arXiv:2608.21041v1 [cs.CV] 21 Aug 2026

CoST: Semantic-Aware Urban Understanding via Spatial-Temporal Alignment

Yutian Jiang*
yjiang194@connect.hkust-gz.edu.cn
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Jiabo Liu*
jliu933@connect.hkust-gz.edu.cn
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Xixuan Hao
xhao390@connect.hkust-gz.edu.cn
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Yuxuan Liang
yuxliang@outlook.com
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Abstract

Geospatial representation learning from satellite imagery is a fundamental problem for large-scale urban analysis and real-world applications. Despite recent advances, current methods struggle with cross-region generalization and semantic interpretability due to their reliance on region-specific auxiliary data and the neglect of semantic alignment within multi-temporal urban imagery. Therefore, we present CoST, a novel Contrastive-based Spatial-Temporal framework that aligns spatial context with multi-temporal semantics to extract universal geographic regularities shared across regions. Specifically, CoST explicitly models spatial correlations to capture transferable geographic structures and exploits multi-year urban change semantics to align learned representations with high-level geo-semantics. Extensive experiments demonstrate that CoST consistently achieves superior performance across various downstream tasks and in unseen scenario, yielding an average relative gain of 8.7% over the strongest competing methods across eight city-indicator settings. The code is available in this repo.

CCS Concepts

• **Information systems** → **Multimedia information systems**.

Keywords

Remote Sensing, Contrastive Learning, Urban Indicator Prediction

ACM Reference Format:
Yutian Jiang, Jiabo Liu, Xixuan Hao, and Yuxuan Liang. 2026. CoST: Semantic-Aware Urban Understanding via Spatial-Temporal Alignment. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3799682.3841155

*Yutian Jiang and Jiabo Liu contributed equally to this research.

This work is licensed under a Creative Commons Attribution 4.0 International License.
CIKM '26, Rome, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3841155

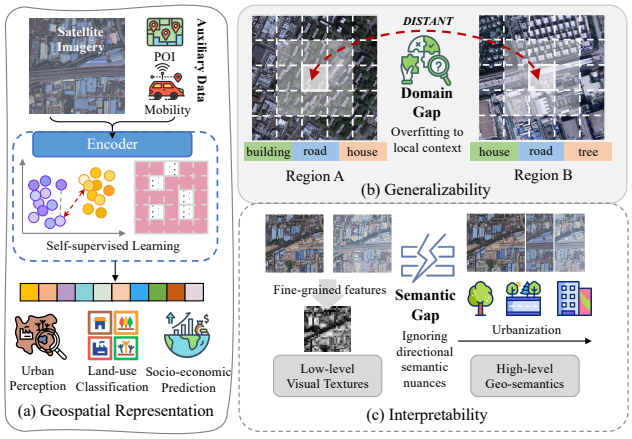

Figure 1: Geospatial representation and its challenges.

1 Introduction

Geospatial representation learning aims to project multi-source data into a unified latent space, effectively capturing intrinsic spatial correlations and temporal dynamics to facilitate robust urban understanding [35]. As illustrated in Figure 1 (a), a meaningful geospatial representation should capture not only local visual patterns but also the higher-level geographic regularities that shape urban structure and its evolution over time [14]. The learned geospatial representations can support a wide range of downstream applications, such as urban perception and socio-economic assessment [31]. Moreover, the quality of these representations fundamentally determines their ability to generalize and transfer across diverse geographic regions and tasks. Therefore, learning transferable and semantically meaningful geospatial representations has become a core problem in data-driven urban system understanding [44].

Recently, satellite imagery has emerged as a primary modality for geospatial representation learning that provides large-scale, geographically consistent observations of the Earth's surface [11]. Existing research on satellite-based representation learning generally follows two paradigms: (1) *Multimodal Semantic Alignment*, which incorporates socio-economic semantics by aligning satellite imagery with auxiliary data, such as POIs [16, 28] or mobility data [41, 43]. While effective for region-specific profiling and real-world

socio-economic prediction, these methods depend heavily on external data sources whose availability, quality, and distribution vary substantially across cities and countries, which limits scalability and weakens transferability. (2) *Self-supervised Visual Pre-training*, which focuses on extracting visual features by mining invariant textural patterns directly from raw imagery. Although such methods produce strong visual encoders for low-level tasks, they primarily emphasize appearance consistency and texture discrimination, with limited explicit grounding in high-level geo-semantics. As a result, the learned features are often effective as visual descriptors but less informative for semantically demanding urban analysis.

These limitations are not independent, pointing to a more fundamental gap in current geospatial representation learning: a representation should generalize across regions with different visual styles, development patterns, and functional compositions, while also remaining semantically interpretable enough to reflect meaningful urban transitions. This leads to two central challenges in learning universal geospatial representations: (i) generalization under cross-region distribution shifts, and (ii) semantic interpretability under weak supervision.

- **Cross-region Generalization**. Cross-region transfer remains a key challenge in urban representation learning [11, 13, 46]. Methods relying on auxiliary signals, such as points of interest (POIs) or mobility data, suffer from limited generalization due to substantial distributional discrepancies across cities, causing learned representations to become entangled with region-specific characteristics. A model trained in one source city may therefore suffer severe performance degradation when transferred to visually and functionally dissimilar cities, or when deployed on downstream tasks that differ from those seen during training. As shown in Figure 1 (b), models trained on dense metropolitan environments such as New York City may struggle to generalize to suburban regions because their representations become biased toward city-specific distributions. Moreover, current multimodal approaches require data recollection and model retraining for each new region [8], highlighting the need for representations that capture transferable geographic structures rather than region-specific correlations.
- **Semantic Interpretability**. Multi-temporal satellite imagery naturally contains rich semantic transition cues across time and space. Such transitions vary across locations (e.g., *land* → *road* versus *land* → *building*) and exhibit temporal continuity at the same location, such as *land* → *road* → *building*. Modeling these dynamic visual patterns together with explicit geo-semantics is therefore critical for learning interpretable geospatial representations. Yet this remains difficult because dense semantic annotations over long temporal horizons are scarce. Although prior work has incorporated temporal information, most existing methods reduce temporal variation to a binary change/no-change signal, which provides only weak supervision and fails to characterize the underlying semantic transitions. As a result, the learned representations are ambiguous and lack semantic interpretability, as illustrated in Figure 1(c).

To bridge the gap, we propose **CoST**, a novel Contrastive-based Spatial-Temporal framework for geospatial representation, which couples space and time for cross-region generalization and semantic interpretability. The key idea is that robust urban representations should be shaped by local spatial proximity and temporal consistency of semantic transitions in neighboring regions. To improve cross-region generalization, CoST introduces a spatial neighborhood modeling strategy inspired by Tobler's First Law of Geography [30], which posits that geographic proximity implies semantic similarity. This strategy preserves local continuity while remaining sensitive to instance discrimination, enabling the model to capture invariant spatial structures shared across diverse urban environments. To enhance semantic interpretability, CoST treats multi-temporal imagery as a semantic pretext through encoding temporal dynamics with embeddings that capture both the *semantic type* and *change extent* of transitions, providing richer semantic supervision than binary change detection and enabling more interpretable representations. Finally, CoST introduces a spatial-temporal alignment mechanism that enforces consistency among neighboring regions exhibiting similar change trajectories, thereby linking spatial similarity with temporal semantics and improving the coherence of the learned representation space.

Overall, the contributions can be summarized as follows:

- We propose CoST, a unified framework for learning geospatial representations from satellite imagery by jointly modeling spatial neighborhood structure and temporal semantic transitions, with the goal of improving both cross-region generalization and semantic interpretability.
- We introduce three complementary components: *spatial neighborhood modeling* to capture transferable local geographic structure, *temporal semantic guidance* to encode multi-year semantic transitions with both change type and change extent, and *spatial-temporal alignment* to couple spatial proximity with temporal consistency in a shared latent space.
- We curate a spatial-temporal aligned satellite imagery dataset spanning 11 years and multiple major cities, and conduct extensive experiments across cross-city and cross-task settings. The results show that CoST consistently outperforms strong baselines and yields more semantically structured embeddings for urban analysis.

2 Related Works

Multimodal Alignment. This line of research learns urban or regional representations by injecting external semantic knowledge into satellite imagery through feature fusion or cross-modal alignment [49]. The core idea is that auxiliary modalities such as POIs, human mobility, and textual descriptions provide complementary semantics about land use, human activity, and urban dynamics [15]. Representative studies align satellite imagery with POI semantics [18, 37], mobility patterns [17, 42], or graph-structured urban context [1, 34, 48] to learn semantically enriched region embeddings. More recent methods further adopt contrastive alignment objectives to associate satellite imagery with textual or functional descriptions, improving region profiling and urban indicator prediction [3, 21, 39]. Despite their strong performance, these methods fundamentally depend on auxiliary data sources whose coverage and quality vary substantially across cities. As a result, their learned

representations are entangled with region-specific external knowledge, which limits scalability and weakens transferability to new regions where such signals are sparse, noisy, or unavailable.

Visual Pre-training. Visual pre-training aims to learn generic representations directly from images. Recent geospatial foundation models exploit self-supervised objectives such as contrastive learning and masked image modeling to capture invariant visual patterns from large-scale image data. Contrastive approaches learn instance discrimination and augmentation invariance from raw imagery, while Masked Autoencoder (MAE) based methods reconstruct missing image patches to model local texture and spatial structure. Representative examples include SatMAE [5], SatMAE++ [22], and ScaleMAE [24], which have shown strong transferability across land-usage classification or segmentation tasks. However, their training signals are usually defined by visual consistency within imagery itself, without explicitly constraining the representation space to reflect urban semantics. Despite learning powerful generic visual features, these methods typically produce embeddings that are less structured for high-level urban understanding and socio-economic inference.

Spatial-Temporal Modeling. Temporal imagery provides a natural signal for modeling the inherent dynamics of urban regions. Contrastive learning approaches [19, 20] learn temporal invariance by treating observations from different dates as related views, improving robustness for downstream visual tasks, and MAE-based methods [9] incorporate multi-temporal observations to enhance large-scale visual understanding. Recent studies in geospatial foundation modeling, such as Prithvi [29], also show that temporal stacks can substantially improve adaptation across different tasks by pre-training over multi-temporal satellite sequences. Nevertheless, most existing temporal methods still treat geospatial tiles as isolated instances, overlooking the coupling of spatial-temporal dimensions. This oversight causes a disconnect between low-level visual textures and high-level geospatial semantics, leading to a performance trade-off: models may excel at low-level vision tasks but struggle with socio-economic inference. Consequently, they fail to achieve a unified representation space that is robust for various downstream applications.

3 Methodology

3.1 Preliminaries

Satellite Image. We formally define geospatial representation learning from satellite imagery. Let $\mathcal{D} = \{I_{i,t}\}$ denote a collection of satellite images, where $i \in \{1, \dots, N\}$ indexes discrete geospatial regions and $t \in \mathcal{T} = \{1, \dots, T\}$ indexes time. For a given region i, a satellite image observed at time t is denoted as $I_{i,t} \in \mathbb{R}^{H \times W \times C}$, where H and W are the image height and width, and C is the number of channels. Accordingly, the multi-temporal satellite observations of region i are represented as a sequence $\mathcal{X}_i = \{I_{i,1}, I_{i,2}, \dots, I_{i,T}\}$ over the temporal period $\mathcal{T}$.

Geospatial Regions. Each satellite tile is associated with a geographic coordinate $\mathbf{l}_i \in \mathbb{R}^2$. Based on geographic distance, we define the spatial neighborhood of region i as $\mathcal{N}_i = \{j_1, j_2, \dots, j_K\}$, where each j_k denotes one of its K nearest neighboring regions. Therefore, each sample is characterized by both its temporal observations $\mathcal{X}_i$ and its spatial context $\mathcal{N}_i$. This formulation provides the basic spatial-temporal structure used in our framework.

Problem Statement. Given a satellite tile $I_{i,t}$, the goal is to learn a generalizable visual encoder $\mathcal{E}_\theta$ that maps it to a representation vector $\mathbf{z}_{i,t} = \mathcal{E}_\theta(I_{i,t}) \in \mathbb{R}^d$. During pre-training, the encoder is optimized without downstream labels. After pre-training, the encoder is transferred to downstream tasks with frozen representations, and a lightweight task-specific head $\mathcal{H}_\phi$ is trained to produce predictions $\hat{y}_{i,t} = \mathcal{H}_\phi(\mathbf{z}_{i,t})$. For temporal downstream tasks, the head can further operate on the representation sequence $\mathbf{Z}_i = \{\mathbf{z}_{i,1}, \dots, \mathbf{z}_{i,T}\}$. The objective is to learn transferable geospatial representations that support diverse downstream tasks across regions.

3.2 Overview

Figure 2 presents the overall paradigm of **CoST**, a unified spatial-temporal representation learning framework designed to learn geospatial embeddings that are both transferable across regions and semantically interpretable for urban analysis. Rather than modeling satellite tiles as isolated visual instances, CoST jointly exploits two inherent structures of urban observations: *spatial dependency* across neighboring regions and *temporal evolution* within the same region over time. To this end, the framework consists of three complementary components. First, *spatial neighborhood modeling* captures local geographic dependency by encouraging semantically related neighboring regions to form consistent representations, thereby improving cross-region generalization. Second, *temporal semantic guidance* introduces supervision from multi-temporal urban changes, enabling the encoder to distinguish not only whether a region changes, but also the semantic direction and magnitude of such change. Third, *spatial-temporal alignment* couples the above two views by constraining neighboring regions with similar temporal transition patterns to remain aligned in the latent space, yielding a more coherent representation structure. After pre-training, the learned encoder is transferred to multiple downstream tasks with lightweight task-specific heads, including socio-economic indicator prediction, land-use classification, and change detection.

3.3 Spatial Neighborhood Modeling

Standard contrastive learning treats regions as independent instances, ignoring that geographically proximate areas share underlying environmental contexts [12]. To improve generalization across heterogeneous landscapes, we explicitly model spatial structure by aggregating nearby regions into coherent neighborhoods, following Tobler's First Law of Geography [30]. Formally, for each anchor I_i, we define a spatial neighborhood $\mathcal{N}_i = \{I_{n_1}, I_{n_2} \dots, I_{n_k}\}$ consisting of its k nearest geographic neighbors. Within a batch, we consider two types of positive signals: the instance-level positive i^+ and neighborhood-level positives drawn from $\mathcal{N}_i$. To incorporate this hierarchical structure into the representation space, we construct a target distribution Q_i to serve as a supervisory signal. For each sample j, we assign a target score $q_{i,j}$ that reflects its spatial proximity to the anchor I_i:

$$q_{i,j} = \begin{cases} 1, & \text{if } j = i^+ \\ \epsilon, & \text{if } j \in \mathcal{N}_i \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

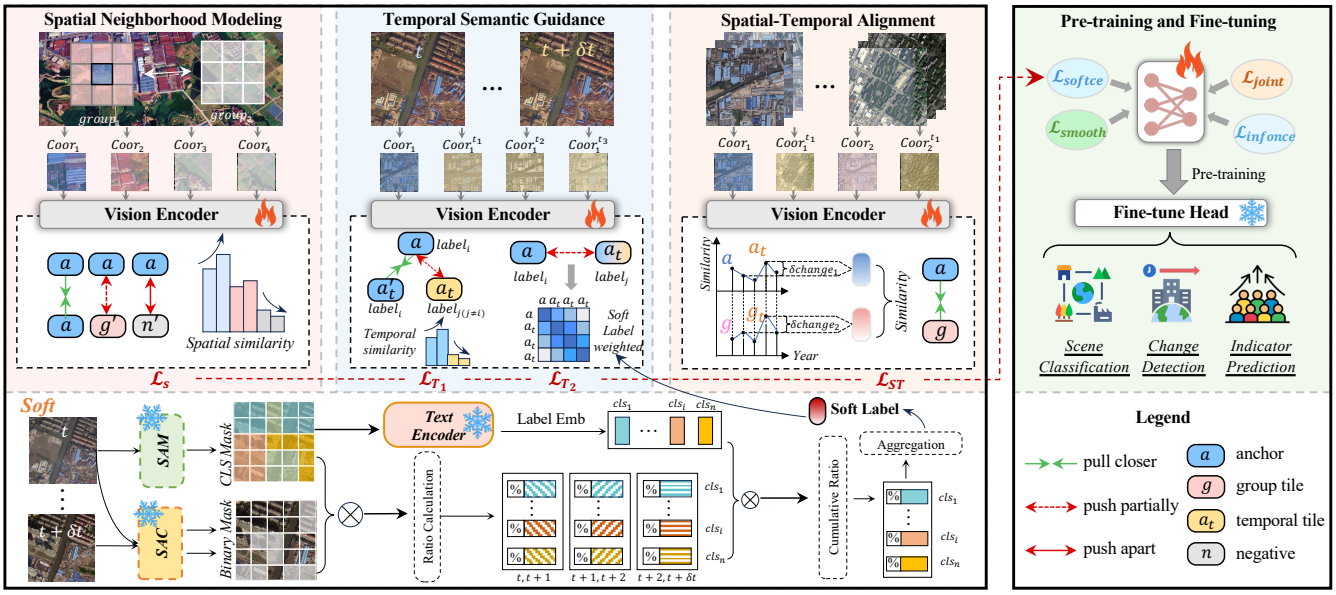

Figure 2: Overall framework of the proposed CoST, aligning geographical neighborhood structures with multi-year urban change semantics to learn generalizable and interpretable geospatial representations for various urban analysis tasks.

where i^+ denotes the augmented view of the anchor, and $\epsilon \in (0, 1)$ modulates the strength of spatial supervision. This distribution reflects a hierarchy in which the anchor itself provides the primary learning signal, while spatial neighbors offer auxiliary supervision to capture shared contexts. Instead of using a one-hot target, the model is trained to align its predicted similarity distribution with this target distribution Q_i by minimizing the following spatial loss:

$$\mathcal{L}_s = -\sum_{j\in\mathcal{B}} q_{i,j} \cdot \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j/\tau)}{\sum_{k\in\mathcal{B}} \exp(\mathbf{z}_i \cdot \mathbf{z}_k/\tau)}, \tag{2}$$

where $\mathcal{B}$ denotes the batch, $\mathbf{z}$ represents the normalized embedding, and τ is the temperature parameter. This objective encourages geographically proximate regions to be embedded nearby, preserving shared spatial semantics and improving transferability to unseen regions.

3.4 Temporal Semantic Guidance

Visual representations learned solely from pixel-level reconstruction or discrimination lack alignment with high-level geo-semantics because they prioritize low-level textures over functional meaning [26]. This limitation is pronounced in multi-temporal satellite imagery, which contains rich semantic transitions (Figure 1). Although explicit labels are unavailable during pre-training, these historical transitions provide semantic supervision. Therefore, we propose **Temporal Semantic Guidance**, which encourages the distance between temporal representations $\mathbf{z}_t$ and $\mathbf{z}_{t+\delta t}$ to reflect both the *extent* and *type* of actual urban changes. This constraint organizes the embedding space along meaningful semantic directions, enabling interpretable, high-level geo-semantic representations.

3.4.1 Cumulative Semantic Signal Construction. Before pre-training, we derive a semantic change metric in three steps: (1) extract visual entities with vision foundation models, (2) accumulate intermediate urban transitions, and (3) project these transitions into a semantic latent space. Pseudocode is provided in Appendix B.

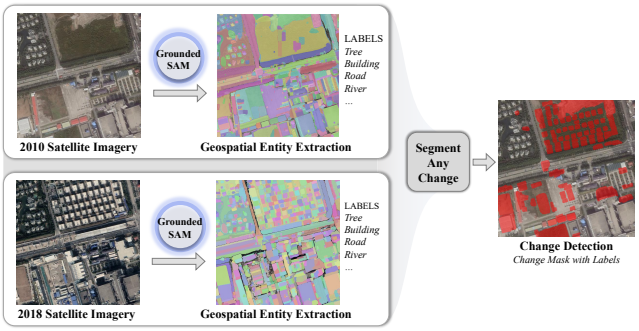

Figure 3: Joint modeling spatial and temporal dimensions from the perspectives of spatial proximity and temporal heterogeneity.

Step 1: Yearly change extraction. For a geospatial tile observed in year t, we first generate a semantic map $\mathcal{S}_t \in \mathbb{R}^{H\times W}$ over land-use classes C (e.g., vegetation, water, buildings) using Grounded-SAM [25]. Based on the detected change regions, we compute a yearly semantic change vector $\boldsymbol{\rho}^{(t,t+1)} \in \mathbb{R}^{|C|}$ that quantifies the proportion of change for each semantic category. For each category $c \in C$, the change ratio is obtained by measuring the overlap between the semantic map and the binary change mask, and normalizing by the total pixel area $|\Omega|$, as shown in Eq 3:

$$\rho_c^{(t,t+1)} = \frac{1}{|\Omega|} \sum_{p\in\Omega} \left(\text{Selector}_c\left(\mathcal{S}_t(p)\right) \cdot \mathcal{M}^{(t,t+1)}(p)\right), \tag{3}$$

where Ω denotes the pixel space, p denotes the pixel, and $\text{Selector}_c(\cdot)$ denotes the selected mask for category c.

Step 2: Semantic Accumulation. Urban region change is a continuous and cumulative process, where intermediate states contain

crucial information (e.g., *barren* → *construction* → *building*), while direct comparison between two time points misses these nuances and results in high computational complexity $O(T^2)$. To capture the total semantic changes over an arbitrary time interval $[t, t+\delta t]$, we aggregate the yearly change vectors $\boldsymbol{\rho}^{(t,t+1)}$ to model intermediate transitions, with linear complexity $O(T)$, obtaining the cumulative change vector $\boldsymbol{\rho}^{(t,t+\delta t)}$, as shown in Eq. 4:

$$\boldsymbol{\rho}^{(t,t+\delta t)} = \sum_{k=t}^{t+\delta t-1} \boldsymbol{\rho}^{(k,k+1)}, \tag{4}$$

Step 3: Semantic Projection. To bridge the gap between low-level pixels and high-level semantics, we project this cumulative change vector into a semantic latent space to represent the accumulative semantic change magnitude. Let $\mathbf{C} \in \mathbb{R}^{|C|\times d}$ denote the class embeddings generated by a frozen text encoder [7]. We compute the change embedding $\mathbf{e}^{(t,t+\delta t)}$ by performing a weighted aggregation of class embeddings based on their change ratios:

$$\mathbf{e}^{(t,t+\delta t)} = \boldsymbol{\rho}^{(t,t+\delta t)} \cdot \mathbf{C} \in \mathbb{R}^d. \tag{5}$$

3.4.2 Dual-Objective Temporal Optimization. To make embedding distances reflect the direction and magnitude of real-world semantic changes, we use a dual-objective strategy. It aligns feature similarities with quantitative change signals while preserving instance discrimination for stable training.

1) Temporal Semantic Alignment. To enhance interpretability, we constrain the latent distance between two temporal views $\mathbf{z}_t$ and $\mathbf{z}_{t+\delta t}$ to be aligned with the cumulative semantic change magnitude $\|\mathbf{e}^{(t,t+\delta t)}\|_2$. Therefore, we define a soft label $y_{i,j} \in (0,1)$ for each temporal pair based on the norm of change embeddings, using a hyperbolic tangent function to map change embeddings into a similarity space:

$$y_{i,j} = 1 - \tanh(\|\mathbf{e}^{(t,t+\delta t)}\|_2). \tag{6}$$

The soft label $y_{i,j}$ defines the target similarity: $y_{i,j} \approx 1$ indicates temporal stability and pulls the representations closer, whereas $y_{i,j} \to 0$ indicates substantial change and pushes them apart. We align visual similarity with these labels using the binary cross-entropy objective in Eq. 7. Consequently, latent distances reflect the magnitude of real-world semantic changes.

$$\mathcal{L}_{T_1} = -\sum_{(i,j)\in\mathcal{P}} \big[y_{i,j} \log \sigma(\mathbf{z}_i \cdot \mathbf{z}_j/\tau) + (1-y_{i,j}) \log(1-\sigma(\mathbf{z}_i \cdot \mathbf{z}_j/\tau)) \big]. \tag{7}$$

2) Temporal Instance Discrimination. To ensure robust learning, representations should maintain instance-level discriminability and stability. Empirical evidence indicates that training solely with a BCE-style objective leads to instability and slow convergence [47], and this issue is exacerbated by noise in the foundation model-derived embeddings $\mathbf{e}$. Thus, we introduce the InfoNCE loss [38] to leverage stable and discriminative gradients, thereby ensuring reliable convergence. In this contrastive task, for an anchor $\mathbf{z}_i$, the positive sample $\mathbf{z}_i^+$ is selected via augmentation to enforce representation invariance. Crucially, we introduce a set of temporal negatives $\mathcal{Z}_i^- = \left\{ \mathbf{z}_{i,t'}^- \mid \|\mathbf{e}^{(t,t')}\|_2 \neq 0 \right\}$, representing the same location at different time steps but exhibiting distinct semantic changes. By contrasting against these hard temporal negatives, the model is compelled to recognize significant urban changes. This objective is minimized by:

$$\mathcal{L}_{T_2} = -\sum_{i=1}^{\mathcal{B}} \log \frac{\exp\left(\langle \mathbf{z}_i, \mathbf{z}_i^+ \rangle\right)}{\exp\left(\langle \mathbf{z}_i, \mathbf{z}_i^+ \rangle\right) + \sum_{\mathcal{Z}_i^-} \exp\left(\langle \mathbf{z}_i, \mathbf{z}_i^- \rangle\right)}. \tag{8}$$

The integration of *Temporal Semantic Alignment* and *Temporal Instance Discrimination* establishes a structurally robust representation space, enabling the model to learn discriminative features with faithful semantics.

3.5 Spatial-Temporal Alignment

While *Spatial Neighborhood Modeling* enhances generalization by capturing spatial correlations and *Temporal Semantic Guidance* improves interpretability by characterizing semantic shifts, optimizing them independently may introduce *false positives* in the embedding space. In real-world urban systems, visually similar regions may possess distinct latent semantics, which are only revealed through long-term urban development rather than static observation. An inherent dependency therefore exists: spatial similarity should be conditioned on whether regions exhibit consistent semantic changes over time, while modeling spatial and temporal dimensions in isolation overlooks this dependency. To resolve this, we propose the *Spatial-Temporal Alignment* as a corrective mechanism that refines spatial similarity using temporal semantic changes.

Specifically, for a given anchor I_i, we identify its most similar neighbor $I_j \in \mathcal{N}_i$ based on cosine similarity, i.e., $j = \arg\max_{j\in\mathcal{N}_i} \mathbf{z}_i \cdot \mathbf{z}_j$. We then identify the timestamp t' with the maximal semantic divergence from the current state and construct normalized semantic change vectors $\boldsymbol{\Delta}_i = \text{norm}(\mathbf{z}_i - \mathbf{z}_i^{t'})$ and $\boldsymbol{\Delta}_j = \text{norm}(\mathbf{z}_j - \mathbf{z}_j^{t'})$. Spatial similarity is penalized when these change directions are misaligned, thereby encouraging discrimination between visually similar but semantically divergent regions (Eq. 9).

$$\mathcal{L}_{st} = \frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} (1 - \langle \boldsymbol{\Delta}_i, \boldsymbol{\Delta}_j \rangle) \cdot \langle \mathbf{z}_i, \mathbf{z}_j \rangle. \tag{9}$$

3.6 Pre-training & Fine-Tuning

Pre-training Stage. CoST adopts a contrastive learning framework for pre-training, initialized from scratch. The encoder $\mathcal{E}$ is optimized via a weighted multi-objective loss:

$$\mathcal{L}_{\text{pretrain}} = \alpha \mathcal{L}_{\text{S}} + \beta_1 \mathcal{L}_{\text{T}_1} + \beta_2 \mathcal{L}_{\text{T}_2} + \gamma \mathcal{L}_{\text{ST}}, \tag{10}$$

where the coefficients α, β_1, β_2, and γ are tuned via grid search to balance the training objectives. Jointly modeling spatial and temporal dynamics yields generalizable and interpretable representations.

Fine-Tuning Stage. We adopt a linear probing strategy where the pre-trained encoder remains frozen. We attach the lightweight task-specific heads to evaluate the representations across three downstream tasks:

- **Static Indicator Prediction**. For socio-economic indicator prediction from a single target year, the target representation $\mathbf{z}_T$ is fed into an MLP regression head:

$$\hat{y} = \text{MLP}_{\text{reg}}(\mathbf{z}_T). \tag{11}$$

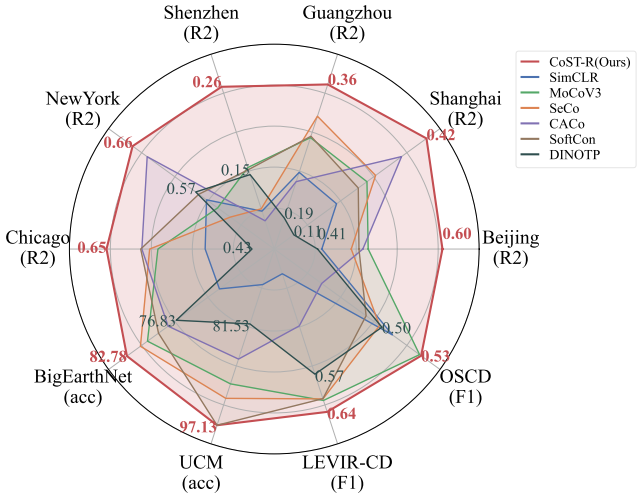

Figure 4: Overall performance of CoST-R across various downstream tasks.

- **Dynamic Indicator Prediction**. To incorporate historical urban dynamics, we further use the temporal representation sequence $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_{T-1}\}$ as auxiliary input. An LSTM encoder summarizes the historical sequence, and its output is combined with the target-year representation to predict the indicator:

$$\hat{y} = \mathrm{MLP}_{\mathrm{base}}(\mathbf{z}_T) + \mathrm{MLP}_{\mathrm{res}}\big(\mathrm{LSTM}(\mathbf{Z})\big). \tag{12}$$

- **Land-Use Classification**. For scene-level land-use recognition, the representation $\mathbf{z}$ is fed into a lightweight MLP classifier followed by a softmax layer, as shown in Eq. 13. The classifier is optimized using cross-entropy loss, evaluating whether the learned representation preserves discriminative semantic cues for region categorization.

$$\hat{y} = \mathrm{Softmax}(\mathrm{MLP}_{\mathrm{cls}}(\mathbf{z})). \tag{13}$$

- **Change Detection**. For bi-temporal change detection, we extract representations from two timestamps and concatenate them as $[\mathbf{z}_t; \mathbf{z}_{t'}]$. The fused feature is then passed to a lightweight U-Net decoder to predict the binary change map, as shown in Eq. 14. The decoder is trained with binary cross-entropy loss to assess whether the learned representations retain fine-grained temporal change information.

$$\hat{y} = \mathcal{U}([\mathbf{z}_t; \mathbf{z}_{t'}]). \tag{14}$$

4 Experiments

In this section, we evaluate the proposed CoST to investigate the following research questions (**RQs**):

- **RQ1:** Does CoST outperform state-of-the-art self-supervised methods and demonstrate *generalizability* across diverse tasks and real-world applications in complex urban environments?
- **RQ2:** Do the learned representations possess semantic *interpretability*, where the latent space effectively captures and aligns with meaningful spatial-temporal transitions?
- **RQ3:** What are the contributions of individual components to the overall performance of the CoST?
- **RQ4:** How do temporal dynamics regularize semantic consistency and improve domain invariance across cities?

Table 1: Dataset statistics. The first five datasets are used for pre-training, and the *Chicago* dataset is held out from pre-training for a fair generalization test.

Dataset	Coverage in Geospace		Satellite Image
	Bottom-left	*Top-right*	
Beijing	39.91°N, 116.20°E	40.24°N, 116.53°E	4,218 × 11
Guangzhou	22.78°N, 113.15°E	23.36°N, 113.51°E	3,726 × 11
Shanghai	30.87°N, 121.24°E	31.40°N, 121.68°E	3848 × 11
Shenzhen	22.52°N, 113.84°E	22.67°N, 114.25°E	2,772 × 11
New York	40.52°N, 73.95°W	40.80°N, 73.79°W	1,728 × 11
Chicago	41.84°N, 87.74°W	41.96°N, 87.57°W	754 × 11

Pre-training Dataset. We pre-train CoST on a spatial-temporal satellite imagery dataset collected from Google Earth, covering five major cities: *Beijing*, *Shanghai*, *Guangzhou*, *Shenzhen*, and *New York*, over the period from 2010 to 2020. Each sample corresponds to a geospatial region with aligned multi-year observations, which enables the model to jointly learn spatial structure and temporal dynamics. Dataset statistics are summarized in Tab. 1. To evaluate cross-city generalization, we further construct a held-out *Chicago* dataset, which is excluded from pre-training and used only for downstream evaluation.

Downstream Datasets. We evaluate CoST on three types of downstream tasks to assess its effectiveness across socio-economic inference, semantic classification, and temporal change understanding.

- **Socio-economic Indicator Prediction.** We evaluate urban indicator prediction using two representative socio-economic signals: population density data from WorldPop[1] and GDP data from a publicly available global dataset[2]. As both datasets provide multi-year observations, they support evaluations under both static prediction and dynamic prediction settings using historical representation sequences. This task assesses whether the learned representations capture cross-sectional socio-economic semantics at a given time point and their temporal dynamics over time.
- **Land-Use Classification.** We adopt two benchmark datasets for semantic scene understanding: UC Merced Land-Use [40], which contains aerial scene categories for single-label classification, and BigEarthNet [27], a large-scale multi-label remote sensing benchmark with diverse land-cover types. This setting evaluates whether the learned embeddings retain discriminative semantics for region categorization.
- **Change Detection.** We evaluate temporal sensitivity on two widely used change detection datasets, OSCD [6] and LEVIR-CD [2]. Both datasets provide bi-temporal observations with pixel-level supervision, enabling us to assess whether the learned representations encode fine-grained change information for downstream segmentation.

Baselines. We compare CoST with self-supervised and geospatial representation learning baselines using both CNN and ViT encoders. These include general visual pre-training methods (SimCLR [4], MoCoV3 [38], DINOv2 [23], and SoftCon [32]), remote-sensing methods (SeCo [20], CACo [19], READ [10], PG-SimCLR [36], SatMAE [5], and ScaleMAE [24]), and temporal transformer variants

[1] https://hub.worldpop.org/
[2] https://www.nature.com/articles/s41597-022-01322-5

Table 2: Performance on urban indicator prediction (2020). Bold and underline denote the best and second-best results, respectively. Both variants of our model (CoST-R and CoST-V) are highlighted in bold when they achieve the top two positions.

Model	Beijing				Guangzhou				New York				*Chicago*			
	GDP		Population		GDP		Population		GDP		Population		GDP		Population	
	R^2	$MSE\downarrow$	R^2	$MSE\downarrow$	R^2	$MSE\downarrow$	R^2	$MSE\downarrow$	R^2	$MSE\downarrow$	R^2	$MSE\downarrow$	R^2	$MSE\downarrow$	R^2	$MSE\downarrow$
SimCLR	0.420	0.584	0.412	0.570	0.248	0.684	0.163	0.467	0.552	0.363	0.608	0.353	0.501	0.561	0.467	0.522
MoCoV3	0.488	0.460	0.504	0.510	0.293	0.781	0.390	0.447	0.535	0.434	0.594	0.391	0.570	0.367	0.597	0.447
DINOv2	0.346	0.605	0.456	0.544	0.151	0.845	0.296	0.771	0.593	0.435	0.622	0.327	0.487	0.499	0.661	0.357
SeCo	0.463	0.551	0.462	0.497	0.320	0.728	0.472	0.491	0.516	0.502	0.606	0.375	0.583	0.342	0.531	0.415
CACo	0.480	0.543	0.477	0.489	0.237	0.714	0.229	0.600	0.642	0.413	0.666	0.361	0.594	0.431	0.669	0.291
SoftCon	0.474	0.424	0.472	0.507	0.295	0.862	0.267	0.837	0.564	0.382	0.624	0.379	0.596	0.438	0.601	0.256
DiNOTP	0.413	0.585	0.364	0.706	0.190	0.774	0.211	1.215	0.569	0.394	0.543	0.408	0.433	0.727	0.463	0.640
READ	0.217	0.805	0.182	0.805	0.291	0.679	0.327	0.610	-	-	-	-	-	-	-	-
PG-SimCLR	0.418	0.572	0.468	0.439	0.266	0.616	-	-	-	-	-	-	-	-	-	-
SatMAE	0.489	0.593	0.513	0.465	0.339	1.005	0.382	0.872	0.639	**0.320**	0.612	0.371	0.607	0.297	0.580	0.360
ScaleMAE	0.498	0.479	0.511	0.501	0.208	0.941	0.333	0.771	0.639	0.389	0.629	0.319	0.588	0.399	0.628	0.383
CoST-R	**0.596**	**0.393**	**0.565**	**0.410**	**0.361**	**0.498**	**0.553**	**0.429**	**0.664**	0.339	**0.698**	**0.273**	**0.646**	**0.276**	**0.691**	**0.253**
CoST-V	**0.561**	0.477	**0.545**	**0.430**	**0.378**	0.738	0.409	0.460	0.630	0.370	**0.678**	0.322	**0.642**	0.331	0.655	0.355

(DiNOTP [33], DiNOTP-ViT, and SoftCon-ViT). Appendix C provides brief descriptions.

Metrics. We adopt task-specific evaluation metrics following standard practice. For socio-economic indicator prediction, we report R^2 and mean squared error (MSE) to measure regression accuracy. For land-use classification, we report Top-1 accuracy. For change detection, we report the F1 score to evaluate the quality of predicted change regions.

Implementation Details. We implement two variants of CoST: **CoST-R**, based on a ResNet50 backbone, and **CoST-V**, based on a ViT-B/16 backbone. Both variants use a feature dimension of $d = 128$ and follow the same pre-training and downstream evaluation protocol unless otherwise specified. All models are pre-trained for 3,000 epochs on a server equipped with two NVIDIA RTX A6000 GPUs. Detailed optimization settings, loss hyperparameters, and downstream fine-tuning configurations are provided in Appendix A.1 and Appendix A.2.

4.1 Overall Performance (RQ1)

To evaluate the effectiveness and generalizability of CoST, we conduct extensive experiments against baselines across diverse tasks. Fig 4 presents a holistic overview of CoST-R's performance in three distinct domains. Detailed quantitative comparisons are provided in Tab. 2 and Tab. 3.

Urban Indicator Prediction. CoST (CoST-R & CoST-V) achieves state-of-the-art performance in urban indicator prediction across multiple cities. In New York City, CoST-R attains an R^2 of 0.664 for GDP and 0.698 for population, outperforming prior methods by over 5%. Similar improvements are observed in Shanghai and Shenzhen (see Appendix A). To evaluate cross-region generalization under domain shift, we further test on *Chicago*, which is excluded from pre-training. As shown in Tab. 2, CoST achieves strong transfer results, with GDP and population R^2 of 0.646 and 0.691, respectively, surpassing region-dependent baselines and demonstrating robust generalization across urban environments.

Table 3: Performance of different models in classification and change detection tasks. The best results are bold, and the second-best results are underlined.

Methods	Classification		Change Detection	
	UCM	BigEarthNet	OSCD	LEVIR-CD
	top-1 Acc	*mAP*	*F1 Score*	*F1 Score*
SimCLR	75.47	71.62	0.5069	0.3944
MoCoV3	90.76	80.30	0.5315	0.6220
SeCo	92.99	81.15	0.4953	0.6192
CACo	86.94	77.81	0.4429	0.4881
SoftCon	97.13	78.97	0.4831	0.6188
DiNOTP	81.53	76.83	0.4974	0.5748
SatMAE	92.36	76.07	–	–
ScaleMAE	98.41	82.57	–	–
CoST-R	97.13	**82.78**	**0.5329**	**0.6425**
CoST-V	**98.73**	**83.55**	–	–

Classification & Change Detection. We evaluate the learned representations on land-use classification and change detection (Tab. 3). For classification, CoST-R and CoST-V achieve 97.13% and 98.73% top-1 accuracy on UCM, respectively, while remaining competitive on the multi-label BigEarthNet dataset. For change detection, CoST-R surpasses specialized bi-temporal methods [19, 20], achieving an F1 score of 0.6425 on LEVIR-CD, demonstrating robust cross-task generalization of CoST.

4.2 Representation Interpretability (RQ2)

To evaluate the semantic interpretability of the learned representations, we investigate whether the latent space captures high-level concepts through vector arithmetic, as shown in Figure 5(a). If visual semantics are aligned with the embedding space, algebraic

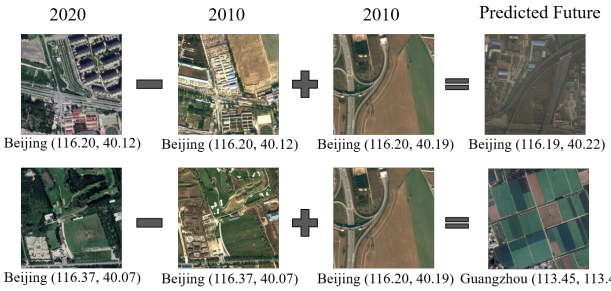

Figure 5: Visualization of embedding arithmetic. It shows that CoST captures high-level semantic concepts via vector operations. Predictable semantic transformations demonstrate that the learned latent space is disentangled and interpretable.

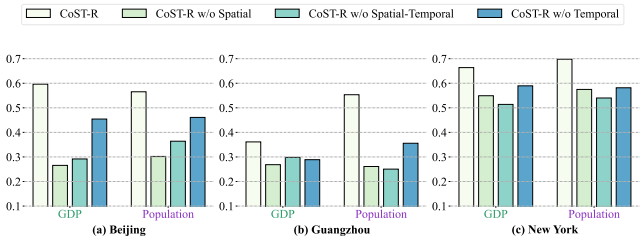

Figure 6: Ablation study of three CoST components: (a) *w/o Spatial*, (b) *w/o Temporal*, and (c) *w/o ST*.

operations should induce predictable semantic changes. Specifically, subtracting one attribute (e.g., built-up areas) and adding another (e.g., farmland) retrieves images in which the target semantic components change while the surrounding context is preserved. This consistency indicates that CoST captures meaningful semantic directions rather than memorizing pixel-level patterns, demonstrating that the learned representations are semantically disentangled and interpretable.

4.3 Ablation Study (RQ3)

To evaluate the contribution of each component in CoST, we conduct an ablation study on socio-economic prediction tasks with three variants: (a) *w/o Spatial*, (b) *w/o Temporal*, and (c) *w/o ST*, as shown in Fig. 6. Removing spatial neighborhood modeling (*w/o Spatial*) leads to a pronounced performance drop, with R^2 decreasing by 15–30% across cities, highlighting the importance of spatial aggregation in capturing shared urban contexts following Tobler's first law [30]. The performance decline in *w/o Temporal* further demonstrates that temporal semantics provide critical supervision for representation learning. Finally, the performance of *w/o ST* suggests that spatial–temporal alignment is essential for avoiding the aggregation of visually similar yet functionally distinct regions, enabling accurate indicator regression.

4.4 Qualitative Analysis (RQ4)

Dynamic Urban Profiling. To assess CoST's ability to model long-term urban development, we perform dynamic urban profiling by leveraging historical representation sequences (2010–2019) to predict 2020 indicators. Detailed task descriptions and experimental

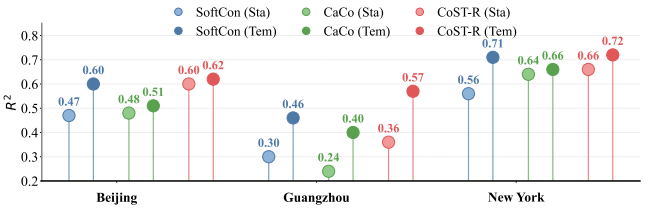

Figure 7: Urban indicator prediction using time-series satellite imagery in Beijing, Guangzhou, and New York City. *Sta* and *Temp* denote static and dynamic prediction, respectively.

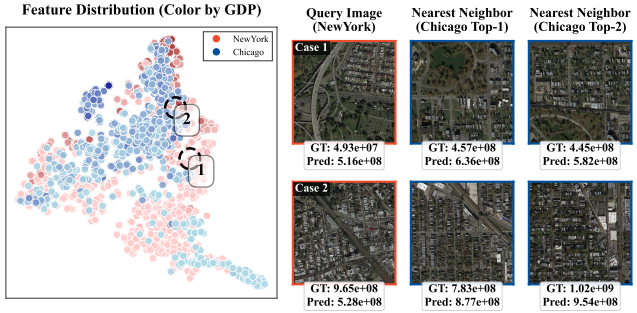

Figure 8: Domain adaptation of CoST between the source city and target city (Chicago). Color saturation represents GDP levels from light (low GDP) to dark (high GDP).

settings are provided in Sec. 3.6 and Appendix A.2. As shown in Figure 7, incorporating temporal dynamics (Tem) consistently outperforms static snapshots (Sta) in terms of R^2 across all cities. Notably, CoST-R achieves the best performance among temporal-aware baselines, reaching R^2 scores of 0.62 in Beijing, 0.57 in Guangzhou, and 0.72 in New York. These results highlight the importance of historical sequences for accurate urban indicator estimation, as temporal accumulation enables the model to capture socio-economic development rather than relying on a single visual snapshot, leading to more coherent and predictive representations over time.

Universal Domain Shift. To understand how CoST handles visual disparities across urban environments, we perform the case-based retrieval analysis in Figure 8. Despite substantial domain shifts, the latent representations align by socio-economic level, and the resulting clusters reflect GDP levels even for unseen data. In both suburban and urban cases, New York queries retrieve semantically consistent Chicago regions with matching ground-truth GDP profiles. This alignment indicates that CoST captures geographic structures that transfer to unseen regions.

5 Conclusion and Future Work

In this work, we have presented CoST, a unified framework designed to learn generalizable and interpretable geospatial representations from satellite imagery, effectively bridging the gap between visual features and high-level geo-semantics. Extensive experiments demonstrate that CoST achieves state-of-the-art performance across diverse downstream tasks. Building upon the continuous multi-year imagery used in this work, future research could further explore predictive representations for urban change prediction and image generation tasks, enabling more proactive urban planning and scenario simulation.

Acknowledgments

This work is supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011994), the National Natural Science Foundation of China (No. 62402414), Guangdong Provincial Project 2025D03J0014, Guangzhou Municipal Science and Technology Project (No. 2023A03J0011), the Guangzhou Industrial Information and Intelligent Key Laboratory Project (No. 2024A03J0628), and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007).

A Experiment Results

We provide additional urban indicator results on Shanghai and Shenzhen, complementing the main-paper evaluation. The results show that CoST remains effective across both cities and both indicators.

Table 4: Performance Comparison on Shanghai Dataset. Bold and underline denote the best and second-best results, respectively. Both variants of our model (CoST-R and CoST-V) are highlighted in bold when they achieve the top two positions.

Model	Backbone	GDP		Population	
		R^2	MSE↓	R^2	MSE↓
SimCLR	R50	0.213	0.955	0.284	0.711
MoCoV3	R50	0.283	**0.523**	0.379	**0.443**
DINOv2	R50	0.179	0.692	0.293	0.736
SeCo	R50	0.302	0.542	0.298	0.919
CACo	R50	0.362	0.860	0.305	0.734
SoftCon	R50	0.263	0.857	0.332	0.685
DiNOTP	R50	0.115	0.984	0.222	0.679
SatMAE	ViT-L	0.404	**0.515**	0.338	0.712
ScaleMAE	ViT-B	0.351	0.916	0.311	0.607
CoST-R	R50	**0.419**	0.675	**0.434**	0.550
CoST-V	ViT-B	**0.442**	0.723	0.372	0.505

We provide additional urban indicator results on the Shanghai and Shenzhen datasets, further supporting the claims in Section 4.2 and demonstrating CoST's robust profiling capability across diverse urban contexts. As shown in Table 4, CoST-V achieves a GDP R^2 of 0.442 in Shanghai, improving over SatMAE (0.404), while CoST-R leads population prediction with an R^2 of 0.434. In Shenzhen, CoST-V obtains the highest GDP R^2 of 0.273, and CoST-R achieves the best population performance with an R^2 of 0.463 and an MSE of 0.496, indicating balanced performance across both indicators and cities. We further compare CoST-R with the remote-sensing foundation models Prithvi-100M and DOFA. As shown in Fig. 9, CoST-R consistently achieves lower GDP and population MSE than both models in Beijing, Guangzhou, and the unseen Chicago setting, despite using a substantially smaller ResNet50 backbone.

A.1 Pre-training Details

A frozen BERT encoder generates embeddings for 11 urban land-use categories: `dense houses`, `trees`, `grass`, `river`, `barren land`, `sidewalks`, `farmland`, `tall buildings`, `soil`, `crop fields`, `roads`. We use Adam ($\beta = (0.9, 0.999)$, weight decay 1×10^{-4}) with a target learning rate of 3×10^{-4}, linearly warmed up over 20 epochs and then held constant. We use 32 groups per GPU (64 groups in total), with four elements per group. We set $num_{neg} = 6$, $\tau = 0.07$, and $\epsilon = 0.1$, and weight the total loss with $(\alpha, \beta_1, \beta_2, \gamma) = (1.0, 0.2, 0.2, 0.3)$.

Table 5: Performance Comparison on Shenzhen Dataset. Bold and underline denote the best and second-best results, respectively. Both variants of our model (CoST-R and CoST-V) are highlighted in bold when they achieve the top two positions.

Model	Backbone	GDP		Population	
		R^2	MSE↓	R^2	MSE↓
SimCLR	R50	0.099	0.832	0.296	0.623
MoCoV3	R50	0.155	0.998	0.382	0.569
DINOv2	R50	0.074	0.943	0.316	0.709
SeCo	R50	0.102	0.894	0.274	0.653
CACo	R50	0.086	0.959	0.291	0.578
SoftCon	R50	0.152	0.951	0.352	0.668
DiNOTP	R50	0.146	0.791	0.320	0.718
READ	R18	0.129	0.889	0.284	0.972
PG-SimCLR	R18	0.123	1.058	-	-
SatMAE	ViT-L	0.176	**0.623**	0.325	0.773
ScaleMAE	ViT-B	0.163	0.796	0.329	0.725
CoST-R	R50	**0.258**	0.967	**0.463**	**0.496**
CoST-V	ViT-B	**0.273**	0.856	**0.359**	**0.591**

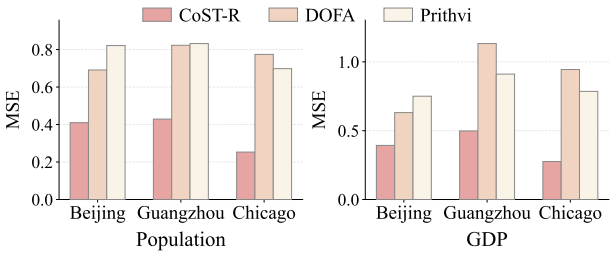

Figure 9: Performance comparison with RS foundation models across indicator predictions. CoST-R, Prithvi-100M, and DOFA are compared using GDP MSE and population MSE.

A.2 Fine-tuning details

Fine-tuning. We adopt linear probing with the pre-trained encoder frozen. Models are trained for 100 epochs with a batch size of 128 using AdamW, with a learning rate of 3×10^{-3} and weight decay of 1×10^{-4}. The task-specific heads are defined as follows:

- **Classification:** A linear projection head is used for both single-label (UCM) and multi-label (BigEarthNet) classification, optimized with Cross-Entropy Loss and Multi-Label Soft Margin Loss, respectively.
- **Static Urban Indicator Prediction:** A 3-layer MLP regressor predicts GDP and population density from the frozen features using the MSE loss.

Algorithm 1 Temporal Semantic Guidance

Step 1: Change Extraction
Requires: Sequence of satellite images $\{I_1, I_2, \ldots, I_T\}$; class labels $C = \{c_1, c_2, \ldots, c_{10}\}$
for each year $t \in \{1, \ldots, T\}$ **do**
  $S_t \leftarrow \mathcal{G}_{SAM}(I_t)$ {Generate semantic maps for land-use classes C}
end for
for each consecutive pair $(t, t+1)$ **do**
  $M^{(t,t+1)} \leftarrow \mathcal{A}_{Any}(I_t, I_{t+1})$ {Detect binary change masks at pixel level}
  for each category $c \in C$ **do**
    $\rho_c^{(t,t+1)} \leftarrow \frac{1}{|\Omega|} \sum_{p \in \Omega} (\text{Selector}_c(S_t(p)) \cdot M^{(t,t+1)}(p))$ {Eq. 3: Intersection of change and class}
  end for
  $\vec{\rho}^{(t,t+1)} \leftarrow [\rho_1, \ldots, \rho_{|C|}]$ {Construct yearly semantic change vector}
end for
Step 2: Semantic Accumulation
Initialize $\vec{\rho}^{(t,t+\delta t)} \leftarrow \vec{0}$
for $k = t$ **to** $t + \delta t - 1$ **do**
  $\vec{\rho}^{(t,t+\delta t)} \leftarrow \vec{\rho}^{(t,t+\delta t)} + \vec{\rho}^{(k,k+1)}$ {Eq. 4: Linear accumulation of intermediate transitions}
end for
Step 3: Semantic Projection and Signal Generation
$e^{(t,t+\delta t)} \leftarrow \vec{\rho}^{(t,t+\delta t)} \cdot \mathbf{C}$ {Eq. 5: Weighted aggregation of class embeddings}
$y \leftarrow 1 - \tanh(\|e^{(t,t+\delta t)}\|_2)$ {Eq. 6: Map change magnitude to soft similarity target}
return $e^{(t,t+\delta t)}, y$

- **Dynamic Urban Indicator Prediction:** We use a Visual-LSTM architecture. Features from 2010–2020 are projected into a 256-dimensional space and processed by a single-layer LSTM with hidden dimension 256. The prediction is given by $\hat{y} = \text{MLP}_{\text{base}}(\mathbf{z}_T) + \text{MLP}_{\text{res}}(\mathbf{h}_{T-1})$, where the second term models the temporal development trend.
- **Change Detection:** A lightweight U-Net head performs pixel-level change segmentation using multi-stage bi-temporal features. Features are fused by concatenation or absolute difference, followed by upsampling blocks with skip connections and double convolutions, optimized with Binary Cross-Entropy Loss.

B Temporal Semantic Guidance

The *Temporal Semantic Guidance* organizes the embedding space into interpretable semantic dimensions by leveraging directional urban change signals. The following pseudo code extend the core methodology presented in the main paper and provide additional clarity on how the supervision signal is implemented.

- **Step 1:** We use frozen Grounded-SAM [25] to generate a semantic map S_t, assigning each pixel p to one of $|C|$ urban land-use categories. As defined in Section 3.3, Selector_c isolates the mask of class c. Its pixel-wise product with the binary change mask $M^{(t,t+1)}$, detected by `Segment Any Change` [45], captures changes within each functional zone and yields the semantic change vector $\vec{\rho}$.
- **Step 2:** Instead of exhaustively comparing all year pairs, we compute yearly transition vectors $\vec{\rho}^{(k,k+1)}$. Changes over any interval $[t, t+\delta t]$ are then obtained by summing category-wise transitions, reducing the complexity from $O(T^2)$ to $O(T)$. This also captures intermediate development stages, such as soil evolving into construction and then tall buildings.
- **Step 3:** We embed the 11 urban categories listed in Appendix B using a frozen BERT encoder, producing label embeddings $\mathbf{C}$. The final change embedding $e^{(t,t+\delta t)}$ is computed as a weighted aggregation of these embeddings. The tanh function in Eq. 6 maps the L_2 norm of the semantic shift to a soft similarity target $y \in (0, 1)$, encouraging latent visual distances to reflect real-world land-use changes.

C Baselines

To evaluate the effectiveness of CoST, we compare it against a comprehensive suite of baseline methods, encompassing general-purpose self-supervised learning frameworks and specialized remote sensing (RS) foundation models.

- **SimCLR** [4]: A fundamental contrastive learning framework that learns representations by maximizing the agreement between different augmented views of the same data sample.
- **MoCoV3** [38]: An advanced momentum-based contrastive learning approach optimized for Vision Transformers (ViT) to improve training stability and feature quality.
- **DINOv2** [23]: A state-of-the-art vision foundation model trained with self-distillation. It generates high-quality, discriminative features that exhibit strong zero-shot performance across a wide range of downstream vision tasks.
- **SeCo** [20]: A seasonal contrastive learning pipeline designed for remote sensing, which utilizes temporal revisits of the same geographical location to learn time-invariant representations.
- **CACo** [19]: A change-aware contrastive learning method that leverages temporal signals from satellite imagery to distinguish transient seasonal variations.
- **SoftCon** [32]: A soft contrastive learning framework that incorporates multi-label land-cover information.
- **DiNOMC** [33]: An extension of the DINO framework for Earth observation that employs multi-sized local crops and global-local view alignment to capture multi-scale object features.
- **READ** [10]: A lightweight representation learning method tailored for district-level economic scale estimation.
- **PG-SimCLR** [36]: A geography-aware model that enhances SimCLR by incorporating Point-of-Interest (POI) data as signals.
- **SatMAE** [5]: A masked autoencoder (MAE) framework specifically adapted for satellite imagery, incorporating temporal and spectral positional encodings to process multi-spectral data.
- **Scale-MAE** [24]: A scale-aware masked autoencoder that explicitly models relationship between spatial resolutions by encoding Ground Sample Distance into positional embeddings.

GenAI Usage Disclosure

We state that AI-assisted technologies were used strictly for language polishing and grammar correction. The authors take full responsibility for the integrity of the work, including the validity of all experimental results and the accuracy of all references.

References

[1] Mingfei Cai, Yanbo Pang, and Yoshihide Sekimoto. 2024. Explainable Hierarchical Urban Representation Learning for Commuting Flow Prediction. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*. 573–576.

[2] Hao Chen and Zhenwei Shi. 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote sensing* 12, 10 (2020), 1662.

[3] Meng Chen, Zechen Li, Weiming Huang, Yongshun Gong, and Yilong Yin. 2024. 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*. 319–328.

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

[5] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. 2022. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems* 35 (2022), 197–211.

[6] Rodrigo Caye Daudt, Bertr Le Saux, Alexandre Boulch, and Yann Gousseau. 2018. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. Ieee, 2115–2118.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[8] Ziquan Fang, Dongen Wu, Lu Pan, Lu Chen, and Yunjun Gao. 2022. When Transfer Learning Meets Cross-City Urban Flow Prediction: Spatio-Temporal Adaptation Matters.. In *IJCAI*, Vol. 22. 2030–2036.

[9] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. 2024. Skysense: A multimodal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27672–27683.

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

[11] Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. 2025. Urbanvlp: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 28061–28069.

[12] Xixuan Hao, Wei Chen, Xingchen Zou, and Yuxuan Liang. 2025. Nature makes no leaps: Building continuous location embeddings with satellite imagery from the web. In *Proceedings of the ACM on Web Conference 2025*. 2799–2812.

[13] Xixuan Hao, Yutian Jiang, Jiabo Liu, Yihang Yang, Guangyin Jin, Song Gao, and Yuxuan Liang. 2026. Multi-Agent Collaborative Reasoning with Tool-Augmented Evidence for Urban Region Profiling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 1590–1601.

[14] Xixuan Hao, Yutian Jiang, Xingchen Zou, Jiabo Liu, Yifang Yin, Song Gao, Flora Salim, Tianrui Li, and Yuxuan Liang. 2025. Geospatial Representation Learning: A Survey from Deep Learning to The LLM Era. *arXiv preprint arXiv:2505.09651* (2025).

[15] Xixuan Hao, Guicheng Li, Daiqiang Wu, Xusen Guo, Yumeng Zhu, Zhichao Zou, Peng Zhen, Yao Yao, and Yuxuan Liang. 2026. Enhancing Ride-Hailing Forecasting at DiDi with Multi-View Geospatial Representation Learning from the Web. In *Proceedings of the ACM Web Conference 2026*. 8200–8211.

[16] Porter Jenkins, Ahmad Farag, Suhang Wang, and Zhenhui Li. 2019. Unsupervised representation learning of spatial data via multimodal embedding. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1993–2002.

[17] Qingxiang Liu, Sheng Sun, Min Liu, Yuwei Wang, and Bo Gao. 2024. Online spatio-temporal correlation-based federated learning for traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems* 25, 10 (2024), 13027–13039.

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

[19] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. 2023. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5261–5270.

[20] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. 2021. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9414–9423.

[21] Dominik J Mühlematter, Lin Che, Ye Hong, Martin Raubal, and Nina Wiedemann. 2025. UrbanFusion: Stochastic Multimodal Fusion for Contrastive Learning of Robust Spatial Representations. *arXiv preprint arXiv:2510.13774* (2025).

[22] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. 2024. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27811–27819.

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[24] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. 2023. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4088–4099.

[25] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159* (2024).

[26] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9243–9252.

[27] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. 2019. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*. IEEE, 5901–5904.

[28] Fengze Sun, Yanchuan Chang, Egemen Tanin, Shanika Karunasekera, and Jianzhong Qi. 2025. FlexiReg: Flexible Urban Region Representation Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 2702–2713.

[29] Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Orsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique De Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. 2025. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing* (2025).

[30] Waldo R Tobler. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography* 46, sup1 (1970), 234–240.

[31] Qiongyan Wang, Xingchen Zou, Yutian Jiang, Haomin Wen, Jiaheng Wei, Qingsong Wen, and Yuxuan Liang. 2025. Urban-R1: Reinforced MLLMs Mitigate Geospatial Biases for Urban General Intelligence. *arXiv preprint arXiv:2510.16555* (2025).

[32] Yi Wang, Conrad M Albrecht, and Xiao Xiang Zhu. 2024. Multi-label Guided Soft Contrastive Learning for Efficient Earth Observation Pretraining. *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[33] Xinye Wanyan, Sachith Seneviratne, Shuchang Shen, and Michael Kirley. 2024. Extending global-local view alignment for self-supervised learning with remote sensing imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2443–2453.

[34] Ya Wen and Yulun Zhou. 2024. Demo2vec: Learning region embedding with demographic information. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*. 71–74.

[35] Nemin Wu, Qian Cao, Zhangyu Wang, Zeping Liu, Yanlin Qi, Jielu Zhang, Joshua Ni, Xiaobai Yao, Hongxu Ma, Lan Mu, et al. 2024. Torchspatial: A location encoding framework and benchmark for spatial representation learning. *Advances in Neural Information Processing Systems* 37 (2024), 81437–81460.

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

[37] Congxi Xiao, Jingbo Zhou, Yixiong Xiao, Jizhou Huang, and Hui Xiong. 2024. Refound: Crafting a foundation model for urban region understanding upon language and visual foundations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3527–3538.

[38] Chen Xinlei, Xie Saining, and He Kaiming. 2021. An empirical study of training self-supervised visual transformers. *arXiv preprint arXiv:2104.02057* 8, 7 (2021), 2.

[39] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference 2024*. 4006–4017.

[40] Yi Yang and Shawn Newsam. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. 270–279.

[41] Xixian Yong and Xiao Zhou. 2024. MuseCL: predicting urban socioeconomic indicators via multi-semantic contrastive learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 7536–7544.

[42] Ruixing Zhang, Bo Wang, Tongyu Zhu, Leilei Sun, and Weifeng Lv. 2025. Urban In-Context Learning: Bridging Pretraining and Inference through Masked Diffusion for Urban Profiling. *arXiv preprint arXiv:2508.03042* (2025).

[43] Yaya Zhao, Kaiqi Zhao, Zixuan Tang, Xiaoling Lu, Yuanyuan Zhang, and Yalei Du. 2025. GraphJCL: A Dual-Perspective Graph-Based Framework for Urban Region Representation via Joint Contrastive Learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 37–53.

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

[45] Zhuo Zheng, Yanfei Zhong, Liangpei Zhang, and Stefano Ermon. 2024. Segment any change. *arXiv preprint arXiv:2402.01188* (2024).

[46] Siru Zhong, Xixuan Hao, Yibo Yan, Ying Zhang, Yangqiu Song, and Yuxuan Liang. 2024. Urbancross: Enhancing satellite image-text retrieval with cross-domain adaptation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 6307–6315.

[47] Ke Zhu, Minghao Fu, and Jianxin Wu. 2023. Multi-label self-supervised learning with scene images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6694–6703.

[48] Xingchen Zou, Jiani Huang, Xixuan Hao, Yuhao Yang, Haomin Wen, Yibo Yan, Chao Huang, Chao Chen, and Yuxuan Liang. 2025. Space-aware socioeconomic indicator inference with heterogeneous graphs. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*. 244–256.

[49] Xingchen Zou, Yibo Yan, Xixuan Hao, Yuehong Hu, Haomin Wen, Erdong Liu, Junbo Zhang, Yong Li, Tianrui Li, Yu Zheng, et al. 2025. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Information Fusion* 113 (2025), 102606.