UrbanMFM: Spatial Graph-based Multiscale Foundation Models for Learning Generalized Urban Representation

Zhaoqi Zhang, Miao Xie, Pasquale Balsebre, Weiming Huang, Siqiang Luo *Member, IEEE*, and Gao Cong

***Abstract*—As geospatial data from web platforms becomes increasingly accessible and regularly updated, urban representation learning has emerged as a critical research area for advancing urban planning. Recent studies have developed foundation model-based algorithms to leverage this data for various urban-related downstream tasks. However, current research has inadequately explored deep integration strategies for multiscale, multimodal urban data in the context of urban foundation models. This gap arises primarily because the relationships between micro-scale (e.g., individual points of interest and street view imagery) and macro-scale (e.g., region-wide satellite imagery) urban features are inherently implicit and highly complex, making traditional interaction modeling insufficient. This paper introduces a novel research problem – how to learn multiscale urban representations by integrating diverse geographic data modalities and modeling complex multimodal relationships across different spatial scales. To address this significant challenge, we propose UrbanMFM, a spatial graph-based multiscale foundation model framework explicitly designed to capture and leverage these intricate relationships. UrbanMFM utilizes a self-supervised learning paradigm that integrates diverse geographic data modalities, including POI data and urban imagery, through novel contrastive learning objectives and advanced sampling techniques. By explicitly modeling spatial graphs to represent complex multiscale urban relationships, UrbanMFM effectively facilitates deep interactions between multimodal data sources. Extensive experiments on datasets from Singapore, New York, and Beijing demonstrate that UrbanMFM outperforms the strongest baselines significantly in four representative downstream tasks. By effectively modelling spatial hierarchies with diverse data, UrbanMFM provides a more comprehensive and adaptable representation of urban environments.**

***Index Terms*—Geospatial data mining, urban regions, foundation model, representation learning**

I. Introduction

URBAN areas are dynamic centres of human activity, and gaining a comprehensive understanding of their complex structures and functions is critical for urban planning. Scholars and policymakers have traditionally relied on manual surveys to collect metropolitan statistics, but this approach faces prohibitive labour and time costs [1], [2]. The increasing availability of large-scale urban datasets, combined with advancements in computational methodologies, has opened up new possibilities for extracting critical urban features from diverse spatial data using machine learning techniques [3], [4]. However, a key limitation is that the models and learned representations are often tailored to a set of specific tasks, and do not generalize well to others.

Recently, Pre-trained Foundation Models (PFMs) have emerged as a promising solution, and there have been studies utilizing urban spatial data from various online platforms, such as OpenStreetMap, Uber, and Google Maps, to develop such models [5], [6]. This approach has two crucial advantages: the training phase is carried out in a self-supervised fashion, without the need of human supervision; and the generated representations can be employed across various tasks. PFMs can, in fact, eliminate the requirement for large amounts of labelled datasets and can efficiently transform urban regions into vector embeddings that can be used for a wide range of downstream tasks, including forecasting socioeconomic indicators [7]–[9] and optimizing traffic management [10]–[14].

Most existing studies on urban foundation models have primarily focused on individual data sources. Language-based models [15]–[18] process urban-related textual data, such as traffic reports, social media posts, and geo-tagged texts. Vision-based models [19]–[22] focus on visual urban data, including satellite imagery, street view images, and surveillance footage. Meanwhile, trajectory-based models [23]–[26] analyze spatiotemporal data, capturing patterns in human mobility, vehicle trajectories, or GPS data. A few very recent approaches have attempted to integrate multiple data modalities to achieve a more holistic understanding of urban regions. For instance, UrbanCLIP [27] employs a large language model (LLM) to generate textual descriptions for satellite images, using contrastive objective during the training phase, while CityFM [28] designs a vision-language contrastive learning objective, utilizing OpenStreetMap data, to learn multimodal representations of geospatial entities.

Despite considerable progress in using multimodal data to model urban spaces, existing studies still face two key limitations, hindering the comprehensive understanding of urban structure and the full exploitation of urban data: (1) **Multiscale Urban Structure.** In urban studies, multi-scale refers to the hierarchical spatial organization of urban elements, capturing

This work was supported in part by the National Research Foundation, Singapore under its Frontier CRP under Grant NRF-F-CRP-2024-0005 and in part by NTU SUG-NAP under Grant 022029-00001. (*Corresponding author: Zhaoqi Zhang*)

Z. Zhang, P. Balsebre, S. Luo and G. Cong are with College of Computing and Data Science, Nanyang Technological University, Singapore, 639798. (E-mail: {zhaoqi001@e., pasquale001@e., siqiang.luo@, gaocong@}ntu.edu.sg)

M. Xie is with College of Information and Electrical Engineering, China Agricultural University, Beijing, China, 100083. (E-mail: xiemiao@cau.edu.cn)

W. Huang is with School of Geography, University of Leeds, Leeds, U.K, LS2 9JT. (E-mail: W.Huang@leeds.ac.uk)

Manuscript received June 09, 2025; revised November 17, 2025

interactions from fine-grained micro-level entities to broader global-level structures. However, existing methods often flatten such spatial hierarchies during early-stage fusion. For instance, in OSM data, both upscale restaurants and budget eateries are labelled with the same tag ("amenity": "restaurant"). When models use pooling techniques to aggregate this data, they homogenize different establishments, leading to the loss of meaningful structural information. (2) **Complex Global Multimodal Modeling.** Given a historical monument that has descriptive textual tags, rich visual features from street view images, and specific spatial boundaries from its polygon. Conventional pairwise approaches might miss insights like how the monument's architecture (visual) relates to its historical significance (textual) within its urban surroundings (spatial), leading to an incomplete representation.

This paper introduces a novel research problem – how to learn multiscale urban representations by integrating diverse geographic data modalities and modeling their complex relationships across various spatial scales. To address this problem, we propose UrbanMFM, a spatial graph-based **M**ultiscale **F**oundation **M**odel designed to leverage multimodal geospatial data within a selected city for pre-training. We define multiscale as the hierarchical spatial structure of urban environments, encompassing four distinct levels: micro (POIs and street view imagery), local (inter-POI proximity), macro (satellite images and regional attributes), and global (inter-region context). Specifically, the local scale is defined by the set of neighbouring POIs within a fixed radius around each target POI; the macro scale corresponds to pre-defined administrative or functional regions provided as ground-truth boundaries rather than self-partitioned ones; and the global scale captures contextual relations between the target region and its directly adjacent regions.This four-level decomposition enables us to capture urban patterns ranging from fine-grained functional semantics to broad spatial contexts.UrbanMFM simultaneously modelling both micro and macro scales. Unlike previous works that focus on region-level representations or rely on pairwise interactions, our approach tackles the challenge of jointly interacting these data sources in a way that preserves both local and global spatial relationships. This multiscale approach enhances the adaptability and effectiveness of UrbanMFM across a wide range of downstream tasks in urban analysis, generating comprehensive urban representations that capture the complex urban structures inherent in urban environments. In adherence to ensure that the model can be widely promoted, UrbanMFM utilizes both OSM data, which is freely available and accessible globally, as well as urban imagery, which can be captured through various Map APIs. Additionally, if more data sources were available, POI-level features such as user reviews and region-level features like human trajectory, UrbanMFM can also flexibly incorporate such data sources into corresponding scales, e.g. from micro scale learning for POI-level information to global scale learning for region-level information, to enhance the generalization ability of our framework. The key contributions of this paper can be summarized as follows:

1) We introduce UrbanMFM, a novel spatial graph-based multiscale foundation model for urban representation learning. UrbanMFM leverages multimodal geospatial data, including OSM data and urban imagery, to jointly model urban environments across micro, local, macro, and global spatial scales. To the best of our knowledge, this is the first framework that systematically unifies multiscale structure and multimodal fusion for urban analysis.
2) We design two spatial graph-based contrastive objectives at the local and global scales, enabling structured and scalable fusion across heterogeneous modalities. Combined with feature-level alignment at the micro and macro scales, this design allows UrbanMFM to capture complex cross-scale and cross-modal dependencies, yielding more robust and comprehensive urban representations.
3) We conduct extensive experiments on POI- and region-level downstream tasks, using data from Singapore, New York and Beijing. The results demonstrate the advantages of our method over established baselines, on the most challenging and sparse Beijing dataset, UrbanMFM achieves up to **28.7%** MAE reduction, **16.3%** RMSE reduction, and **19.1%** R^2 improvement over the strongest baseline, showing its robustness in addressing the previously identified challenges. Source code is available at **Github**.

II. Related Work

Urban Representation Learning. Urban representation learning has significantly evolved, driven by the need to better capture and predict urban dynamics through various urban data. However, a key challenge remains: how to effectively extract and integrate features from diverse data sources. Early efforts [29]–[32] focused on integrating these diverse urban data to construct comprehensive embeddings. For instance, Wang et al. [33] proposed an unsupervised multimodal framework that pioneered the incorporation of street view images and POIs, which utilized convolutional neural networks to extract visual features from street view images while adopting a bag-of-words model for POI data. Zhang et al. [4] designed a multi-view joint learning model that unified various urban data sources, such as human mobility patterns and inherent region characteristics, into a graph-based representation, incorporating a joint learning module to facilitate cross-view information sharing. Park et al. [34] explored the fusion of multi-level geospatial data, including satellite imagery, POIs, and socioeconomic metrics, to predict economic indicators. While these pioneering studies effectively leveraged multimodal data to improve urban representation, they often relied on unsupervised approaches and lacked the modality-specific embeddings needed to fully exploit each data type's unique characteristics.

Subsequent studies [35], [36] addressed these limitations using contrastive learning, which enhances feature learning by comparing related data modalities. Li et al. [37] focused on constructing positive sample pairs based on geographic proximity and data augmentation, while Xi et al. [38] incorporated spatial adjacency and POI-based similarity metrics to ensure that images with similar POI distributions have closer visual representations. Although these works introduced modality-

specific enhancements, their reliance on single-modal contrastive learning limited their ability to capture complementary information across different urban data sources.

Recent advancements in the field have highlighted the potential of multimodal contrastive learning to exploit synergies across diverse urban data, capturing complementary information from multiple modalities and addressing the limitations of earlier single-modal approaches. Liu et al. [39] introduced a knowledge graph-based framework for modelling spatial and mobility knowledge and developed an image-to-knowledge graph inter-modal contrastive learning approach. Cepeda et al. [40] extended vision–language alignment to global-scale geo-localization, enabling effective pretraining of geospatial embeddings through contrastive objectives. Li et al. [41] presented an inter- and intra-view contrastive learning framework that combined human mobility data and region attributes, resulting in more comprehensive urban region embeddings. Yong et al. [42] proposed a multi-semantic contrastive learning framework, which employs an Attentional Fusion Module to combine street view and satellite imagery before contrasting the fused feature with POI embeddings. Yan et al. [27] leveraged LLMs to generate descriptive text for street view images and applied contrastive learning to integrate textual and visual data for urban region profiling.

Beyond contrastive learning, recent studies [43], [44] have moved LLM-enhanced urban understanding. Xiao et al. [45] integrate language and visual representations to improve urban region understanding. Hao et al. [46] introduced continuous location embeddings derived from web-scale satellite imagery, bridging semantic and spatial continuity. Feng et al. [47] developed UrbanLLaVA, a multimodal LLM capable of spatial reasoning and visual–textual alignment for urban analysis.

While these recent approaches have shown promising results, they often lack the ability to explicitly model spatial structures across different urban scales and overlook the structural alignment and semantic interplay between heterogeneous modalities. This limits their capacity to preserve critical spatial context and to exploit the full potential of multimodal data synergy in complex urban environments.

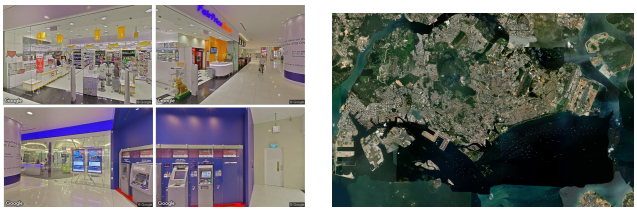
III. Preliminary

In this section, we provide a detailed description of the urban data used in this study and a formal problem definition.

Definition 3.1 (*Urban Region*): An urban region u refers to an urban area with a geographic central position, $u.g = (lat_u, lon_u)$.

Definition 3.2 (*Point of Interest*): A POI refers to a point with a geographic position, $p.g = (lat_p, lon_p)$. It can be associated with a set of tags $p.t = \{t_1, ..., t_i\}$, where each tag t_n is a key-value pair $k_n : v_n$. Some larger POIs, such as buildings, universities and airports are associated with an ordered list of nodes $p.n = [n_1, ..., n_j]$, where $n_1 = n_j$, that defines its size and shape as a polygon.

Definition 3.3 (*Way*): A Way w is defined by an ordered list of nodes $w.n = [n_1, \ldots, n_k]$ and a length $w.l$, which outlines its shape as a polyline. Similar to nodes, a Way can also be associated with a set of tags $w.t$. Polyline-type Ways are used to represent linear features such as roads, bridges, and rivers.



(a) Example of street view group. (b) Singapore satellite image.

Fig. 1: Example of street view group and satellite image.

Definition 3.4 (*Street View Imagery Group*): A street view imagery group sv refers to a set of street view images captured in a sampling point with a geographic position, $sv.g = (lat_{sv}, lon_{sv})$ and contains street view images from the four directions: east, south, west, and north, denoted as $sv.i = (s_e, s_s, s_w, s_n)$.

Definition 3.5 (*Satellite Imagery*): A satellite image si refers to a photo of the Earth's surface taken by satellites orbiting the Earth, with a central geographic position, $si.g = (lat_{si}, lon_{si})$ and the image $si.i$ is cropped to correspond to a specific urban region.

Examples of POI, Way, Street View Imagery and Satellite Image are provided in Table I, and Figure 1.

Problem Statement. Given a set of disjoint urban regions $\mathcal{U} = \{u_1, ..., u_k\}$, each urban region u is associated with multimodal geospatial data distributed across multiple spatial scales, including a segmented satellite image $SI = \{si.i\}$, a series of street view imagery groups $SV = \{sv_1, ..., sv_m\}$ captured within, and a set of POIs $P = \{p_1, ..., p_n\}$. Our objective is to model urban environments across multiple spatial scales to effectively integrate these multimodal geospatial data across different spatial scales into a unified urban representation, which is both generalizable and effective across a wide range of downstream urban analysis tasks.

IV. Methodology

UrbanMFM is a flexible framework for pre-training urban representation models using diverse multimodal data, including but not limited to OSM data and urban imagery. The architecture of UrbanMFM is summarized in Figure 2.

A. Encoder Layer

1) Visual Encoder.: Our proposed UrbanMFM allows encoding urban images to extract rich and meaningful features from urban imagery, enabling effective representation learning for various urban-related tasks. We utilize ResNet18 [48] as the encoder $f^{image}(\cdot)$ to each image in the group to generate high-dimensional feature representations. Formally, given a satellite image si or a street view imagery group sv:

$$E_{si} = f^{image}(si.i) = \text{ResNet18}(si.i) \tag{1}$$

$$E_e, E_s, E_w, E_n = f^{image}(sv.i) = \text{ResNet18}(sv_e, sv_s, sv_w, sv_n) \tag{2}$$

TABLE I: Examples of OSM data.

Type	Example

POI	type : node,
	id : 368278973,
	lat : 1.3939865,
	lon : 103.8892811,
	tags : {
	brand: 7-Eleven,,
	brand:wikidata: Q259340,
	brand:wikipedia: en:7-Eleven,
	shop: convenience,
	wheelchair: limited
	}

POI (polygon)	type : polygon,
	id : 543994414,
	lat : 1.4176167473370995,
	lon : 103.84822884762922,
	tags : {
	addr: housenumber: 556,
	building: apartments
	}
	area: 495,
	points : [
	[1.4168693,103.8490036],
	[1.4168419,103.8489909],
	...
	[1.4168693,103.8490036]
	]

Way	type : way, id : 9585621, nodes : [74368120, 6873115634, ... 74368117] tags : { highway: primary, maxspeed: 50, lanes: 5, name: Paterson Road, oneway: yes, ...} points : [[1.3050396,103.8318399], [1.3049007,103.8317163], ... [1.3038595,103.8308472]] length: 171

2) Textual Encoder.: The textual encoder aims to produce semantic representations of textual tags. We leverage pre-trained Language Models (LMs) to generate dense embeddings that encapsulate semantic information, thereby reducing dimensionality and preserving relationships between similar categories. Formally, given a POI p, we combine the textual tags and employ BERT [49] as textual encoder $f^{text}(\cdot)$ to capture its textual representation E_t:

$$E_t = f^{text}(p.t) = BERT(\text{[CLS] } t_1, t_2, ..., t_i \text{ [SEP]}) \quad (3)$$

where [CLS] represents the entire input sequence, while [SEP] is used to separate different segments.

3) Polygon Encoder.: Since larger POIs are associated with node lists, we characterize their shape and size as polygons. Following [28], we leverage a polygon encoder $f^{poly}(\cdot)$ to generate a high-dimensional representation that captures both the surface area and the rasterized features of the POI:

$$E_p = f^{poly}(p.n) = \frac{1}{2}(\text{ResNet18}(Raster(p.n) + \frac{Surface(p.n)}{max_a}) \quad (4)$$

where $Raster$ is a rasterization function that maps a closed polygon, represented as an ordered list of nodes, to a binary image; $Surface$ is a function to compute the surface area of a polygon in m^2, and max_a is the maximum polygon area.

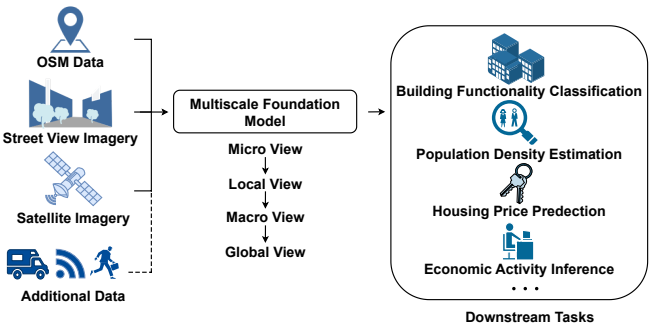


Fig. 2: An overview framework of UrbanMFM.

4) Graph Encoder.: To capture the spatial context inherent in the network topology of surrounding neighbours and thereby create a more contextualized representation for each POI or urban region, we separately construct local and global spatial graphs. The local spatial graph is constructed based on POIs and their neighbouring POIs as well as the spatial distance, while the global spatial graph is built using regions and their adjacent regions. The specific details of the local and global spatial graph will be discussed in Section IV-B2 and Section IV-B4. To effectively process these graphs, we employ the Graph Attention Network (GAT) [50] as the backbone of our graph encoder $f^{graph}(\cdot)$, which is designed to address the complexities of both local and global spatial relationships. We also compared the performance of GAT and GATv2 [51], and the experimental results indicate that the choice of model has a negligible impact on the overall performance. Formally, given a graph $\mathcal{G}$:

$$E_{\mathcal{G}} = f^{graph}(\mathcal{G}) = \text{GAT}(\mathcal{G}) \quad (5)$$

B. Multiscale Foundation Model

The core of the framework is Multiscale Foundation Model, as illustrated in Figure 3, which integrates five interconnected objectives to address fine-grained spatial details and multi-modal interactions from data point, local, region to global scales: (1) *Micro Scale Learning* which which brings spatially close POIs together in the feature space, ensuring that functionally similar POIs are properly aligned and ensures the street view images remain semantically consistent across scales by incorporating visual diversity from different angles at the same sampling point; (2) *Local Scale Learning* builds local spatial graphs where POI text, polygon, and street view features are treated as modality-specific nodes, with a triplet contrastive objective models their interactions and enhance local representations by aggregating neighbourhood information; (3) *Macro Scale Learning* captures macro-level urban features, integrating similar road network topology and POI distribution to enhance the overall region-level representation; (4) *Global Scale Learning* constructs global spatial graphs, macro representations derived from satellite imagery are contrastively aligned with micro representations aggregated from POI and street view features, capturing cross-region spatial dependencies and multi-scale semantic consistency.

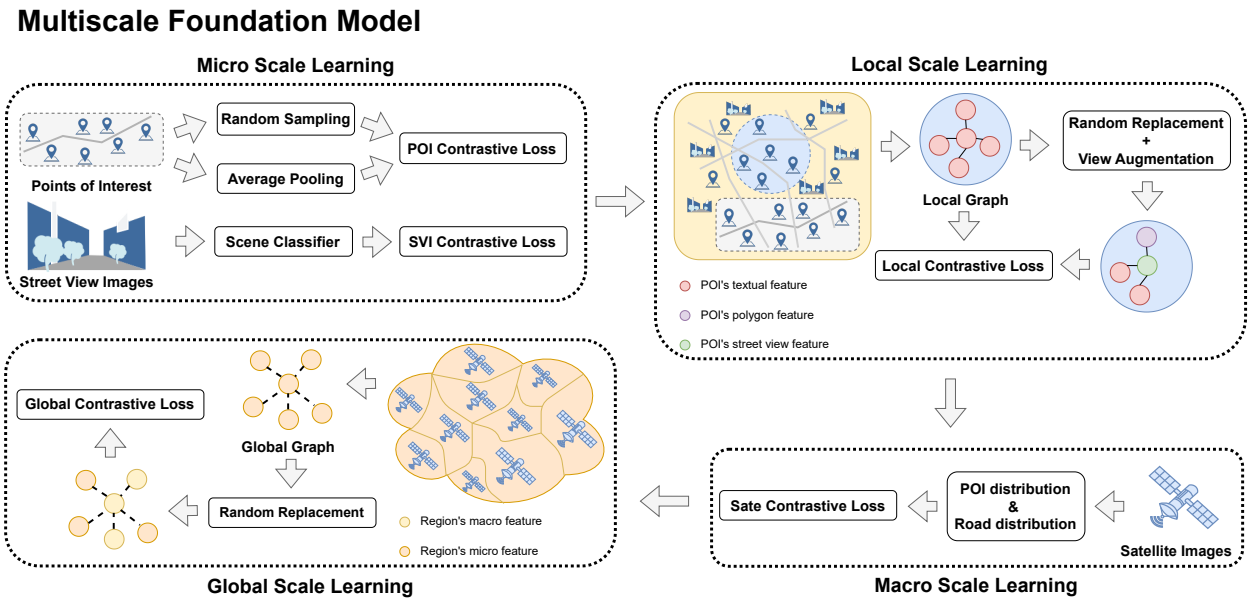


Fig. 3: An overview framework of Multiscale Foundation Model.

1) Micro Scale Learning.: The most valuable aspect of a POI's textual annotation lies in the information that denotes its functionality. This contrastive objective is designed around the expectation that *POIs located on the same road, show a stronger correlation* [52]. To achieve this, we group POIs that are geographically close, specifically those located on the same road segment, and maximize the similarity between the road representation, obtained by aggregating the POIs in a road context using an average-pooling layer, and a randomly sampled POI:

$$E_{avg} = Avg(E_t^1, E_t^2, ..., E_t^n) \tag{6}$$

where n represents the number of POIs on the road. For the implementation, we employ an information noise contrastive estimation (InfoNCE) [53] loss function:

$$\mathcal{L}_{poi} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(E_t^i, E_{avg}^i)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(E_t^i, E_{avg}^j)/\tau)} \tag{7}$$

where B is the mini-batch size, E_t is the initial textual representation for POI p, τ (= 0.5) is a temperature parameter.

Street view images are essential for capturing detailed, ground-level information about urban environments, providing insights into the visual and functional characteristics of streetscapes and built environments. A key challenge in utilizing street view images is to accurately capture the inherent visual diversity from different perspectives at the same location while maintaining semantic consistency across views. This diversity can reveal critical urban details, but inconsistent or biased sampling may hinder the model's ability to generalize effectively across different urban scenarios. To address the challenges of spatial consistency and directional variation in street view imagery, we propose a sampling method that selects images from the same location but different directions as positive pairs while accounting for the significant content variation in images taken from different directions. This preserves semantic coherence while capturing rich visual diversity. To enhance alignment across directions, we train a scene classifier using LLMs. Specifically, we generate textual descriptions from image embeddings of each multi-view street view group sv, which are used as prompts in LLAMA2 [54] to produce scene-level embeddings E_{scene}. The classifier, validated on 500 manually labeled samples, achieves over 85% accuracy, confirming the reliability of the generated scene representations.

$$E_{scene} = f^{text}(\text{LLAMA2}(f^{image}(sv.i))) \tag{8}$$

We then utilize different street view images within $sv.i$ as positive samples to ensure that the positive samples share a consistent geographical context while capturing a wide range of visual diversity. The InfoNCE loss is given by:

$$\mathcal{L}_{svi} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}([E_\alpha^i, E_{scene}^i], [E_\beta^i, E_{scene}^i])/\tau)}{\sum_{j=1}^{N}\sum_{\gamma}\exp(\text{sim}([E_\alpha^i, E_{scene}^i], [E_\gamma^j, E_{scene}^i])/\tau)} \tag{9}$$

where $\alpha, \beta, \gamma \in \{e, s, w, n\}$ and $\alpha \neq \beta$, B is the mini-batch size, τ (= 0.5) is a temperature parameter. Compared to previous methods [31], [33], [37], which often relied on geographical similarity or data augmentation, our approach addresses the limitations of purely proximity-based sampling and artificial augmentations. By focusing on spatial and visual consistency, we provide more robust and generalizable representations of urban streetscapes.

2) Local Scale Learning.: In urban environments, effectively capturing interactions across multiple data modalities is essential for accurate representation learning. Traditional methods often focus on pairwise interactions, overlooking the complex relationships among modalities such as POIs, spatial layouts, and street view imagery. To address this, we propose a spatial graph-based triplet inter-modal contrastive learning approach that jointly models three key modalities: functional POI

data, spatial polygons, and visual street views. By constructing a local spatial graph to integrate these modalities, our method preserves spatial relationships and enables more comprehensive feature extraction. This triplet framework ensures that each modality contributes complementary information, leading to richer and more spatially aware urban representations than pairwise methods. However, POIs do not have a direct one-to-one match with street view images. A common workaround is to pair each POI with its nearest street view sampling point, but this can be unreliable—especially in complex urban settings—due to image quality issues, inconsistent relevance, and physical obstructions that may block the POI from view.

In OSM data, for each POI p, we can encode its textual tags to obtain a POI textual information E_t. A subset of POIs also includes polygonal shape and size information, allowing us to extract spatial features E_p. However, POIs do not directly correspond to street view images. Common practice involves pairing POIs with the nearest street view sampling point, but this can be unreliable—especially in complex urban settings—due to image quality issues, inconsistent relevance, and physical obstructions that may block the POI from view. To overcome these issues, we propose *Spatial-Aware Semantic Consistency* to capture the appropriate street view representation for each POI. For each POI p, we collect a set of street view imagery groups $SV = \{sv_1, ..., sv_k\}$ if the following condition holds:

$$dist = Haversine(lat_p, lat_{sv}, lon_p, lon_{sv}, r) \leq max_{dist} \tag{10}$$

where the $Haversine$ formula calculates the great-circle distance between two points given their latitudes and longitudes, r is the radius of the Earth, and max_{dist} is the threshold to 1 km. We compute the angle between the coordinates of a sampling point and the POI, designating the street view image in the direction of the angle as the anchor, and the images in the other three directions as the context (e.g., s_e as the anchor image and s_s, s_w, s_nas the context images):

$$E^a_{sv} = E_e,\ E^c_{sv} = Avg(E_s, E_w, E_n) \tag{11}$$

We apply a gated fusion mechanism to fusion E^a_{sv} and E^c_{sv} as the final street view imagery group representation E_{sv}:

$$E_{sv} = softmax([E^a_{sv}, E^c_{sv}] \cdot W_g + b_g) \cdot [E^a_{sv}, E^c_{sv}]^\top \tag{12}$$

where W_g and b_g are learnable parameters. All the embeddings of obtained street view image groups are served as street view memory $SV = \{E^1_{sv}, ..., E^M_{sv}\}$ and corresponding distance memory $D = \{dist^1, ..., dist^M\}$. We generate a corresponding street view embedding E_v that is semantically similar to the POI textual embedding E_t:

$$E_v = \sum_{i=1}^{M} \frac{\exp(\text{sim}(E_t, E^i_{sv}) \cdot b_i/\tau)}{\sum_{j=1}^{M} \exp(\text{sim}(E_t, E^j_{sv}) \cdot b_j/\tau)} \cdot E^i_{sv} \tag{13}$$

$$b_i = \frac{max_{dist}}{dist^i + max_{dist}} \tag{14}$$

where $\tau(= 0.5)$ is the temperature hyperparameter. By dynamically absorbing information from memories based on spatial-aware semantic similarity to the POI textual embeddings E_t, we can generate more diverse and accurate street view embeddings E_v. For each POI p, we collect a set of surrounding POIs $N = \{n_1, ..., n_i\}$ as spatial context if the following condition is met:

$$dist = Haversine(lat_p, lat_n, lon_p, lon_n, r) \leq max_{dist} \tag{15}$$

We then construct a local spatial graph $\mathcal{G}_l$ for each POI with the following principles:

- Consider the target POI as a node of type *POI*, its neighbours in P_n as nodes of type *neighbour*, and the initial value of each node based on the textual tags provided by OSM,
- Connect the *POI* node with all *neighbour* nodes, where the weight of each edge is the inverse of the distance between the connected nodes and the *POI* node.

To construct a positive sample graph $\mathcal{G}^+_l$, we randomly replace the original representation of each *POI* and *neighbour* node with either its street view or polygon embedding. The key idea is that the semantic (textual), spatial (polygon), and visual (street view) representations of a POI should be close in the embedding space. To improve the robustness of our graph-based triplet contrastive learning objective, we further apply view augmentation by selectively removing nodes from the local spatial graph. The InfoNCE loss is defined as:

$$\mathcal{L}_{local} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(E^i_{\mathcal{G}_l}, E^{i+}_{\mathcal{G}_l})/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(E^i_{\mathcal{G}_l}, E^{j+}_{\mathcal{G}_l})/\tau)} \tag{16}$$

where B is the mini-batch size, $E^i_{\mathcal{G}_l}$ and $E^{i+}_{\mathcal{G}_l}$ are the spatial-aware representations of a local spatial graph and its positive sample, computed as in Eq. 5, τ (= 0.5) is a temperature parameter. Compared to prior methods [27], [39], [41] that focus only on two-modality interactions, our approach captures the deeper and more complex relationships between multiple modalities, improving the model's ability to represent the spatial and functional aspects of urban areas.

3) Macro Scale Learning.: Satellite imagery plays a crucial role in capturing macro-level urban features and spatial patterns, providing a comprehensive view of a city's structure and landscape. However, a significant challenge is effectively capturing the complex spatial relationships within regions, which include not only the distribution of POIs but also the intricate layout of road networks. To address this, our method leverages POI and road distributions to better capture spatial and topological structures. To alleviate these issues, we propose a POI and Road distribution similarity-based sampling method that captures more complex spatial relationships and topological structures. Instead of using a fixed-size grid division, we divide the city into fine-grained regions based on the road networks. For each urban region u, we first collect POIs $P = \{p_1, ..., p_m\}$ within the region, where each $P_c = \{p_1, ..., p_{c_i}\}$ denotes the set of POIs belonging to the c-th POI category. Given the non-uniform distribution of POIs, we apply Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [55] to extract POI distribution features for different categories:

$$C_c = \text{DBSCAN}(P_c, \epsilon_c, \text{MinPts}_c) \tag{17}$$

where ϵ_c (= 0.05) is the maximum distance between two samples, and MinPts_c (= 2) represents the minimum number of points required to form a dense cluster. All POIs, including those identified as outliers by DBSCAN, are retained in the learning process to maintain full spatial coverage and avoid information loss in sparsely populated areas. The set of clusters for the c-th POI category is denoted as $C_c = \{C_{c_1}, ..., C_{c_k}\}$, where each cluster $C_{c_i} = \{p_1, ..., p_{m_i}\}$ contains a collection of POIs belonging to the c-th category. We then extract the distribution feature f_c for the c-th POI category as follows:

$$lat_c, lon_c = \left(\frac{1}{k} \sum_{j=1}^{k} \frac{1}{|C_{c_j}|} \sum_{p_i \in C_{c_j}} lat_{p_i}, \frac{1}{k} \sum_{j=1}^{k} \frac{1}{|C_{c_j}|} \sum_{p_i \in C_{c_j}} lon_{p_i} \right) \tag{18}$$

$$f_c = [\sqrt{\frac{\pi}{a_u}} \max_{p_j \in C_c} \sqrt{(lat_{p_j} - lat_c)^2 + (lon_{p_j} - lon_c)^2}, \ \frac{|P_c|}{a_u}] \tag{19}$$

where r_c and d_c respectively denote radium and density features, we utilize Equivalent Circle Radius to normalize radium, and a_u is the area of the region u. Then we integrate into region POI distribution feature $F_p = [f_1, ..., f_m]^\top$. Additionally, we gather ways $W = \{w_1, ..., w_n\}$ that traverse or are contained within the area. We consider road length density and road intersection density to compare road distribution similarity F_r:

$$F_r = [\frac{\sum_{i=1}^{n} w_i.l}{a_u}, \ \frac{inter_u}{a_u}] \tag{20}$$

where $inter_u$ represents the number of road intersections within the region u. We measure the POI and road distribution between two regions based on cosine similarity. Two regions are deemed to be a positive pair if the following holds:

$$cosine(F_p^i, F_p^j) \geq 0.95 \wedge cosine(F_r^i, F_r^j) \geq 0.95 \tag{21}$$

The threshold value 0.95 balances selecting sufficiently similar region pairs while maintaining sample diversity. We also explored several alternative formulations, including a learnable similarity threshold, a margin-based objective, and hard-negative mining. These variants yielded results with only about a 1% improvement over the fixed-threshold formulation; therefore, we adopt the simpler and more stable design in the final model. We utilize InfoNCE to maximize the similarity between the positive pairs:

$$\mathcal{L}_{sate} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\text{sim}(E_{si}^i, E_{si}^{i+})/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(E_{si}^i, E_{si}^{j+})/\tau)} \tag{22}$$

where E_{si}^i and E_{si}^{i+} denote the satellite embeddings of the i-th urban region and its corresponding positive sample, computed as in Eq.1, τ (= 0.5) is a temperature parameter.

4) Global Scale Learning.: In urban representation learning, integrating diverse data sources is imperative for capturing the complex characteristics of urban environments. In order to do so, we propose a novel contrastive objective for multi-modal data fusion. Unlike previous methods that treat the representations of each data source independently, our approach generates macro representations from satellite imagery while modelling each POI using OSM data and street view imagery, which are pooled to form micro representations. The key insight is to align the macro and micro representations of a region, reflecting the complementary relationship between these spatial scales. Furthermore, we construct a global spatial graph that enables the model to capture spatial dependencies across regions, facilitating more comprehensive and consistent representation learning across multiple urban scales.

For each urban region u, we collect a set of POIs within it, denoted as $P = \{p_1, ..., p_n\}$ and derive the macro representation E_{macro} from the corresponding satellite image. Then we model each POI p separately, computing its POI representation E_t^j. To extract the street view representation E_v^j, we apply the *spatial-aware semantic consistency* method described in Section IV-B2. We then fuse the textual and street view features for each POI. Initially, we propose a Residual Multi-Head Attention mechanism to compute the interaction between E_p^j and E_v^j:

$$E_{poi}^j, E_{svi}^j = \text{RMHA}(E_p^j, E_v^j) \tag{23}$$

where RMHA consists of a 2-layer multi-head attention mechanism and a 1-layer residual connection. The fusion of the POI and street view representations is achieved through a weighted feature fusion mechanism:

$$g_j = \text{Sigmoid}(\text{ReLU}([E_{poi}^j, E_{svi}^j] \cdot W_j + b_j) \tag{24}$$

$$E_j = g_j \cdot E_{poi}^j + (1 - g_j) \cdot E_{svi}^j \tag{25}$$

For POIs with polygon representations E_P^j, as outlined in Eq. 6, we integrate the fused feature E_j with the polygon features E_P^j as above. Thereby obtaining an updated representation, also denoted as E_j. Finally, the micro representation E_{micro} for region u is derived by aggregating the POI representations within the region as:

$$E_{micro} = Avg(E_1, ..., E_n) \tag{26}$$

We construct a global spatial graph $\mathcal{G}_g$, based on the following principles:

- Consider the target region as a node of type *region*, with its geographically adjacent regions as nodes of type *neighbour*, and the value assigned to each node corresponds to its macro representation E_{macro},
- Connect the *region* node with all *neighbour*-type node, where the weight of each edge is the inverse of the distance between the centre coordinates of the connected regions.

To generate a positive sample $\mathcal{G}_g^+$, we randomly replace the initial macro value with the corresponding micro representation E_{micro}. The InfoNCE loss is defined as:

$$\mathcal{L}_{global} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\text{sim}(E_{\mathcal{G}_g}^i, E_{\mathcal{G}_g}^{i+})/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(E_{\mathcal{G}_g}^i, E_{\mathcal{G}_g}^{j+})/\tau)} \tag{27}$$

where B is the mini-batch size, $E_{\mathcal{G}_g}^i$ and $E_{\mathcal{G}_g}^{i+}$ are the spatial-aware representations of a global spatial graph and its positive sample, τ (= 0.5) is a temperature parameter. In comparison to previous methods [56]–[58], which often use grid-based divisions [38], our approach ensures that spatial boundaries are defined by actual urban structures, avoiding the artificial splitting of functional areas. By integrating both POI distribution and road network information, we provide a more holistic and nuanced representation of urban regions.

TABLE II: Statistics on the datasets used in the experiments. The columns $Region$, SVI and POI separately represent the number of urban regions, street view imagery groups and points of interest in the corresponding dataset. The column $POLY(\%)$ represents the number and percentage of points of interest with polygons in the corresponding dataset.

City	Land Area(km^2)	Region	SVI	POI	#POLY (%)
Singapore	728	3,056	34,078	64,215	40,594 (63.22%)
New York	1,214	31,449	63,559	93,051	49,502 (53.20%)
Beijing	16,410	1,010	16,152	59,989	43,588 (72.66%)

V. Experiments

The objective of the experimental study is fivefold: (1) Demonstrating Effectiveness Across Representative Tasks, (2) Ablation Studies for Multiscale Foundation Model Analysis, (3) Verifying Adaptability, (4) Conducting Parameter Sensitivity Analysis, and (5) Presenting Qualitative Case Studies.

A. Experimental Setups

1) Datasets: In our data collection process, we use the OpenStreetMap Overpass API to download and preprocess POIs from Singapore, New York, and Beijing, and we apply a preprocessing step to the raw OSM data to remove redundant and non-informative tags, ensuring consistent and reliable POI semantics across cities. We segment each city into regions based on its road networks. Using the Google Maps API for Singapore and New York, and the Baidu Map API for Beijing, we systematically generate evenly distributed sampling points, capturing street view images in all four cardinal directions at each point. Additionally, we obtain satellite imagery through the Bing Maps API, referencing each city's vector boundaries. The extensive scale and careful selection of these representative cities ensure our datasets are robust and comprehensive, facilitating meaningful and generalizable analyses across diverse urban environments. As shown in Table II, the dataset statistics cover multiple variables such as the number of regions, average area, POI count, and imagery samples. Notably, the number of regions and their corresponding areas are not proportional across cities, reflecting heterogeneous spatial granularities that allow UrbanMFM to be evaluated across diverse geographic scales. The dataset on housing price prediction consists of 6,073 communities, scraped from Beike, which provides the housing prices of the collected communities, and then calculate the average housing price in the area through road network segmentation. The dataset on GDP, with a spatial resolution of 1 km, is sourced from Resource and Environmental Science Data Platform. The dataset on building functionality classification task consists of 64,384 polygons that belong to one of 8 classes, obtained from Urban Redevelopment Authority.

2) Downstream Tasks: We aim to validate the superiority of our proposed method by comparing it against the best baseline models specifically designed for four representative tasks: Population Density Estimation, Housing Price Prediction, GDP Density Estimation and Building Functionality Classification. These comparisons span multiple countries and culturally diverse regions, highlighting the generalizability of our approach. Notably, the first two are region-level tasks, while the last is a POI-level task. This diversity of downstream tasks demonstrates that our model can effectively generalize across different spatial granularities, capturing both macroscopic socioeconomic attributes and microscopic functional characteristics of urban areas. Building functionality data is available for only a limited number of cities, and within our datasets, such information was accessible exclusively for Singapore. Despite this constraint, Singapore is a major metropolitan area with diverse building types and comprehensive infrastructure, making it a representative case for our study. Housing price and GDP density data in our datasets were limited to Beijing. As one of the world's leading megacities and the capital of China, Beijing's real estate market encompasses a wide spectrum of housing types, pronounced spatial economic disparities, and complex urban structure, providing a representative context for studying regional socioeconomic patterns. This diversity provides valuable insights, making it a suitable focus for our analysis despite the geographical limitation.

3) Baselines: We compare UrbanMFM with seven strong baselines, including three unimodal baselines: GeoVectors [59], SpaBERT [60], HGI [9], and four multimodal baselines: Urban2vec [33], MuseCL [42], CityFM [28] and UrbanCLIP [27]. Among these multimodal baselines, Urban2Vec includes POI and street view images, while CityFM leverages POI and polygon data from OSM and additional road-network information corresponding to the micro and local scales in our framework. UrbanCLIP, in contrast, focuses on satellite imagery and employs LLM-generated textual captions, primarily activating the macro and global scale through enhanced semantic supervision. This limitation highlights a fundamental disparity: none of these baselines fully align with UrbanMFM's design and research motivation, which integrates multiple modalities to capture urban complexity comprehensively. Notably, MuseCL adopts the same input modalities as UrbanMFM but differs substantially in its modeling strategy and objectives.

- **Urban2Vec** [33] learns a low-dimensional representation of each urban region utilizing urban images and POIs inside.
- **GeoVector** [59] uses the corpus to obtain the location encoding for untagged polygons, while tag embeddings represent the entities located in their spatial proximity.
- **SpaBERT** [60] is a pre-trained model designed to encode entities within each region by leveraging LMs and aggregating by calculating the average representations.
- **HGI** [9] employs Hierarchical Graph Infomax with POI data to learn representations at the POI- and region-levels, effectively capturing hierarchical urban dynamics.
- **MuseCL** [42] aggregates street view and satellite imagery

TABLE III: Estimation and ablation study of population density, expressed in thousands of people per square kilometre. The row highlighted in bold represents the performance of UrbanMFM, while the rows above it show the baseline models, distinguished by whether they adopt a foundation model (✓) or not (✗), and further categorized by modality as either unimodal (u) or multimodal (m). Below, we provide the results from the ablation study.

Foundation	Mobility	Model	Singapore			New York			Beijing		
			RMSE ↓	MAE ↓	R^2 ↑	RMSE ↓	MAE ↓	R^2 ↑	RMSE ↓	MAE ↓	R^2 ↑
✗	u	GeoVectors	6.38 (± 0.23)	5.2 (± 0.27)	0.51 (± 0.06)	7.56 (± 0.3)	6.09 (± 0.22)	0.58 (± 0.01)	7.56 (± 0.3)	6.09 (± 0.22)	0.58 (± 0.01)
✗	u	SpaBERT	6.89 (± 0.68)	4.85 (± 0.43)	0.49 (± 0.05)	8.08 (± 0.19)	5.69 (± 0.16)	0.51 (± 0.03)	9.84 (± 0.73)	7.83 (± 0.46)	0.27 (± 0.05)
✗	u	HGI	5.31 (± 0.57)	4.06 (± 0.31)	0.57 (± 0.05)	6.15 (± 0.23)	4.18 (± 0.17)	0.72 (± 0.02)	8.88 (± 0.50)	6.25 (± 0.36)	0.37 (± 0.07)
✗	m	Urban2Vec	5.36 (± 0.24)	4.16 (± 0.18)	0.64 (± 0.04)	6.17 (± 0.16)	5.28 (± 0.20)	0.69 (± 0.02)	8.94 (± 0.34)	5.35 (± 0.24)	0.40 (± 0.03)
✗	m	MuseCL	4.31 (± 0.60)	3.00 (± 0.42)	0.82 (± 0.04)	5.92 (± 0.28)	4.20 (± 0.17)	0.72 (± 0.02)	6.83 (± 0.84)	4.88 (± 0.49)	0.62 (± 0.06)
✓	m	CityFM	4.68 (± 0.45)	3.20 (± 0.23)	0.79 (± 0.05)	5.90 (± 0.21)	4.21 (± 0.14)	0.73 (± 0.02)	7.16 (± 0.90)	4.92 (± 0.54)	0.58 (± 0.08)
✓	m	UrbanCLIP	4.58 (± 0.40)	3.24 (± 0.32)	0.82 (± 0.03)	6.05 (± 0.18)	4.30 (± 0.16)	0.72 (± 0.02)	7.25 (± 0.45)	5.12 (± 0.31)	0.63 (± 0.03)
✓	m	**UrbanMFM**	**3.88** (± 0.27)	**2.64** (± 0.19)	**0.87** (± 0.02)	**5.45** (± 0.15)	**3.74** (± 0.13)	**0.77** (± 0.01)	**5.72** (± 0.62)	**3.48** (± 0.30)	**0.75** (± 0.04)
-	-	w/o Micro Scale	4.53 (± 0.39)	2.95 (± 0.15)	0.85 (± 0.04)	5.88 (± 0.25)	4.12 (± 0.21)	0.74 (± 0.02)	6.02 (± 0.88)	3.71 (± 0.35)	0.72 (± 0.05)
-	-	w/o Local Scale	4.14 (± 0.24)	2.74 (± 0.18)	0.86 (± 0.02)	5.60 (± 0.15)	3.81 (± 0.13)	0.76 (± 0.01)	5.79 (± 0.66)	3.62 (± 0.29)	0.74 (± 0.05)
-	-	w/o Macro Scale	4.21 (± 0.54)	2.79 (± 0.30)	0.83 (± 0.04)	6.06 (± 0.29)	4.26 (± 0.17)	0.71 (± 0.01)	6.84 (± 0.81)	4.17 (± 0.49)	0.64 (± 0.06)
-	-	w/o Global Scale	4.68 (± 0.45)	3.20 (± 0.23)	0.79 (± 0.05)	6.01 (± 0.17)	4.15 (± 0.17)	0.71 (± 0.02)	7.04 (± 0.71)	4.15 (± 0.53)	0.63 (± 0.09)

with an attentional fusion module, then utilizes a contrastive mechanism for integrating POI features.

- **CityFM** [28] leverages geospatial entities extracted from OpenStreetMap and employs contrastive learning with three objectives to pre-train a foundation model: a mutual information-based text-to-text objective, a vision-language objective, and a road-based context-to-context objective.
- **UrbanCLIP** [27] generates detailed textual descriptions for each satellite image using LLMs and trains on the resulting image-text pairs, combining contrastive loss and language modelling loss for joint optimization.

To achieve a fair and spatially consistent comparison across models, we construct a unified setting using only POI and polygon data as inputs. For baselines, we include Urban2Vec (replacing street-view features with polygon embeddings) and CityFM (disabling the road-network branch), ensuring that all models are trained on identical data sources. UrbanMFM is correspondingly restricted to its POI–polygon components, activating the micro and local scales within our framework. This controlled setup enables a direct and fair evaluation of model architectures under a consistent spatial context, while still capturing the intrinsic multiscale structure of urban spaces.

4) Experimental Settings: UrbanMFM is pre-trained using OSM data and urban imagery, on the five contrastive objectives presented in Section IV-B. We follow CityFM [28], each objective is optimized in an independent stage with the model weights initialized from the previous stage. The pretraining loss can be expressed as the sum of stage-wise objectives:

$$\mathcal{L}_{pretrain} = \mathcal{L}_{poi} + \mathcal{L}_{svi} + \mathcal{L}_{local} + \mathcal{L}_{sate} + \mathcal{L}_{global}$$

The model is trained until convergence with a learning rate of 1e-4, a batch size of 256, and a linearly decreasing learning rate scheduler with warm-up. We utilize the *bert-base-uncased* model with a hidden size of 768 to encode textual information. We preprocess the images to 224x224 pixels for urban imagery, ensuring the generated rasterized images are also 224x224. We then employ ResNet18 with a hidden size of 512 for image encoding. Following the pre-training phase, the model's parameters are frozen, and UrbanMFM is utilized to generate comprehensive representations for the geospatial entities and urban imagery involved in the different downstream applications. All the experiments involving deep learning frameworks are executed on a V100-SXM2 GPU. Under this setting, the pre-training phase requires 16–18 hours across our three datasets. Peak GPU memory usage is around 18 GB.

B. Performance Analysis

For the Population Density Estimation and Housing Price Prediction task, we report evaluation metrics including absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2), with mean and standard deviation over 10 independent runs, for each algorithm. For the Building Functionality Classification task, we randomly partition the dataset into 50%, 25%, and 25% as the training, validation, and testing sets, and maintain a consistent ratio of building functionalities during the splitting process.

1) Population Density Estimation: Table III demonstrates that UrbanMFM consistently outperforms all state-of-the-art baselines across diverse urban regions. The performance gain

TABLE IV: Population Density Estimation under Unified Setting.

Model	Singapore			New York			Beijing		
	RMSE ↓	MAE ↓	R^2 ↑	RMSE ↓	MAE ↓	R^2 ↑	RMSE ↓	MAE ↓	R^2 ↑
Urban2Vec	6.30 (±0.30)	4.62 (±0.21)	0.61 (±0.03)	7.10 (±0.33)	5.12 (±0.23)	0.56 (±0.03)	9.50 (±0.35)	5.60 (±0.27)	0.33 (±0.04)
CityFM	5.95 (±0.27)	4.15 (±0.19)	0.65 (±0.02)	6.70 (±0.30)	4.60 (±0.20)	0.60 (±0.03)	8.10 (±0.32)	5.05 (±0.23)	0.52 (±0.03)
UrbanMFM	**5.55** (±0.25)	**3.78** (±0.18)	**0.67** (±0.02)	**6.28** (±0.28)	**4.30** (±0.19)	**0.63** (±0.02)	**7.80** (±0.31)	**4.85** (±0.22)	**0.58** (±0.03)

is particularly striking on the Beijing dataset, where RMSE is reduced by **16.25%**, MAE by **28.69%**, and R^2 is improved by **19.05%** compared to the strongest baseline. Notably, this performance is achieved despite Beijing being the most challenging setting, where the dataset contains the fewest POIs and street view sampling points and the largest urban area among all cities. Such improvements under sparse and imbalanced multimodal conditions clearly demonstrate the robustness and adaptability of our approach.

Among unimodal baselines without foundation model support, GeoVectors, SpaBERT and HGI perform suboptimally due to their limited ability to capture complex spatial and multimodal patterns. While HGI is specifically designed for region-level estimation, its unimodal nature hinders its ability to model urban heterogeneity. Multimodal baselines without foundation models, Urban2Vec and MuseCL show moderate gains, benefiting from the integration of spatial and visual features. However, the absence of pre-trained semantic priors limits their representational expressiveness and generalizability. Models that incorporate foundation models, CityFM and UrbanCLIP achieve better performance by leveraging large-scale pre-trained knowledge to extract richer semantics from textual and visual data, yet still struggle to capture fine-grained cross-modal interactions and spatial hierarchies. By contrast, UrbanMFM achieves consistent and substantial improvements across all metrics and regions. Its advantage lies in combining multiscale representation learning with foundation models, enabling fine-grained multimodal fusion. This allows UrbanMFM to capture both local detail and global context, resulting in more robust and generalizable urban representations.

As shown in Table IV, under the unified POI–polygon setting, overall performance drops compared with the full multimodal configuration, as expected due to the removal of imagery and road-network cues. Nevertheless, UrbanMFM consistently outperforms both CityFM and Urban2Vec across all three cities. These results indicate that UrbanMFM's multiscale design provides substantial gains.

In addition, we assess robustness to different region granularities by merging adjacent 2–3 macro regions. In Beijing, where the original number of regions is small, merging reduces training units and results in a moderate performance drop of around 5%. In contrast, New York contains far more regions, and the same merging operation leads to only about 1% fluctuation. These results indicate that UrbanMFM remains stable under reasonable variations of the region scale.

2) Housing Price Prediction: Table V demonstrates that UrbanMFM outperforms all state-of-the-art baselines, achieves RMSE reduction of **11.64%**, MAE reduction of **27.60%**, and R^2 improvement of **5.26%** over the strongest baseline.

TABLE V: Experimental Comparison of Housing Price.

Foundation	Mobility	Model	RMSE ↓	MAE ↓	R^2 ↑
✗	u	GeooVector	27.56 (± 2.88)	20.94 (± 1.73)	0.64 (± 0.06)
✗	u	SpaBERT	28.25 (± 2.51)	21.62 (± 1.55)	0.58 (± 0.08)
✗	u	HGI	21.14 (± 2.06)	14.76 (± 0.74)	0.76 (± 0.05)
✗	m	Urban2Vec	24.67 (± 4.45)	19.04 (± 2.23)	0.66 (± 0.06)
✗	m	MuseCL	21.37 (± 3.25)	15.54 (± 1.25)	0.74 (± 0.04)
✓	m	CityFM	22.97 (± 3.36)	17.14 (± 1.94)	0.72 (± 0.06)
✓	m	UrbanCLIP	23.64 (± 2.98)	18.09 (± 1.67)	0.74 (± 0.05)
✓	m	**UrbanMFM**	**18.68** (± 1.47)	**10.69** (± 0.49)	**0.80** (± 0.03)

TABLE VI: Experimental Comparison of GDP Density.

Foundation	Mobility	Model	RMSE ↓	MAE ↓	R^2 ↑
✗	u	GeoVector	7.43 (± 0.59)	4.92 (± 0.26)	0.63 (± 0.05)
✗	u	SpaBERT	7.18 (± 0.52)	4.75 (± 0.23)	0.65 (± 0.05)
✗	u	HGI	6.91 (± 0.68)	4.07 (± 0.34)	0.70 (± 0.02)
✗	m	Urban2Vec	6.66 (± 0.50)	3.98 (± 0.29)	0.72 (± 0.04)
✗	m	MuseCL	5.64 (± 0.46)	3.71 (± 0.22)	0.78 (± 0.03)
✓	m	CityFM	5.95 (± 0.48)	3.80 (± 0.18)	0.78 (± 0.04)
✓	m	UrbanCLIP	5.68 (± 0.44)	3.62 (± 0.17)	0.79 (± 0.03)
✓	m	**UrbanMFM**	**5.48** (± 0.42)	**3.46** (± 0.16)	**0.82** (± 0.03)

All baselines—except for HGI and our proposed model—are limited in their ability to capture the holistic spatial environment, leading to suboptimal performance. This limitation is particularly detrimental in real-world housing price estimation, where both the semantic information of nearby POIs and the structural characteristics of surrounding regions play a crucial role. Although HGI performs relatively well due to its region-level design, it still lacks the capacity to integrate complementary modalities. UrbanMFM overcomes these limitations by integrating foundation models with multiscale multimodal fusion, enabling it to model both fine-grained local semantics (individual POI functions) and macro-scale spatial dependencies (neighbourhood composition). This multiscale modeling framework allows UrbanMFM to more accurately capture the mutual influence between local features and regional context, which is essential for reliable housing price prediction.

3) GDP Density Estimation: Table VI showcases that **UrbanMFM** achieves the best overall performance on the GDP

TABLE VII: Building Functionality Statics and Experimental Comparison of F1 Scores Across Different Functionalities.

Functionality	Number	Percentage	GeoVectors	SpaBERT	CityFM	UrbanMFM
Residential	43,224	67.1%	76.97%	82.15%	**96.03%**	95.83%
Industrial	10,431	16.2%	70.57%	74.25%	94.65%	**94.79%**
Commercial	5,190	8.1%	88.22%	66.66%	87.07%	**88.34%**
Commercial & Residential	1,645	2.5%	21.08%	18.14%	46.29%	**63.86%**
Educational	1,427	2.2%	57.98%	46.05%	73.00%	**79.56%**
Civic & Community Institution	1,205	1.9%	13.82%	12.46%	24.92%	**39.94%**
Sports & Recreation	751	1.2%	31.73%	38.62%	69.76%	**78.55%**
Transport	511	0.8%	21.84%	17.09%	61.53%	**71.12%**

density prediction task, outperforming the strongest baseline, with RMSE reduction of **2.8%**, MAE reduction of **4.4%**, and R^2 improvement of **3.8%**.

MuseCL performs well by leveraging contrastive multimodal alignment between POI and visual features, enabling it to capture semantic relationships relevant to urban economic patterns. UrbanCLIP further improves overall consistency through large-scale multimodal pretraining, which enhances cross-modal representation quality. However, both models lack explicit mechanisms to model spatial hierarchies and contextual dependencies across different urban scales. UrbanMFM overcomes these limitations through a foundation-model-based multiscale architecture that jointly integrates POI, street-view, and satellite features. By capturing both micro-level functional semantics and macro-level spatial dependencies, it achieves a more comprehensive understanding of regional economic structures, leading to superior GDP density estimation.

4) Building Functionality Classification: We specifically selected GeoVectors, SpaBERT and CityFM for comparison, as they represent the best performance baseline models, with GeoVectors and SpaBERT excelling in unimodal and CityFM in multimodal foundation models, both derived from volunteered geospatial information sources. Table VII showcases the F1-scores of UrbanMFM and baselines for each building functionality category. The experimental results indicate that UrbanMFM not only maintains its outstanding performance in Residential, Industrial, and Commercial categories but also delivers substantial gains in challenging multifunctional categories (Commercial & Residential and Civic & Community Institution), with the highest improvement being **17.57%**. UrbanMFM excels in these difficult scenarios, substantially outperforming existing approaches.

C. Ablation Study

To understand the contributions of each scale within UrbanMFM, we conduct ablation studies on population density estimation task that examine the interactions between the various scales. This analysis provides insights into the influence of each module and verifies how well they work together. The comprehensive results in Table III indicate that our model has substantial enhancements owing to the incorporation of four scales. To delve into specifics, the local view emerges as a strong predictor, with the global view also having a significant impact, highlighting the importance of spatial context and multimodal interaction in urban understanding. The macro scale captures broad regional characteristics from satellite imagery that are crucial for predicting urban density on a macro scale. Lastly, the micro scale proves to be the weakest contributor, focusing on intra-modal contrastive learning between POI and street view imagery group. While valuable for capturing localized spatial details, it has a smaller overall impact compared to the broader contextual insights provided by the other scales.

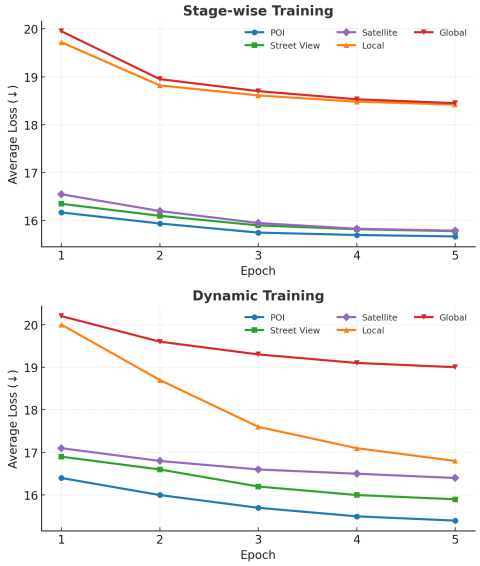


Fig. 4: Training loss of the five contrastive objectives.

TABLE VIII: Adaptability on Population Density Estimation.

Model	RMSE ↓	MAE ↓	R^2 ↑
UrbanMFM	3.88 (± 0.27)	2.64 (± 0.19)	0.87 (± 0.02)
UrbanMFM + trajectory	3.69 (± 0.29)	2.56 (± 0.23)	0.88 (± 0.02)

We compare the training loss between the stage-wise and dynamic strategies. As shown in Figure 4, stage-wise training achieves smooth and balanced convergence across all five contrastive objectives, showing no loss domination. In contrast, dynamic joint training causes the Local objective (orange) to converge faster, indicating an imbalance during optimization.

D. Adaptability Study

UrbanMFM is a flexible framework that seamlessly integrates diverse urban data sources. We conduct case studies using Singapore data to showcase the scalability of UrbanMFM to introduce additional datasets. Building on the original framework, we further integrate region-level features, trajectory data from Uber, on the Singapore dataset. This data

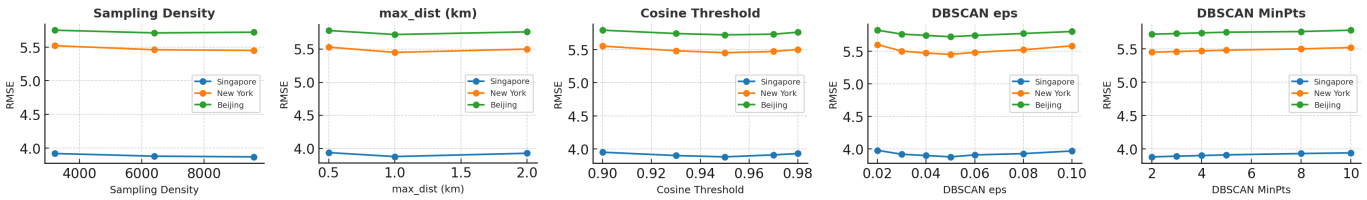
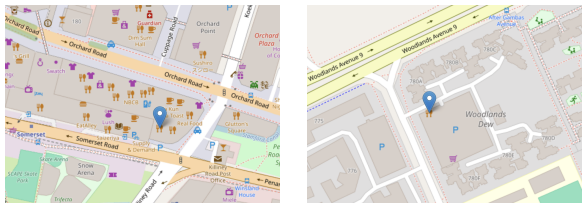


Fig. 5: Parameter sensitivity analysis of UrbanMFM across different datasets.



(a) Restaurant in Business Area. (b) Restaurant in Resident Area.

Fig. 6: Two real-world entities from different areas.

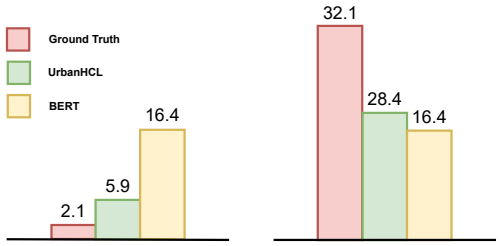


Fig. 7: Comparison of prediction results with ground truth.

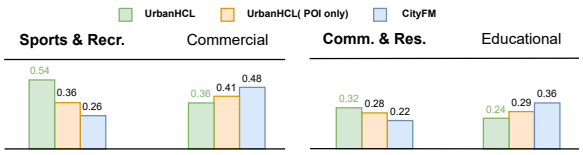


Fig. 8: Comparison of classification results.

source is incorporated into the Global Contrastive Objective during the training phase. As shown in Table VIII, UrbanMFM can seamlessly support new data inputs while continuing to improve performance, showcasing the framework's flexibility and adaptability to varying datasets.

E. Parameters Analysis

We evaluate the sensitivity of UrbanMFM to five key parameters: sampling density during training, maximum neighbour distance max_{dist} that controls local graph size, cosine similarity threshold that controls the number of satellite positive pairs, and DBSCAN settings ϵ and MinPts that control the neighbourhood radius and the minimum number of points required to form a cluster. As shown in Figure 5, our parameter configuration achieves the best overall performance. A too-small sampling density limits training stability, while an overly large one brings only marginal improvement at the cost of significantly higher computational expense. An overly small spatial range prevents the model from capturing adequate spatial context, whereas an excessively large range introduces noise. Cosine value threshold of 0.95 ensures semantic consistency without over-filtering, and the DBSCAN configuration balances regional granularity and spatial coherence.

F. Case Studies

To further illustrate the effectiveness of UrbanMFM, we present two representative case studies. The first case highlights the model's ability to capture spatial context in data-scarce environments, demonstrating its robustness through hierarchical representation learning. The second case focuses on the model's capacity to associate visual cues with functional semantics, showcasing the advantages of multimodal fusion.

1) Verify Ability To Capture Spatial Context: Figure 6 shows two real-world entities, described solely by the tag "amenity": "restaurant". Notably, these two POIs are the only ones within their respective regions in the OSM data. We respectively encode them using BERT and UrbanMFM, and report their corresponding predicted values in Figure 7. As shown, UrbanMFM demonstrates robustness in data-scarce or low-density environments, thanks to its hierarchical design. Even when POIs and street views are sparsely distributed, the inclusion of satellite imagery ensures comprehensive spatial coverage, allowing the framework to capture broader macro-level features.

2) Verifying Capability To Associate Visual Characteristics With Corresponding Functionality: We report the top 2 probability values of two real-world buildings' functionality calculated by UrbanMFM and CityFM (the best-perform baseline), with the ground truth highlighted in bold font. We evaluate the performance of both models under two scenarios: (1) when only POI data (incorporating both textual tags and polygons) is used, and (2) when UrbanMFM leverages a combination of POI data and street view imagery. In contrast, CityFM solely relies on POI data. As shown in Figure 8, even when both models are restricted to using only POI data, the merits of UrbanMFM are readily discernible. Furthermore, when UrbanMFM incorporates street view images into its input, its classification accuracy is further boosted. This superiority in performance underscores UrbanMFM's superior multimodal fusion capabilities and flexibility to incorporate a wider range of data sources.

VI. Conclusions and Future Work

This paper presents a novel problem – how to learn multiscale urban representations by integrating diverse geographic data modalities and modeling complex multimodal relationships across different spatial scales. To address this, we propose UrbanMFM, a scalable framework built upon a Spatial Graph-based Multi-scale Foundation Model, which explicitly models hierarchical spatial relationships among urban entities. By leveraging spatial graphs to encode the structural dependencies across micro, local, macro, and global urban scales, the framework effectively highlights complex interactions among multimodal data sources. This structured representation enables deep cross-modal fusion and contributes to the model's superior effectiveness in modeling diverse urban environments.

While UrbanMFM demonstrates strong performance across diverse urban tasks, several limitations remain. Coverage bias is an inherent issue in open-source urban datasets, as spatial coverage may vary across regions. We mitigate this limitation by integrating multiple complementary data sources to provide balanced multimodal coverage. Moreover, model outputs should be interpreted as analytical aids rather than prescriptive guidance for real-world planning or policy-making.

Looking ahead, extending UrbanMFM to incorporate temporal dynamics represents a promising direction. While this study intentionally focuses on spatial structures—reflecting the nature of many core urban tasks—numerous urban phenomena (e.g., mobility flows, POI evolution, and land-use transitions) exhibit inherent temporal patterns. Future work will explore spatio-temporal pretraining objectives and dedicated temporal modules, such as dynamic graph encoders or trajectory-informed temporal heads, to broaden the framework's applicability to dynamic prediction and forecasting tasks. Moreover, enhancing cross-city generalization offers another meaningful avenue. Given the substantial variations in POI semantics, imagery sources, and land-use conventions across countries, incorporating semantic harmonization strategies or domain adaptation techniques may further improve the transferability of UrbanMFM across heterogeneous urban environments.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its Frontier CRP Grant (NRF-F-CRP-2024-0005), and NTU SUG-NAP (022029-00001). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

References

[1] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke, "Using publicly available satellite imagery and deep learning to understand economic well-being in africa," *Nature communications*, vol. 11, no. 1, p. 2583, 2020.

[2] I. A. Pissourios, "Survey methodologies of urban land uses: An oddment of the past, or a gap in contemporary planning theory?" *Land Use Policy*, vol. 83, pp. 403–411, 2019.

[3] Z. Zhang, P. Balsebre, S. Luo, Z. Hai, and J. Huang, "Structam: Enhancing address matching through semantic understanding of structure-aware information," in *LREC-COLING*, 2024, pp. 15 350–15 361.

[4] M. Zhang, T. Li, Y. Li, and P. Hui, "Multi-view joint graph representation learning for urban region embedding," in *IJCAI*, 2021, pp. 4431–4437.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*. PMLR, 2020, pp. 1597–1607.

[6] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *TKDE*, vol. 35, no. 1, pp. 857–876, 2021.

[7] S. Han, D. Ahn, S. Park, J. Yang, S. Lee, J. Kim, H. Yang, S. Park, and M. Cha, "Learning to score economic development from satellite imagery," in *KDD*, 2020, pp. 2970–2979.

[8] N. Johnson, W. Treible, and D. Crispell, "Opensentinelmap: A large-scale land use dataset using openstreetmap and sentinel-2 imagery," in *CVPR*, 2022, pp. 1333–1341.

[9] W. Huang, D. Zhang, G. Mai, X. Guo, and L. Cui, "Learning urban region representations with pois and hierarchical graph infomax," *ISPRS*, vol. 196, pp. 134–145, 2023.

[10] J. Hu, C. Guo, B. Yang, and C. S. Jensen, "Stochastic weight completion for road networks using graph convolutional networks," in *ICDE*. IEEE, 2019, pp. 1274–1285.

[11] M.-x. Wang, W.-C. Lee, T.-y. Fu, and G. Yu, "Learning embeddings of intersections on road networks," in *SIGSPATIAL*, 2019, pp. 309–318.

[12] T. S. Jepsen, C. S. Jensen, and T. D. Nielsen, "Relational fusion networks: Graph convolutional networks for road networks," *T-ITS*, vol. 23, no. 1, pp. 418–429, 2020.

[13] H. Yuan, G. Li, Z. Bao, and L. Feng, "Effective travel time estimation: When historical trajectories over road networks matter," in *SIGMOD*, 2020, pp. 2135–2149.

[14] Y. Chen, X. Li, G. Cong, Z. Bao, C. Long, Y. Liu, A. K. Chandran, and R. Ellison, "Robust road network representation learning: When traffic patterns meet traveling semantics," in *CIKM*, 2021, pp. 211–220.

[15] J. Huang, H. Wang, Y. Sun, Y. Shi, Z. Huang, A. Zhuo, and S. Feng, "Ernie-geol: A geography-and-language pre-trained model and its applications in baidu maps," in *KDD*, 2022, pp. 3029–3039.

[16] R. Ding, B. Chen, P. Xie, F. Huang, X. Li, Q. Zhang, and Y. Xu, "Mgeo: Multi-modal geographic language model pre-training," in *SIGIR*, 2023, pp. 185–194.

[17] Z. Li, W. Zhou, Y.-Y. Chiang, and M. Chen, "Geolm: Empowering language models for geospatially grounded language understanding," *arXiv preprint arXiv:2310.14478*, 2023.

[18] C. Deng, T. Zhang, Z. He, Q. Chen, Y. Shi, Y. Xu, L. Fu, W. Zhang, X. Wang, C. Zhou *et al.*, "K2: A foundation language model for geoscience knowledge understanding and utilization," in *WSDM*, 2024, pp. 161–170.

[19] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang *et al.*, "Ringmo: A remote sensing foundation model with masked image modeling," *TGRS*, vol. 61, pp. 1–22, 2022.

[20] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell, "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *ICCV*, 2023, pp. 4088–4099.

[21] F. Yao, W. Lu, H. Yang, L. Xu, C. Liu, L. Hu, H. Yu, N. Liu, C. Deng, D. Tang *et al.*, "Ringmo-sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling," *TGRS*, 2023.

[22] G. Mai, N. Lao, Y. He, J. Song, and S. Ermon, "Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations," in *ICML*. PMLR, 2023, pp. 23 498–23 515.

[23] X. Li, K. Zhao, G. Cong, C. S. Jensen, and W. Wei, "Deep representation learning for trajectory similarity computation," in *ICDE*. IEEE, 2018, pp. 617–628.

[24] D. Yao, C. Zhang, Z. Zhu, J. Huang, and J. Bi, "Trajectory clustering via deep representation learning," in *IJCNN*. IEEE, 2017, pp. 3880–3887.

[25] J. Jiang, D. Pan, H. Ren, X. Jiang, C. Li, and J. Wang, "Self-supervised trajectory representation learning with temporal regularities and travel semantics," in *ICDE*. IEEhuang2023learningE, 2023, pp. 843–855.

[26] F. Zhou, Y. Dai, Q. Gao, P. Wang, and T. Zhong, "Self-supervised human mobility learning for next location prediction and trajectory classification," *Knowledge-Based Systems*, vol. 228, p. 107214, 2021.

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

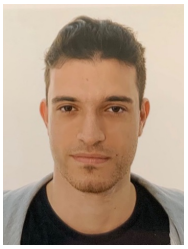
[28] P. Balsebre, W. Huang, G. Cong, and Y. Li, "City foundation models for learning general purpose representations from openstreetmap," in *CIKM*, 2024, pp. 87–97.

[29] W. Wang, S. Yang, Z. He, M. Wang, J. Zhang, and W. Zhang, "Urban perception of commercial activeness from satellite images and streetscapes," in *WWW*, 2018, pp. 647–654.

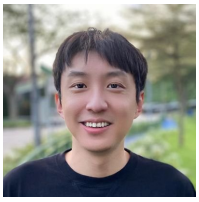
[30] J. Lee, D. Grosz, B. Uzkent, S. Zeng, M. Burke, D. Lobell, and S. Ermon, "Predicting livelihood indicators from community-generated street-level imagery," in *AAAI*, vol. 35, no. 1, 2021, pp. 268–276.
[31] S. Wu, X. Yan, X. Fan, S. Pan, S. Zhu, C. Zheng, M. Cheng, and C. Wang, "Multi-graph fusion networks for urban region embedding," *arXiv preprint arXiv:2201.09760*, 2022.
[32] W. Chan and Q. Ren, "Region-wise attentive multi-view representation learning for urban region embedding," in *CIKM*, 2023, pp. 3763–3767.
[33] Z. Wang, H. Li, and R. Rajagopal, "Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding," in *AAAI*, vol. 34, no. 01, 2020, pp. 1013–1020.
[34] S. Park, S. Han, D. Ahn, J. Kim, J. Yang, S. Lee, S. Hong, J. Kim, S. Park, H. Yang *et al.*, "Learning economic indicators by aggregating multi-level geospatial information," in *AAAI*, vol. 36, no. 11, 2022, pp. 12 053–12 061.
[35] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *TGRS*, vol. 59, no. 3, pp. 2598–2610, 2020.
[36] T. Li, Y. Xi, H. Wang, Y. Li, S. Tarkoma, and P. Hui, "Learning representations of satellite imagery by leveraging point-of-interests," *TIST*, vol. 14, no. 4, pp. 1–32, 2023.
[37] T. Li, S. Xin, Y. Xi, S. Tarkoma, P. Hui, and Y. Li, "Predicting multi-level socioeconomic indicators from structural urban imagery," in *CIKM*, 2022, pp. 3282–3291.
[38] Y. Xi, T. Li, H. Wang, Y. Li, S. Tarkoma, and P. Hui, "Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests," in *WWW*, 2022, pp. 3308–3316.
[39] Y. Liu, X. Zhang, J. Ding, Y. Xi, and Y. Li, "Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction," in *WWW*, 2023, pp. 4150–4160.
[40] V. Vivanco Cepeda, G. K. Nayak, and M. Shah, "Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization," *NeurIPS*, vol. 36, pp. 8690–8701, 2023.
[41] Z. Li, W. Huang, K. Zhao, M. Yang, Y. Gong, and M. Chen, "Urban region embedding via multi-view contrastive prediction," in *AAAI*, vol. 38, no. 8, 2024, pp. 8724–8732.
[42] X. Yong and X. Zhou, "Musecl: Predicting urban socioeconomic indicators via multi-semantic contrastive learning," *arXiv preprint arXiv:2407.09523*, 2024.
[43] R. Manvi, S. Khanna, G. Mai, M. Burke, D. Lobell, and S. Ermon, "Geollm: Extracting geospatial knowledge from large language models," *arXiv preprint arXiv:2310.06213*, 2023.
[44] J. He, T. Nie, and W. Ma, "Geolocation representation from large language models are generic enhancers for spatio-temporal learning," in *AAAI*, vol. 39, no. 16, 2025, pp. 17 094–17 104.
[45] C. Xiao, J. Zhou, Y. Xiao, J. Huang, and H. Xiong, "Refound: Crafting a foundation model for urban region understanding upon language and visual foundations," in *KDD*, 2024, pp. 3527–3538.
[46] X. Hao, W. Chen, X. Zou, and Y. Liang, "Nature makes no leaps: Building continuous location embeddings with satellite imagery from the web," in *WWW*, 2025, pp. 2799–2812.
[47] J. Feng, S. Wang, T. Liu, Y. Xi, and Y. Li, "Urbanllava: A multi-modal large language model for urban intelligence with spatial reasoning and understanding," *arXiv preprint arXiv:2506.23219*, 2025.
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[49] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[50] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.
[51] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" *arXiv preprint arXiv:2105.14491*, 2021.
[52] G. Giannopoulos, K. Alexis, N. Kostagiolas, and D. Skoutas, "Classifying points of interest with minimum metadata," in *SIGSPATIAL Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising*, 2019, pp. 1–4.
[53] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
[54] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.
[55] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.
[56] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, "Tile2vec: Unsupervised representation learning for spatially distributed data," in *AAAI*, vol. 33, no. 01, 2019, pp. 3967–3974.
[57] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," in *ICCV*, 2021, pp. 10 181–10 190.
[58] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *CVPR*, 2021, pp. 9414–9423.
[59] N. Tempelmeier, S. Gottschalk, and E. Demidova, "Geovectors: a linked open corpus of openstreetmap embeddings on world scale," in *CIKM*, 2021, pp. 4604–4612.
[60] Z. Li, J. Kim, Y.-Y. Chiang, and M. Chen, "Spabert: A pretrained language model from geographic data for geo-entity representation," *arXiv preprint arXiv:2210.12213*, 2022.

Zhaoqi Zhang received a Bachelor's degree from Dalian Technological University in 2017. He is currently working toward the PhD degree in the College of Computing and Data Science, Nanyang Technological University. His current research interests include spatial-temporal data mining, geospatial foundation models, and causal inference for urban data.

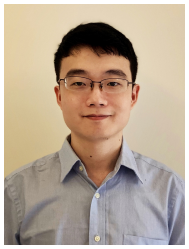
Miao Xie received the joint PhD degree from the Chinese Academy of Sciences and Nanyang Technological University in 2016. From 2016 to 2024, he was a senior algorithm expert and business lead with Huawei, Alibaba, and Kuaishou. He is currently a professor with the College of Information and Electrical Engineering, China Agricultural University, Beijing, China. His research interests include spatial-temporal data mining and recommendation systems.



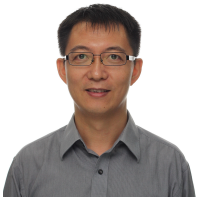
Pasquale Balsebre is currently a NLP Research Engineer in Bosch, Tokyo, Japan. He received Bachelor's and Master's degrees from Politecnico di Torino, in 2018 and 2020, respectively. He received the PhD degree in Nanyang Technological University in 2024. His current research interests include spatial-temporal data mining and data management.



Weiming Huang received the PhD degree in geographical information science from Lund University, Sweden. He was a Wallenberg postdoctoral fellow with Nanyang Technological University and Lund University. He is currently a lecturer with the School of Geography, University of Leeds, U.K. His research interests include spatial data mining and geospatial foundation models.



Siqiang Luo (Member, IEEE) received the PhD degree in computer science from the University of Hong Kong in 2019. He is currently an assistant professor with the College of Computing and Data Science, Nanyang Technological University. He was a postdoc at Harvard University from 2019 to 2020. His research interest includes scalable data structures and systems, as well as graph analytics and mining.



Gao Cong (Member, IEEE) is currently a Professor in the College of Computing and Data Science at Nanyang Technological University (NTU). He previously worked at Aalborg University, Denmark, Microsoft Research Asia, and the University of Edinburgh. He received his PhD degree from the National University of Singapore in 2004. His current research interests include spatial data management, ML4DB, spatial-temporal data mining, and recommendation systems.