbf{z}^1_{I_{sv}}, ..., \mathbf{z}^m_{I_{sv}}), \mathrm{Aggr}(\mathbf{z}^1_L, ..., \mathbf{z}^m_L)), \tag{4}$$

where $\mathbf{z}_{I_{st}}$ denotes the satellite visual features for the query region, $\mathbf{z}^k_{I_{sv}}$ is the k-th street-view visual feature in that region, and $\mathbf{z}^k_L$ represents the k-th street-view location feature in the same region. m indicates the count of street-view scenes within the region. $\mathrm{Aggr}(\cdot)$ denotes the aggregation method. The function f signifies the feature fusion method.

To establish a global-level semantic correspondence between the satellite features and their associated textual features, we employ a joint contrastive optimization approach for the image and text encoders in the satellite branch. For the i-th satellite image-text pair (I^{st}_i, T^{st}_i) in a mini-batch, we contrast the image-text pairs against others within the samples, which aims to maximally preserve the mutual information between the pairs in latent space. The global contrastive loss function $\mathcal{L}_{CG}$ is composed of two terms: $\mathcal{L}_{CG}^{Image \to Text}$ and $\mathcal{L}_{CG}^{Text \to Image}$, which measures the similarity between the visual and textual embeddings, respectively, defined as:

$$\mathcal{L}_{CG} = -\frac{1}{N}\left(\sum_i^N \log \frac{\exp(I_{g_i}^\top T_{g_i}/\tau)}{\sum_{j=0}^N \exp(I_{g_i}^\top T_{g_j}/\tau)} + \sum_i^N \log \frac{\exp(T_{g_i}^\top I_{g_i}/\tau)}{\sum_{j=0}^N \exp(T_{g_i}^\top I_{g_j}/\tau)}\right), \tag{5}$$

where I_{g_i} and T_{g_j} are the normalized embedding of satellite representation $\mathbf{z}_{g_{st}}$ in the i-th pair and that of textual representation $\mathbf{z}_{T_{st}}$ in the j-th pair, respectively. Besides, N is the batch size, and τ is the temperature for contrastive learning.

The street-view branch aims to enrich region embedding with fine-grained local-level information. However, previous approaches like (Radford et al. 2021; Jia et al. 2021) rely exclusively on global feature similarity within each modality, disregarding the necessity for fine-grained alignment, such as the correspondence between visual objects and textual tokens (Yao et al. 2022). To address this challenge, we leverage a fine-grained interaction mechanism to implement cross-modal alignment at the local level. Specifically, we utilize token-level maximum similarity between visual and tex-

tual tokens to direct the contrastive objective. We first compute the similarity between each visual token and all textual tokens, and then leverage the maximum value to calculate the average similarity of all image tokens to textual tokens. The similar approach is also applied to text-to-image process.

$$\mathrm{SIM}(v_i, t_i) = \frac{1}{l_1}(\sum_{k_1=1}^{l_1} \mathrm{argmax}_{k_2 \in [0,l_2)}(v_{ik_1}^\top t_{ik_2})), \quad (6)$$

$$\mathrm{SIM}(t_i, v_i) = \frac{1}{l_2}(\sum_{k_2=1}^{l_1} \mathrm{argmax}_{k_2 \in [0,l_1)}(t_{ik_1}^\top v_{ik_2})), \quad (7)$$

where v_i and t_j denote the normalized embedding of street-view representation $\mathbf{z}_{I_{sv}}$ in the i-th pair and that of textual representation $\mathbf{z}_{T_{sv}}$ in the j-th pair, respectively. The fine-grained token-level representation can be optimized via:

$$\mathcal{L}_{CL} = -\frac{1}{N}\left(\sum_i^N \log \frac{\exp(\mathrm{SIM}(v_i^\top, t_i)/\tau)}{\sum_{j=0}^N \exp(\mathrm{SIM}(v_i^\top, t_j)/\tau)} + \sum_i^N \log \frac{\exp(\mathrm{SIM}(t_i^\top, v_i)/\tau)}{\sum_{j=0}^N \exp(\mathrm{SIM}(t_i^\top, v_j)/\tau)}\right). \quad (8)$$

Pretraining & Fine-Tuning

Pretraining Stage. The overall objective of UrbanVLP can be defined as the joint optimization of the above two losses:

$$\mathcal{L}_{Total} = \alpha\mathcal{L}_{CG} + \beta\mathcal{L}_{CL}. \quad (9)$$

where α and β are hyperparameters for a trade-off. Through backpropagation optimization, we achieve multi-granularity cross-modal alignment, resulting in robust encoders.

Fine-Tuning Stage. As depicted in Figure 3 Right, during the fine-tuning stage, we employ a linear probing approach (He et al. 2022) which begins by extracting $\mathbf{e}_{st}, \mathbf{e}_{sv}, \mathbf{e}_p$ features from the pretrained encoder for satellite images, street-view images, and street-view positions. Subsequently, these features are fused together, and a minimalist classifier (MLP) is trained on top to fine-tune the prediction of urban metrics, denoted as $\mathbf{Y}_i = MLP(\mathbf{e}_{st}, \mathbf{e}_{sv}, \mathbf{e}_p)$. It is noteworthy that in downstream tasks, textual information is unnecessary since the knowledge embedded in the text has already been imparted to other modality encoders through pretraining, and additional text generation time hinders real-time application.

Experiments

Experimental Setup

Dataset & Task Description. Given the current lack of open-sourced datasets in the research community, we introduce a new benchmark dataset named `CityView`, which will be released upon paper notification. `CityView` is unique in that it comprises a dual-category structure encompassing both satellite and street-view image components, each paired with corresponding high-quality textual descriptions. We adhere to (Yan et al. 2024) to cover core area data of four cities in China. A comprehensive overview of `CityView`, along with relevant statistics, is provided in Appendix C.

Baselines. Following the established practice (Yan et al. 2024; Xi et al. 2022; Liu et al. 2023) in this area, we compare our method with seven recent baselines in the field of imagery-based USI prediction. The single-granularity urban imagery-based methods include: **ViT** (Dosovitskiy et al. 2021), **PG-SimCLR** (Xi et al. 2022), **UrbanCLIP** (Yan et al. 2024), **UrbanCLIP-SV** (Yan et al. 2024).

We also apply multiple baselines for multi-granularity urban imagery-based USI prediction: **Vision-LSTM** (Huang et al. 2023) and **StructuralUrban** (Li et al. 2022). They are all trained using data from our `CityView` dataset. The in-depth overview of baselines is presented in Appendix E.

Evaluation Metrics. We follow (Yan et al. 2024) to evaluate our method in terms of: the coefficient of determination R^2, the root mean squared error (RMSE), and the mean absolute error (MAE). An increase in R^2, along with a decrease in RMSE and MAE values, signifies improved model accuracy.

Implementation Details. The implementation details are provided in Appendix F.

Performance Evaluation

To evaluate our UrbanVLP framework, we conduct comparisons with existing state-of-the-art methods on our proposed `CityView` datasets. Table 1 presents the overall results, from which we can obtain the following findings:

1) UrbanVLP significantly outperforms the baselines which employ single/multi-granularity imagery. It can be seen that UrbanVLP surpasses the best baseline (StructuralUrban) by 3.3%, 2.3%, 2.5%, 4.9%, 1.0% and 2.4% in terms of R^2 for all six indicators: Carbon, Population, GDP, Night Light, House Price and POI in Beijing. Similarly, we can witness consistent improvements in the other three cities. In addition, the average reduction of UrbanVLP in terms of RMSE and MAE on the Beijing dataset are 1.8% and 2.1%, respectively. For a more intuitive display, we visualize the overall performance on Beijing and Shenzhen in Figure 2, where the results further prove the versatility of our framework for USI prediction.

2) Integrating multi-granularity imagery can lead to superior performance against baselines. This merit can be rationalized by the incorporation of additional fine-grained information derived from street-view imagery. Meanwhile, UrbanCLIP-SV underperforms UrbanCLIP, revealing that street-view imagery (while rich in details) lacks a macro view, resulting in suboptimal outcomes when used in isolation. This also underscores the irreplaceability of satellite imagery serving as a macro visual modality.

3) Compared to other modalities (e.g., POIs), the textual modality facilitates a more comprehensive understanding of the region. We include various baselines for comparison, such as PGSimCLR which incorporates POI distributions to integrate external information and enhance regional representations. Despite their promising results, our UrbanVLP illustrates the efficacy of integrating textual data as a information-compact modality (He et al. 2022), which fully leverages the inherent knowledge of LLMs and explicitly enhances interpretability in the training phase.

Ablation Studies

As shown in Figure 5, we conduct ablation studies to examine each component in UrbanVLP on the `CityView`-Beijing

	Methods	UrbanVLP			StructuralUrban			Vision-LSTM			UrbanCLIP-SV			UrbanCLIP			PG-SimCLR			ViT		
	Metric	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Beijing	Carbon	**0.769**	**0.477**	**0.369**	0.736	0.518	0.405	0.674	0.519	0.433	0.489	0.713	0.548	0.703	0.541	0.539	0.430	0.797	0.632	0.411	0.810	0.607
	Population	**0.714**	**0.523**	**0.411**	0.691	0.545	0.427	0.640	0.591	0.518	0.435	0.734	0.581	0.655	0.576	0.459	0.476	1.228	0.963	0.442	0.861	0.635
	GDP	**0.537**	**0.684**	**0.416**	0.512	0.694	0.426	0.497	0.717	0.469	0.188	0.910	0.568	0.514	0.694	0.445	0.270	1.679	1.067	0.265	1.073	0.730
	Night Light	**0.470**	**0.668**	**0.403**	0.421	0.696	0.459	0.354	0.747	0.475	0.304	0.769	0.483	0.377	0.741	0.468	0.367	0.728	0.404	0.358	0.733	0.523
	House Price	**0.503**	**0.644**	**0.482**	0.493	0.649	0.485	0.471	0.658	0.495	0.451	0.674	0.515	0.501	0.647	0.486	0.341	0.718	0.555	0.332	0.719	0.569
	POI	**0.299**	**0.723**	**0.374**	0.275	0.725	0.380	0.235	0.728	0.388	0.226	0.729	0.407	0.251	0.734	0.385	-	-	-	0.215	0.742	0.411
Shanghai	Carbon	**0.718**	0.520	**0.381**	0.678	0.550	0.432	0.571	0.606	0.455	0.576	**0.465**	0.450	0.673	0.560	0.426	0.270	0.742	0.551	0.254	0.758	0.522
	Population	**0.589**	**0.609**	**0.476**	0.545	0.631	0.488	0.518	0.659	0.522	0.400	0.793	0.592	0.533	0.650	0.511	0.288	1.051	0.831	0.279	0.826	0.627
	GDP	0.323	0.805	0.593	0.312	0.840	0.613	0.311	1.008	0.661	0.323	1.214	0.741	**0.334**	**0.804**	**0.586**	0.270	1.679	1.067	0.263	1.221	0.735
	Night Light	**0.435**	**0.683**	**0.490**	0.411	0.690	0.497	0.352	0.734	0.550	0.274	0.765	0.556	0.384	0.710	0.515	0.228	0.756	0.495	0.209	0.768	0.561
	POI	**0.381**	**0.680**	**0.386**	0.375	0.685	0.392	0.246	0.704	0.390	0.240	0.711	0.402	0.265	0.710	0.395	-	-	-	0.234	0.713	0.406
Guangzhou	Carbon	**0.701**	**0.513**	**0.372**	0.655	0.550	0.504	0.554	0.564	0.457	0.425	0.520	0.473	0.585	0.603	0.440	0.413	0.951	0.716	0.374	0.665	0.528
	Population	**0.655**	**0.573**	**0.454**	0.631	0.587	0.459	0.614	0.619	0.509	0.491	0.693	0.622	0.627	0.596	0.473	0.294	1.046	2.112	0.263	0.821	0.661
	GDP	0.442	0.755	**0.541**	0.429	0.781	0.549	0.412	0.997	0.624	0.327	0.998	0.690	**0.446**	**0.752**	0.546	0.263	1.157	1.553	0.249	1.122	0.719
	Night Light	**0.548**	**0.594**	**0.428**	0.508	0.607	0.448	0.381	1.021	0.629	0.314	0.692	0.524	0.463	0.651	0.483	0.436	0.649	0.432	0.357	0.642	0.505
	POI	**0.438**	**0.732**	**0.333**	0.357	0.758	0.359	0.175	0.811	0.401	0.169	0.826	0.405	0.185	0.802	0.401	-	-	-	0.136	0.904	0.427
Shenzhen	Carbon	**0.625**	0.605	**0.464**	0.587	**0.593**	0.481	0.534	0.609	0.516	0.460	0.621	0.529	0.541	0.607	0.526	0.234	0.975	0.746	0.221	0.727	0.574
	Population	**0.790**	**0.452**	**0.348**	0.724	0.476	0.386	0.624	0.562	0.471	0.589	0.604	0.484	0.727	0.510	0.392	0.294	1.046	2.112	0.280	0.832	0.654
	GDP	**0.533**	**0.682**	**0.447**	0.489	0.697	0.471	0.462	0.723	0.517	0.433	0.755	0.577	0.508	0.693	0.464	0.294	1.046	2.112	0.254	1.226	0.779
	Night Light	**0.457**	**0.667**	**0.459**	0.421	0.694	0.502	0.404	0.702	0.508	0.320	0.717	0.531	0.387	0.709	0.511	0.247	0.934	0.738	0.254	0.715	0.530
	POI	**0.461**	**0.752**	**0.370**	0.321	0.780	0.389	0.254	0.795	0.412	0.147	0.849	0.453	0.185	0.838	0.436	-	-	-	0.137	0.954	0.455
	1st Count	56			1			0			1			5			0			0		

Table 1: Socioeconomic indicators prediction results in four datasets. The best results are in bold and the second-best results are underlined.

dataset, including the Satellite branch (ST branch), Street-View branch (SV branch), and Location Encoding branch (LE branch). More discussion can be found in Appendix J.

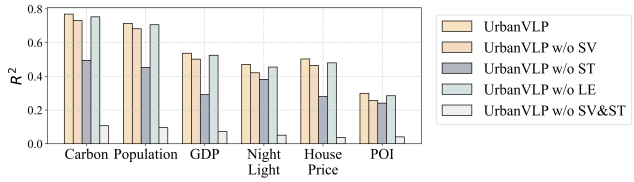

Figure 5: Ablation study on `CityView-Beijing`.

Effects of Street-View Branch. One of our contributions is the incorporation of multi-granularity information that reflects the urban spatial hierarchy. It can be observed that the incorporation of street-view branch results in an average improvement of 3.56% in terms of R^2. This enhancement is attributed to the street-view branch's ability to capture rich, detailed information, thereby improving modeling precision.

Effects of Textual Modality. In Table 1, we demonstrate the function of the textual modality by comparing UrbanVLP with a standard ViT-based model, which shares the same configuration as the unimodal visual encoder of UrbanVLP. We then utilize the visual features extracted without textual augmentation to predict downstream USI. As we can see, the lack of textual information leads to substantial performance degradation, underscoring the critical role of textual modalities in attaining a comprehensive visual representation. Similar findings were also reported in (Yan et al. 2024).

Effects of Location Coordinates. In Figure 5, we evaluate the performance of the model without the Location Encoding branch. Although not as impactful as visual modalities, the inclusion of geospatial locations has yielded an average improvement of 1.5% in R^2. It is noteworthy that we utilize a well-pretrained location encoder and freeze it within our framework. Consequently, its generalizability is ensured, preventing overfitting to specific cities in `CityView` dataset.

Qualitative Analysis

We further investigate the quality of generated texts and the predictive performance of UrbanVLP in practice. *More empirical analysis can be found in Appendix J.*

Illustration of the quality of our generated descriptions. To visually illustrate the quality of the generated description, we present an example in Figure 6. The generated description of the street-view image (a) is depicted in (b). Subsequently, we employ GPT4V (Yang et al. 2023) to assess the quality score of the generated description, which is rated as 7 out of 10, indicating the effectiveness of the generated text. Furthermore, GPT4V also provides specific areas for improvement in (c). In (d), we utilize GPT4V to generate an image based on the description in (b). It is evident that the generated image bears a resemblance to the original one.

Case Study for Predicted Results. Here we also show some predicted results in Figure 7 on `CityView-Beijing` dataset. Satellite images (a) and (b), though similar in layout, differ significantly in land use. (a) encompasses residential and campus areas, whereas (b) represents industrial zones, leading to entirely different socioeconomic characteristics. The carbon emissions and GDP of campus and residential area are significantly lower than those of industrial parks, whereas the disparities in population are not as pronounced, possibly due to the unique characteristics of school district housing near educational institutions.

As observed, UrbanCLIP's predictions fail to effectively differentiate between the two, as the downstream metrics in Figure 7(b) still lean towards predicting a homogenized so-

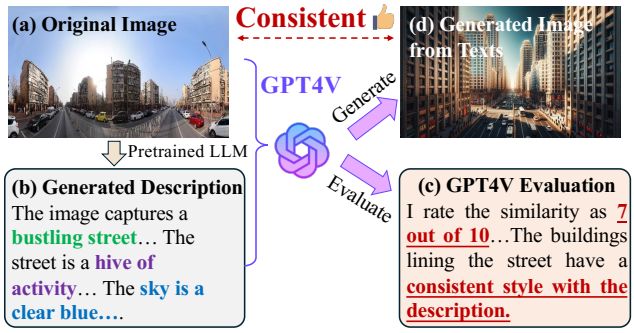

Figure 6: An example of generated text descriptions.

cioeconomic area similar to (a). In contrast, the results from UrbanVLP are capable of distinctly distinguishing the socioeconomic attributes of the two in terms of carbon emission and GDP.

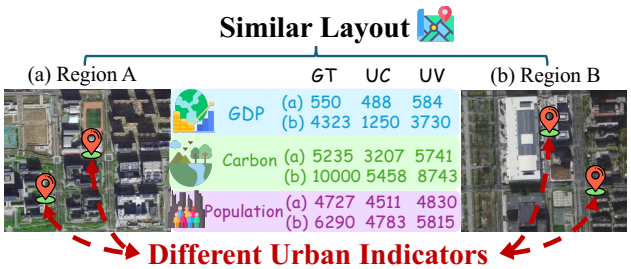

Figure 7: Case study of UrbanCLIP (UC) and our UrbanVLP (UV).

Visualization of Region Representations

In this section, we map the region representations learned in UrbanVLP into a two-dimensional space using the PCA algorithm in Figure 8. As we can see, the three images on the left and on the right belong to different clusters, each exhibiting intra-cluster similarities and inter-cluster differences. Therefore, UrbanVLP could effectively model regions into high-dimensional space, wherein satellite images with similar architectural layouts demonstrate spatial similarity. Our framework learns a reasonable, accurate, and semantically rich region representation for USI prediction.

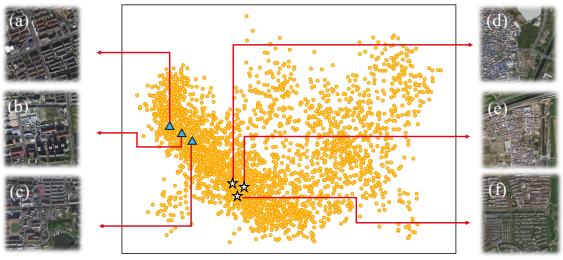

Figure 8: Representation space visualization.

Transferability Study

In this study, we explore the practical application of USI prediction within a transfer learning context (Park et al. 2022; Liu et al. 2023). Specifically, We use visual encoders trained on source city data and fine-tune them on target city data. We experiment with different pairs of source and target cities and present the carbon emission prediction in Figure 9. As we can see, on 16 source-target city pairs, UrbanVLP achieves an average R^2 of 0.574, while that of UrbanCLIP is 0.495. Additionally, the R^2 values show greater similarity between Beijing and Shanghai, as well as between Shenzhen and Guangzhou. This observation aligns with the geographical proximity of Beijing to Shanghai and Guangzhou to Shenzhen. These findings confirm the robust transferability of our UrbanVLP model in urban areas, despite the previously mentioned differences in geological and demographic characteristics among the selected cities.

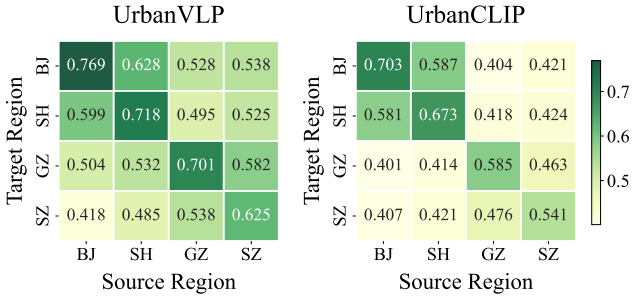

Figure 9: Transferability test on R^2 between UrbanVLP and UrbanCLIP, on the Carbon indicator across 4 cities.

Conclusion and Future Work

USI prediction plays a significant role for understanding societal patterns and dynamics. UrbanVLP, for the first time, explores the differences between street-view and satellite images from a semantic granularity perspective, as well as their roles in modeling urban region representation. A text generation and calibration mechanism is also proposed to ensure high-quality description generation. It has achieved state-of-the-art results on the constructed `CityView` dataset. Future research could incorporate additional modalities such as POIs, road networks, and building footprints to enhance information richness and introduce a broader perspective.

Acknowledgments

This work is mainly supported by the National Natural Science Foundation of China (No. 62402414). This work is also supported by the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2024A03J0620), Guangzhou Municipal Science and Technology Project (No. 2023A03J0011), the Guangzhou Industrial Information and Intelligent Key Laboratory Project (No. 2024A03J0628), and a grant from State Key Laboratory of Resources and Environmental Information System, and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007).

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Ayush, K.; Uzkent, B.; Burke, M.; Lobell, D.; and Ermon, S. 2020. Generating interpretable poverty maps using object detection in satellite images. *arXiv preprint arXiv:2002.01612*.

Ayush, K.; Uzkent, B.; Tanmay, K.; Burke, M.; Lobell, D.; and Ermon, S. 2021. Efficient poverty mapping from high resolution remote sensing images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 12–20.

Bai, L.; Huang, W.; Zhang, X.; Du, S.; Cong, G.; Wang, H.; and Liu, B. 2023. Geographic mapping with unsupervised multi-modal representation learning from VHR images and POIs. *ISPRS Journal of Photogrammetry and Remote Sensing*, 201: 193–208.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. *arXiv preprint arXiv:2311.12793*.

Chen, W.; Hao, X.; Liang, Y.; et al. ???? Terra: A Multimodal Spatio-Temporal Dataset Spanning the Earth. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chen, W.; Huang, C.; Yu, Y.; Jiang, Y.; and Dong, J. 2024. Trajectory-User Linking via Hierarchical Spatio-Temporal Attention Networks. *ACM Transactions on Knowledge Discovery from Data*, 18(4): 1–22.

CSAILVision. 2024. GitHub - CSAILVision/semantic-segmentation-pytorch: Pytorch implementation for Semantic Segmentation/Scene Parsing on MIT ADE20K dataset — github.com. https://github.com/CSAILVision/semantic-segmentation-pytorch. [Accessed 22-05-2024].

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

Fan, L.; Krishnan, D.; Isola, P.; Katabi, D.; and Tian, Y. 2023a. Improving CLIP Training with Language Rewrites. In *NeurIPS*.

Fan, Z.; Zhang, F.; Loo, B. P.; and Ratti, C. 2023b. Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proceedings of the National Academy of Sciences*, 120(27): e2220417120.

Fu, C.; Zhang, R.; Lin, H.; Wang, Z.; Gao, T.; Luo, Y.; Huang, Y.; Zhang, Z.; Qiu, L.; Ye, G.; et al. 2023. A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv preprint arXiv:2312.12436*.

Gao, J.; Chen, B.; Zhao, X.; Liu, W.; Li, X.; Wang, Y.; Zhang, Z.; Wang, W.; Ye, Y.; Lin, S.; et al. 2024. LLM-enhanced Reranking in Recommender Systems. *arXiv preprint arXiv:2406.12433*.

Han, S.; Ahn, D.; Park, S.; Yang, J.; Lee, S.; Kim, J.; Yang, H.; Park, S.; and Cha, M. 2020. Learning to score economic development from satellite imagery. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2970–2979.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, Z.; Yang, S.; Zhang, W.; and Zhang, J. 2018. Perceiving commerial activeness over satellite images. In *Companion Proceedings of the The Web Conference 2018*, 387–394.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*.

Hu, A.; Chen, S.; Zhang, L.; and Jin, Q. 2023. InfoMetIC: An Informative Metric for Reference-free Image Caption Evaluation. *arXiv preprint arXiv:2305.06002*.

Huang, Y.; Zhang, F.; Gao, Y.; Tu, W.; Duarte, F.; Ratti, C.; Guo, D.; and Liu, Y. 2023. Comprehensive urban space representation with varying numbers of street-level images. *Computers, Environment and Urban Systems*, 106: 102043.

Jaiswal, A.; Babu, A. R.; Zadeh, M. Z.; Banerjee, D.; and Makedon, F. 2020. A survey on contrastive self-supervised learning. *Technologies*, 9(1): 2.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Jiang, H.; Hao, X.; Huang, Y.; Ma, C.; Zhang, J.; Pan, Y.; and Zhang, R. 2024. Advancing Medical Radiograph Representation Learning: A Hybrid Pre-training Paradigm with Multilevel Semantic Granularity. *arXiv preprint arXiv:2410.00448*.

Lee, H.; Yoon, S.; Dernoncourt, F.; Kim, D. S.; Bui, T.; and Jung, K. 2020. Vilbertscore: Evaluating image caption using vision-and-language bert. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, 34–39.

Li, L.; Ye, Y.; Jiang, B.; and Zeng, W. 2024a. GeoReasoner: Geo-localization with Reasoning in Street Views using a Large Vision-Language Model. In *Forty-first International Conference on Machine Learning*.

Li, S.; Chen, W.; Wang, B.; Huang, C.; Yu, Y.; and Dong, J. 2024b. MCN4Rec: Multi-level Collaborative Neural Network for Next Location Recommendation. *ACM Trans. Inf. Syst.*, 42(4).

Li, T.; Xin, S.; Xi, Y.; Tarkoma, S.; Hui, P.; and Li, Y. 2022. Predicting multi-level socioeconomic indicators from structural urban imagery. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3282–3291.

Li, Y.; Liang, F.; Zhao, L.; Cui, Y.; Ouyang, W.; Shao, J.; Yu, F.; and Yan, J. 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.

Liang, Y.; Ke, S.; Zhang, J.; Yi, X.; and Zheng, Y. 2018. Geoman: Multi-level attention networks for geo-sensory time series prediction. In *IJCAI*, volume 2018, 3428–3434.

Liang, Y.; Ouyang, K.; Sun, J.; Wang, Y.; Zhang, J.; Zheng, Y.; Rosenblum, D.; and Zimmermann, R. 2021. Fine-grained urban flow prediction. In *Proceedings of the Web Conference 2021*, 1833–1845.

Liu, Y.; Zhang, X.; Ding, J.; Xi, Y.; and Li, Y. 2023. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. In *Proceedings of the ACM Web Conference 2023*, 4150–4160.

M Rustowicz, R.; Cheong, R.; Wang, L.; Ermon, S.; Burke, M.; and Lobell, D. 2019. Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 75–82.

Manvi, R.; Khanna, S.; Mai, G.; Burke, M.; Lobell, D.; and Ermon, S. 2023. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*.

Martinez, J. A. C.; La Rosa, L. E. C.; Feitosa, R. Q.; Sanches, I. D.; and Happ, P. N. 2021. Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 171: 188–201.

Park, S.; Han, S.; Ahn, D.; Kim, J.; Yang, J.; Lee, S.; Hong, S.; Kim, J.; Park, S.; Yang, H.; et al. 2022. Learning economic indicators by aggregating multi-level geospatial information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12053–12061.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rawte, V.; Sheth, A.; and Das, A. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Rußwurm, M.; and Körner, M. 2020. Self-attention for raw optical satellite time series classification. *ISPRS journal of photogrammetry and remote sensing*, 169: 421–435.

Tsimpoukelli, M.; Menick, J. L.; Cabi, S.; Eslami, S.; Vinyals, O.; and Hill, F. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34: 200–212.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vivanco, V.; Nayak, G. K.; and Shah, M. 2023. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization.

Wang, Z.; Li, H.; and Rajagopal, R. 2020. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1013–1020.

Xi, Y.; Li, T.; Wang, H.; Li, Y.; Tarkoma, S.; and Hui, P. 2022. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests. In *Proceedings of the ACM Web Conference 2022*, 3308–3316.

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

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1).

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2022. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *International Conference on Learning Representations*.

Yeh, C.; Meng, C.; Wang, S.; Driscoll, A.; Rozi, E.; Liu, P.; Lee, J.; Burke, M.; Lobell, D. B.; and Ermon, S. 2021. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. *arXiv preprint arXiv:2111.04724*.

Zhang, M.; Li, T.; Li, Y.; and Hui, P. 2021. Multi-view joint graph representation learning for urban region embedding. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 4431–4437.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhang, W.; Han, J.; Xu, Z.; Ni, H.; Liu, H.; and Xiong, H. 2024. Towards Urban General Intelligence: A Review and Outlook of Urban Foundation Models. *arXiv preprint arXiv:2402.01749*.

Zou, X.; Yan, Y.; Hao, X.; Hu, Y.; Wen, H.; Liu, E.; Zhang, J.; Li, Y.; Li, T.; Zheng, Y.; et al. 2024. Deep Learning for Cross-Domain Data Fusion in Urban Computing: Taxonomy, Advances, and Outlook. *arXiv preprint arXiv:2402.19348*.