UrbanMMCL: Urban Region Representations via Multi-Modal and Multi-Graph Self-Supervised Contrastive Learning

Jinzhou Cao[a], Jiashi Chen[a], Xiangxu Wang[a], Weiming Huang[c], Dongsheng Chen[d], Tianhong Zhao[a,*], Wei Tu[b], Qingquan Li[b]

[a]*Shenzhen Technology University, School of Artificial Intelligence, Shenzhen, Guangdong, China*
[b]*Shenzhen University, Department of Urban Informatics & Guangdong Key Laboratory of Urban Informatics & Guangdong–Hong Kong-Macau Joint Laboratory for Smart Cities, Shenzhen, Guangdong, China*
[c]*School of Geography, University of Leeds, Leeds, UK*
[d] *Technical University of Munich, Chair of Cartography and Visual Analytics, Munich, Germany*

Abstract

Urban region representation learning has emerged as a fundamental approach for diverse urban analytics tasks, where each neighborhood is encoded as a dense embedding vector for effective downstream applications. However, existing approaches suffer from insufficient multi-modal alignment and inadequate spatial relationship modeling, limiting their representation quality and generalizability. To address these challenges, we propose UrbanMMCL, a novel self-supervised framework that integrates multi-modal multi-view contrastive pre-training with unified fine-tuning for comprehensive urban representation learning. UrbanMMCL employs a dual-stage architecture. First, cross-modal contrastive learning aligns diverse data modalities

*Corresponding author
Email address: zhaotianhong@sztu.edu.cn (Tianhong Zhao)

including remote sensing imagery, street view imagery, location encodings, and Vision-Language Model (VLM)-generated textual descriptions. Second, multi-view adaptive graph contrastive learning captures complex spatial relationships across human mobility, functional similarity, and geographic distance perspectives. The framework then integrates the learned representations through a dedicated fusion mechanism for effective adaptation to downstream tasks. Comprehensive experiments demonstrate that UrbanMMCL consistently outperforms state-of-the-art methods across pollutant emission prediction, population density estimation, and land use classification with minimal fine-tuning requirements, thereby advancing foundation model development for diverse Geo-AI applications.

Keywords: Urban Region Representation Learning, Contrastive Learning, Graph Learning, Multimodal Fusion, Urban Foundation Model

1. Introduction

Urban region representation learning extracts compact features from heterogeneous data to capture spatial, social, and economic characteristics essential for urban tasks like pollution prediction (He and Huang, 2025), socioeconomic estimation (Cao et al., 2025b), and land-use classification (Cao et al., 2025a). These applications provide valuable contributions to urban planning and environmental management in increasingly complex urban environments driven by the rapid urbanization process.

Conventional region representation learning approaches encounter substantial difficulties in efficiently incorporating multi-source multi-modal data to generate unified representations (Zhang et al., 2025). Urban regions are in-

herently heterogeneous, characterized by diverse physical attributes, dynamic socioeconomic activities, and complex interregional interactions. Thus, urban region representations require sophisticated modeling to capture this multidimensional nature (Wang et al., 2024, 2025).

Fusing multi-perspective visual data has emerged as a promising approach to comprehensively reveal heterogeneous urban characteristics. Remote sensing images (RSIs) provide foundational macroscopic perspectives with extensive coverage (Bai et al., 2023; Zhou et al., 2021), capturing urban morphology and land use patterns (Bai et al., 2025), while street view images (SVIs) offer complementary micro-level details by documenting street environments and building facades (Zhao et al., 2025; Zhang et al., 2019, 2024b). Nevertheless, the fusion of multi-perspective visual data presents unique challenges (Zou et al., 2025), as traditional approaches have treated these data independently or employed simplistic fusion mechanisms, failing to preserve complementary information (Gao et al., 2020).

Despite detailed physical characteristics, visual data alone lack semantic depth for complete regional description. While existing representation learning approaches rely on Point-of-Interest (POI) data for textual semantics (Li et al., 2023a), POI data's spatial sparsity and uneven distribution frequently result in inconsistent representation quality (Qin et al., 2025). Vision-Language Models (VLMs) offer promising alternatives by generating comprehensive semantic descriptions from visual content (Huang et al., 2024), transforming RSIs and SVIs into rich textual annotations such as 'high-density residential area' and 'busy transportation hub with commercial activities'. However, current methodologies underutilise these descriptions,

treating them as rudimentary labels rather than leveraging their semantic intricacy for profound visual-textual alignment (Liu et al., 2024).

The modeling inherent spatial relationships between urban regions is imperative for the effective representation learning. Urban regions exhibit multi-faceted spatial interdependencies characterised by geographical adjacency, mobility patterns, and functional similarity (Wang et al., 2024). While recent multi-view frameworks incorporate these dependencies (Li et al., 2019; Wu et al., 2022), they typically process views independently or use simple aggregation strategies (Zhang et al., 2020; Chan and Ren, 2023), missing synergistic information across relational perspectives. The utilization of graph contrastive learning in urban spatial modeling represents a potentially fruitful yet underexplored research avenue (Zhang et al., 2023d; Liu et al., 2025).

To address these challenges, we propose **UrbanMMCL**, a **Urban** **M**ulti-Modal and **M**ulti-View dual **C**ontrastive **L**earning framework that establishes a self-supervised pre-training and fine-tuning paradigm for comprehensive region representation learning. **Pre-training Stage** consists of two synergistic components: (1) **multi-modal vision-language contrastive learning** that aligns RSI, SVI, location encodings, and semantic textual descriptions through specialized encoders and multi-level contrastive objectives; (2) **adaptive multi-view graph contrastive learning** that models complex spatial relationships through dynamic graph structure optimization across multiple relational views. This stage learns generalizable urban representations from unlabeled multi-modal data without requiring task-specific annotations. **Fine-tuning Stage** integrates the pre-trained multi-modal and multi-view representations through dedicated fusion mechanisms, en-

abling effective knowledge transfer to diverse downstream urban analytics tasks including population estimation, pollutant emission monitoring, and land use classification with minimal labeled data requirements.

Our key innovations are fourfold:

1. A systematic dual-stage framework that simultaneously addresses multi-modal data and multi-view relationships, overcoming prior works' single-focus limitation in urban representation learning.
2. An comprehensive multi-modal alignment mechanism that unifies RSI-SVI-Location-Text data through triple contrastive learning, establishing deep semantic alignment while preserving semantic richness and spatial context.
3. Adaptive multi-view spatial modeling that captures complex urban dependencies (proximity, mobility, demographic similarity) through dynamic graph structure learning, enabling effective integration of multiple relational perspectives without requiring predefined graph topologies.
4. A domain-specific self-supervised pre-training paradigm with superior transferability across diverse urban analytics tasks, providing extensive analysis of how different modalities, fusion strategies, and training approaches contribute to representation quality in resource-constrained deployment scenarios.

Section 2 reviews related work on multimodal contrastive learning, graph contrastive learning, and urban representation learning. Section 3 details the UrbanMMCL framework. Section 4 presents experiments and evaluations. Section 5 analyzes model components, training paradigms, and limitations.

Section 6 concludes the study.

2. Related works

2.1. Multimodal contrastive learning

Self-supervised learning (SSL) has emerged as a powerful paradigm that leverages unlabeled data to learn generalizable representations, eliminating the need for costly manual annotations. Among SSL approaches, contrastive learning stands out as a particularly effective technique that learns representations by maximizing similarity between positive pairs while minimizing similarity with negative samples (Dai et al., 2025; Zhang et al., 2023c). Methods such as InstDis (Wu et al., 2018), SimCLR (Chen et al., 2020), and MoCo series (He et al., 2020; Chen et al., 2021) have proven to be effective in learning robust representations from unlabeled data.

Multimodal contrastive learning extends this paradigm by integrating information from different data modalities to create unified representations that capture complementary cross-modal correspondences (Wang et al., 2025; Yong and Zhou, 2024). Vision-language contrastive learning represents a particularly promising approach, combining rich spatial information from imagery with semantic descriptions (Bao et al., 2022). CLIP (Radford et al., 2021b) demonstrates the power of joint image-text representations through contrastive training, enabling enhanced cross-modal understanding. Similar approaches such as ALIGN (Jia et al., 2021) have expanded to billion-level image-text pairs.

In urban analytics, multimodal approaches are particularly crucial due to the inherently complex nature of urban environments, which generate di-

verse data types including RSIs, SVIs, POIs, and textual descriptions (Zhou et al., 2023b; Shen et al., 2023). Recent works have explored this direction in geospatial domains (Weng et al., 2025). GeoCLIP (Cepeda et al., 2023) applies contrastive learning for image-based geolocalization, while SatCLIP (Klemmer et al., 2025) extends CLIP to RSIs, learning representations that bridge RSIs with natural language descriptions. UrbanCLIP (Huang et al., 2024; Yan et al., 2024) specifically targets urban region understanding by integrating satellite imagery or street-view images with textual descriptions, and other works (Liu et al., 2023) have explored vision-language modeling and knowledge-infused contrastive frameworks for enhanced geographic understanding.

However, multimodal contrastive learning for urban region representation remains underexplored. Existing methods typically focus on single visual modalities with limited integration and lack effective adaptation of vision-language models for urban contexts. They treat geographical coordinates as auxiliary features rather than fundamental organizing principles for multimodal alignment. These highlight the need for specialized frameworks tailored to urban representation requirements.

2.2. Graph contrastive learning

Graph Neural Networks (GNNs) have revolutionized urban analysis by modeling urban regions as graph-structured data (Khoshraftar and An, 2024; Cao et al., 2025c). However, most GNN models rely on supervised training requiring substantial labeled data (Ju et al., 2024), which may be unavailable in many urban scenarios. To address these limitations, self-supervised graph contrastive learning (GCL) has emerged as a promising alternative that can

learn meaningful representations without labeled supervision.

GCL integrates both structural and attribute information by maximizing agreement between disparate versions of the same graph while contrasting with negative samples through the implementation of sophisticated architectures and augmentation strategies (Wu et al., 2023; Sun et al., 2020a). Two primary paradigms have emerged: global-local methods such as Deep Graph Infomax (DGI) (Veličković et al., 2018), MVGRL (Hassani and Khasahmadi, 2020), and InfoGraph (Sun et al., 2020b) that contrast node-level with graph-level representations, and local-local approaches such as GRACE (Zhu et al., 2020), GraphCL (You et al., 2020) with its variants (You et al., 2021; Suresh et al., 2021), and GCA (Zhu et al., 2021) that maximize agreement between node embeddings across augmented graph views.

Multi-view graph contrastive learning integrates multiple graph perspectives to capture diverse urban relationships (He et al., 2025). Urban applications have constructed complementary views including POI co-occurrence networks (Huang et al., 2023; Zhang et al., 2023a), trajectory-based mobility graphs (Zhang et al., 2024a), and spatial adjacency graphs (Luo et al., 2022). However, contemporary multi-view GCL methods encounter critical limitations. Zhang et al. (Zhang et al., 2023d) propose a multi-view framework using triplet loss, but their node-level approach with static view construction misses subgraph-level patterns that characterize urban functional areas. Their method relies on fixed topologies and simple augmentation strategies that cannot adapt to dynamic urban spatial relationships. This highlights the need for sophisticated multi-view GCL frameworks that integrate heterogeneous urban data through adaptive augmentation strategies while preserving

semantic coherence of urban functional regions.

2.3. Urban representation learning

Urban region representation learning aims to generate low-dimensional embeddings that reflect urban regional attributes and interregional relationships while preserving spatial and semantic structures. A effective learning requires mining intrinsic correlations among heterogeneous data sources, including geographic topology, urban visual imagery and human mobility (Wang et al., 2026; Guan et al., 2024). This paradigm enables effective analysis across diverse urban applications from sociodemographic prediction to land use classification.

Early methods primarily relied on single modalities such as POI features (Zhai et al., 2019; Sun et al., 2021), human mobility patterns (Zhou and Huang, 2018), or visual imagery (Li et al., 2023b). While achieving task-specific success, single-modal approaches fail to capture multi-dimensional urban characteristics (Zou et al., 2025). Recent advancements focus on multi-modal fusion, integrating spatial, visual and textual data for comprehensive regional characterization (Zou et al., 2025). Representative works include RegionEncoder (Jenkins et al., 2019) for joint encoding of POIs, mobility flows, and RSIs, and Urban2Vec (Wang et al., 2020) combining SVIs with POI descriptions. However, existing approaches predominantly rely on simple concatenation and attention mechanisms, lacking sophisticated semantic alignment and hierarchical adaptive fusion strategies.

The field has evolved from traditional techniques including matrix factorization (Belkin and Niyogi, 2001) and network embedding methods such as DeepWalk and Node2Vec (Perozzi et al., 2014; Grover and Leskovec, 2016)

to GNNs (Xu et al., 2022). Traditional approaches heavily depends on task-customized supervised paradigms (Gao et al., 2020) targeting specific objectives like poverty assessment (Jean et al., 2016; Yeh et al., 2020) and urban function classification (Cao et al., 2020). Recent advances embrace self-supervised learning (Chen et al., 2025), with notable approaches including ReCP (Li et al., 2024) and GraphST (Zhang et al., 2023b), and multiview graph learning such as MVURE (Zhang et al., 2020) and CGAP (Xu and Zhou, 2024). Notwithstanding the advances that have been made, the design of self-supervised pre-training tasks for universal urban representation remains a critical challenge.

3. Methodology

3.1. Preliminaries

We formalize the urban region representation problem through the following key components.

Definition 1. Urban Spatial Partitioning. Given a city divided into N non-overlapping grid regions $\mathcal{R} = \{r_i\}_{i=1}^{N}$, each region r_i is associated with multi-modal urban data.

Definition 2. Remote Sensing Imagery. Remote sensing imagery $\mathcal{I}^{\mathrm{RS}}$ captures aerial views of the earth's surface, providing insights into building distributions and land use patterns. For each region r_i, an orthorectified image patch $\mathcal{I}_i^{\mathrm{RS}} \in \mathbb{R}^{h \times w}$ is used, where h and w are the dimensions of the grid.

Definition 3. Street View Imagery. Street view imagery $\mathcal{I}^{\mathrm{SV}}$ provides ground-level views of urban areas. For each region r_i, multi-directional street

view images are collected as:

$$\mathcal{I}_i^{\mathrm{SV}} = \bigcup_{j=1}^{n} I_{i,j}^{\theta}, \tag{1}$$

where $\{s_{i,j}\}_{j=1}^{n}$ represents the uniformly distributed n sampling points along the road network within the region r_i, and $I_{i,j}^{\theta}$ denotes the image captured at the point $s_{i,j}$. This collection approach ensures comprehensive coverage of urban streetscapes from multiple viewpoints.

Definition 4. VLM-Enhanced Textual Description. Textual descriptions of a region r_i include satellite-derived text $\mathcal{T}_i^{\mathrm{RS}}$ and street-view-derived text $\mathcal{T}_i^{\mathrm{SV}}$. These descriptions are generated through advanced visual language models (VLMs). They provide contextual insights into urban morphology, infrastructure, and functional attributes, complementing visual data.

Definition 5. Multi-view Urban Graph. The urban system is modeled as a collection of multiple view graphs $\mathcal{G} = \{\mathcal{G}^{(k)}\}_{k=1}^{K}$, where each view $\mathcal{G}^{(k)} = (\mathcal{V}, \mathbf{A}^{(k)})$ shares the set of common nodes $\mathcal{V} = \{v_i\}_{i=1}^{N}$ representing the regions of the urban grid, but has distinct adjacency matrices $\mathbf{A}^{(k)} \in \mathbb{R}^{N \times N}$. Each view graph captures a specific type of urban relationship (e.g. POI-based functional similarity, mobility flow, or spatial proximity). This structure enables comprehensive modeling of the urban system through complementary perspectives while maintaining consistent regional representation across views.

Definition 6. Urban Region Representation Learning. Given a set $\mathcal{R}$ of urban regions and K modal feature matrices $\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^K$ derived from multi-modal data sources (e.g. imagery $\mathcal{I}$, textual descriptions $\mathcal{T}$), we

aim to learn a mapping function $F : (r_i, \mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^V) \rightarrow \mathbf{h}_i$ that transforms a region $r_i \in \mathcal{R}$, described by its feature vectors $\mathbf{x}_i^k \in \mathbf{X}^k$ $(1 \leq j \leq K)$, into a d-dimensional representation $\mathbf{h}_i \in \mathbb{R}^d$, where d is a small constant. The resulting region embeddings $\mathcal{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N\}$ should preserve essential urban characteristics across all modalities, enabling their effective application to a wide range of downstream tasks $\mathbf{Y} \in \mathbb{R}^{N \times K}$ across N regions for K different socioeconomic and environmental attributes.

3.2. Overview

The proposed UrbanMMCL framework (Figure 1) enriches urban region representations through a dual-stage contrastive learning approach that establishes a self-supervised pre-training paradigm for urban tasks.

Stage 1: Multi-Modal Multi-View Contrastive Pre-training combines two complementary learning paradigms to establish comprehensive urban representations. The *cross-modal contrastive learning* leverages VLMs to generate semantic descriptions for both RSIs and SVIs, employing specialized encoders (textual, visual, and location) with multiple contrastive objectives including RSI-text alignment, SVI-text alignment, and location-image correspondence. Simultaneously, the *multi-view graph contrastive learning* captures complex spatial dependencies through three distinct view graphs representing mobility patterns (Mob-view), functional similarities based on POI attributes (Fun-view), and spatial distance relationships (Dis-view). Using adaptive graph encoders with independent processing pathways, this component dynamically learns optimized graph structures while capturing both intra-view dependencies and inter-view correlations for comprehensive spatial relationship modeling.

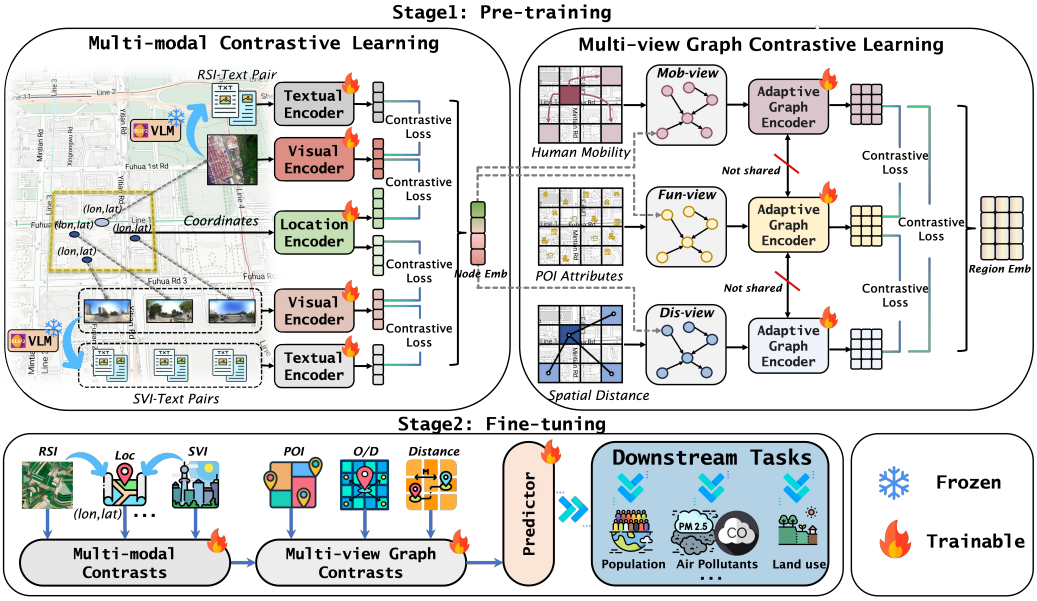


Figure 1: Overview of the UrbanMMCL framework. The framework operates in two stages: (1) **Pre-training Stage** employs multi-modal contrastive learning (aligning RSI, SVI, VLM-generated textual descriptions, and geographical coordinates) alongside multi-view graph contrastive learning across heterogeneous graph views (mobility, functional, and distance) using adaptive graph encoders; (2) **Fine-tuning Stage** integrates the pre-trained multi-modal contrasts and multi-view graph contrasts through dedicated fusion mechanisms, employing trainable predictors for efficient knowledge transfer to downstream urban analytics tasks.

Stage 2: Unified Fine-tuning integrates the pre-trained multi-modal and multi-view representations through dedicated fusion mechanisms, enabling efficient knowledge transfer to diverse downstream urban analytics tasks including population estimation, pollutant emission prediction ($PM_{2.5}$, CO), and land use classification with minimal computational overhead and labeled data requirements. Details are provided in Sections 3.3, 3.4, and 3.5.

3.3. Multi-modal contrastive learning

3.3.1. VLM knowledge distillation for text generation

We employ BLIP-2, a state-of-the-art vision-language model, to perform knowledge distillation from large-scale pre-trained models, extracting rich semantic information from urban imagery. While advanced models like GPT-4V or Gemini possess extensive knowledge capabilities, their prohibitive costs make them impractical for our dataset of tens of thousands of images. BLIP-2 provides an efficient alternative through knowledge distillation via its lightweight Querying Transformer (Q-Former) architecture, which bridges a frozen image encoder and a frozen language model without requiring end-to-end fine-tuning, significantly reducing computational demands while distilling comprehensive knowledge into high-quality textual descriptions. For each RSI or SVI, BLIP-2 processes the input with prompts to generate descriptive text that distills general knowledge into urban-specific semantic representations. Figure 2 illustrates these pairs of prompt descriptions.

3.3.2. Vision-language-location feature encoders

Using VLM-enhanced text generation, we create a dataset of visual-textual pairs $(\mathcal{I}, \mathcal{T})$, where $\mathcal{I}$ represents RSIs I^{RS} or SVIs $\mathcal{I}^{SV}$, and $\mathcal{T}$ includes the corresponding textual descriptions $\mathcal{T}^{RS}$ or $\mathcal{T}^{SV}$. We implement a factorized encoder architecture with dedicated visual, textual, and location encoders, enabling each to capture modality-specific characteristics while establishing the foundation for multimodal alignment.

Visual encoder. We deploy the Vision Transformer (ViT) architecture (Dosovitskiy et al., 2021) to process urban RSIs and SVIs. Recognizing that standard pre-trained models are optimized for general scenes rather

Figure 2: Examples of prompts and corresponding BLIP-2 generated descriptions for RSI and SVI.

than urban environments, we fine-tune this encoder to better capture the unique structural patterns and spatial relationships characteristic of urban landscapes.

Our encoding process begins by dividing each input image $\mathcal{I}_i$ into p non-overlapping patches. Each patch $\mathbf{P}_j$ is flattened and projected into a d-dimensional embedding space with positional encodings:

$$\mathbf{z}_j^{\text{vis}} = \mathbf{E}^{\text{vis}} \cdot \text{Flatten}(\mathbf{P}_j) + \mathbf{p}_j^{\text{vis}}, \quad j = 1, 2, \ldots, p, \tag{2}$$

where $\mathbf{E}^v \in \mathbb{R}^{(h \cdot w) \times d}$ is a learnable projection matrix and $\mathbf{p}_j^v \in \mathbb{R}^d$ is the

positional embedding.

These patch embeddings $\{\mathbf{z}_j^{\text{vis}}\}_{j=1}^{p}$, prepended with a [CLS] token, are processed through L Transformer layers. Each layer applies multi-head self-attention (MSA) followed by a multi-layer perceptron (MLP):

$$\mathbf{z}^{\text{vis}'} = \text{MSA}(\text{LN}(\mathbf{z}^{\text{vis}})) + \mathbf{z}^{\text{vis}} \tag{3}$$

$$\mathbf{z}^{\text{vis}''} = \text{MLP}(\text{LN}(\mathbf{z}^{\text{vis}'})) + \mathbf{z}^{\text{vis}'} \tag{4}$$

The self-attention mechanism allows each patch to attend to all others.

$$\text{Attention}(\mathbf{Q}^{\text{vis}}, \mathbf{K}^{\text{vis}}, \mathbf{V}^{\text{vis}}) = \text{softmax}\left(\frac{\mathbf{Q}^{\text{vis}}(\mathbf{K}^{\text{vis}})^{\top}}{\sqrt{d_k}}\right)\mathbf{V}^{\text{vis}} \tag{5}$$

where $\mathbf{Q}^{\text{vis}} = \mathbf{W}_Q^{\text{vis}}\mathbf{z}^{\text{vis}}$, $\mathbf{K}^{\text{vis}} = \mathbf{W}_K^{\text{vis}}\mathbf{z}^{\text{vis}}$, $\mathbf{V}^{\text{vis}} = \mathbf{W}_V^{\text{vis}}\mathbf{z}^{\text{vis}}$ are linear projections.

After processing through all transformer layers, we obtain the following:

$$\mathbf{X}_i^{\text{vis}} = \phi_{\text{vis}}(\mathcal{I}_i) = \text{Transformer}(\{\mathbf{z}_j^{\text{vis}}\}_{j=1}^{p}), \tag{6}$$

where $\mathbf{X}_i^{\text{vis}} = [\boldsymbol{x}_{\text{CLS}}^{\text{vis}_i}, \boldsymbol{x}_1^{\text{vis}_i}, \boldsymbol{x}_2^{\text{vis}_i}, \cdots, \boldsymbol{x}_p^{\text{vis}_i}] \in \mathbb{R}^{d\times(p+1)}$ contains embeddings for the [CLS] token and all image patches. We extract the [CLS] token embedding $\boldsymbol{x}_{\text{CLS}}^{\text{vis}_i}$ as the global image representation:

$$\boldsymbol{v}_i := \boldsymbol{x}_{\text{CLS}}^{\text{vis}_i}, \quad \boldsymbol{v}_i \in \mathbb{R}^d. \tag{7}$$

The resulting vector $\boldsymbol{v}_i^{\text{rs}}$ or $\boldsymbol{v}_i^{\text{sv}}$ serves as our visual feature for subsequent cross-modal alignment.

Textual encoder. Concurrently, we employ a transformer encoder architecture (Vaswani et al., 2017) to process textual descriptions generated by BLIP-2 for our urban imagery.

Given a text sequence $\mathcal{T}_j$ with tokens n, we tokenize it and map each token to an embedding vector using a learnable embedding matrix $\mathbf{E}^t \in \mathbb{R}^{V\times d}$, where V is the vocabulary size and d is the embedding dimension. Positional embeddings are added to preserve sequential information:

$$\mathbf{z}_i^{\text{text}} = \mathbf{E}^{\text{text}}[t_i] + \mathbf{p}_i^{\text{text}}, \quad i = 1, 2, \ldots, n. \tag{8}$$

The sequence passes through L transformer layers:

$$\mathbf{X}_j^{\text{text}} = \phi_{\text{text}}(\mathcal{T}_j) = \text{Transformer}(\{\mathbf{z}_i^{\text{text}}\}_{i=1}^n), \tag{9}$$

where $\mathbf{X}_j^{\text{text}} = [\mathbf{h}_1^{\text{text}_j}, \mathbf{h}_2^{\text{text}_j}, \ldots, \mathbf{h}_{\text{EOS}}^{\text{text}_j}] \in \mathbb{R}^{d\times(n+1)}$. We extract the [EOS] token embedding as the global text representation:

$$\boldsymbol{t}_j = \mathbf{h}_{\text{EOS}}^{\text{text}_j}, \quad \boldsymbol{t}_j \in \mathbb{R}^d. \tag{10}$$

This global text embedding $\boldsymbol{t}_j^{\text{rs}}$ or $\boldsymbol{t}_j^{\text{sv}}$ captures the semantic content of the entire description, enabling alignment with visual features in subsequent stages.

Locational encoder. We integrate geospatial context by encoding the precise coordinates associated with each visual image. For RSIs, we encode the center coordinates, while for SVIs, we encode the exact sampling locations. Inspired by GeoCLIP (Vivanco et al., 2023), our location encoder transforms geographic coordinates (lon, lat) into meaningful semantic representations, producing high-dimensional embeddings that capture spatial context.

$$\boldsymbol{l}_i = \phi_l(\text{lon}, \text{lat}) \tag{11}$$

These locational embeddings are also further associated with corresponding outputs derived from RSI ($\boldsymbol{l}_i^{\text{rs}}$) and SVI ($\boldsymbol{l}_i^{\text{sv}}$). They can capture crucial

geographic context that complements visual and textual features, enabling our model to understand location-specific urban patterns within a unified representation that preserves both semantic content and spatial relationships.

3.3.3. Cross-modality alignment

We implement a cross-modality alignment framework that integrates visual, textual, and location features through specialized contrastive objectives. For each RSI, we aggregate features from all corresponding SVIs via average pooling, creating comprehensive representations that preserve both spatial correspondence and semantic coherence across diverse urban data modalities.

Vision-language alignment. We establish bidirectional alignment between visual and textual modalities through contrastive learning, optimizing the relationships between visual features and their corresponding textual features. The principle is that representations of the same region across modalities converge in semantic space while remaining distinct from other regions.

For each visual-textual pair, we jointly optimize encoders by contrasting matched pairs against others within the batch through a dual-directional contrastive loss:

$$\begin{aligned}\mathcal{L}_{\text{con}}^{\text{vis-text}} = -\frac{1}{N}\Bigg(&\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(\boldsymbol{v}_i,\boldsymbol{t}_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(\boldsymbol{v}_i,\boldsymbol{t}_j)/\tau\right)}\\&+\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(\boldsymbol{t}_i,\boldsymbol{v}_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(\boldsymbol{t}_i,\boldsymbol{v}_j/\tau\right)}\Bigg),\end{aligned} \tag{12}$$

where $\boldsymbol{v}_i$ represents visual features (either $\boldsymbol{v}_i^{\text{sv}}$ or $\boldsymbol{v}_i^{\text{rs}}$), $\boldsymbol{t}_i$ denotes the corresponding textual features $\boldsymbol{t}^{\text{sv}}$ and $\boldsymbol{t}^{\text{rs}}$, $\text{sim}(\cdot,\cdot)$ represents cosine similarity and

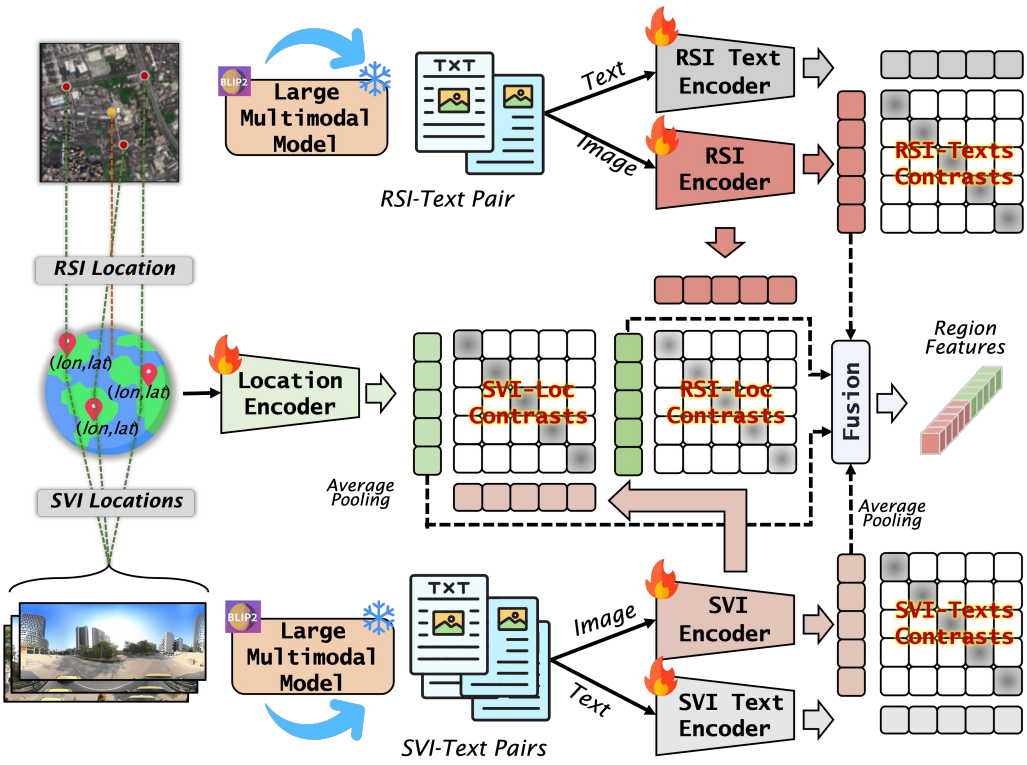


Figure 3: Cross-modal contrastive learning framework integrating visual, textual, and location features through RSI-text, SVI-text, and location-image contrastive objectives.

τ is a temperature hyperparameter that controls the similarity distribution. The first term optimizes image-to-text retrieval, while the second addresses text-to-image retrieval, creating a unified semantic space for effective cross-modal understanding.

Visual-location alignment. To incorporate spatial context, we align

visual characteristics $\boldsymbol{v}_i$ with location features $\boldsymbol{l}_i$ using contrastive learning:

$$\mathcal{L}_{\text{con}}^{\text{vis-loc}} = -\frac{1}{N}\left(\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(\boldsymbol{v}_i,\boldsymbol{l}_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(\boldsymbol{v}_i,\boldsymbol{l}_j)/\tau\right)} + \sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(\boldsymbol{l}_i,\boldsymbol{v}_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(\boldsymbol{l}_i,\boldsymbol{v}_j)/\tau\right)}\right), \tag{13}$$

where $\boldsymbol{l}_i$ denotes the corresponding location features (either $\boldsymbol{l}_i^{\text{sv}}$ or $\boldsymbol{l}_i^{\text{rs}}$), and τ is the temperature hyperparameter. Bidirectional loss optimizes both location retrieval from visual features and visual content retrieval from coordinates, associating street-level features with precise locations while capturing broader spatial relationships in overhead views.

Overall objective. The complete multi-modal contrastive learning loss integrates all cross-modal alignments:

$$\mathcal{L}_{\text{mmcl}} = \mathcal{L}_{\text{con}}^{\text{rsi-text}} + \mathcal{L}_{\text{con}}^{\text{rsi-loc}} + \mathcal{L}_{\text{con}}^{\text{svi-text}} + \mathcal{L}_{\text{con}}^{\text{svi-loc}}. \tag{14}$$

This unified optimization creates a shared embedding space where visual, textual, and spatial information are semantically coherent and mutually reinforcing.

Fusion of multi-modal region features. To construct comprehensive region representations, we integrate features from multiple modalities. Given that each region's RSI typically encompasses multiple SVIs, we first consolidate the SVI features using an averaging operation:

$$\bar{\boldsymbol{v}}_i^{\text{sv}} = \frac{1}{k}\sum_{j=1}^{k}\boldsymbol{v}_j^{\text{sv}}, \quad \bar{\boldsymbol{l}}_i^{\text{sv}} = \frac{1}{k}\sum_{j=1}^{k}\boldsymbol{l}_j^{\text{sv}} \tag{15}$$

where k is the number of SVIs within the region i.

The final region feature representation is formed by concatenating RSI visual features, RSI location features, aggregated SVI features, and aggregated SVI location features:

$$\mathbf{X}_i = [\boldsymbol{v}_i^{\mathrm{rs}}; \boldsymbol{l}_i^{\mathrm{rs}}; \bar{\boldsymbol{v}}_i^{\mathrm{sv}}; \bar{\boldsymbol{l}}_i^{\mathrm{sv}}] \tag{16}$$

This unified embedding seamlessly combines aerial and ground-level visual information with their corresponding geographic contexts, providing a holistic representation that serves as the initial node feature for subsequent region graph construction.

3.4. Multi-view graph contrastive learning

To enhance urban region representations with complex spatial relationships, we employ adaptive multi-view graph contrastive learning that captures intricate interdependencies between urban regions. This component leverages heterogeneous graph views that represent spatial proximity, functional similarity, and mobility flows, using adaptive graph encoders with VGAEs and GCNs to dynamically learn optimized graph structures rather than relying on fixed topologies. The framework incorporates random walk-based subgraph sampling and employs inter-view contrastive learning to model spatial dependencies while preserving view-specific characteristics. This multi-view approach transcends single-view limitations by learning optimal graph representations that demonstrate enhanced robustness against noise and data sparsity in urban spatial interactions.

3.4.1. Multi-view region graph construction

We model geographic regions through a multi-view graph $\mathcal{G} = \{\mathcal{G}^{(k)} = (\mathcal{V}, \mathbf{A}^{(k)}) | k \in \mathcal{K}\}$, where $\mathcal{V}$ represents region nodes, $\mathbf{A}^{(k)}$ denotes the adja-

cency matrix of the relationship type k, and $\mathcal{K} = \{P, M, D\}$ specifies relationship type sets. Our multi-view representation integrates three complementary urban relationships: **human mobility** flows between regions, **POI category similarity** reflecting functional characteristics, and **geographical distance** capturing spatial proximity. The adjacency matrix $\mathbf{A}^{(k)} \in \mathbb{R}^{N \times N}$ for each relationship type $k \in \mathcal{K}$ defines the pairwise connections between the regions, where $\mathbf{A}_{ij}^{(k)}$ quantifies the strength of the connection between the regions v_i and v_j under the corresponding relationship.

Function-aware Region Graph $\mathcal{G}^{(P)} = (\mathcal{V}, \mathbf{A}^{(P)})$: We characterize urban functionality through POIs using a distribution matrix $\mathcal{P} \in \mathbb{R}^{N \times C}$, where C denotes POI categories (restaurants, hotels, hospitals, etc.). Each element $p_{i,c}$ counts places in the region r_i belonging to the c-th POI category, with the functionality of each region encoded as a vector $\mathbf{p}_i \in \mathbb{R}^{1 \times C}$. The adjacency matrix $\mathbf{A}^{(P)} = [a_{ij}^p] \in \mathbb{R}^{N \times N}$ encodes functional similarity through cosine similarity: $a_{ij}^p = \text{sim}(\mathbf{p}_i, \mathbf{p}_j)$, allowing information flow between functionally similar areas regardless of geographical distance.

Mobility-based Region Graph $\mathcal{G}^{(M)} = (\mathcal{V}, \mathbf{A}^{(M)})$: Human movement patterns are captured through trajectory records in format (r_s, r_d, m_{sd}), capturing source/destination regions and departure/arrival times. These trajectories are aggregated into an origin-destination flow matrix $\mathcal{M} = [m_{ij}] \in \mathbb{R}^{N \times N}$, where m_{ij} measures trips from region r_i to r_j. The adjacency matrix $\mathbf{A}^{(M)} = [a_{ij}^m] \in \mathbb{R}^{N \times N}$ is defined as $a_{ij}^m = \frac{\log(1+m_{ij})}{\sum_{k=1}^{N} \log(1+m_{ik})}$, using logarithmic normalization to balance flow variations in regions of different populated regions while preserving movement patterns.

Distance-based Region Graph $\mathcal{G}^{(D)} = (\mathcal{V}, \mathbf{A}^{(D)})$: Spatial proximity

relationships are encoded in a distance matrix $\mathcal{D} \in \mathbb{R}^{N\times N}$ based on Euclidean distances between the centroids of the region. The adjacency matrix $\mathbf{A}^{(D)} = [a_{ij}^d] \in \mathbb{R}^{N\times N}$ is calculated as $a_{ij}^d = 1/d_{ij}$, creating stronger connections between the physically proximate regions. This structure facilitates the propagation of information between adjacent or nearby areas that typically share similar urban characteristics due to their spatial proximity.

3.4.2. Variational graph auto-encoder

We employ Variational Graph Auto-Encoder (VGAE) as the first component of our multi-view graph contrastive learning framework. VGAE's probabilistic nature effectively models variability across all three views, while providing regularized latent representations that prevent overfitting and enable meaningful interpolation between region embeddings.

For each view-specific region graph $\mathcal{G}^{(k)}$ with adjacency matrix $\mathbf{A}^{(k)}$ and feature matrix $\mathbf{X}$, VGAE employs graph convolutional networks (GCNs) to encode graph structure into latent space parameters, specifically the mean vector $\boldsymbol{\mu}$ and the diagonal covariance vector $\boldsymbol{\sigma}$. For notational simplicity, we omit the superscript (k):

$$\boldsymbol{\mu} = \mathrm{GCN}_{\mu}(\mathbf{A}, \mathbf{X}), \quad \boldsymbol{\sigma} = \mathrm{GCN}_{\sigma}(\mathbf{A}, \mathbf{X}). \tag{17}$$

Latent representations are sampled using the reparameterization trick. For each node i, we have:

$$\mathbf{z}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \mathbf{I}), \tag{18}$$

where $\mathbf{z}_i \in \mathbb{R}^d$ is the embedding for node i, and $\boldsymbol{\epsilon}_i$ is random noise. The complete embedding matrix $\mathbf{Z} = [\mathbf{z}_1; \mathbf{z}_2; \ldots; \mathbf{z}_N] \in \mathbb{R}^{N\times d}$ contains all node embeddings.

The decoder reconstructs the adjacency matrix from the latent node representations. For each pair of nodes (i, j), the edge probability is:

$$\tilde{A}_{ij} = \text{Softplus}(\mathbf{z}_i^\top \mathbf{z}_j), \tag{19}$$

where $\text{Softplus}(\cdot)$ is the activation function.

The VGAE is trained by optimizing both reconstruction loss and KL-divergence. The reconstruction loss minimizes the difference between original and reconstructed adjacency matrices:

$$\mathcal{L}_{\text{recon}} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(A_{ij} - \tilde{A}_{ij} \right)^2. \tag{20}$$

The KL divergence loss regularizes the latent space by minimizing divergence between the learned latent distribution $q(\mathbf{z}_i|\mathbf{X}, \mathbf{A}) = \mathcal{N}(\boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}_i^2))$ and a prior distribution $p(\mathbf{z}_i) = \mathcal{N}(0, \mathbf{I})$:

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{d} \left(1 + \log \sigma_{ij}^2 - \mu_{ij}^2 - \sigma_{ij}^2 \right), \tag{21}$$

This process is applied independently to each view $\mathcal{G}^{(k)}$, yielding respective reconstructed adjacency matrices $\tilde{\mathbf{A}}^{(k)}$. These view-specific embeddings capture different aspects of urban region relationships for subsequent multi-view contrastive learning.

3.4.3. Random walk-based subgraph generation

To efficiently handle complex urban graph structures and enhance robustness against data skewness, we employ adaptive random walks on the reconstructed view-specific graphs $\tilde{\mathcal{G}}^{(P)}$, $\tilde{\mathcal{G}}^{(M)}$, and $\tilde{\mathcal{G}}^{(D)}$. For each node i in a given graph $\tilde{\mathcal{G}}$, we perform a single random walk of fixed length L to capture

both local and global structural properties while ensuring equal contribution from each node.

Starting from node i, the walker transitions from the current node v_t to neighboring node v_{t+1} based on adaptive transition probabilities:

$$P(v_{t+1}|v_t) = \frac{\tilde{A}_{v_t,v_{t+1}}}{\sum_{v_k \in \mathcal{N}(v_t)} \tilde{A}_{v_t,v_k}}, \tag{22}$$

where $\mathcal{N}(v_t)$ denotes neighbors of v_t, and $\tilde{A}_{v_t,v_{t+1}}$ is the learned edge weight. These adaptive allow transition probabilities to prioritize stronger connections while reducing influence of weaker relationships.

The walker continues for L steps, with probabilistic rather than deterministic sampling to introduce variability and enhance robustness. If a node lacks neighbors, the walk terminates early. Upon completion, the visited node sequence $[v_0, v_1, \ldots, v_L]$ forms subgraph $\hat{\mathcal{G}}_i$, including all visited nodes and their interconnecting edges.

This sampling strategy effectively addresses noisy edges by utilizing VGAE-learned weights to guide walks toward relevant connections. For each subgraph, we learn node-level representations through:

$$\mathbf{H}_i = \mathrm{GCN}(\hat{\mathcal{G}}_i, \mathbf{X}[\mathcal{V}_i]), \tag{23}$$

where $\mathbf{H}_i \in \mathbb{R}^{|\mathcal{V}_i| \times d}$ contains learned node embeddings, and $\mathbf{X}[\mathcal{V}_i]$ represents features restricted to the subgraph nodes.

We then aggregate subgraph information into a single comprehensive vector through readout operations:

$$\boldsymbol{s}_i = \mathrm{READOUT}(\mathbf{H}_i), \tag{24}$$

where $\mathbf{s}_i \in \mathbb{R}^d$ encapsulates structural and semantic information of the neighborhood centered on node i.

The process is repeated independently for each view, generating three sets of subgraph representations $\boldsymbol{s}_i^{(k)}$, where $k \in \mathcal{K} = \{P, M, D\}$, each capturing view-specific structural properties. Unlike rule-based methods, this approach adaptively captures region dependencies through learnable parameters, providing enhanced resilience against data skewness and noise.

3.4.4. Inter-view contrastive learning

After generating view-specific subgraphs and their embeddings, we design a multi-view contrastive learning objective that maximizes mutual information between different view representations of the same node while minimizing it between different nodes. This core intuition is that subgraphs centered around the same urban region should exhibit semantic similarities despite structural differences across views. This contrastive mechanism achieves automatic feature selection through the InfoNCE loss structure, where the numerator enforces cross-view alignment while the denominator ensures inter-node discrimination. This design amplifies features that are consistently discriminative across multiple views while suppressing view-specific noise that fails to maintain cross-view consistency. Consequently, the model automatically prioritizes meaningful urban relationships while filtering out noisy connections, resulting in robust representations that capture fundamental regional patterns.

For any pair of views (k_1, k_2), the contrastive loss is formulated as:

$$\mathcal{L}_{\text{con}}^{k_1,k_2} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{s}_i^{(k_1)}, \mathbf{s}_i^{(k_2)})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{s}_i^{(k_1)}, \mathbf{s}_j^{(k_2)})/\tau)}, \tag{25}$$

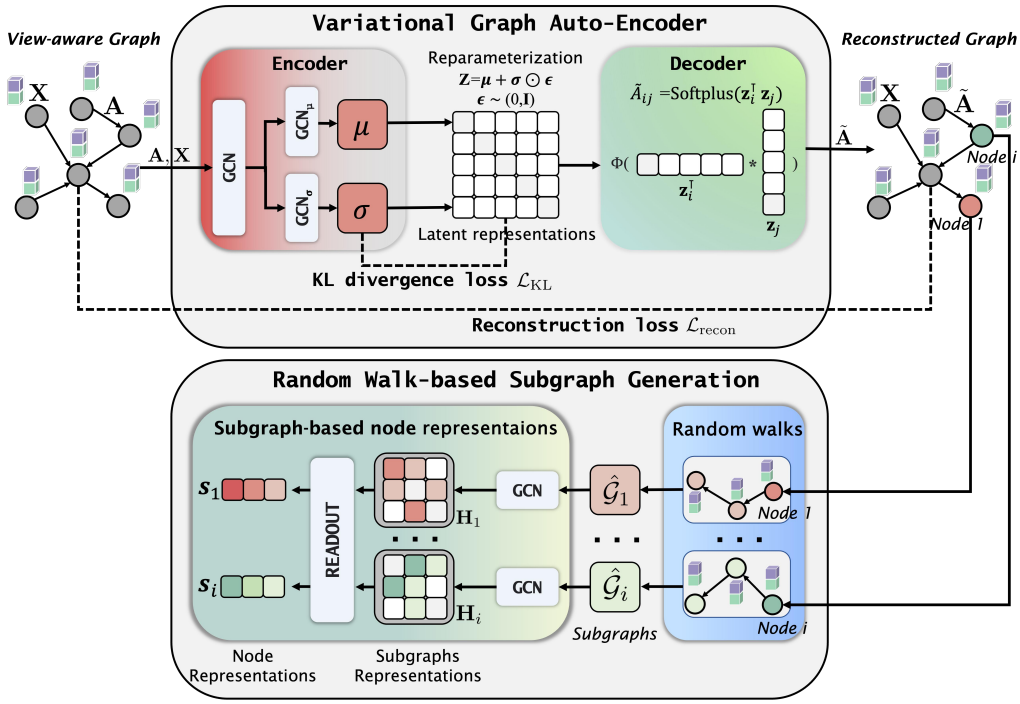


Figure 4: Architecture of the adaptive graph encoder framework. The top component shows the Variational Graph Auto-Encoder (VGAE) that learns latent graph representations through an encoder-decoder structure with reparameterization, incorporating both KL divergence and reconstruction losses for adaptive graph structure learning. The bottom component illustrates the random walk-based subgraph generation process, where random walks from each node create diverse subgraphs that are processed through GCNs to generate enhanced node representations.

where $\mathbf{s}_i^{(k_1)}$ and $\mathbf{s}_i^{(k_2)}$ are the embeddings of node i in views k_1 and k_1, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ is a temperature parameter.

Overall objective. The complete multi-view graph contrastive learning loss combines view-specific graph reconstruction with cross-view alignment:

$$\mathcal{L}_{\text{mvgcl}} = \sum_{k \in \mathcal{K}} \left(\mathcal{L}_{\text{recon}}^{(k)} + \mathcal{L}_{\text{KL}}^{(k)} \right) + \sum_{(k_1,k_2) \in \mathcal{K} \times \mathcal{K}} \mathcal{L}_{\text{con}}^{k_1,k_2}. \tag{26}$$

where $\mathcal{K} \times \mathcal{K} = \{(P, M), (P, D), (M, D)\}$ represents all view pairs.

The first term maintains view-specific structural information through VGAE reconstruction and regularization losses, ensuring each view preserves its inherent graph properties. The second term enforces cross-view consistency by aligning representations across pairwise views.

3.5. Urban Region Representation Task

3.5.1. Pre-training stage

Our framework employs a unified pre-training strategy that jointly optimizes multi-modal contrastive learning and multi-view graph contrastive learning. This approach aligns heterogeneous data modalities into a coherent feature space while learning view-invariant representations across different urban graph perspectives. The complete pre-training objective combines both learning stages:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{mmcl}} + \beta \mathcal{L}_{\text{mvgcl}}, \tag{27}$$

where α and β are hyperparameters to balance cross-modal alignment and inter-view consistency.

This self-supervised framework generates complementary learning signals through modality fusion and view integration, producing robust representations that are both semantically meaningful and structurally consistent. After pre-training, we obtain region embeddings $\mathbf{Z}^* \in \mathbb{R}^{N \times d}$, providing a strong foundation for downstream urban analysis tasks.

3.5.2. Fine-tuning for downstream tasks

We employ a task-specific fine-tuning approach to tailor our region embeddings to various urban prediction tasks. In this process, all model parameters are updated, and the generated embeddings $\mathbf{Z}^*$ are used as input features for a lightweight MLP classifier or regressor: $Y_i = MLP(\mathbf{Z}^*)$. This strategy capitalizes on the rich representations learned during pre-training to demonstrate how effectively our framework captures essential urban patterns across various applications.

We evaluated our framework on three downstream tasks: pollutant emission prediction (PEP), population density estimation (PDE), and land use classification (LUC). PEP involves predicting environmental pollutant levels, which tests the model's ability to capture environmental and spatial factors that influence emissions. PDE is a regression task that estimates regional population density, assessing how well the embeddings capture demographic patterns. LUC is a multi-class classification task categorizing regions into specific land use types (residential, commercial, industrial, etc.), evaluating the model's capacity to identify distinct urban functional patterns. This fine-tuning approach demonstrates efficient transfer of learned representations to diverse urban applications spanning environmental, demographic, and land-use domains with minimal additional training.

4. Experiments and analyses

4.1. Study areas and datasets

As illustrated in Figure 5, our research focuses on Shenzhen, a rapidly developing metropolitan city in China with distinct urban characteristics. he

study leverages diverse datasets for both pre-training and fine-tuning phases to conduct comprehensive experiments and analyses.

The Pre-training Data. The pre-training phase utilizes multi-modal urban data including RSI, SVI, POI data, and human mobility data to learn comprehensive urban representations. RSI, sourced from the GaoFen-2 satellite via Tianditu, features 1.0 meter spatial resolution and three spectral bands (RGB). These images were segmented into 1 km $\times$ 1 km tiles to align with the grid-based analysis framework. Complementing this, 224,826 high-resolution panoramic SVIs of Shenzhen (4096 $\times$ 1036 pixels), obtained from Baidu Maps at approximately 15-meter intervals along the road network, provide comprehensive 360-degree ground-level visual coverage. POI data, sourced from AMap[1], includes 1,064,085 points of interest categorized into 23 primary classes such as Life Services, Corporate Entities and Mixed-use Commercial and Residential Areas. Furthermore, human mobility data from China Unicom[2] consist of 34,960,199 hourly movement records, aggregated to daily origin-destination flows at a 1 km $\times$ 1 km resolution.

The Fine-tuning Data. The fine-tuning phase employs task-specific datasets for downstream urban analysis applications. Population density data from WorldPop[3] uses random forest-based dasymetric mapping to deliver high-resolution estimates (30 arc seconds, approximately 1 km at the equator) of people per km^2, resampled to align with our grid structure. Pollutant emission data include key air pollutants—carbon monoxide (CO,

[1]https://lbs.amap.com
[2]http://www.smartsteps.com/
[3]https://hub.worldpop.org

mg/m^3) and particulate matter ($PM_{2.5}$, μg/m^3)—sourced from the National Tibetan Plateau Scientific Data Center (Wei et al., 2023; Wei and Li, 2024), providing high-quality 1 km × 1 km resolution raster data. Land use classification data are derived from SinoLC-1 (Li et al., 2022, 2023c), China's first national-scale 1 meter resolution land cover map developed using deep learning techniques, with the Shenzhen portion specifically utilized for our analysis.

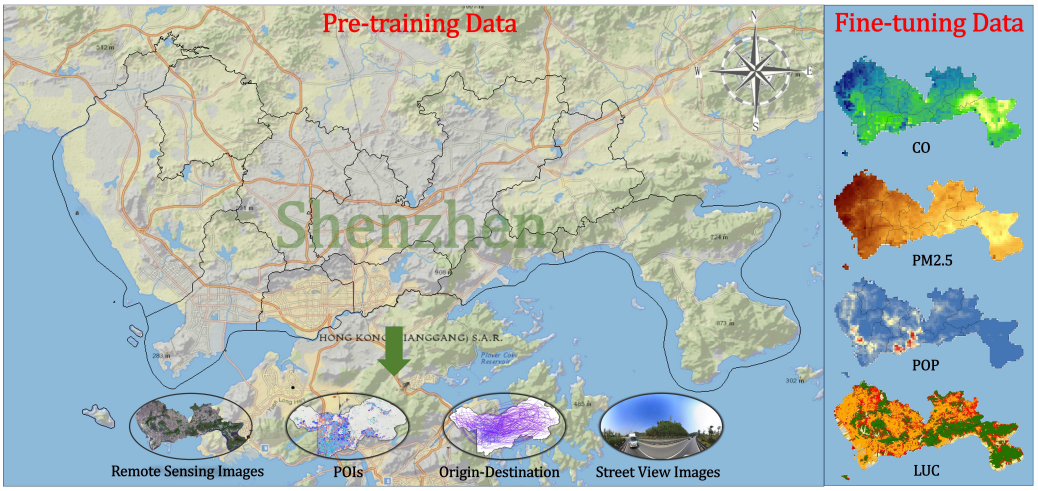


Figure 5: Overview of the datasets used in this study.

4.2. Experiment setup

4.2.1. Baselines

To comprehensively evaluate our model, we compare with six recent baselines in two categories: (1) **Vision-based methods**, including ViT, PG-SimCLR, and UrbanVLP, where the latter two employ contrastive learning strategies. (2) **Graph-based methods**, including MVURE, HREP and ReMVC:

- **ViT (Dosovitskiy et al., 2021).** ViT adapts transformers to computer vision by partitioning images into fixed-size patches and demonstrates strong performance with sufficient pre-training data. In our experiments, we employ ViT-B as the baseline image encoder, concatenating extracted features from different modalities for final embeddings.

- **PG-SimCLR (Xi et al., 2022).** A contrastive learning framework that adapts SimCLR(Chen et al., 2020) for urban region representation using satellite imagery by incorporating geographic proximity constraints and POI category distributions, allowing the model to learn representations that respect both spatial relationships and functional similarities.

- **UrbanVLP (Hao et al., 2025).** A multi-granularity vision-language pretraining framework that combines RSI, SVI, and high-quality textual descriptions to predict urban socioeconomic indicators through cross-modal alignment and automatic text calibration.

- **MVURE (Zhang et al., 2020).** Leverages human mobility data and urban region attributes (POI and check-in data) to construct multi-view correlations through graph attention networks, enabling cross-view information sharing and adaptive fusion for comprehensive urban region embeddings. In our experiments, we did not use check-in data.

- **HREP (Zhou et al., 2023a).** A relation-aware graph-based approach using human mobility, POI information, and geographic neighbor data, combined with prompt learning to capture intra-region and

inter-region correlations for robust region embeddings.

- **ReMVC (Zhang et al., 2023a).** Employs multi-view contrastive learning with POI data and human mobility records to extract robust region embeddings by capturing intra-view distinctions and cross-view correlations.

4.2.2. Evaluation metrics

To quantitatively evaluate the performance of our method, we employ standard metrics for regression and classification tasks (↓ indicates lower is better, and ↑ indicates higher is better):

Metrics for Regression Tasks. For regression tasks, the goal is to predict continuous variables. We assess model performance using three complementary metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2).

Metrics for Classification Tasks. For classification tasks, our objective is to evaluate the ability of the model to correctly classify samples into their respective categories. We employ two complementary metrics: F1 Score and Recall. These metrics provide a comprehensive assessment of the accuracy and robustness of our method across different task types.

4.2.3. Implementation details

Pre-training Setup. We divide the dataset into 60% training, 20% validation and 20% testing sets. For image inputs, we apply data augmentation techniques that include random cropping, flipping, and normalization, following the methodology described in Radford et al. (2021a). During the multi-modal contrastive learning stage, we generate textual descriptions for

each image using the BLIP-2 model, leveraging OPT-2.7b (a large language model with 2.7 billion parameters). Text descriptions are limited to 77 tokens. For the visual encoder, we use ViT-B/32 with a hidden dimension of 768 and an output dimension of 512. For the textual encoder, both the hidden and output dimensions are 512. For the location encoder, we adopt the architecture and configuration settings proposed by Vivanco et al. (2023). For feature aggregation, SVI features within the same RSI coverage area are aggregated using average pooling. The aggregated SVI features are then added element-wise to the RSI features to form unified regional representations. In the multi-view graph contrastive learning stage, we construct three graph structures based on population flow, POI similarity, and spatial distance. VGAE is used to reconstruct these graphs. VGAE uses a one-layer GCN encoder with an output latent space size of 64, and LeakyReLU activation, and Adam optimizer with learning rate 1e-4. Next, random walks with walk length 20 are performed on the reconstructed graphs to generate subgraphs. Finally, GCNs with hidden layer size 128, output size 64, and LeakyReLU activation are applied for graph representation embedding. The temperature parameter τ for contrastive learning is 0.5. The Adam optimizer is used with a learning rate of 1e-6 and weight decay of 1e-4. The model is trained for 1000 epochs with early stopping based on validation loss.

Fine-tuning Setup. During fine-tuning, we update both the pre-trained encoders and the task-specific components. For regression tasks, we use mean squared error loss, while for classification tasks, we use cross-entropy loss with accuracy as the evaluation metric. The training runs for 1000 epochs using the Adam optimizer, with learning rate 1e-6 and weight decay 1e-4.

Early stopping is applied based on validation loss to prevent overfitting. All experiments were performed on NVIDIA A6000 GPUs with 48GB memory.

4.3. Model performance

4.3.1. Hyperparameter sensitivity analysis

To understand the interaction between multi-modal and multi-view graph contrastive learning, we conduct sensitivity analysis of hyperparameters α and β in our unified objective function $\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{mmcl}} + \beta\mathcal{L}_{\text{mvgcl}}$ with constraint $\alpha + \beta = 1$. Table 1 shows the optimal configuration occurs at $\alpha = 0.5, \beta = 0.5$ across all tasks, indicating that indicating equal weighting between multi-modal contrastive learning and spatial learning yields best performance.

The results also reveal asymmetric degradation patterns. Pure graph learning ($\alpha = 0$) causes dramatic performance drops, while pure multi-modal learning ($\alpha = 1.0$) shows moderate decreases. This asymmetry indicates that multi-modal information provides fundamental semantic grounding, while spatial learning offers crucial structural guidance. The consistent optimal ratio ($\alpha : \beta = 1 : 1$) across diverse urban tasks suggests that effective urban representation learning requires balanced integration of semantic richness and urban spatial structure, , rather than over-relying on either component alone.

4.3.2. Comparison with baseline methods

We conduct comprehensive comparisons with state-of-the-art methods. Table 2 presents the overall results, from which we can derive the following

α	β	PEP(CO)			PEP($PM_{2.5}$)			PDE			LUC	
		MAE ↓	**MSE** ↓	**R^2** ↑	**MAE** ↓	**MSE** ↓	**R^2** ↑	**MAE** ↓	**MSE** ↓	**R^2** ↑	**F1** ↑	**Recall** ↑
1.0	0.0	0.0382	0.0018	0.4025	1.4009	3.2739	0.5019	4134.67	14969395.03	0.4817	0.4106	0.4235
0.8	0.2	0.0248	0.0012	0.6034	1.2431	2.1900	0.6756	3816.31	15647112.25	0.4841	0.4189	0.4673
0.6	0.4	0.0250	0.0012	0.5927	0.8591	1.4778	0.7654	3622.39	13598495.15	0.6556	0.4504	0.4957
0.5	**0.5**	**0.0212**	**0.0009**	**0.7417**	**0.6785**	**1.0212**	**0.8481**	**1931.64**	**9943831.52**	**0.7670**	**0.6058**	**0.5994**
0.4	0.6	0.0247	0.0012	0.6857	0.8140	1.3593	0.7895	3656.10	16647168.90	0.6777	0.4844	0.5229
0.2	0.8	0.0238	0.0011	0.6290	1.0394	1.7856	0.7234	3860.65	19548932.31	0.5648	0.4738	0.5068
0.0	1.0	0.0327	0.0018	0.4341	1.4808	3.6995	0.4824	3824.24	17542438.41	0.5194	0.4223	0.4445

Table 1: Hyperparameter sensitivity analysis for α and β. The best results are highlighted in **boldface**.

Methods	PEP(CO)			PEP($PM_{2.5}$)			PDE			LUC	
	MAE ↓	**MSE** ↓	**R^2** ↑	**MAE** ↓	**MSE** ↓	**R^2** ↑	**MAE** ↓	**MSE** ↓	**R^2** ↑	**F1** ↑	**Recall** ↑
ViT	0.0252	0.0012	0.5569	1.0062	1.7733	0.7225	2086.0500	12767905.19	0.6654	0.5136	0.5235
PG-SimCLR	0.0358	0.0020	0.4158	1.4840	3.6337	0.4280	2194.7627	13386507.27	0.7116	0.4690	0.4555
UrbanVLP	0.0214	0.0011	0.6875	0.7573	1.1847	0.8240	1954.4890	**9754879.86**	0.7635	0.5811	0.5873
MVURE	0.0239	0.0011	0.5669	0.8121	1.0288	0.7933	2046.9664	11701711.17	0.6933	0.5450	0.5446
HREP	0.0237	0.0012	0.5604	0.8860	1.5067	0.7677	2089.4557	13625186.03	0.6429	0.4299	0.4291
ReMVC	0.0260	0.0013	0.5217	1.3249	3.0701	0.5106	2242.8103	16310952.66	0.5725	0.4821	0.4718
UrbanMMCL	**0.0212**	**0.0009**	**0.7417**	**0.6785**	**1.0212**	**0.8481**	**1931.64**	9943831.52	**0.7670**	**0.6058**	**0.5994**
Improvement(%)	**0.93**	**18.18**	**7.88**	**10.41**	**0.74**	**2.92**	**1.17**	**-1.94**	**0.46**	**4.25**	**2.06**

Table 2: Performance comparison of different methods on pollutant emission prediction (PEP) (CO, $PM_{2.5}$), population density estimation (PDE), and land use classification (LUC). The best results are in **boldface**, and the second-best results are underlined. Improvement(%) shows the relative improvement of our method over the second-best baseline.

key findings.

(1) UrbanMMCL achieves superior performance across most metrics, demonstrating the effectiveness of our dual contrastive learning approach. Our framework outperforms the best baselines in 10 of 11 metrics, with an average R^2 improvement of 3.75% in regression tasks and a 4. 25% improvement in the F1 score for classification compared to the second best method (UrbanVLP). The only exception is the PDE MSE metric, where our method shows a marginal difference of 1. 94%. This slight

discrepancy stems from the long-tail distribution of population density data, where extreme values disproportionately influence the squared error metric. Our superior MAE and R^2 scores demonstrate robustness across the majority of urban regions.

(2) Text-enhanced vision-language methods significantly outperform single-modality and POI-enhanced approaches. UrbanVLP consistently outperforms both the vision-only ViT model and the POI-enhanced PG-SimCLR across all tasks, with notable improvements in $PM_{2.5}$ prediction (R^2: 0.8240 vs 0.4280 for PG-SimCLR). This confirms that rich textual descriptions provide more contextually relevant information than structured POI data alone. UrbanMMCL further advances this paradigm by effectively integrating visual features with geographical coordinates and adaptive graph relationships.

(3) Adaptive graph contrastive learning significantly outperforms static graph-based methods. Unlike existing graph-based methods (MVURE, HREP and ReMVC) that rely on predetermined region relationships, UrbanMMCL uses VGAE and adaptive random walks to automatically learn and refine meaningful region connections. Our method achieves a remarkable 30.8% R^2 improvement in CO prediction over MVURE, highlighting how our adaptive approach addresses the limitations of fixed graph structures in complex urban environments.

(4) The synergy between multi-modal integration and multi-view graph modeling creates generalizable urban representations. The integration of RSI, SVI, geographical positions, and textual descriptions through dual-stage contrastive learning allows UrbanMMCL to capture both

Methods	PEP(CO)			PEP($PM_{2.5}$)			PDE			LUC	
	MAE ↓	**MSE** ↓	$\mathbf{R^2}$ ↑	**MAE** ↓	**MSE** ↓	$\mathbf{R^2}$ ↑	**MAE** ↓	**MSE** ↓	$\mathbf{R^2}$ ↑	**F1** ↑	**Recall** ↑
RSI-CLIP	0.0313	0.0019	0.4573	1.4185	2.7707	0.5969	2615.4236	16563126.19	0.5844	0.4952	0.5114
SVI-CLIP	0.0288	0.0015	0.5435	1.1461	2.1668	0.6376	2404.1780	14664183.37	0.5596	0.4597	0.4394
w/o Text	0.0246	0.0013	0.6372	0.8683	1.4384	0.7751	1996.8511	10413018.36	0.6957	0.5550	0.5536
w/o MCL	0.0327	0.0018	0.4341	1.4808	3.6995	0.4824	3824.2488	17542438.41	0.5194	0.4223	0.4445
UrbanMMCL	**0.0212**	**0.0009**	**0.7417**	**0.6785**	**1.0212**	**0.8481**	**1931.64**	**9943831.52**	**0.7670**	**0.6058**	**0.5994**

Table 3: Ablation on multimodal components. The best results are highlighted in **boldface**.

Methods	PEP(CO)			PEP($PM_{2.5}$)			PDE			LUC	
	MAE ↓	**MSE** ↓	$\mathbf{R^2}$ ↑	**MAE** ↓	**MSE** ↓	$\mathbf{R^2}$ ↑	**MAE** ↓	**MSE** ↓	$\mathbf{R^2}$ ↑	**F1** ↑	**Recall** ↑
w/o $\mathcal{G}^{(P)}$	0.2284	0.0011	0.5342	1.1219	1.9569	0.6983	2527.02	17422814.99	0.5768	0.3852	0.3952
w/o $\mathcal{G}^{(M)}$	0.0202	0.0009	0.5910	1.2310	2.5737	0.6111	2629.2102	18889712.73	0.5679	0.4702	0.5124
w/o $\mathcal{G}^{(D)}$	0.0276	0.0014	0.3575	1.0027	1.8363	0.6967	2454.5840	17698496.25	0.4292	0.4450	0.4373
w/o VGAE	0.0263	0.0014	0.6020	1.0447	1.9016	0.6771	1935.23	13555443.7517	0.6531	0.4424	0.4428
w/o RW	0.0220	0.0010	0.6554	0.7185	1.2764	0.7947	1912.9005	11225293.52	0.6808	0.4680	0.4851
w/o GCL	0.0335	0.0019	0.4104	1.8679	3.5980	0.5078	2095.8574	13879595.94	0.6071	0.3966	0.4123
UrbanMMCL	**0.0212**	**0.0009**	**0.7417**	**0.6785**	**1.0212**	**0.8481**	**1931.64**	**9943831.52**	**0.7670**	**0.6058**	**0.5994**

Table 4: Ablation on multi-view graph components. The best results are highlighted in **boldface**.

fine-grained visual details and macro-scale spatial relationships. This comprehensive modeling creates generalizable features that maintain consistent performance across both regression and classification tasks, from environmental monitoring (CO, $PM_{2.5}$) to socioeconomic analysis (PDE, LUC).

4.3.3. Cross-city generality

To assess the generalization capability of UrbanMMCL, we conducted validation studies in Beijing and Chengdu, two cities with contrasting developmental and geographic profiles. Our evaluation adopts hierarchical transfer learning leveraging UrbanMMCL's modular design, where multimodal encoders trained on Shenzhen are directly transferred while graph components

Cities	Models	PEP(CO)	PEP($PM_{2.5}$)	PDE	LUC
		R^2 ↑	**R^2** ↑	**R^2** ↑	**F1** ↑
Beijing	PG-SimCLR	0.4929	0.4301	0.6109	0.4380
	ReMVC	0.5488	0.5157	0.5433	0.4791
	UrbanMMCL	**0.7032**	**0.7811**	**0.6317**	**0.5538**
	Improvement(%)	+28.15%	+51.46%	+3.40%	+15.59%
Chengdu	PG-SimCLR	0.5171	0.4540	0.5583	0.4581
	ReMVC	0.5728	0.5384	0.5679	0.4904
	UrbanMMCL	**0.6912**	**0.7508**	**0.6420**	**0.5395**
	Improvement(%)	+20.67%	+39.45%	+13.04%	+10.01%

Table 5: Cross-city transfer learning performance comparison in Beijing and Chengdu.

are re-initialized for city-specific spatial relationships.

Table 5 demonstrates impressive cross-city performance. Both cities achieve achieve strong performance with R^2 scores of 0.63-0.78 across regression tasks and competitive F1 scores of 0.54-0.55 for land use classification, maintaining remarkably consistent results despite diverse urban contexts. UrbanMMCL consistently outperforms baseline methods PG-SimCLR and ReMVC by 10-51%, confirming robust generalization across diverse urban environments.

4.3.4. Ablation studies

To validate our design principles, we conduct comprehensive ablation studies addressing two key questions: (1) What are the essential multi-modal components and integration strategies for effective urban representation learning? (2) What are the essential graph perspectives and learning mechanisms for effective urban spatial relationship modeling?

Ablation on multimodal components. We design four variants to test specific hypotheses: (1) **RSI-CLIP** and **SVI-CLIP** replace our domain-specific encoders with general-purpose pre-trained CLIP model (ViT-B/32) to assess the necessity of domain specialization; (2) ***w/o* Text** eliminates textual enhancement to quantify semantic information contribution; (3) ***w/o* MCL** removes cross-modal contrastive learning and initializes the encoders with their original weights while preserving multimodal fusion to isolate the impact of explicit cross-modal alignment.

Table 3 reveals three key findings that validate our design choices. First, replacing specialized components with general CLIP causes substantial degradation, demonstrating that urban scene understanding requires architectural adaptations beyond general vision models. Second, removing text consistently decreases performance by 14.1% across tasks, with environmental monitoring particularly affected, showing that semantic descriptions capture abstract urban characteristics invisible to visual features alone. Third, eliminating contrastive learning causes the most severe degradation, confirming that explicit cross-modal alignment is essential for coherent multimodal representations. These results demonstrate that each component addresses specific urban representation challenges, and their synergistic integration is critical for optimal performance across diverse urban tasks.

Ablation on multi-view graph components. We design six variants testing: (1) individual graph view contributions (***w/o*** $\mathcal{G}^{(P/M/D)}$); (2) ***w/o* VGAE**: replacing the VGAE with standard GCN for graph encoding to assess probabilistic graph structure learning; (3) ***w/o* RW**: eliminating random walk-based subgraph generation and using full graphs to evaluate local struc-

ture sampling effectiveness; and (4) ***w/o* GCL**: removing graph contrastive learning while retaining basic graph encoders to isolate cross-view alignment impact.

Table 4 reveals critical insights into urban spatial modeling. First, distance-based graph removal causes the most severe degradation, particularly impacting the CO prediction, while mobility-based graph and function-aware graph removal results in 24.8% and 23.5% average decrease, demonstrating that geometric relationships serve as fundamental structural foundation with all three perspectives capturing distinct spatial aspects. Second, replacing VGAE with standard GCN decreases performance by 17.9%, while removing random walk sampling causes a 9.7% drop, demonstrating that both probabilistic structure learning and local sampling contribute to effective spatial modeling. Third, eliminating graph contrastive learning results in the largest performance decline with 35.2% average R^2 decrease, confirming that learning coherent multi-perspective representations requires explicit alignment mechanisms. These results demonstrate that effective urban spatial understanding requires integrated design of multiple graph perspectives, adaptive structure learning, and cross-view contrastive alignment.

4.4. Analysis of learned representations

4.4.1. Representation visualizations

To validate that our model learns meaningful representations that effectively distinguish different urban area types, we examine learned region representations through a t-SNE dimensionality reduction, which maps the high-dimensional embeddings learned by UrbanMMCL into an interpretable two-dimensional space. Figure 6 reveals three distinct clusters with high

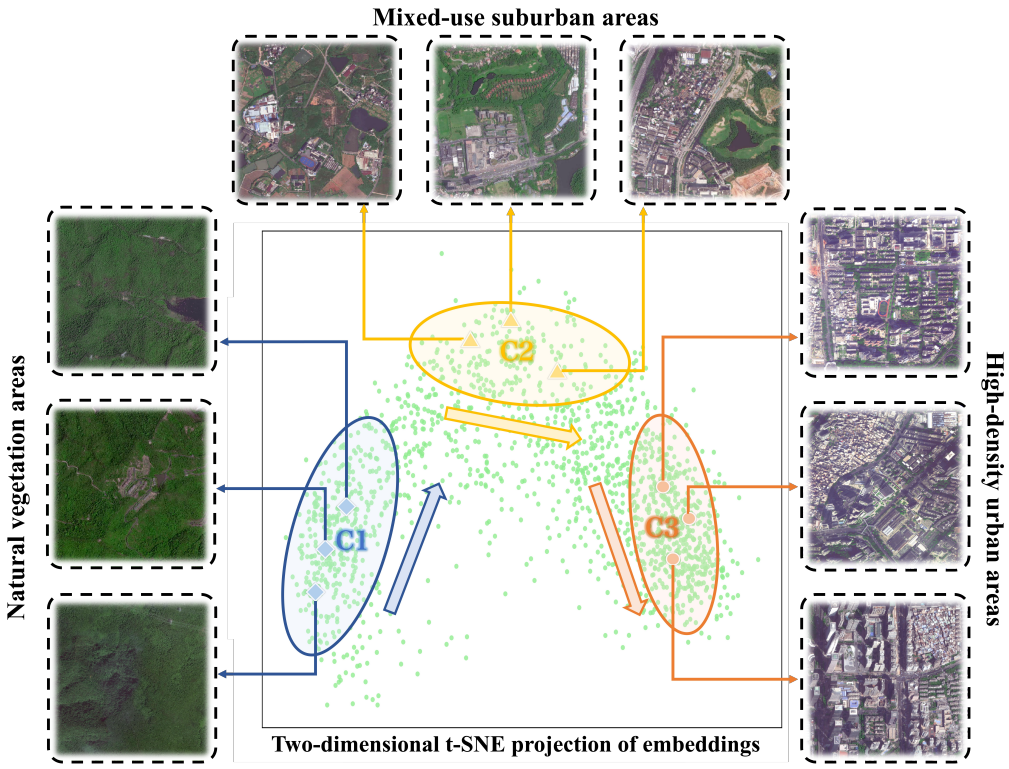


Figure 6: t-SNE visualization of region representations showing three distinct clusters corresponding to different urbanization levels: natural vegetation areas (left), mixed-use suburban regions (center), and high-density urban areas (right).

intra-cluster similarity. The clusters exhibit a progressive urbanization gradient from left to right: areas dominated by natural vegetation, mixed-use suburban regions, and high-density urban areas. This clustering pattern validates our approach successfully captures subtle yet critical geographical differences and maps regions with similar architectural layouts and land use patterns into proximate embedding positions.

To validate the necessity and effectiveness of multi-view fusion over single-view approaches, Figure 7 demonstrates our multi-view approach through

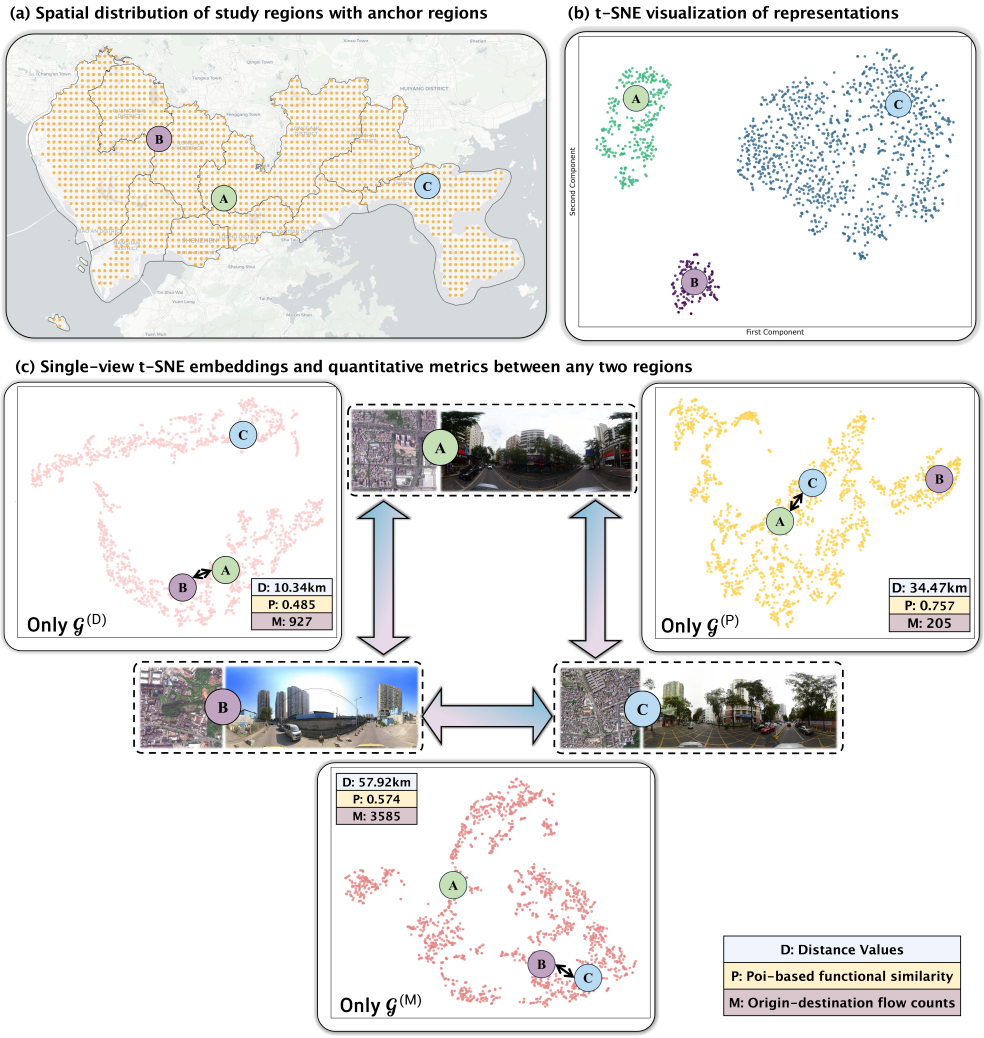


Figure 7: Multi-view region representation analysis. (a) Spatial distribution of study regions with anchor regions A, B, C. (b) UrbanMMCL embedding space showing integrated clustering. (c) Single-view embeddings and quantitative metrics revealing individual graph limitations.

comparative analysis of three anchor regions. Panel (c) reveals limitations of single-view embeddings through individual embedding spaces and quan-

titative relationship metrics (D, P, M values). The only distance-based embedding places regions A-B closely while positioning A-C far apart despite their strong functional similarity. The only POI-based embedding brings functionally similar regions A-C together but inappropriately positions A-B and B-C by neglecting spatial and mobility constraints. The only mobility-based embedding clusters regions B-C closely due to strong movement connections while under-representing A-B and A-C relationships. These positioning biases highlight the the limitations of single-view approaches in capturing comprehensive urban relationships. In contrast, Panel (b) demonstrates our UrbanMMCL embedding space where regions achieve balanced clustering through consensus-based optimization integrating all perspectives. This integrated approach produces robust representations that position regions appropriately by balancing functional similarity, spatial proximity, and mobility connectivity in a unified embedding space.

4.4.2. Geographic mapping of clustered representations

To validate that our UrbanMMCL framework captures meaningful urban structures, we apply hierarchical clustering to the learned embeddings. Figure 8 presents clustering results for k k=2 to 6, with the dendrogram distances indicating cluster distinctiveness. Clustering analysis demonstrates a clear hierarchical organization of urban spaces. At k=2, a fundamental binary partition emerges: built-up areas (pink) and natural areas (cyan), separating urban development zones from mountainous regions and water bodies. As k increases to 3, the urban domain subdivides into high-density cores (Futian, Luohu, Nanshan) and lower-density periphery (Longgang, Guangming, Pingshan), while natural areas remain cohesive. Higher k values (4-6)

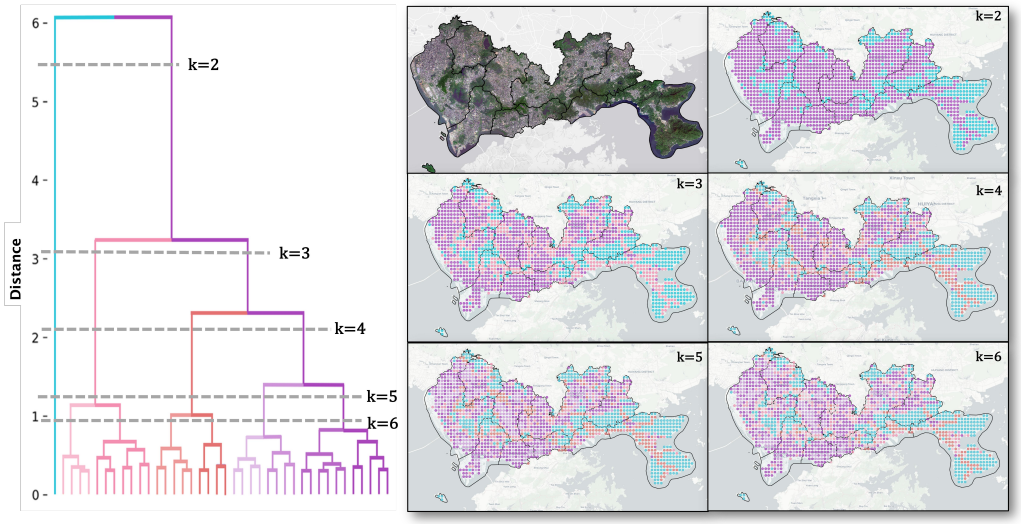


Figure 8: Geographic mapping of clustered representations across different cluster numbers (k=2 to k=6).The dendrogram (left) shows the hierarchical structure of learned embeddings, with dashed lines indicating cut heights for different k values. The satellite image (top center) provides the geographic context of Shenzhen. The cluster maps (right) visualize the spatial distribution of clusters for each k value.

demonstrate progressive refinement within urban areas while maintaining stable natural clusters. This pattern indicates that our embeddings successfully encode urban heterogeneity, as they capture development intensity variations and functional zones while recognizing the homogeneity of natural landscapes. Such spatially coherent clustering demonstrates the practical utility of our framework for automated urban region categorization.

4.4.3. Predictive performance analysis

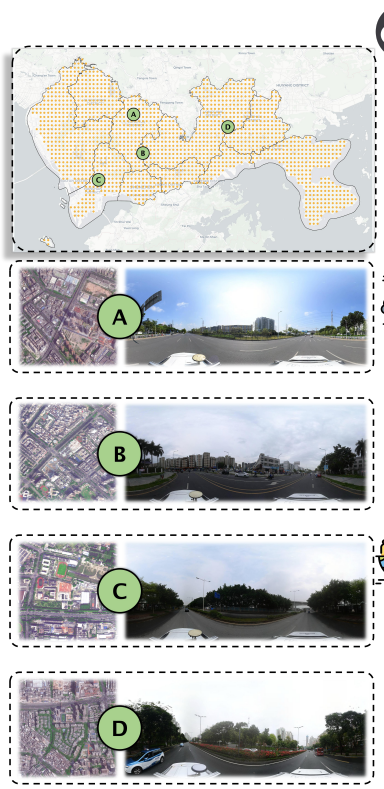
To demonstrate the practical effectiveness of our method, we conduct comprehensive case studies examining prediction performance in representative four urban regions. Our analysis includes: (1) regression analysis with

characteristics for CO, $PM_{2.5}$ emissions, and population density prediction, and (2) classification analysis for land use.

Regression Analysis. We select four representative regions that provide an ideal testbed for evaluating multi-modal and multi-view necessity. Critically, region pairs (A-B and C-D) exhibit similar visual appearances but substantial differences in urban indicators, creating challenging discrimination scenarios.

Figure 9 presents regression results across three urban indicators. Our complete UrbanMMCL framework consistently achieves the closest approximations to ground truth. When individual modalities are removed, systematic degradation emerges. Eliminating RSI or SVI causes predictions to converge toward averaged values, losing spatial discrimination. For example, without RSI, CO predictions become nearly uniform (0.33-0.39), failing to capture the actual variation (0.63-0.96). View-specific ablation reveals distinct dependency patterns for different urban indicators. Environmental indicators show greater sensitivity to structural patterns capturing physical processes, while socio-economic indicators correlate more strongly with human behavior modeling and functional interactions. These heterogeneous patterns validate our multi-view approach by demonstrating that different urban processes operate through distinct channels, and no single structural perspective adequately captures urban system complexity.

Land Use Classification Analysis. We examine prediction performance across six primary land cover categories: tree cover, building, shrubland, cropland, traffic route, and grassland. The spatial distribution analysis(Figure 10 a-b) shows that UrbanMMCL predictions closely align with



CO

	A	B	C	D
Ground Truth	0.73	0.96	0.63	0.92
UrbanMMCL	0.67	0.84	0.49	0.89
w/o RSI	0.33	0.37	0.36	0.39
w/o SVI	0.35	0.32	0.32	0.33
w/o $\mathcal{G}^{(P)}$	0.49	0.71	0.24	0.49
w/o $\mathcal{G}^{(M)}$	0.51	0.68	0.36	0.55
w/o $\mathcal{G}^{(D)}$	0.41	0.66	0.32	0.54

PM 2.5

	A	B	C	D
Ground Truth	584	1018	326	1069
UrbanMMCL	507	925	295	912
w/o RSI	832	759	799	747
w/o SVI	770	784	773	729
w/o $\mathcal{G}^{(P)}$	424	881	224	637
w/o $\mathcal{G}^{(M)}$	477	892	245	704
w/o $\mathcal{G}^{(D)}$	393	751	237	653

	A	B	C	D
Ground Truth	5096	6829	5707	7507
UrbanMMCL	4903	6021	4287	6373
w/o RSI	7982	8039	7291	7400
w/o SVI	7437	7803	7910	7723
w/o $\mathcal{G}^{(P)}$	3375	4689	3203	4390
w/o $\mathcal{G}^{(M)}$	3044	3927	2193	3742
w/o $\mathcal{G}^{(D)}$	4099	4934	3358	4554

Figure 9: Case Study analysis comparing prediction performance across four representative regions for three urban indicators. The left panel displays the spatial distribution and corresponding SVI/RSI data, while the right panel presents compares ground truth with UrbanMMCL predictions and ablation configurations.

ground truth patterns. Our framework successfully captures complex spatial organization and maintains clear boundaries between natural areas and built environments. The confusion matrix (Figure 10 c) reveals varying performance across categories. Grassland achieves the highest accuracy at 85.3%, followed by shrubland at 75.0%, building at 64.8%, and traffic routes at 59.7%. Tree cover and cropland show more challenging classification at 37.7% and 36.4% respectively, likely due to seasonal variations and spectral simi-

869 larity with other vegetation types.

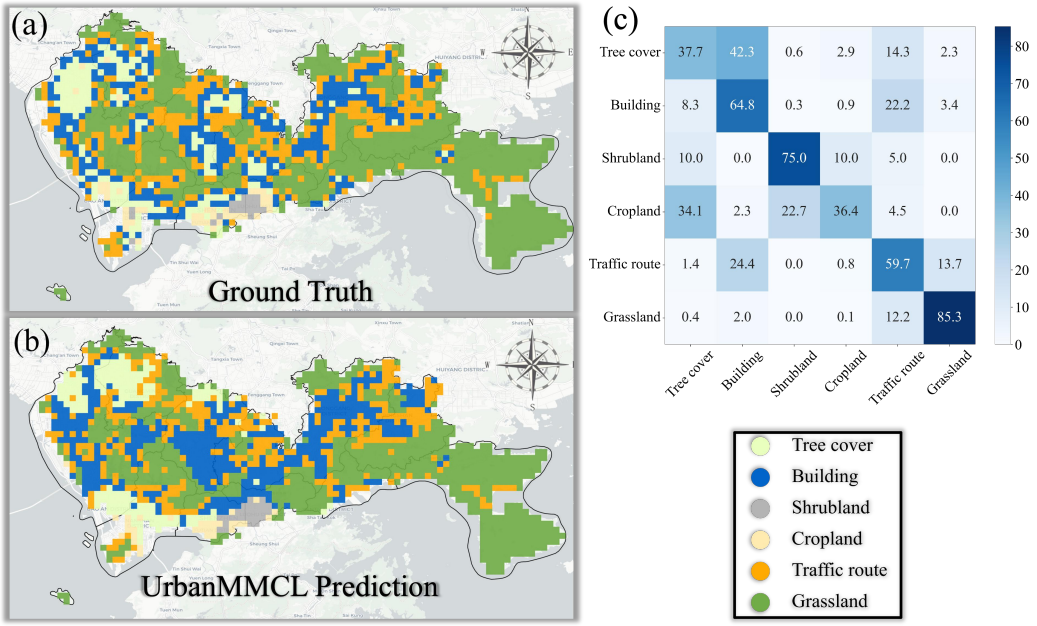


Figure 10: Land use classification analysis comparing UrbanMMCL predictions with ground truth data. (a) Ground truth spatial distribution. (b) UrbanMMCL prediction results. (c) Confusion matrix quantifying classification accuracy for each land cover category.

870 Our complete UrbanMMCL framework consistently achieves the closest
871 approximations to ground truth across all regions and indicators in both
872 regression and classification tasks. The multimodal data integration and
873 multi-view framework ensures that when visual similarities mask functional
874 differences, complementary perspectives provide the discriminative power
875 necessary for accurate urban dynamics prediction and land use classification.

5. Discussions

5.1. Multi-modal contribution analysis

Understanding how disparate urban data modalities contribute to representation learning provides insights into feature complementarity and information hierarchies in multi-modal urban analysis. Our ablation experiments reveal distinct roles for each modality in capturing different aspects of urban complexity, with results presented in Figure 11.

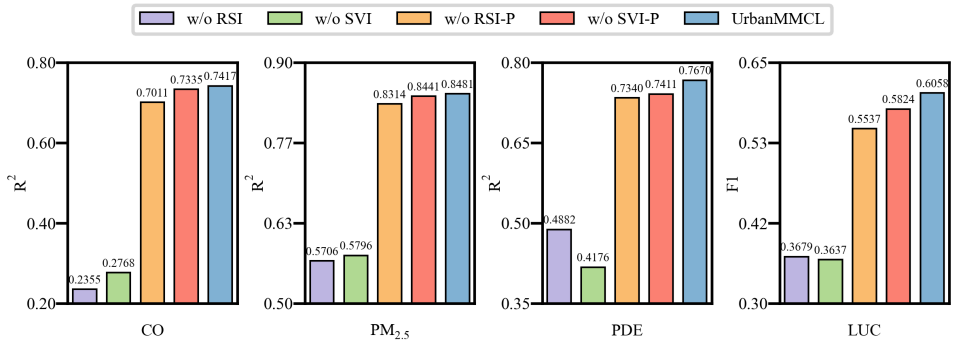


Figure 11: Multi-modal ablation study showing the contribution of different modalities to urban representation learning.

Fundamental Role of Remote Sensing Imagery. RSI emerges as the most fundamental modality, with its removal causing severe performance degradation across all tasks. This dominance stems from RSI's ability to capture spatial patterns and urban morphological features at scale. For PEP, substantial performance drops reveal that RSI encodes critical spatial dependencies correlating with environmental phenomena. RSI's high information

density enables learning of rich spatial representations that serve as foundational embeddings for other modalities.

Complementary Value of Street-View Imagery. SVI contributes fine-grained environmental features through local context augmentation. Performance improvements from SVI inclusion demonstrate its role in capturing micro-environmental variations invisible in overhead imagery. SVI functions as local environmental validators that refine broad spatial patterns captured by RSI, particularly evident in LUC where ground-level visual cues help distinguish functionally similar areas.

Spatial Context Enhancement Through Positional Encoding. Geographical coordinates provide modest but consistent contributions as spatial relationship encoders. The relatively small impact when removing positional encoding suggests that visual features carry majority predictive information, while coordinates primarily enhance spatial coherence and topological consistency in learned representations.

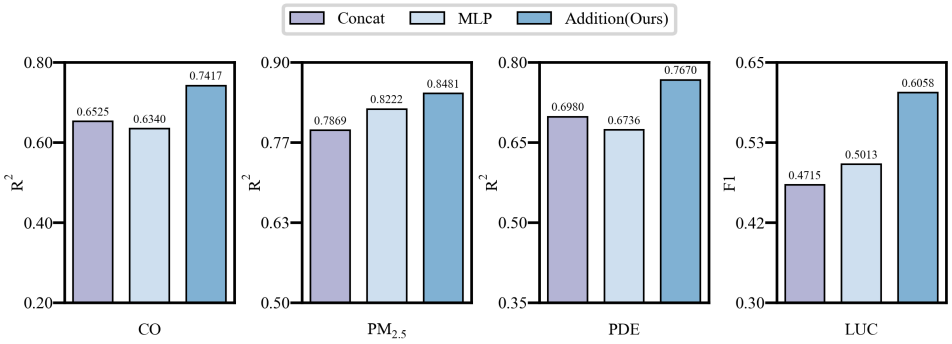


Figure 12: Performance comparison of different multi-modal fusion strategies.

Implications for Urban Representation Learning. These findings establish fundamental principles for effective urban AI systems. Performance differentials when removing individual modalities demonstrate that each modality contributes unique, irreplaceable information, establishing a clear modality hierarchy where RSI provides foundational spatial structure, SVI adds critical environmental detail, and positional encoding serves as spatial regularization (Figure 11). More critically, the superior performance of specialized encoders over generic CLIP-based alternatives (Table 3) and the critical role of multi-view contrastive learning (Table 4) demonstrate that urban environments require domain-specific architectures and multi-perspective integration rather than universal approaches. This advocates for specialized multi-modal urban AI systems that embrace complexity through tailored encoders and multi-perspective integration.

5.2. Multi-modal feature fusion strategy analysis

Multi-modal feature fusion significantly impacts the model's ability to leverage complementary information from heterogeneous urban data sources. The choice of fusion strategy is therefore crucial for maximizing the benefits of multi-modal urban data integration. We compared three fusion strategies: (1) concatenation of visual and location features from RSI and aggregated SVIs; and (2) MLP-based fusion with multi-layer perceptrons; and (3) element-wise addition. As shown in Figure 12, our addition method consistently achieves superior performance across all metrics despite its simplicity.

The superior performance of element-wise addition can be attributed to its ability to preserve original feature distributions while enabling direct correspondence between spatially aligned features from different modalities.

Unlike concatenation, which introduces feature redundancy and increased dimensionality, or MLP fusion, which adds parameters and optimization complexity, addition fusion maintains the semantic integrity of individual modalities while creating meaningful cross-modal interactions. This validates our design choice and demonstrates that simpler fusion strategies can be more effective.

5.3. Training paradigms and efficiency analysis

To comprehensively evaluate our model's representation capabilities and training efficiency, we examine two additional training paradigms that represent different approaches to leveraging pre-trained knowledge for urban downstream tasks. **Pretrain-finetune** first optimizes the encoder on a large-scale, task-agnostic urban data to learn general representations, followed by fine-tuning on downstream tasks; **Linear probing** freezes the pretrained encoder and trains only a linear head, providing an efficient assessment of representation quality with minimal computational resources; **End-to-end training** initializes with pre-trained weights but allows unrestricted parameter updates throughout the entire architecture.

Figure 13 presents comparative analysis of three training paradigms across prediction accuracy, runtime per epoch, and epochs to convergence. The bars represent average performance while scattered points show individual task values for CO, $PM_{2.5}$, PDE, and LUC tasks. The results reveals critical insights into the performance-efficiency trade-offs inherent in each training approach. Pretrain-finetune emerges as the optimal strategy, achieving the highest average accuracy (0.7409) with computational efficiency (16.7

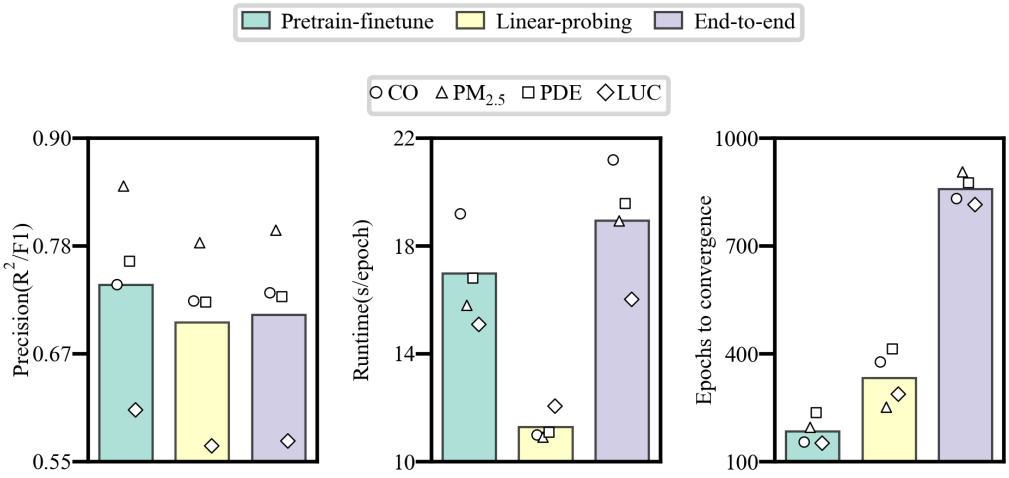


Figure 13: Comparative analysis of three training paradigms across prediction accuracy, runtime per epoch, and epochs to convergence across urban tasks. Bars represent average performance across all urban tasks, while scattered points show individual task performance (circles: CO prediction, triangles: $PM_{2.5}$ prediction, squares: Population Density Estimation, diamonds: Land Use Classification).

s/epoch, 184 epochs). This paradigm preserves lspatial reasoning capabilities through selective parameter adaptation, making it ideal for resource-constrained urban monitoring applications. Linear probing shows the fastest per-epoch computation (11.3s) but suffers from limited representation adaptability. The frozen encoder prevents overfitting but results in systematic accuracy degradation, particularly in complex regression tasks. Despite faster iterations, it requires nearly twice as many epochs to converge (348 vs 184), offsetting its computational advantage. End-to-end training incurs prohibitive costs (18.9s/epoch, 857 epochs to converge) while achieving only marginal improvements over linear probing and falling 2.53% short of pretrain-finetune, making it viable only with abundant resources. Overall, the results validate

pretrain-finetune as the optimal training paradigm, effectively balancing predictive performance with computational efficiency.

5.4. Limitations and future directions

While UrbanMMCL demonstrates significant advances in urban representation learning, several limitations warrant acknowledgment and present opportunities for future research.

First, our framework's reliance on high-quality textual descriptions generated by BLIP-2 introduces a potential bottleneck, as variations in text generation quality across different urban scenes could lead to inconsistent performance, particularly in challenging scenarios where visual content is ambiguous or degraded. Future work should explore more robust text generation methods or develop alternative approaches to incorporate semantic information less dependent on generative model.

Second, the static nature of our graph construction methods may not fully capture dynamic temporal patterns inherent in urban systems, such as daily traffic patterns or seasonal environmental changes. Additionally, the choice of graph construction criteria may not be optimal for all urban tasks. Future directions should explore temporal modeling capabilities, dynamic graph learning approaches, and task-adaptive graph construction strategies.

Finally, while our framework demonstrates cross-city transferability from Shenzhen to Beijing and Chengdu, complete zero-shot generalization remains limited. The graph structure components require re-initialization and adaptation for city-specific spatial relationships, indicating that spatial modeling still needs localized fine-tuning. Future research should investigate developing fully generalizable urban foundation models that can achieve complete

zero-shot inference without requiring any component re-training, potentially through learning universal spatial relationship patterns or developing city-agnostic graph construction strategies that can adapt automatically to new urban environments.

6. Conclusions

This paper presents UrbanMMCL, a novel self-supervised dual-stage contrastive learning framework that advances urban representation learning through innovative integration of multi-modal fusion and adaptive graph learning. Our approach establishes a comprehensive pre-training paradigm that learns generalizable urban representations without requiring task-specific labels, addressing the critical challenge of limited annotated data in urban analysis.

Comprehensive experimental validation demonstrates that UrbanMMCL consistently outperforms state-of-the-art methods across environmental monitoring, population estimation, and land use classification tasks. Cross-city transfer experiments further validate the generalizability of our learned representations across different urban environments. The framework's success stems from its principled integration of RSI, SVI, textual descriptions, and geographical coordinates through contrastive learning, while adaptive graph learning captures dynamic inter-regional relationships that static approaches cannot model.

UrbanMMCL represents a significant advancement toward urban foundation models by demonstrating how multi-modal pre-training can learn transferable urban knowledge that generalizes across different tasks and cities. This work bridges the gap between domain-specific urban analysis tools and

the broader vision of unified urban AI systems, laying the groundwork for more comprehensive urban foundation models that can support evidence-based urban planning, sustainable development, and smart city initiatives at unprecedented scale and sophistication.

CRediT authorship contribution statement

Jinzhou Cao: Methodology, Conceptualization, Investigation, Funding acquisition, Writing - original draft, Writing – review & editing. **Jiashi Chen**: Formal analysis, Data curation, Visualization, Writing - original draft, Writing – review & editing. **Xiangxu Wang**: Writing - original draft, Writing – review & editing. **Weiming Huang**: Writing – review & editing. **Dongsheng Chen**: Writing – review & editing. **Tianhong Zhao**: Writing – review & editing. **Wei Tu**: Investigation, Writing – review & editing. **Qingquan Li**: Resources, Writing – review & editing.

Acknowledgements

This research was supported in part by Shenzhen Science and Technology Program (No. JCYJ20240813113300001, 20231127180406001); the National Natural Science Foundation of China (No. 42401553); Natural Science Foundation of Top Talent of SZTU (No. GDRC202415).

References

Bai, L., Huang, W., Zhang, X., Du, S., Cong, G., Wang, H., Liu, B., 2023. Geographic mapping with unsupervised multi-modal representation learn-

ing from VHR images and POIs. ISPRS Journal of Photogrammetry and Remote Sensing 201, 193–208.

Bai, L., Zhang, X., Wang, H., Du, S., 2025. Integrating remote sensing with OpenStreetMap data for comprehensive scene understanding through multi-modal self-supervised learning. Remote Sensing of Environment 318, 114573.

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O.K., Aggarwal, K., Som, S., Piao, S., Wei, F., 2022. VLMo: Unified vision-language pre-training with mixture-of-modality-experts, in: Advances in Neural Information Processing Systems, pp. 32897–32912.

Belkin, M., Niyogi, P., 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Advances in Neural Information Processing Systems.

Cao, J., Wang, X., Chen, G., Tu, W., Shen, X., Zhao, T., Chen, J., Li, Q., 2025a. Disentangling the hourly dynamics of mixed urban function: A multimodal fusion perspective using dynamic graphs. Information Fusion 117, 102832.

Cao, J., Wang, X., Chen, J., Tu, W., Li, Z., Yang, X., Zhao, T., Li, Q., 2025b. Urban representation learning for fine-grained economic mapping: A semi-supervised graph-based approach. ISPRS Journal of Photogrammetry and Remote Sensing 226, 317–331.

Cao, J., Wang, X., Chen, J., Zhang, B., Ma, Y., Zhao, T., 2025c. SemiGPS: GraphGPS-based Semi-supervised Graph Learning for Sector-

Specific GDP Mapping, in: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5.

Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., Zhang, Q., Li, Q., Qiu, G., 2020. Deep learning-based remote and social sensing data fusion for urban region function recognition. ISPRS Journal of Photogrammetry and Remote Sensing 163, 82–97.

Cepeda, V.V., Nayak, G.K., Shah, M., 2023. GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. pp. 8690–8701.

Chan, W., Ren, Q., 2023. Region-Wise Attentive Multi-View Representation Learning For Urban Region Embedding, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA. pp. 3763–3767.

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

Chen, X., Xie, S., He, K., 2021. An empirical study of training self-supervised vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9640–9649.

Chen, Y., Huang, W., Zhao, K., Jiang, Y., Cong, G., 2025. Self-supervised representation learning for geospatial objects: A survey. Information Fusion 123, 103265.

Dai, G., Yi, W., Cao, J., Gong, Z., Fu, X., Zhang, B., 2025. CRRL: Contrastive Region Relevance Learning Framework for Cross-city Traffic Prediction. Information Fusion 122, 103215.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations.

Gao, J., Li, P., Chen, Z., Zhang, J., 2020. A Survey on Deep Learning for Multimodal Data Fusion. Neural Computation 32, 829–864.

Grover, A., Leskovec, J., 2016. Node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA. pp. 855–864.

Guan, Q., Wang, J., Ren, S., Gao, H., Liang, Z., Wang, J., Yao, Y., 2024. Predicting short-term pm2. 5 concentrations at fine temporal resolutions using a multi-branch temporal graph convolutional neural network. International Journal of Geographical Information Science 38, 778–801.

Hao, X., Chen, W., Yan, Y., Zhong, S., Wang, K., Wen, Q., Liang, Y., 2025. UrbanVLP: Multi-granularity vision-language pretraining for urban

socioeconomic indicator prediction. Proceedings of the AAAI Conference on Artificial Intelligence 39, 28061–28069.

Hassani, K., Khasahmadi, A.H., 2020. Contrastive multi-view representation learning on graphs, in: International conference on machine learning, PMLR. pp. 4116–4126.

He, J., Huang, B., 2025. Estimating global anthropogenic CO2 emissions through satellite observations. Environmental Research 279, 121767.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738.

He, L., Cheng, D., Zhang, G., Zhang, S., 2025. Leveraging long-range nodes in multi-view graph contrastive learning. Information Fusion 122, 103186.

Huang, W., Wang, J., Cong, G., 2024. Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. International Journal of Geographical Information Science 38, 1414–1442.

Huang, W., Zhang, D., Mai, G., Guo, X., Cui, L., 2023. Learning urban region representations with POIs and hierarchical graph infomax. ISPRS Journal of Photogrammetry and Remote Sensing 196, 134–145.

Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. Science 353, 790–794.

Jenkins, P., Farag, A., Wang, S., Li, Z., 2019. Unsupervised representation learning of spatial data via multimodal embedding, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA. pp. 1993–2002.

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

Ju, W., Fang, Z., Gu, Y., Liu, Z., Long, Q., Qiao, Z., Qin, Y., Shen, J., Sun, F., Xiao, Z., Yang, J., Yuan, J., Zhao, Y., Wang, Y., Luo, X., Zhang, M., 2024. A Comprehensive Survey on Deep Graph Representation Learning. Neural Networks 173, 106207.

Khoshraftar, S., An, A., 2024. A survey on graph representation learning methods. ACM Transactions on Intelligent Systems and Technology 15, 19:1–19:55.

Klemmer, K., Rolf, E., Robinson, C., Mackey, L., Rußwurm, M., 2025. SatCLIP: Global, general-purpose location embeddings with satellite imagery, in: Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence: AAAI-25 Technical Tracks, pp. 4347–4355.

Li, T., Xi, Y., Wang, H., Li, Y., Tarkoma, S., Hui, P., 2023a. Learning Representations of Satellite Imagery by Leveraging Point-of-Interests. ACM Transactions on Intelligent Systems and Technology 14, 1–32.

Li, Y., Huang, W., Cong, G., Wang, H., Wang, Z., 2023b. Urban Region Representation Learning with OpenStreetMap Building Footprints, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA. pp. 1363–1373.

Li, Y., Yang, M., Zhang, Z., 2019. A Survey of Multi-View Representation Learning. IEEE Transactions on Knowledge and Data Engineering 31, 1863–1883.

Li, Z., He, W., Cheng, M., Hu, J., Yang, G., Zhang, H., 2023c. Sinolc-1: the first 1 m resolution national-scale land-cover map of china created with a deep learning framework and open-access data. Earth System Science Data 15, 4749–4780.

Li, Z., Huang, W., Zhao, K., Yang, M., Gong, Y., Chen, M., 2024. Urban region embedding via multi-view contrastive prediction, in: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI Press. pp. 8724–8732.

Li, Z., Zhang, H., Lu, F., Xue, R., Yang, G., Zhang, L., 2022. Breaking the resolution barrier: A low-to-high network for large-scale high-resolution land-cover mapping using low-resolution labels. ISPRS Journal of Photogrammetry and Remote Sensing 192, 244–267.

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J.,

2024. RemoteCLIP: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing 62, 1–16.

Liu, S., Zhang, T., Fu, N., Huang, Y., 2025. Fine-grained graph representation learning for heterogeneous mobile networks with attentive fusion and contrastive learning. Proceedings of the AAAI Conference on Artificial Intelligence 39, 18933–18942.

Liu, Y., Zhang, X., Ding, J., Xi, Y., Li, Y., 2023. Knowledge-infused Contrastive Learning for Urban Imagery-based Socioeconomic Prediction, in: Proceedings of the ACM Web Conference 2023, Association for Computing Machinery, New York, NY, USA. pp. 4150–4160.

Luo, Y., Chung, F.l., Chen, K., 2022. Urban Region Profiling via Multi-Graph Representation Learning, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Association for Computing Machinery, New York, NY, USA. pp. 4294–4298.

Perozzi, B., Al-Rfou, R., Skiena, S., 2014. DeepWalk: Online Learning of Social Representations, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 701–710.

Qin, Q., Ai, T., Xu, S., Zhang, Y., Huang, W., Du, M., Li, S., 2025. Learning dual context aware POI representations for geographic mapping. International Journal of Applied Earth Observation and Geoinformation 142, 104683.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021a. Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, PMLR. pp. 8748–8763.

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

Shen, X., Wang, H., Wei, B., Cao, J., 2023. Real-time scene classification of unmanned aerial vehicles remote sensing image based on Modified GhostNet. PLOS ONE 18, e0286873.

Sun, F.Y., Hoffman, J., Verma, V., Tang, J., 2020a. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization, in: International Conference on Learning Representations.

Sun, F.Y., Hoffmann, J., Verma, V., Tang, J., 2020b. InfoGraph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. arXiv:1908.01000.

Sun, Z., Jiao, H., Wu, H., Peng, Z., Liu, L., 2021. Block2vec: An Approach for Identifying Urban Functional Regions by Integrating Sentence Embedding Model and Points of Interest. ISPRS International Journal of Geo-Information 10, 339.

Suresh, S., Li, P., Hao, C., Neville, J., 2021. Adversarial graph augmentation to improve graph contrastive learning, in: Proceedings of the 35th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. pp. 15920–15933.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems.

Veličković, P., Fedus, W., Hamilton, W.L., Liò, P., Bengio, Y., Hjelm, R.D., 2018. Deep graph infomax, in: International Conference on Learning Representations.

Vivanco, V., Nayak, G.K., Shah, M., 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization, in: Advances in Neural Information Processing Systems.

Wang, X., Cao, J., Zhao, T., Zhang, B., Chen, G., Li, Z., Chen, H., Tu, W., Li, Q., 2026. ST-camba: A decoupled-free spatiotemporal graph fusion state space model with linear complexity for efficient traffic forecasting. Information Fusion 126, 103495.

Wang, X., Chen, H., Liu, Y., 2024. Learning place representations from spatial interactions. International Journal of Geographical Information Science 38, 1065–1090.

Wang, X., Cheng, T., Law, S., Zeng, Z., Yin, L., Liu, J., 2025. Multimodal contrastive learning of urban space representations from POI data. Computers, Environment and Urban Systems 120, 102299.

Wang, Z., Li, H., 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, pp. 1013–1020.

Wei, J., Li, Z., 2024. GlobalHighCO: Global daily seamless 1 km ground- level CO dataset over land (2018–present). doi:10.5281/zenodo.14207363.

Wei, J., Li, Z., Lyapustin, A., Wang, J., Dubovik, O., Schwartz, J., Sun, L., Li, C., Liu, S., Zhu, T., 2023. First close insight into global daily gapless 1 km PM2.5 pollution, variability, and health impact. Nature Communications 14, 8349.

Weng, X., Pang, C., Xia, G.S., 2025. Vision-language modeling meets remote sensing: Models, datasets, and perspectives. IEEE Geoscience and Remote Sensing Magazine , 2–50.

Wu, L., Lin, H., Tan, C., Gao, Z., Li, S.Z., 2023. Self-supervised learning on graphs: Contrastive, generative, or predictive. IEEE Transactions on Knowledge and Data Engineering 35, 4216–4235.

Wu, S., Yan, X., Fan, X., Pan, S., Zhu, S., Zheng, C., Cheng, M., Wang, C., 2022. Multi-Graph Fusion Networks for Urban Region Embedding, in: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence.

Wu, Z., Xiong, Y., Yu, S.X., Lin, D., 2018. Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3733–3742.

Xi, Y., Li, T., Wang, H., Li, Y., Tarkoma, S., 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, Association for Computing Machinery, New York, NY, USA. pp. 3308–3316.

Xu, Y., Jin, S., Chen, Z., Xie, X., Hu, S., Xie, Z., 2022. Application of a graph convolutional network with visual and semantic features to classify urban scenes. International Journal of Geographical Information Science 36, 2009–2034.

Xu, Z., Zhou, X., 2024. CGAP: Urban Region Representation Learning with Coarsened Graph Attention Pooling, in: Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Jeju, South Korea. pp. 7518–7526.

Yan, Y., Wen, H., Zhong, S., Chen, W., Chen, H., Wen, Q., Zimmermann, R., Liang, Y., 2024. UrbanCLIP: Learning Text-enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web, in: Proceedings of the ACM Web Conference 2024, Association for Computing Machinery, New York, NY, USA. pp. 4006–4017.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., Burke, M., 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. Nature Communications 11, 2583.

Yong, X., Zhou, X., 2024. MuseCL: Predicting Urban Socioeconomic Indicators via Multi-Semantic Contrastive Learning, in: Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization. pp. 7536–7544.

You, Y., Chen, T., Shen, Y., Wang, Z., 2021. Graph Contrastive Learning Automated, in: Proceedings of the 38th International Conference on Machine Learning, PMLR. pp. 12121–12132.

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., Shen, Y., 2020. Graph contrastive learning with augmentations. Advances in neural information processing systems 33, 5812–5823.

Zhai, W., Bai, X., Shi, Y., Han, Y., Peng, Z.R., Gu, C., 2019. Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs. Computers, Environment and Urban Systems 74, 1–12.

Zhang, F., Wu, L., Zhu, D., Liu, Y., 2019. Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. ISPRS Journal of Photogrammetry and Remote Sensing 153, 48–58.

Zhang, L., Long, C., Cong, G., 2023a. Region Embedding With Intra and Inter-View Contrastive Learning. IEEE Transactions on Knowledge and Data Engineering 35, 9031–9036.

Zhang, M., Li, T., Li, Y., Hui, P., 2020. Multi-View Joint Graph Representation Learning for Urban Region Embedding, in: Proceedings of the

Twenty-Ninth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan. pp. 4431–4437.

Zhang, Q., Huang, C., Xia, L., Wang, Z., Yiu, S., Han, R., 2023b. Spatial-temporal graph learning with adversarial contrastive adaptation, in: Proceedings of the 40th International Conference on Machine Learning, pp. 41151–41163.

Zhang, W., Han, J., Xu, Z., Ni, H., Lyu, T., Liu, H., Xiong, H., 2025. Towards urban general intelligence: A review and outlook of urban foundation models. arXiv:2402.01749.

Zhang, X., Gong, Y., Zhang, C., Wu, X., Guo, Y., Lu, W., Zhao, L., Dong, X., 2023c. Spatio-temporal fusion and contrastive learning for urban flow prediction. Knowledge-Based Systems 282, 111104.

Zhang, Y., Huang, W., Yao, Y., Gao, S., Cui, L., Yan, Z., 2024a. Urban region representation learning with human trajectories: A multi-view approach incorporating transition, spatial, and temporal perspectives. GIScience & Remote Sensing 61, 2387392.

Zhang, Y., Li, Y., Zhang, F., 2024b. Multi-level urban street representation with street-view imagery and hybrid semantic graph. ISPRS Journal of Photogrammetry and Remote Sensing 218, 19–32.

Zhang, Y., Xu, Y., Cui, L., Yan, Z., 2023d. Multi-view graph contrastive learning for urban region representation, in: 2023 International Joint Conference on Neural Networks (IJCNN), pp. 1–8.

Zhao, T., Liang, X., Biljecki, F., Tu, W., Cao, J., Li, X., Yi, S., 2025. Quantifying seasonal bias in street view imagery for urban form assessment: A global analysis of 40 cities. Computers, Environment and Urban Systems 120, 102302.

Zhou, S., He, D., Chen, L., Shang, S., Han, P., 2023a. Heterogeneous Region Embedding with Prompt Learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4981–4989.

Zhou, W., Sun, F., Jiang, Q., Cong, R., Hwang, J.N., 2023b. Wavenet: Wavelet network with knowledge distillation for rgb-t salient object detection. IEEE Transactions on Image Processing 32, 3027–3039.

Zhou, W., Zhu, Y., Lei, J., Wan, J., Yu, L., 2021. Ccafnet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in rgb-d images. IEEE Transactions on Multimedia 24, 2192–2204.

Zhou, Y., Huang, Y., 2018. DeepMove: Learning Place Representations through Large Scale Movement Data, in: 2018 IEEE International Conference on Big Data (Big Data), pp. 2403–2412.

Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., Wang, L., 2020. Deep graph contrastive representation learning. arXiv preprint arXiv:2006.04131 .

Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., Wang, L., 2021. Graph contrastive learning with adaptive augmentation, in: Proceedings of the web conference 2021, pp. 2069–2080.

Zou, X., Yan, Y., Hao, X., Hu, Y., Wen, H., Liu, E., Zhang, J., Li, Y., Li, T., Zheng, Y., Liang, Y., 2025. Deep learning for cross-domain data fusion in

urban computing: Taxonomy, advances, and outlook. Information Fusion 113, 102606.