ISPRS Journal of Photogrammetry and Remote Sensing 234 (2026) 185–204

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Heterogeneous graph neural networks for building attribute prediction from hierarchical urban features and cross-view imagery

Xiucheng Liang [a], Winston Yap [a], Filip Biljecki [a,b,*]

[a] Department of Architecture, National University of Singapore, Singapore
[b] Department of Real Estate, National University of Singapore, Singapore

ARTICLE INFO

Keywords:
Building semantics
Building function
Volunteered geographic information
Crowdsourced data
Multi-modal

ABSTRACT

Data on building properties are essential for a variety of urban applications, yet such information remains scarce in many parts of the world. Recent efforts have leveraged instruments such as machine learning (ML), computer vision (CV), and graph neural networks (GNNs) to assess these properties at scale by leveraging urban features or visual information. However, extracting holistic representations to infer building attributes from multi-modal data across multiple spatial scales and vertical building characteristics remains a significant challenge. To bridge this gap, we present a innovative framework, that captures both hierarchical urban features and cross-view visual information through a heterogeneous graph. First, we construct a heterogeneous graph that incorporates multi-dimensional urban elements — buildings, streets, intersections, and urban plots — to comprehensively represent multi-scale geospatial features. Second, we automatically crop images of individual buildings from both very high-resolution satellite and street-level imagery, and introduce feature propagation on semantic similarity graphs to supplement missing facade information. Third, feature fusion is applied to integrate both morphological and visual features, with holistic representations generated for building attribute prediction. Systematic experiments across three global cities demonstrate that our method outperforms existing CV, ML, and homogeneous GNN-based models, achieving classification accuracies of 86% to 96% across 10 to 12 distinct building types, with mean F1 scores ranging from 0.70 to 0.73. The framework demonstrates robustness to class imbalance and produces more distinctive embeddings for ambiguous categories. In additional task of inferring building age, the method delivers similarly strong performance. This framework advances scalable approaches for filling gaps in building attribute data and offers new insights into modeling holistic urban environments. Our dataset and code are available openly at: https://github.com/seshing/HeteroGNN-building-attribute-prediction.

1. Introduction

Buildings are the dominant components of the urban environment. They define the form of cities and play a crucial role in shaping their social, environmental, and economic qualities and sustainability (Biljecki et al., 2021). Attributes such as building type, age, and number of floors are therefore central to a wide range of applications, including energy modeling (Kumar et al., 2018; Roth et al., 2020), climate adaptation (Creutzig et al., 2019), and disaster impact assessments (Westrope et al., 2014). Despite their importance, reliable information on buildings is often scarce, fragmented, or inconsistent across cities (Biljecki et al., 2023; Lei et al., 2023; Herfort et al., 2023). Many regions lack detailed records, and global coverage of open datasets remains limited. Closing these data gaps would enable more comprehensive and openly accessible geospatial resources, which would not only facilitate downstream applications, but also support nuanced analyses of population distribution (Schug et al., 2021), socio-economic conditions (Feldmeyer et al., 2020), and urban resilience and sustainability (Elmqvist et al., 2019).

In recent years, machine learning has emerged as a promising tool to enrich building datasets, with two main streams of data modalities widely explored. The first relies on urban features, such as descriptors of geometry, topology, and context that capture how buildings are situated in the built environment (Lu et al., 2014; Wurm et al., 2016; Tooke et al., 2014; Biljecki and Sindram, 2017; Rosser et al., 2019; Nachtigall et al., 2023). Such features offer interpretable signals and can be computed directly from geospatial data. The second stream leverages visual information from imagery based on computer vision models. Remote sensing data, particularly very high-resolution (VHR)

* Corresponding author at: Department of Architecture, National University of Singapore, Singapore.
E-mail address: filip@nus.edu.sg (F. Biljecki).

https://doi.org/10.1016/j.isprsjprs.2026.02.016
Received 21 September 2025; Received in revised form 7 January 2026; Accepted 9 February 2026

satellite images, provide wide coverage and increasingly fine detail to estimate height (Wu et al., 2023; Frantz et al., 2021; Florio et al., 2025; Zhu et al., 2025), or classify roof structure (Zhao et al., 2022) and building functions (He et al., 2024). Street view imagery (SVI), in contrast, offers a pedestrian perspective and captures facade-level cues, enabling inference of attributes like use type, architectural style, and materials (Kang et al., 2018; Zhao et al., 2021; Lindenthal and Johnson, 2025; Ramalingam and Kumar, 2023; Li et al., 2025c).

More recently, graph-based methods that model the spatial and relational structure of cities has gained momentum. Unlike traditional approaches that treat buildings as independent samples, graph neural networks (GNNs) operate on networks of buildings, streets or different spatial units, capturing the dependencies and interactions across objects (Liu and Biljecki, 2022; Zhu and Ma, 2025). It offers a principled way to account for objects' relations and have shown promise in predicting missing attributes by propagating information across connected entities (Yan et al., 2019; Zhang et al., 2023; Lei et al., 2024; Wang et al., 2024; Yap et al., 2025; Wang et al., 2025). They extend beyond local features and allow urban structure itself to become a source of predictive power.

Despite progress across these streams, several challenges remain. First, many methods still reduce information to a single, building-level representation, overlooking the hierarchical nature of urban systems and limiting the ability to capture cross-scale context. Second, while both urban features and visual cues have been studied extensively, relatively few frameworks systematically integrate them. Recent work has combined building features with satellite imagery (Wang et al., 2024; Yap et al., 2025), but street-level information — despite its clear potential to capture vertical details — remains largely unexplored at scale. Third, street-level data coverage is uneven that may be missing or partially obstructed (Hou and Biljecki, 2022), which poses constraints when applied to research at the urban scale. These gaps highlight the need for methods that can model urban hierarchies, fuse multiple modalities, and handle incomplete visual information.

To address these challenges, we propose a hierarchical, multi-modal graph neural network framework for predicting building attributes. First, our approach begins by constructing a heterogeneous urban graph that represents buildings, streets, intersections, and urban plots, and encodes with their morphological, topological, functional, socio-demographic, and environmental features. We then embed these urban features into a semantic space using a graph encoder to build a semantic similarity graph, capturing potentially similar structures in urban environment. In parallel, we extract visual features from VHR satellite imagery and street-level imagery, bringing together complementary top–down and facade perspectives. To overcome incomplete coverage of street-level images, we propagate facade features across similar nodes in the semantic similarity graph, ensuring that vertical cues are available even where images are missing. Finally, feature fusion is conducted to integrate these multi-modal embeddings within a heterogeneous GNN that performs relation-aware message passing and outputs predictions for target building attributes.

The primary contributions of this work are threefold:

- We introduce an innovative framework that models the urban environment as a heterogeneous graph, capturing hierarchical relations between buildings and their surrounding elements. This approach outperforms traditional homogeneous graph frameworks that aggregate features only at the building level.
- We propose a systematic method to integrate urban features with both satellite and street-level imagery, supplemented by feature propagation to address incomplete coverage. With this design, our model demonstrates robust and balanced performance in predicting building use types, compared to conventional machine learning approaches and computer vision approach that primarily rely on single modality.
- We validate our approach across multiple global cities and additional building attribute, showing that visual cues substantially enhance building attribute prediction and that our method remains robust across contexts and tasks. By combining the interpretability of urban indicators, the richness of cross-view imagery, and the relational power of GNNs, our work advances scalable approaches for enriching building databases and creates new opportunities for urban analytics in data-scarce environments.

2. Related work

2.1. Predicting building attributes

Acquiring information on buildings, such as usage, construction period, and number of storeys, plays a crucial role across various domains in urban research, including energy modeling (Kumar et al., 2018; Xu et al., 2019; Roth et al., 2020), disaster risk management (Westrope et al., 2014), and environmental planning (Creutzig et al., 2019). Over the past decade, a wide range of studies has focused on inferring these attributes by leveraging morphological features and visual information as key data sources in analytical frameworks.

Urban features. Early studies demonstrated that descriptors of the built form carry substantial predictive power. These indicators typically include geometric properties, such as area, compactness, elongation, or perimeter complexity, as well as contextual urban metrics like block density or spatial configuration. Such features offer indirect yet interpretable proxies for inferring building characteristics like building type (Lu et al., 2014; Wurm et al., 2016) and age (Tooke et al., 2014; Biljecki and Sindram, 2017; Rosser et al., 2019; Nachtigall et al., 2023). For example, Wurm et al. (2016) use a linear discriminant analysis framework to classify building types by analyzing the discriminatory power of a set of 1D (e.g., length), 2D (e.g., area), and 3D (e.g., volume) shape-based features derived from digital building models. More recently, Milojevic-Dupont et al. (2020) broadened this idea by compiling 152 urban-form metrics that incorporate interactions between buildings, streets, and urban blocks, thereby capturing multi-scale effects that single-building predictors miss. Nachtigall et al. (2023) further leverage this framework to infer age for buildings constructed after 1900 by developing 119 feature metrics. Furthermore, Biljecki et al. (2017) demonstrate that adding census features (e.g., average household size) on top of morphological identities enhances the prediction of building height. These studies laid the groundwork for attributing buildings by urban morphological features and urban context without visual information.

Visual information. The proliferation of high-resolution imagery and advances in deep learning have enabled the extraction of visual semantics for building classification. Two principal image sources dominate this space: remote sensing and street-level imagery.

Remote sensing imagery, with its broad coverage and ever-improving resolution, enables large-scale characterization of urban morphology and land cover (Li et al., 2022a). Aerial and satellite data are now routinely used to extract building attributes, including height (Frantz et al., 2021; Wu et al., 2023; Zhu et al., 2025) and functional class (Du et al., 2015; Zhao et al., 2019; Florio et al., 2025). Specifically, Du et al. (2015) use VHR imagery and GIS data to classify urban buildings into seven semantic categories. They achieve this by integrating a two-level segmentation mechanism to obtain spectral, texture, geometric, and distribution features with improved Random Forest classifier. A recent work by He et al. (2024) introduces a super resolution method to enhance satellite imagery for the fine-grained classification of 11 different building types in districts in Hong Kong.

While satellite views offer top–down observations, they inherently lack vertical detail of building facades. To overcome this limitation, recent efforts have focused on utilizing street-level imagery, which

provides a pedestrian-scale view of urban environments (Biljecki and Ito, 2021). SVI has shown promise in inferring building typology (Kang et al., 2018), materials (Ghione et al., 2022; Raghu et al., 2023), and facade style (Lindenthal and Johnson, 2025; Sun et al., 2022; Ogawa et al., 2023) from direct observation. For instance, CNN-based classifiers have been trained to distinguish buildings across architectural periods (Sun et al., 2022) and to categorize usage types (e.g., apartment, office, or retail) (Kang et al., 2018) through visual cues embedded in facades.

Despite the demonstrated effectiveness of morphological and image-based methods, urban environments are inherently complex, with diverse spatial structures and irregular patterns that challenge traditional methods. Buildings do not exist in isolation. They are embedded within a spatial network of streets, plots, and surrounding urban elements (e.g., vegetation, infrastructure). Capturing these interdependencies is critical for nuanced understanding and robust inference of their characteristics, yet remains underexplored in many attribute inference frameworks.

2.2. Graph neural network in urban modeling

Graph neural networks (GNNs) have emerged as a powerful class of models capable for addressing the spatial complexity of urban systems (Liu and Biljecki, 2022). By operating on graph-structured data, GNNs can incorporate both node-level features (e.g., individual building characteristics) and the relational structure of the urban fabric (e.g., street connectivity, proximity to amenities). This makes them particularly suited for urban modeling tasks where spatial interactions matter. Recent studies have demonstrated the potential of GNNs in predicting urban characteristics by leveraging both topological context and feature propagation across spatial entities (Yan et al., 2019; Liu and Biljecki, 2022; Zhang et al., 2023; Wang et al., 2025; Liu et al., 2025a). For instance, by formalizing urban places as nodes and their spatial relationships as edges in a graph, Zhu et al. (2020) use embedded urban visual features from SVI to predict place characteristics in Beijing, and De Sabbata and Liu (2023) leverage census variables to infer geodemographic classifications in Greater London. Another application developed by Zhang et al. (2023) embeds street scene descriptions as node features and combines them with road network topology in a GNN to predict urban functions. This two-layer GNN approach effectively captures both semantic content and spatial dependencies using minimal labeled data. GNNs provide a compelling pathway toward more holistic and spatially-aware urban analytical frameworks.

For building-related research, GNNs have also been increasingly adopted to model spatial relationships (Xu et al., 2022). For example, Yan et al. (2019) model building groups as graphs by using building-level geometric and semantic indices as node features to learn spatial patterns. Tested on large datasets from Guangzhou and Shanghai, the method effectively classifies regular and irregular building patterns, outperforming traditional machine learning methods. Similarly, Lei et al. (2024) construct building graphs and apply GraphSAGE to infer missing attributes such as building storeys and type from surrounding OpenStreetMap (OSM) objects (e.g., points of interest, transport facilities). Moving beyond contextual urban features, Kong et al. (2024) incorporate visual features from SVI (e.g., building view index, sky view factor) and spectral features (i.e., the mean and standard deviation of red-band pixel values within each building) to classify buildings into seven functional categories in Shenzhen. Recent work by Wang et al. (2024) proposes a multi-view GNN framework that captures multi-scale spatial relations and integrates topological context with top–down visual embedding from satellite imagery to estimate building age in Beijing. Other research also leverages the spatial dependence and spatial autocorrelation of GNNs for fine-grained data prediction, with applications including predicting human activity intensity (Wang and Zhu, 2024), downscaling Local Climate Zones (Li et al., 2025b), and modeling time-series energy consumption (Hu et al., 2022).

However, there are key challenges remain. Most existing studies aggregate features and model relationships at a single, building-centric level, which overlooks the inherently hierarchical nature of urban environments and limits the ability to capture holistic context and cross-scale interactions. Furthermore, while data sources such as SVI offer rich vertical visual information of buildings, integrating them into a city-scale GNN framework is challenging due to incomplete coverage, occlusions, and variable image quality. Hence, our research introduces a hierarchical, multi-modal GNN framework that incorporates a street-level information supplementation process to model urban environments and improve building attribute prediction.

3. Methodology

3.1. Urban heterogeneous graph

Urban systems exhibit complex, hierarchical structures that emerge from interactions, feedbacks, and scaling processes across multiple levels of urban form and function (Batty, 2009). To model the spatial relationships between buildings and their surrounding context, we represent the urban environment as a heterogeneous graph with four node types: buildings, street segments, intersections, and urban plots. This extends the traditional homogeneous GNN approach, where nodes typically represent only buildings, by incorporating additional urban element types and defining edges based on their spatial relationships. Urbanity, a network and graph based Python package developed by Yap et al. (2023), is leveraged to construct the urban heterogeneous graph. Fig. 1 illustrates the construction process, which involves three main stages: node generation, edge construction and data integration.

Node generation. To represent a complex urban system in graph space, we construct an undirected heterogeneous graph $G(V, E)$, where nodes V are generated from both building geometry and street network topology derived from OSM. Four distinct node types are considered: building nodes V_B, street segment nodes V_S, intersection nodes V_I and urban plot nodes V_U.

Building nodes V_B represent individual building footprints and are generated by processing raw OSM building geometry. The preprocessing pipeline first ensures topological validity by removing non-polygonal geometries, decomposing MultiPolygons, and discarding invalid polygons. The geometries are locally projected to an appropriate planar coordinate system to enable accurate metric calculations. Buildings with areas smaller than a predefined threshold (30 m^2 in our implementation) are excluded. This threshold is empirically selected to mitigate noise from minor polygons (e.g., sheds, temporary structures, or mapping artifacts) while preserving valid building footprints. The centroid of each retained footprint is taken as the spatial anchor point for the corresponding node.

Street nodes V_S and intersection nodes V_I represent the road segments in the OSM-derived street network. The network's vertices and edges are first obtained using the `pyrosm` library. We then process the network by simplifying intermediate nodes between intersections and removing unconnected subgraphs and self-loops. Each street segment is assigned a unique identifier, and its midpoint geometry is used as the node location. Intersection nodes correspond to topological junctions in the processed network and are identified from the simplified graph as nodes connected to three or more street segments. Their spatial coordinates are taken directly from the network geometry.

Urban plot nodes V_U represent contiguous land parcels enclosed by the street network, analogous to the enclosures or enclosed tessellation cells described in morphological urban studies (Fleischmann and Arribas-Bel, 2022; Tang et al., 2025). By defining polygons bounded by physical barriers, urban plots capture the spatial structure of city

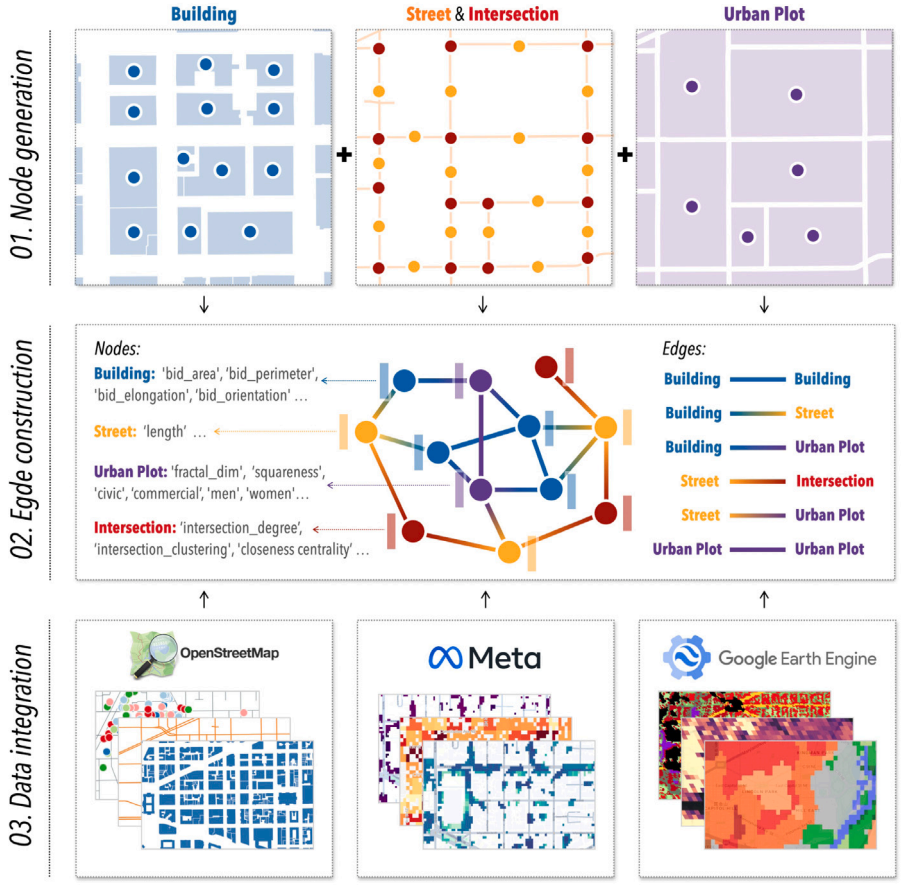

Fig. 1. Overall research framework for constructing the heterogeneous urban graph, including node generation, edge construction, and data integration with four element types (buildings, street segments, intersections, and urban plots).

blocks, which are closely linked to urban form, functions, and human activity patterns. They also provide an intermediate spatial scale between individual buildings and administrative boundaries, allowing features to be analyzed within a broader spatial context while remaining sufficiently granular to represent neighborhoods containing clusters of buildings. Here, we generate these by polygonizing the processed street network within the study area boundary. Specifically, street edges inside the buffered boundary are merged and polygonized, yielding contiguous land parcels. Each polygon is assigned a unique plot identifier and enriched with geometric attributes such as area and perimeter. A minimum area threshold is again applied to remove small samples.

Edge construction. Following the definition of the heterogeneous node set $V = \{V_B, V_S, V_I, V_U\}$, we construct a set of typed edges E that encode spatial, morphological, and topological relationships between node types. All relations are modeled as bidirectional connections to support information exchange in both directions during graph deep learning.

- Building–building proximity (E_{BB}): Edges between building nodes are established based on spatial proximity, connecting each building to its nearest neighbors within the local urban fabric. Here the number of nearest neighbors is set to 5.
- Building–street proximity (E_{BS}): Buildings are connected to their nearest street segments, representing immediate frontage or direct access.
- Building–plot containment (E_{BP}): Each building node is connected to the urban plot polygon that contains it. This hierarchical link integrates building-level information with the broader spatial context provided by the plot, allowing the model to capture parcel-level influences on individual buildings.
- Street-intersection incidence (E_{SI}): Street segment nodes are linked to intersection nodes at their endpoints, reflecting the topological structure of the street network.
- Street–plot adjacency (E_{SP}): Urban plots are connected to the street segments that form their boundaries, to capture the degree of exposure and accessibility of each plot to the surrounding street network.
- Plot–plot adjacency (E_{PP}): Urban plots that share a common boundary segment are connected to form a neighborhood graph. This relation reflects morphological adjacency between parcels, enabling the modeling of spatial diffusion processes and the influence of neighboring land parcels.

Data integration. Together with the graph construction process, an extensive set of urban indicators is computed for each node type based on established literature, encompassing morphological, topological, functional, socio-demographic, and environmental characteristics of the urban environment. Tables 1 and 2 contain the detail breakdown of urban features computed to relevant nodes.

For morphological indicators, metrics are calculated and assigned to building nodes and urban plot nodes, including measures such as area, convexity, orientation, elongation, and shape indices. These metrics capture the geometric form and spatial configuration of buildings as well as the characteristics of the urban plots they occupy (Basaraner and Cetinkaya, 2017; Dibble et al., 2019; Yan et al., 2019; Nachtigall et al., 2023; Yap and Biljecki, 2023; Kong et al., 2024). For topological indicators, network-based measures such as degree, clustering coefficient, betweenness centrality, closeness centrality, and PageRank are

Table 1
Morphological and topological indicators with units and descriptions.

Indicator	Unit	Description
Morphological (Building & Urban plot)		
Area	m^2	Total surface area covered by the footprint.
Perimeter	m	Total boundary length of the footprint.
Complexity	–	Degree of boundary irregularity; higher values indicate more intricate edges.
Circular Compactness	–	How closely the footprint shape approaches a perfect circle.
Convexity	–	Ratio of footprint area to its convex hull; measures concavity.
Rectangularity	–	How well the footprint fills its minimum bounding rectangle.
Squareness	degrees	Average deviation of building corners from right angles.
Square Compactness	–	Similar to compactness but benchmarked to a square.
Shape Index	–	Degree to which shape deviates from a compact form; normalized by footprint size.
Elongation	–	Ratio of the shortest to the longest side of the footprint's bounding rectangle.
Orientation	degrees	Dominant alignment of the footprint's longest axis.
Longest Axis Length	m	Maximum span across the footprint.
Fractal Dimension	–	Complexity of the shape boundary across scales.
Equivalent Rectangular Index	–	Fit of footprint to an equivalent rectangle in both area and perimeter.
No. of Corners	count	Count of significant directional changes along the footprint boundary.
Geometric/Topological (Street or Intersection)		
Street Length	m	Length of a street segment.
Degree	count	Number of directly connected edges at a node.
Clustering Coefficient	–	Tendency of a node's neighbors to be connected with each other.
Weighted Clustering Coefficient	–	Clustering tendency considering connection strengths.
Closeness Centrality	–	Accessibility of a node to all others in the network.
Betweenness Centrality	–	Frequency of a node lying on shortest paths between other nodes.
Eigenvector Centrality	–	Importance of a node based on the importance of its neighbors.
Katz Centrality	–	Influence of a node considering all paths with distance-based decay.
PageRank	–	Relative importance of a node in terms of link structure and probability flow.

Table 2
Functional, socio-demographic and environmental indicators computed as node features in this study.

Indicator	Unit	Subcategories	Description
Environmental indicators (Street)			
Street-level Features	% or index	Green View, Sky View, Building View, Road View, Visual Complexity	Proportion or index values of visible elements from street-level imagery, capturing greenery, sky openness, built-up surfaces, and overall visual diversity.
Functional, socio-demographic and environmental indicators (Urban plot)			
Amenities Count	count	Social, Recreational, Healthcare, Entertainment, Civic, Institutional, Food, Commercial	Number of amenities within a defined spatial unit, categorized by their primary function.
Population	count	Total population, Women, Men, Elderly (aged 60+), Youth (15–24), Children (under 5)	Demographic composition of residents, segmented by age group and gender.
Tree Canopy Height	m	–	Average or maximum tree canopy height within the area.

computed from the street network. These indices quantify the connectivity, accessibility, and relative importance of streets or intersections within the urban fabric (Kirkley et al., 2018; Ozuduru et al., 2021; Xue et al., 2022; Prieto-Curiel et al., 2022; Boeing, 2022; Jia et al., 2019). For functional characteristics, amenity counts are computed for categories including social, recreational, healthcare, entertainment, civic, institutional, food, and commercial services based on points of interest (POI) from OSM. This captures the distribution and diversity of urban functions across the network. Population data is derived from Meta High Resolution Population Density dataset and includes population counts disaggregated by gender and age group (e.g., elderly, youth, children). These indicators describe the demographic composition and potential service needs of local communities. Lastly, environmental indicators are generated using data from multiple sources, including Google Earth Engine[1] and Mapillary[2] (Yap and Biljecki, 2023).

[1] https://earthengine.google.com/
[2] https://www.mapillary.com

3.2. Holistic building representation

To construct a holistic representation of each building, we integrate information from multiple data modalities and spatial contexts. This representation encapsulates the building's intrinsic characteristics, visual appearance, immediate spatial surroundings, and broader position within the urban network. Given the limited coverage of visible buildings in cities (Fan et al., 2025), we introduce a similarity-graph-based approach to propagate building features from structurally similar buildings, thereby mitigating data sparsity. Our approach rests on the assumption that semantically similar nodes are potentially exhibit similar visual features. Building-level urban features are first embedded into a homogeneous graph, where visual information extracted from street-level imagery can be propagated to support property prediction. As illustrated in Fig. 2, the process consists of four major steps: semantic similarity graph generation, deep visual information extraction, vertical feature propagation, and building attribute prediction.

Semantic similarity graph. The first step constructs a semantic similarity graph among building nodes to capture latent relationships that go beyond purely geographic proximity. Following prior work that

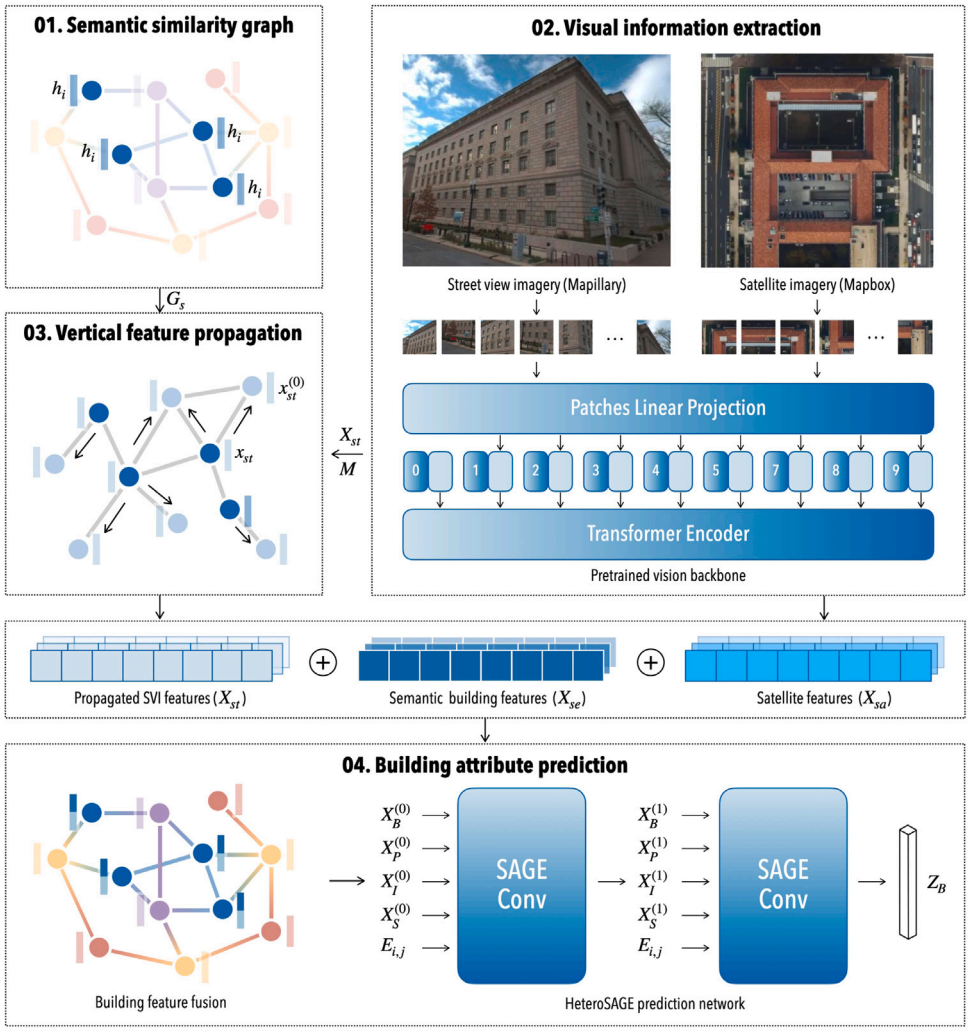

Fig. 2. Illustration of the process for generating holistic building representations to support building attribute prediction. Data: (c) Mapbox, (c) Mapillary contributors.

incorporates semantic structures in addition to spatial adjacency (Wang et al., 2024, 2020), we generate building representations by encoding urban features using the heterogeneous graph described in Section 3.1. Specifically, a two–layer heterogeneous GraphSAGE encoder f_θ (with mean aggregation across relations) maps multi-type urban features into a shared embedding space:

$$h_i = f_\theta(X, G)_i \in R^d, \tag{1}$$

where X denotes the node attributes on the heterogeneous urban graph G, and h_i is the d-dimensional embedding of building i returned by the encoder. These embeddings place buildings with similar form, function, and built environment closer together in a latent feature space. Once the building embeddings are obtained, we connect each building to its k most similar peers using a Gaussian kernel similarity measure:

$$S_{ij} = \exp\left(-\frac{\|h_i - h_j\|^2}{2\sigma^2}\right), \tag{2}$$

where h_i and h_j are the embeddings of buildings i and j, and σ controls the scaling of distances. This process produces a semantic graph $G_s = (V_B, E_s)$, where edges link buildings that are similar, to facilitate propagation between buildings that share meaningful traits but may not be connected in the spatial network.

Visual information extraction. To capture complementary perspectives of the building, we extract visual features from both VHR satellite and street-level imagery. The satellite view provides top–down context, while the street view captures facade details. VHR imagery is clipped using masks generated from building footprints, ensuring that each image corresponds to an individual building. A 10 m buffer is applied to each mask to capture both the building and its immediate surroundings, including extended structures and adjacent environmental context. For street-level imagery of buildings, we leverage the open-source tool OpenFACADES (Liang et al., 2025), which enables the retrieval of individual building images and their association with geospatial locations. For buildings with multiple images, the one with the highest visual completeness (i.e., the largest visible proportion from the observation point) is retained. Fig. 3 provides examples obtained from both data sources.

Then, we employ a pretrained vision backbone F, remove the classification head, and append a linear projection so that F outputs a d-dimensional embedding. The model produces embeddings of both SVI (X_{st}) and satellite imagery (X_{sa}), yielding modality-specific vectors that can be fused downstream. Buildings without SVI are retained for later graph-based imputation.

Vertical feature propagation. Urban data often suffers from incomplete coverage (e.g., missing SVI in certain areas). Inspired by Kong et al. (2024), who transfer visual and socio-economic features by averaging information from neighboring nodes, we employ feature propagation over the constructed semantic graph G_s to impute missing

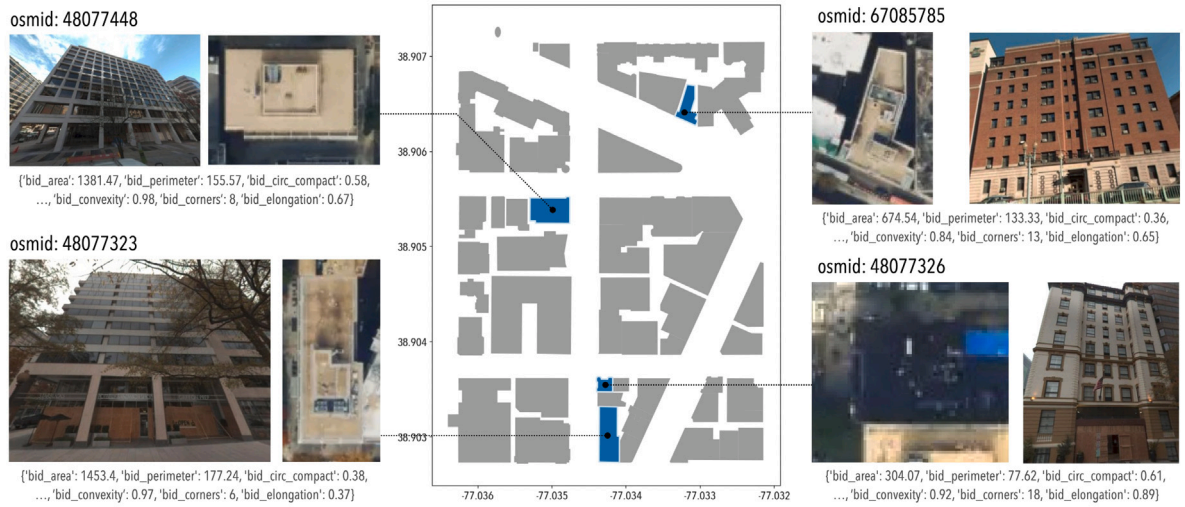

Fig. 3. Examples of extracted morphological features along with individual street-level and satellite images of buildings in Washington D.C. Data: (c) Mapbox, (c) Mapillary contributors, (c) OpenStreetMap contributors.

values (Rossi et al., 2022). This method leverages the graph structure to propagate known features to their missing counterparts. Let X_{st} be the feature matrix of vertical information, and $M \in \{0,1\}$ be the mask indicating observed entries (i.e., $M = 1$ if the feature is missing). The process starts with an initial feature matrix $X_{st}^{(0)}$, where all missing values are set to zero:

$$X_{st}^{(0)} = (1 - M) \cdot X_{st} \tag{3}$$

Then, the features are iteratively updated:

$$X_{st}^{(\ell+1)} = X_{st}^{(0)} + M \cdot (D^{-1/2} A D^{-1/2} X_{st}^{(\ell)}) \tag{4}$$

In each step, at iteration ℓ, a new value is calculated by adding the propagated features from the graph to the initial known features. The normalized adjacency matrix $D^{-1/2}AD^{-1/2}$, which ensures that features are propagated as a weighted average of a node's neighbors. The mask M is used again to ensure that only the missing values are updated.

Building attribute prediction. The final stage applies a heterogeneous GraphSAGE-based network to jointly learn from multi-modal building embeddings and the heterogeneous urban graph (Fig. 4a). First, we employ a deep feature encoder to fuse the semantic building features (X_{se}), propagated SVI features (X_{st}), and satellite features (X_{sa}). This encoder maps the concatenated inputs into a latent building embedding $X_B^{(0)}$ via a two-layer perceptron with batch normalization:

$$X_B^{(0)} = \sigma\Big(W_2 \cdot \sigma\big(W_1 \left[X_{se} \,\|\, X_{st} \,\|\, X_{sa}\right] + b_1\big) + b_2\Big) \tag{5}$$

where W_1, b_1 and W_2, b_2 are learnable parameters of the encoder layers, and σ denotes the ReLU activation function accompanied by batch normalization. This deep fusion step enables the network to capture non-linear correlations and relative importance between modalities before graph propagation.

The fused embeddings are subsequently propagated through stacked heterogeneous GraphSAGE layers, which aggregate information across different relation types via neighborhood message passing. The first GraphSAGE layer refines node embeddings within the same type based on edges R_{intra} (e.g., building–building, plot–plot, street–intersection), enabling localized smoothing (Fig. 4b):

$$X_v^{(1)} = \sigma\left(\underset{r \in R_{intra}}{\mathrm{AGG}}\left(X_v^{(0)} \times W_v^{(1)} + \mathrm{AGG}\big(X_u^{(0)} : u \in \mathcal{N}_r(v)\big) \times W_r^{(1)}\right)\right) \tag{6}$$

where $X_v^{(0)}$ and $X_v^{(1)}$ are the features of node v before and after the update, $N_r(v)$ denotes its same-type neighbors under relation r, W_v and W_r are learnable weights, and $AGG(\cdot)$ represents the aggregator function (e.g., $mean(\cdot)$, $sum(\cdot)$, $\max(\cdot)$). Then, the second GraphSAGE layer integrates contextual information from different node types (i.e., building, urban plot and street) into buildings through heterogeneous relations (Fig. 4c). For each cross relation $r \in R_{cross}$ targeting a building node b (e.g., street-building, plot-building), the final building representation Z_B computed using the embeddings from the previous layer:

$$Z_b = X_b^{(2)} = \sigma\left(\underset{r \in R_{cross}}{\mathrm{AGG}}\left(X_b^{(1)} \times W_v^{(2)} + \mathrm{AGG}\big(X_u^{(1)} : u \in \mathcal{N}_r(b)\big) \times W_r^{(2)}\right)\right) \tag{7}$$

Dropout regularization is applied at each stage to improve generalization. Finally, a fully connected prediction head maps the learned building embeddings to the target property. Specifically, for the task of inferring building type, the computational process of the prediction layer can be formulated as:

$$\hat{Y} = \mathrm{Softmax}\big(W_o Z_B + b_o\big) \tag{8}$$

where Z_B is the final embedding of building nodes B after GraphSAGE layers, W_h, W_o and b_h, b_o are learnable parameters, and $\hat{Y}$ denotes the predicted probability distribution over C building types. Since the building types are inherently imbalance in cities, a weighted cross-entropy loss is employed to handle class imbalance:

$$L(Y, \hat{Y}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} w_c \, y_{i,c} \, \log \hat{y}_{i,c} \tag{9}$$

where w_c denotes the weight for class c, $y_{i,c}$ is the ground-truth indicator, and $\hat{y}_{i,c}$ is the predicted probability for sample i.

4. Implementation

4.1. Data

To implement our methodology, we focus on three cities in this study: Amsterdam, Washington D.C., and Berlin. These study areas are selected based on the availability of diverse, high-quality urban datasets, thereby ensuring both geographic variability and sufficient coverage of the data stock. Table 3 presents a detailed breakdown of the data collected for this study, while Fig. 5 illustrates the spatial distribution of buildings counts with available labels and SVIs. The task undertaken in this study is to predict building type through the integration of multiple data modalities.

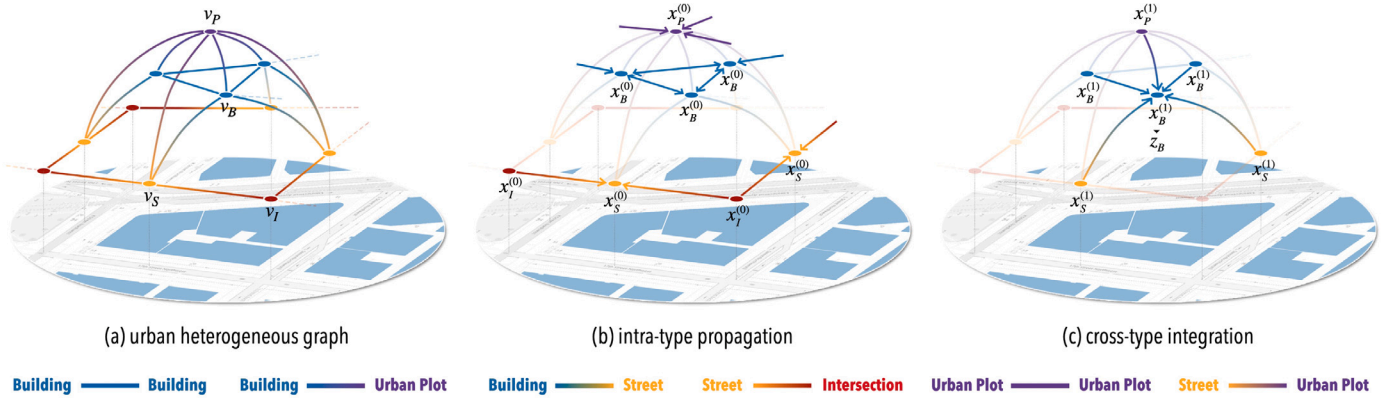

Fig. 4. Illustration of the heterogeneous GraphSAGE framework for building attribute prediction. (a) Example of the constructed urban heterogeneous graph, with GraphSAGE layers propagating information across (b) same-type relations and (c) cross-type relations. Data: (c) OpenStreetMap contributors.

Table 3
Summary of data collected for the three study cities.

City	Urban nodes				Building type			Buildings with SVI	
	Buildings	Streets	Intersections	Urban plots	Count	Classes	%	Count	%
Amsterdam	134,749	32,995	24,255	9,007	114,700	10	85.12%	107,097	79.5%
Washington D.C.	139,076	52,860	37,313	16,110	91,121	11	65.52%	84,456	60.7%
Berlin	408,227	197,818	168,279	30,446	200,743	12	49.17%	131,098	32.1%

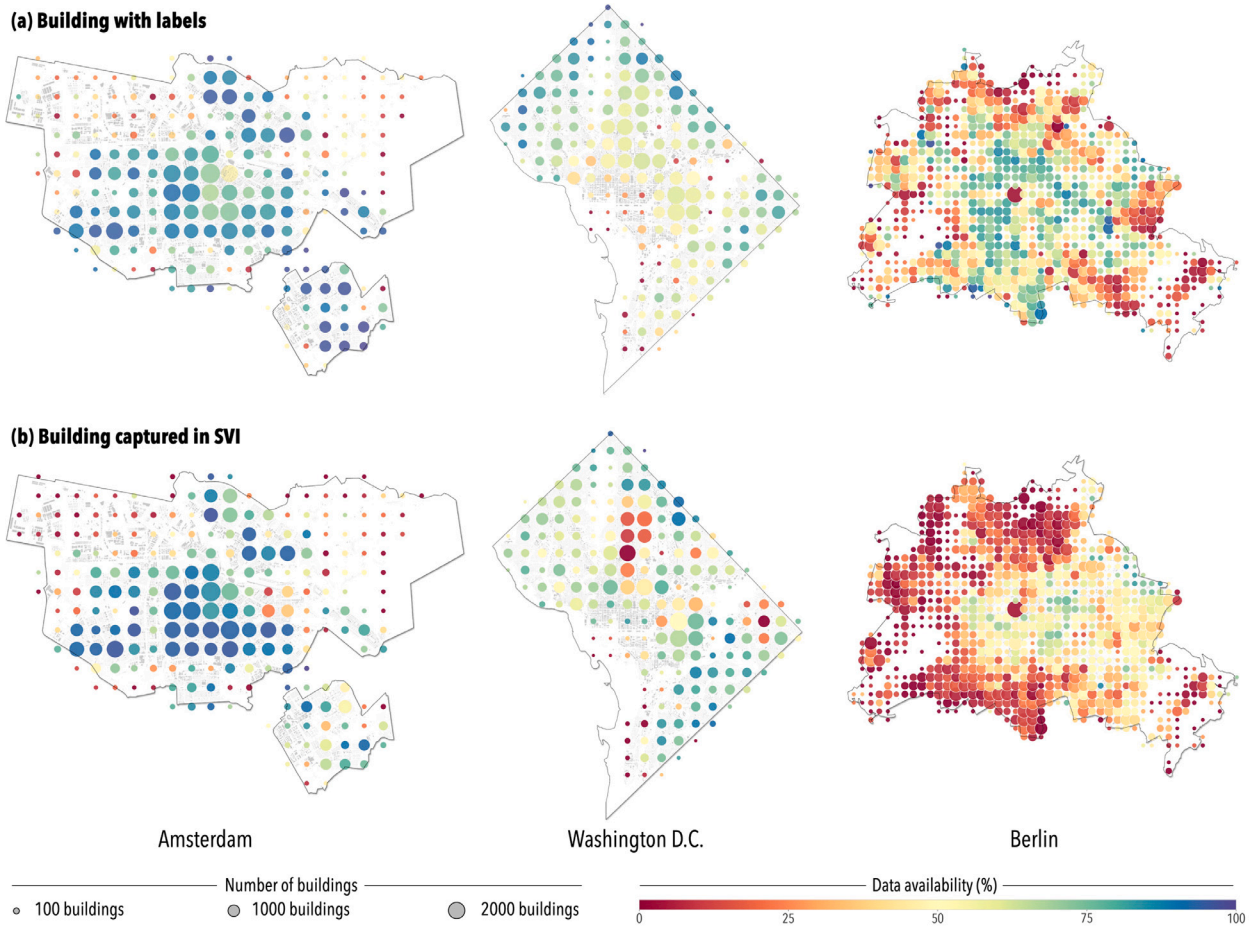

Fig. 5. Spatial distribution of buildings and data completeness across the study areas. Circle size reflects the number of buildings per grid cell, while color encodes the percentage of available (a) building-type labels and (b) street-level imagery (blue = higher completeness, red = lower completeness). Data: (c) OpenStreetMap contributors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Distribution of building types across Amsterdam, Washington D.C., and Berlin.

Building type	Amsterdam		Washington		Berlin	
	count	%	count	%	count	%
house	53,375	46.53	42,676	46.83	38,992	19.42
houseboat	2,762	2.41	–	–	–	–
detached_house	–	–	31,488	34.56	26,674	13.29
semidetached_house	–	–	14,753	16.19	11,664	5.81
allotment_house	–	–	–	–	12,819	6.39
apartments	52,082	45.41	749	0.82	82,688	41.19
commercial	2,843	2.48	397	0.44	8,063	4.02
office	194	0.17	218	0.24	1,980	0.99
public/governmental	63	0.05	68	0.07	1,932	0.96
education	254	0.22	290	0.32	2,554	1.27
religious	83	0.07	122	0.13	663	0.33
industrial	2,979	2.60	36	0.04	3,133	1.56
garage	65	0.06	324	0.36	9,581	4.77

Urban features. As discussed in Section 3.1, a set of urban morphology features is derived from multiple open data platforms. These features capture the spatial context of each building and encompass morphological, topological, functional, socio-demographic, and environmental characteristics. In total, 15 dimensional features are computed for buildings, 6 for streets, 8 for intersections, and 31 for urban plots. The supplementary material provides descriptive statistics for these features across the three cities, including their mean, maximum, and standard deviation values.

Building type. Building footprint and attribute information are obtained from OSM, which provides openly accessible and regularly updated volunteered geographic information. We extract functional tags from the OSM building data using the `key:building` attribute, where available. Because classification criteria vary across cities, we harmonize the commonly available classes among the selected study areas, resulting in 10 classes for Amsterdam, 11 classes for Washington D.C., and 12 classes for Berlin (Table 4). These classes constitute the foundation of our dataset and serve as the ground truth labels for building type classification.

Satellite imagery. We obtain VHR satellite imagery from Mapbox's global raster tileset, which aggregates data from multiple sources such as NASA, the United States Geological Survey, and others.[3] For each study area, image tiles at zoom level 17 are retrieved, and building-level patches are generated by locating the geometric information of each footprint and extracting the corresponding imagery. To account for potential discrepancies, such as roof structures extending beyond the base footprint due to building height or slight misalignments across platforms, each footprint is expanded with a 10 meter buffer. This buffer not only ensures complete coverage of the building but also incorporates a portion of its immediate surroundings, offering valuable contextual cues for building classification.

Street-level imagery. Street-level imagery is obtained from Mapillary, a crowdsourced platform with global coverage that allows free use and adaptation under the CC BY-SA 4.0 license. Using the Mapillary Python SDK, we download geotagged images within the study boundaries and employ the OpenFACADES framework[4] to extract individual observations and align them with the corresponding building footprints. Facade-level images are further filtered based on quality, visibility, and semantic segmentation criteria. Due to the varying availability of panoramic imagery across cities, the proportion of buildings with street-level coverage differs: 79.5% in Amsterdam, 60.7% in Washington D.C., and 32.1% in Berlin (Table 3), offering diverse scenarios for evaluating the framework. These images provide fine-grained visual information about building exteriors, including facade materials, entrance configurations, and stylistic features.

[3] https://docs.mapbox.com/help/glossary/mapbox-satellite/
[4] https://github.com/seshing/OpenFACADES

4.2. Model settings

As detailed in Section 3.2, our model is formulated as a heterogeneous graph approach for building node classification, integrating multi-modal geospatial data. Standardized urban features are extracted for each node type, including buildings, streets, intersections, and urban plots. Using the established heterogeneous graph to extract semantic building embeddings (512 dimensions), we establish a semantic similarity graph among buildings by constructing a 5-nearest neighbor graph. For building nodes, aerial and street-level imagery are encoded using the DINOv3 backbone pretrained with the `dinov3-vitb16-pretrain-lvd1689m` weights, producing 256-dimensional embeddings. To address missing visual features, a feature propagation step is applied on the semantic graph, where embeddings are iteratively diffused across neighboring nodes (10 iterations) under the guidance of a missing-value mask. Fig. 6 illustrates examples of this propagation process for Washington D.C. and Amsterdam. Finally, the visual embeddings are fused with the original building features and processed by a two-layer MLP with batch normalization to refine the input before the graph branch.

The graph learning component is implemented as a two-layer heterogeneous GraphSAGE network. The first layer propagates information across intersections, streets, plots, and buildings, while the second layer focuses on building-related connections. Each layer employs mean aggregation and a hidden dimensionality of 256. The resulting building embeddings are passed through a linear layer to produce the final class predictions. For optimization, we use the AdamW optimizer with a batch size of 128, a learning rate of 5×10^{-3}, and dropout with $p = 0.2$ is applied after each layer to prevent overfitting. The dataset is randomly split into training (60%), validation (20%), and test (20%) subsets. All models are trained for up to 500 epochs with early stopping, and class weights are applied to mitigate the effects of class imbalance. Ablation studies of alternative settings are reported in Section 6.

4.3. Benchmarks

The following inference models are selected as benchmarks for comparison with our heterogeneous GraphSAGE approach, including classical machine learning models and graph-based learning models:

- **Computer vision**: CNN- and Transformer-based architectures are widely used for building classification tasks (Kang et al., 2018; Ghione et al., 2022; Sun et al., 2022; Ogawa et al., 2023). In this study, we implement ResNet-50 (He et al., 2016) and the Swin Transformer, both of which have demonstrated strong performance in prior research on building classification (Raghu et al., 2023; Ogawa et al., 2023; Liang et al., 2024). Models are initialized with pretrained ImageNet-1K weights and trained on remote sensing imagery using a learning rate of 1×10^{-5} for up to 36 epochs, with early stopping based on validation performance. The checkpoint achieving the best validation score is selected for final evaluation on the test set and used as a benchmark for comparison with the proposed method.
- **Machine learning**: We establish two classical machine learning baselines: Random Forest (RF) and Extreme Gradient Boosting (XGBoost). RF is an ensemble approach that aggregates multiple decision trees, and XGBoost represents a gradient boosting framework that have achieved competitive performance in various building-related tasks (Tooke et al., 2014; Du et al., 2015; Biljecki and Sindram, 2017; Milojevic-Dupont et al., 2020; Nachtigall

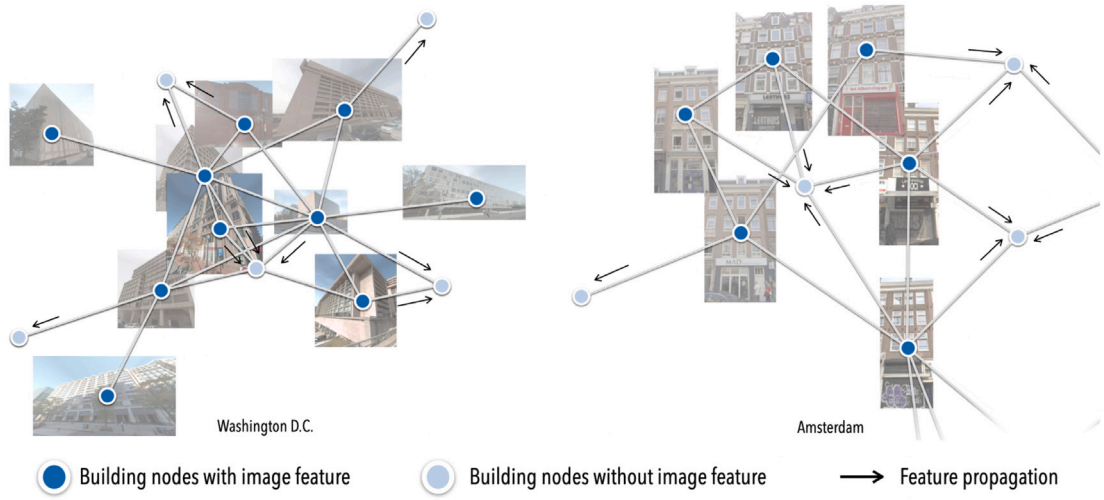

Fig. 6. Examples of feature propagation on the semantic similarity graph for buildings in Washington D.C. and Amsterdam. Data: (c) Mapillary contributors.

et al., 2023). To account for spatial autocorrelation and enable a fair comparison with the GNN-based method, we also implement neighborhood variants in which building footprint features from surrounding neighbors (i.e., the mean of values from adjacent building, street, intersection, and plot nodes) are incorporated as additional inputs. The neighborhood buffer is set to 200 m, and the aggregated features comprise 75 dimensions in total. The RF models are configured with 200 estimators, a minimum split size of 2, a minimum leaf size of 1, automatic feature selection for splits, and the Gini impurity criterion. The XGBoost models are configured with 300 estimators, the `gbtree` booster, a maximum tree depth of 8, a maximum delta step of 0, a minimum child weight of 1, and a learning rate of 0.05. To mitigate class imbalance issue, class weights are applied in both models.

- **Machine learning + visual features**: To evaluate the benefit of SVI and VHR satellite imagery, and to enable a fair comparison between model architectures, we extend XGBoost by adding visual features. We follow the same procedure as described in Section 3.2 to extract, normalize, and concatenate SVI and VHR embeddings with the urban features as inputs to XGBoost. To ensure comparability, the visual inputs for each building are controlled to be identical to those used in the HeteroGraphSAGE setting (e.g., same imagery sources, encoder, and embedding dimensions). The XGBoost configuration is kept the same as described above.
- **Homogeneous graph**: GraphSAGE is applied as another benchmark model, given its efficiency in neighborhood aggregation and its demonstrated capability in inferring building attributes (Lei et al., 2024; Kong et al., 2024). To integrate urban features into building nodes, we aggregate the mean values from adjacent street, intersection, and plot nodes within a 200 meter buffer, resulting in the same feature dimensionality as in our proposed method. In this study, the GraphSAGE architecture consists of three SAGEConv layers with mean aggregation, followed by three fully connected layers and a final classification layer. The model is configured with 256 hidden units and ReLU activations, and trained using the cross-entropy loss function for up to 500 epochs with early stopping. Optimization is performed with the Adam optimizer and a learning rate of 5×10^{-3}.

To compare model performance, we employ widely used classification metrics: Accuracy (Acc), macro-averaged Precision (mPre), Recall (mRec), and F1-score (mF1). These metrics are standard in building classification research (He et al., 2024), providing a balanced assessment of overall predictive performance.

5. Results

5.1. Overall performance

As shown in Table 5, our proposed HeteroGraphSAGE framework achieves consistently more robust and balanced performance than CV baselines, ML models and their multimodal extensions. While Random Forest and XGBoost attain relatively high overall accuracy and precision, their macro-F1 and recall remain notably lower, indicating a strong bias toward majority classes and limited effectiveness on underrepresented building types. Similarly, Swin Transformer, although effective in some scenarios, falls short in producing stable results across all metrics and cities. In contrast, GNN-based approaches capture spatial dependencies among buildings, leading to more balanced classification, especially across underrepresented classes.

When further enriched with multimodal inputs, the extended ML baselines (XGBoost with visual features) show performance improvements, indicating that visual embeddings provide compelling additional cues for this task. However, these gains remain limited compared to the proposed graph-based approach, suggesting that ML models are insufficient to fully exploit the multimodal information. Our HeteroGraphSAGE framework achieves the best overall performance in all three cities. For example, in Amsterdam, the multimodal variant improves accuracy to 93% and macro-F1 to 0.70. Similarly, in Washington, D.C., and Berlin, it delivers the highest accuracy and macro-F1 across 11 and 12 building classes, respectively, demonstrating consistent advantages across diverse urban morphologies. Specifically, VHR yielding substantial improvements and SVI offering complementary gains. Due to the limited spatial coverage of SVI, VHR remains the superior choice when restricted to a single visual source.

In addition to performance metrics, we further compared deep learning methods by visualizing the learned embeddings of our multimodal model against unimodal baselines. Specifically, we projected the high-dimensional test set representations into a 2D space using t-SNE (Fig. 7) to examine the discriminative structure of the latent space. The result shows that the CV and homogeneous GNN baselines exhibit significant manifold overlap, particularly among semantically similar categories, resulting in diffuse and entangled feature distributions. This visual ambiguity aligns with their lower recall values, suggesting these models struggle to differentiate complex urban classes. In contrast, the proposed multimodal HeteroGraphSAGE framework yields a more structured latent space with clearer inter-class separation and tighter intra-class compactness, in line with its higher macro-F1. This qualitative improvement is further supported by the Silhouette Coefficients, which increase markedly across all cities.

Table 5
Performance comparison of different models across Amsterdam, Washington D.C., and Berlin.

Model	Input data	Amsterdam				Washington D.C.				Berlin			
		Acc	mF1	mPre	mRec	Acc	mF1	mPre	mRec	Acc	mF1	mPre	mRec
ResNet50	VHR	86.45%	0.53	0.52	0.55	89.77%	0.55	0.56	0.58	74.16%	0.55	0.54	0.58
Swin-T		88.21%	0.58	0.61	0.57	91.60%	0.63	0.64	0.65	75.92%	0.59	0.58	0.63
Random Forest	urban features (75 dim.)	92.37%	0.49	0.62	0.45	95.37%	0.59	0.68	0.56	80.67%	0.48	**0.74**	0.45
XGBoost	urban features (75 dim.)	92.85%	0.53	0.66	0.48	95.07%	0.64	0.71	0.61	80.92%	0.56	0.72	0.51
	(+) VHR	92.03%	0.56	0.70	0.51	94.90%	0.62	0.65	0.60	81.53%	0.61	0.69	0.58
	(+) SVI	92.46%	0.57	0.67	0.53	94.77%	0.64	0.68	0.62	81.96%	0.62	0.69	0.60
	(+) SVI (+) VHR	92.35%	0.59	0.72	0.54	94.69%	0.67	0.74	0.63	81.93%	0.62	0.71	0.59
GraphSAGE	urban features (60 dim.)	88.95%	0.53	0.51	0.56	92.17%	0.63	0.61	0.66	76.29%	0.58	0.55	0.66
HeteroGraphSAGE	urban features (60 dim.)	89.50%	0.58	0.53	0.67	95.00%	0.67	0.68	0.68	80.69%	0.64	0.61	0.68
	(+) VHR	92.12%	0.65	0.71	0.64	95.54%	0.71	**0.75**	0.70	85.87%	0.68	0.69	0.68
	(+) SVI	92.27%	0.63	0.64	0.63	95.24%	0.70	0.71	0.70	84.98%	0.68	0.69	0.67
	(+) SVI (+) VHR	**93.03%**	**0.70**	**0.73**	**0.69**	**96.51%**	**0.73**	0.73	**0.74**	**86.58%**	**0.71**	0.72	**0.70**

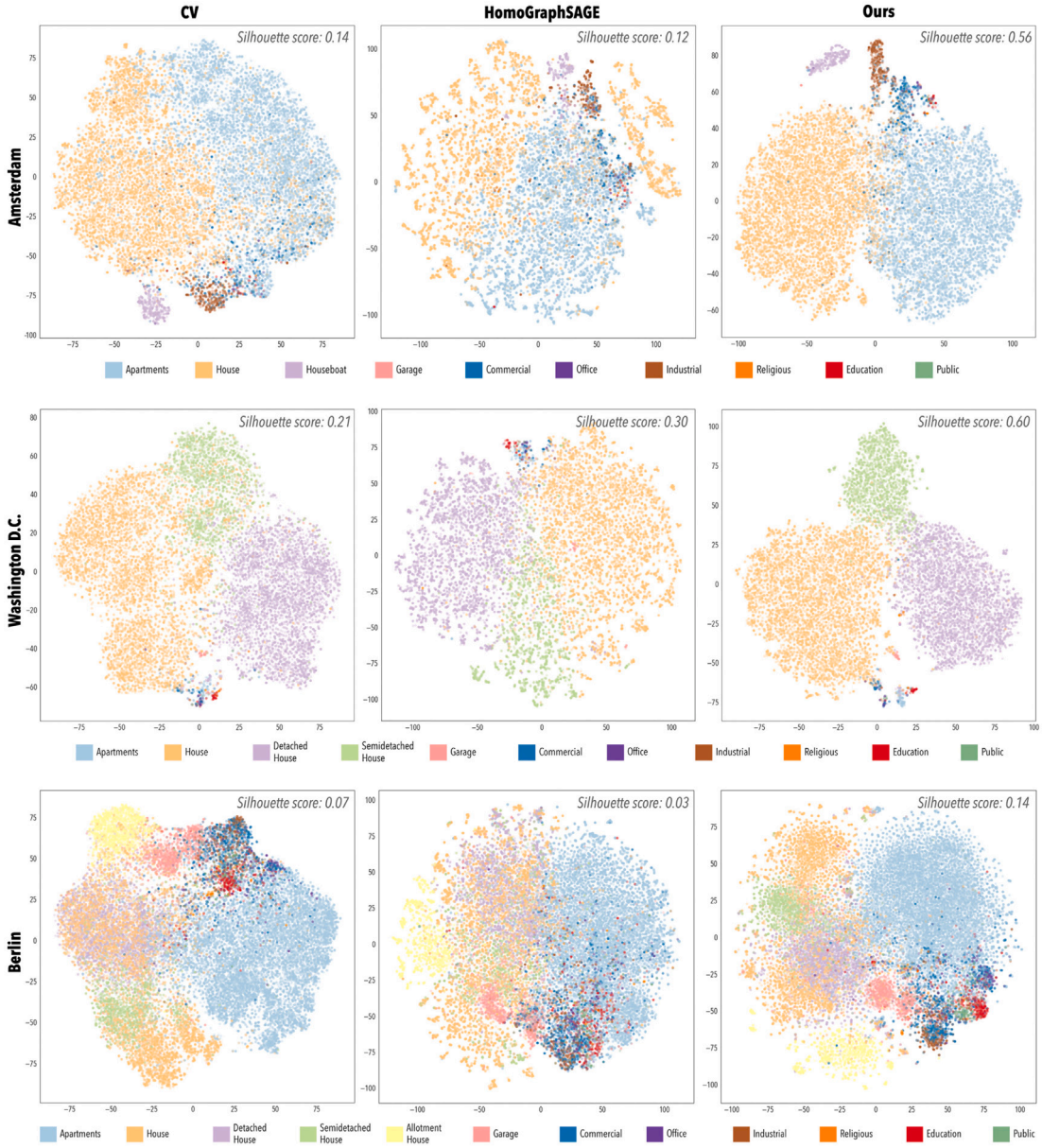

Fig. 7. t-SNE plots of building embeddings generated by the CV baseline (Swin Transformer), GraphSAGE, and our proposed HeteroSAGE for Amsterdam, Washington D.C., and Berlin, with the mean Silhouette Coefficient displayed to quantify the class separability of each model.

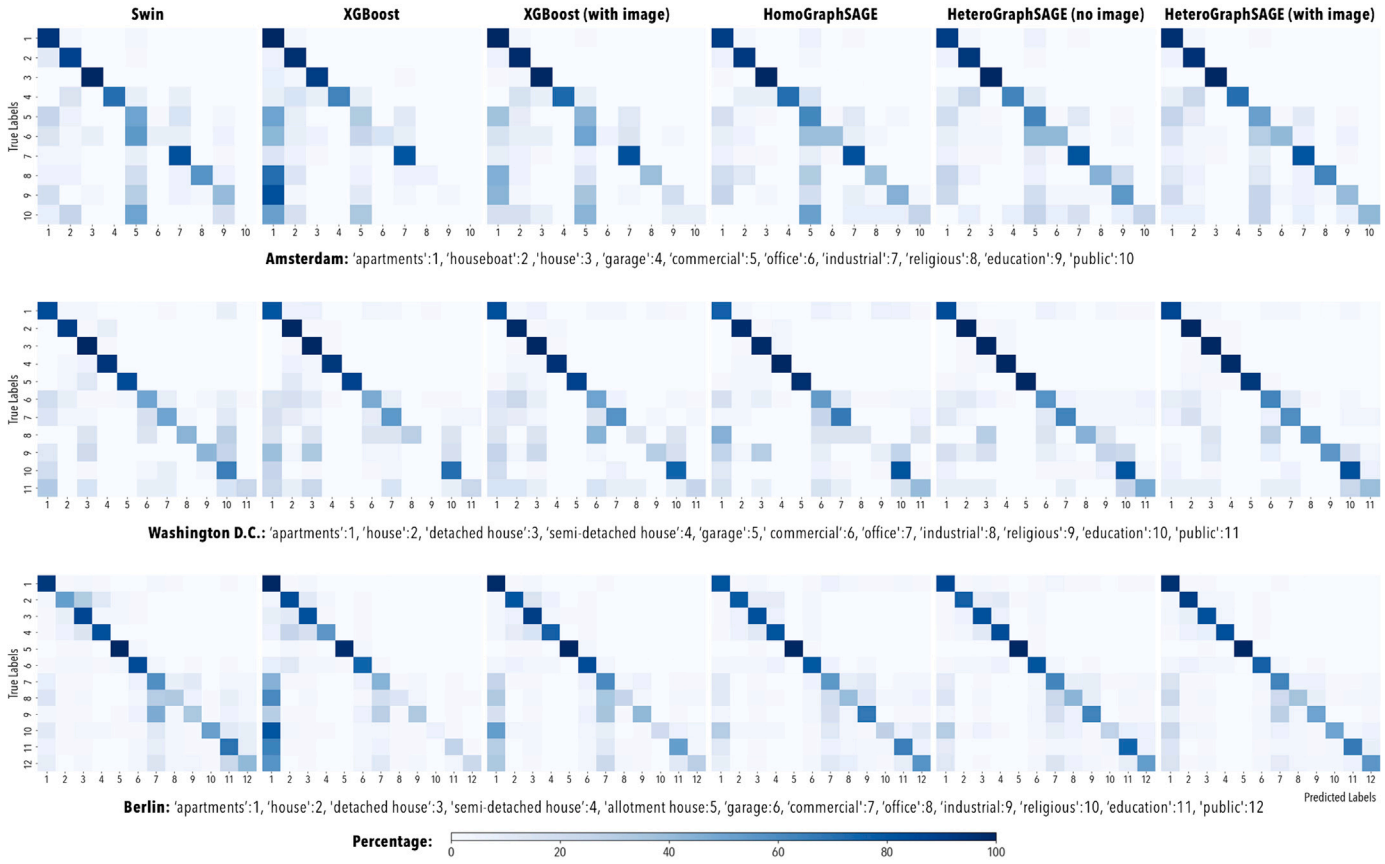

Fig. 8. Confusion matrices of different models in Amsterdam, Washington D.C. and Berlin.

5.2. Performance by categories

Fig. 8 presents the confusion matrices of different models across the three cities, illustrating how classification performance varies by building type. Compared to other baselines, our proposed HeteroGraphSAGE achieves more balanced predictions across categories, consistent with the macro-F1 results discussed earlier. Swin Transformer and XGBoost (with image) also deliver relatively balanced outcomes, yet they remain less stable in predicting minority classes than graph-based approaches. Across all cities, commercial and office, as well as office and industrial, emerge as the most challenging categories to separate, reflecting their appearance and morphological similarities. Notably, incorporating visual features (i.e., XGBoost with image and HeteroGraphSAGE with image) consistently improves performance over morphology-only inputs, particularly by enhancing the recognition of non-residential types relative to residential structures. Nevertheless, our deep graph-based model demonstrates a stronger ability to integrate multimodal inputs, resulting in more coherent and stable predictions across building categories. This underscores the complementary role of visual cues alongside urban morphological features, enabling finer-grained discrimination among ambiguous building classes.

To further assess performance, we compare selected models on unseen data, including both test and unlabeled buildings. Fig. 9 illustrates predictions in a representative area of Washington D.C., with selected examples. In general, all models reproduce the broad spatial patterns of building use, but differences become clear in more ambiguous cases. Religious buildings (1 and 2) illustrate the added value of integrating street-level imagery, as facade elements such as towers and masonry details provide discriminative signals that are absent from morphology or roof appearance alone. Similarly, educational and public buildings (3 and 4) exhibit similar advantages, as institutional facade designs provide additional visual signals. Office buildings (5 and 6), which often overlap with commercial or apartments structures, are identified more consistently when visual and urban-context cues are jointly considered. Despite these improvements, some cases remain difficult. For instance, building 7 , a government court facility, and building 9 , a university department building, are incorrectly labeled by all methods. This is likely due to a combination of factors, including their visually and morphologically generic appearance, conflicting signals across modalities, or insufficient discriminative features to support a confident classification. Mixed-use buildings (8), where commercial functions occupy the ground level, pose an additional challenge. Although common in city centers, they are neither explicitly annotated in OSM nor observable from imagery, resulting in ground-truth inconsistencies that further complicate model prediction.

5.3. Spatial error analysis

To better understand spatial variation in model performance, we compute classification accuracy on the test set for areas containing more than three buildings. As shown in Fig. 10a, performance is generally strong across regions. When compared with Fig. 5, the results demonstrate convincing performance even in areas with limited SVI coverage, reflecting the effectiveness of propagating street-level features.

Nonetheless, city centers often appear as the most challenging areas, likely due to the high heterogeneity of building functions and forms, which complicates prediction. To examine these cases more closely,

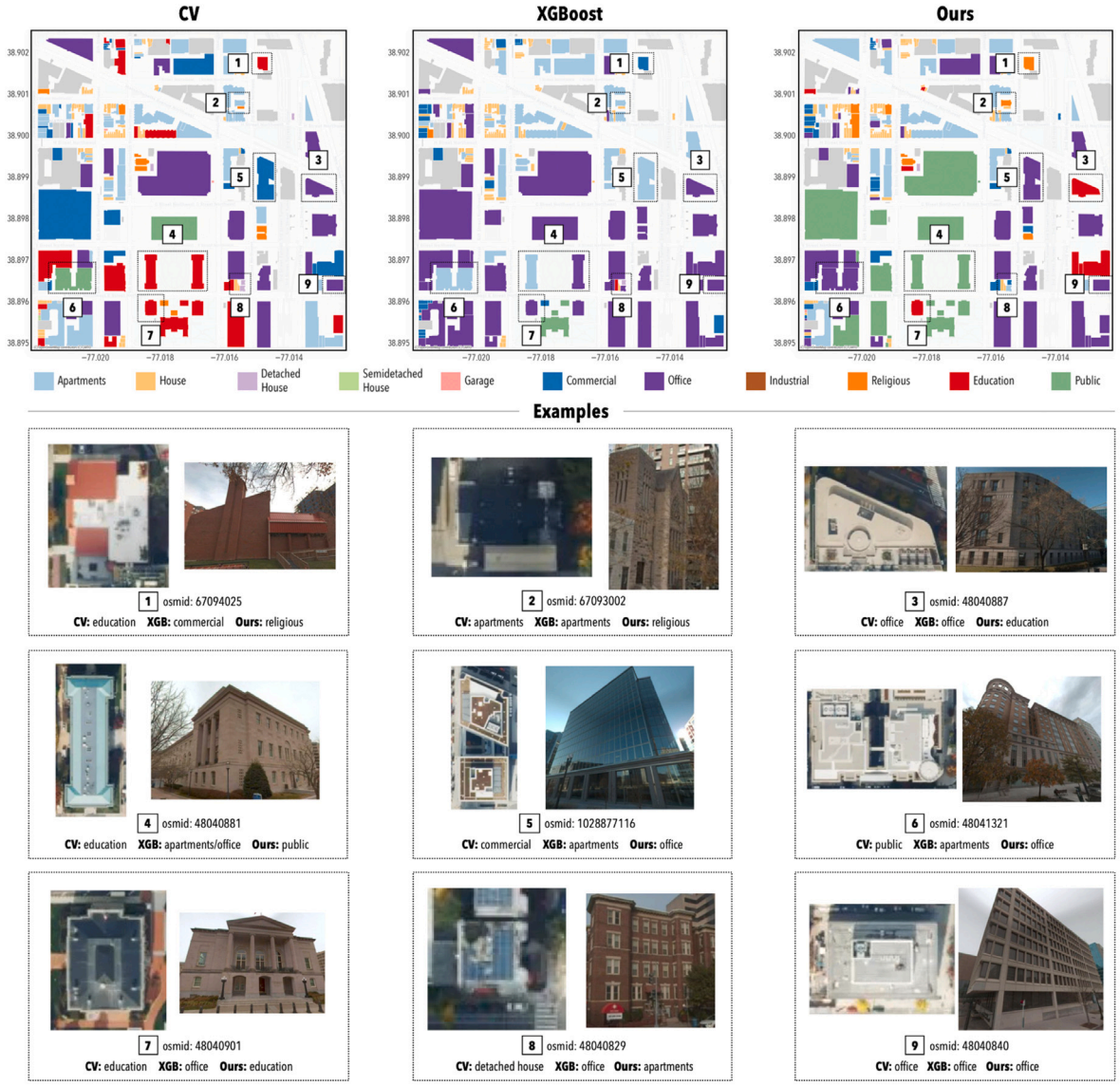

Fig. 9. Predictions across different methods of building use types on unseen data in Washington D.C., with selected examples. Data: (c) Mapbox, (c) Mapillary contributors, (c) OpenStreetMap contributors, (c) CARTO.

we sampled 20 lowest-performing zones in each city and summarized the top five most frequent 'ground truth ≫ prediction' errors, illustrated with examples in Fig. 10b. In general, several typical failure patterns emerge. First, consistent with earlier observations, limitations in the current labeling pipeline for multi-functional buildings (e.g., prediction between apartments and commercial) and inherent ambiguities between certain types (e.g., house ≫ apartments) reduce model reliability. Second, visual cues in imagery can sometimes be insufficient or misleading. For example, spatial gaps between buildings may be underrepresented (e.g., detached house ≫ semi-detached house), certain facade characteristics may be obscured from specific viewpoint (e.g., commercial ≫ office), or some types may be visually confounded (e.g., apartments ≫ detached house). Third, although feature propagation helps to compensate for missing SVI, it remains a supplementary strategy. In certain cases, semantically similar buildings may serve different functions (e.g., office ≫ commercial), and propagation can inadvertently introduce bias. These findings point to several directions for future work: refining building type labels and classification pipelines, incorporating additional identifiers to capture functional complexity, leveraging multi-view street-level imagery, and ensuring sufficient input data.

Taken together, the results yield three main insights. First, by capturing hierarchical features, our proposed heterogeneous graph approach enhances the recognition of minority classes and achieves more effective classification across 10–12 categories. Second, the inclusion of street-level information further improves robustness in cases where categories share similar morphology or top–down appearance. Third, some buildings remain intrinsically challenging due to ambiguous or uncommon forms, hybrid uses, or the absence of distinctive identifiers, underscoring the need for richer datasets and multi-label frameworks in future research.

6. Ablation experiments

This section presents a series of ablation and sensitivity experiments designed to better understand the contributions of different components

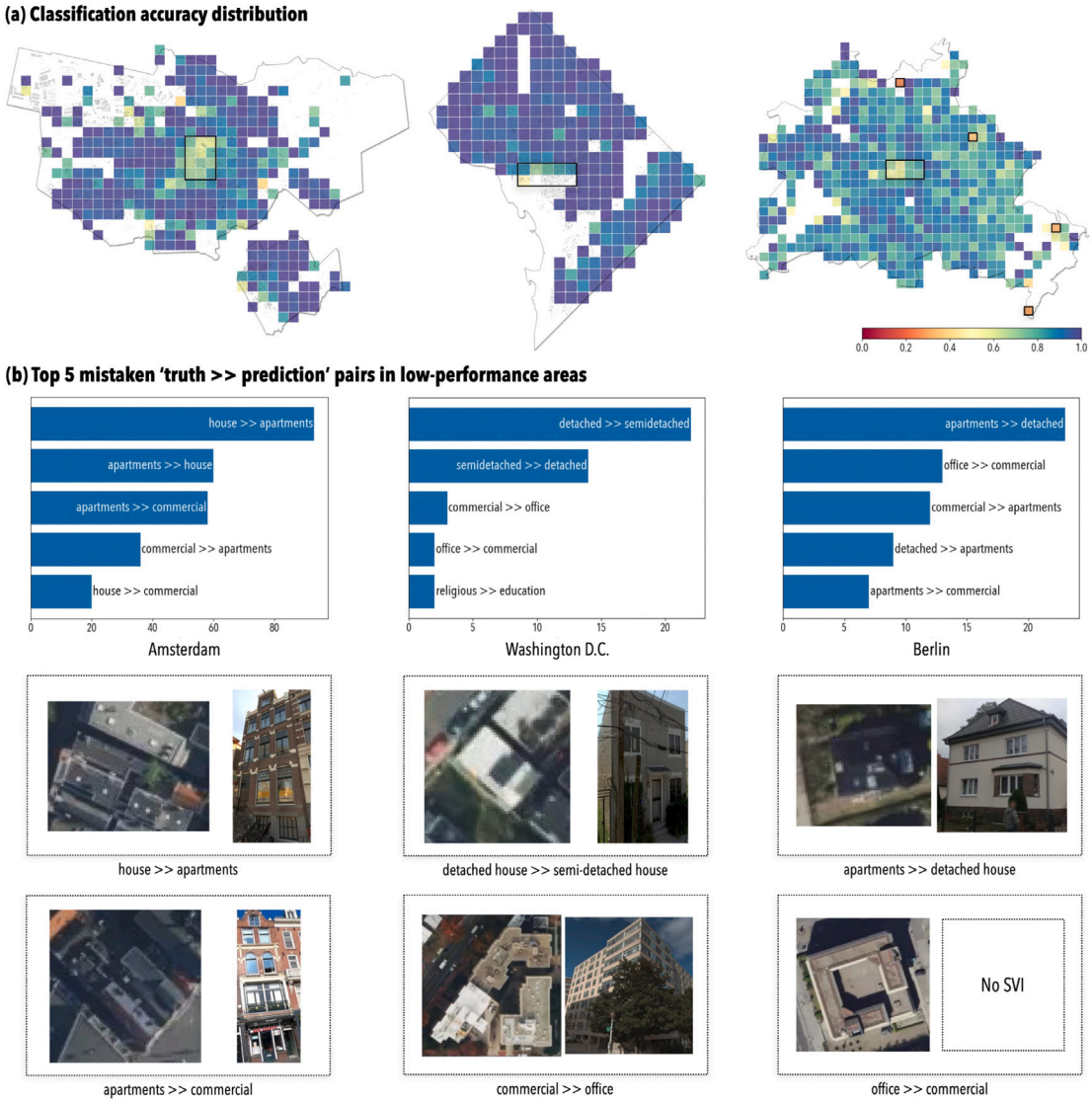

Fig. 10. Spatial performance of building type classification. (a) Accuracy aggregated by local zones (grid size: 1 × 1 km for Amsterdam and Washington D.C.; 2 × 2 km for Berlin), (b) Frequent misclassification patterns from the 20 lowest-performing zones, illustrating typical 'ground truth ≫ prediction' errors and selected examples. Data: (c) Mapbox, (c) Mapillary contributors, (c) OpenStreetMap contributors.

in our framework. Specifically, we examine (1) the contribution of different node types through graph component ablation, (2) the effect of different pretrained vision backbones on multimodal learning; (3) the sensitivity of the model to the completeness of street-level features, and (4) the impact of key hyperparameters on model performance. All models are trained using the AdamW optimizer with default β parameters. All experiments are conducted on a single NVIDIA RTX 4090 GPU. Each configuration is repeated five times under identical settings, and average performance is reported.

6.1. Graph component ablation

To quantify the contribution of cross-scale information integration, we conduct a graph component ablation study that systematically evaluates different node-type combinations in the heterogeneous graph. Specifically, we evaluate three configurations: (1) individual information sources, where only one node type is used; (2) partial integration, where different combinations of two or three node types are included; and (3) full integration, where all four node types are jointly incorporated. Different configurations are implemented by masking node features of excluded types and restricting message passing to edges connecting the selected node types. When the building node is excluded, its features are masked while predictions are still generated at building nodes based solely on information propagated from other node types.

Table 6 reports the results across Amsterdam, Washington D.C., and Berlin. Three key observations emerge. First, full integration consistently achieves the best and most stable performance across all cities, highlighting the complementary nature of cross-scale relational information. Second, building-level features contribute the most to performance, followed by urban plot features, while intersection-only graphs perform weakest. This could be attributed to the limited feature dimensionality of streets and intersections, or indicates that pure topological information alone is insufficient for building type prediction. Third, the magnitude of performance improvement yielded by augmenting building nodes with other node types varies significantly across cities. This suggests that the degree of reliance on broader

Table 6
Performance comparison using different node-type combinations across Amsterdam, Washington D.C., and Berlin.

Node type(s)	Amsterdam		Washington		Berlin	
	Acc	mF1	Acc	mF1	Acc	mF1
Individual source						
building	81.81%	0.49	82.90%	0.52	67.72%	0.50
urban plot	67.84%	0.33	85.49%	0.47	57.80%	0.42
street	63.58%	0.30	66.43%	0.33	45.16%	0.33
intersection	44.79%	0.19	39.16%	0.17	25.49%	0.19
Partial integration						
building, urban plot	87.66%	0.53	92.95%	0.65	79.27%	0.62
building, street	87.95%	0.55	89.76%	0.57	73.79%	0.56
building, intersection	85.96%	0.54	86.17%	0.54	71.06%	0.53
urban plot, street	73.12%	0.39	89.36%	0.54	65.23%	0.49
urban plot, intersection	71.95%	0.37	88.66%	0.52	63.65%	0.47
street, intersection	68.32%	0.34	74.22%	0.39	50.18%	0.37
building, urban plot, street	89.37%	0.56	93.89%	0.66	80.90%	0.63
building, urban plot, intersection	88.82%	0.57	93.70%	0.65	80.50%	0.63
building, street, intersection	88.40%	0.56	91.40%	0.60	74.83%	0.57
urban plot, street, intersection	73.69%	0.39	90.38%	0.55	66.03%	0.50
Full integration						
all	**89.52%**	**0.58**	**94.46%**	**0.67**	**81.10%**	**0.64**

spatial and topological context fluctuates according to the specific urban morphology and planning patterns of each city.

6.2. Vision module selection

To determine the most effective vision module, we evaluated several pretrained vision backbones, including ResNet, ViT-16, Swin Transformer, DINOv2, and DINOv3. All vision models are applied as standalone feature extractors prior to GNN training, producing fixed-length embeddings from 224 × 224 images which are stored for subsequent training and inference. In total, 241,846 images in Amsterdam, 223,532 in Washington, D.C., and 539,325 in Berlin are processed. These embeddings are then loaded and integrated as node features during GNN training, avoiding repeated feature extraction. Consequently, GNN training efficiency remains largely unaffected by backbone complexity, with training times ranging from 16 to 46 s for 500 epochs and inference times between 0.05 and 0.1 s.

As shown in Table 7, self-supervised transformer backbones (DINOv2/v3) consistently outperform the other baselines. Among them, DINOv3 achieves the best overall performance, with the base model (DINOv3-B) yielding the highest accuracy and mF1 in Washington D.C. (96.38%, 0.74) and the highest accuracy in Berlin (86.56%). While the large model (DINOv3-L) achieves the highest accuracy in Amsterdam (93.20%) and mF1 in Berlin (0.71), DINOv3-B maintains comparable performance in the two cities. Considering both efficiency and stability across datasets, we select DINOv3 with base pretrained weights as the vision encoder in our framework.

6.3. Sensitivity to completeness of street-level features

To evaluate the dependence and effectiveness of our framework with respect to street-level feature propagation, we conduct a sensitivity analysis by progressively masking SVI features at ratios ranging from 0.1 to 1.0. Fig. 11 shows the resulting performance in terms of accuracy and macro-F1 across the three cities.

Overall, model performance decreases as the proportion of masked SVI increases, confirming the contribution of street-level information to building attribute prediction. The decline is most evident in Amsterdam and Washington D.C., where SVI coverage is relatively high, with macro-F1 dropping by more than five percentage points when all SVI features are removed. In contrast, the impact is less pronounced in Berlin, where SVI coverage is limited. Importantly, the results show that the model remains relatively stable under partial masking when sufficient SVI is available (e.g., Amsterdam and Washington D.C.), indicating that feature propagation provides resilience against missing facade information. Moreover, even in Berlin, retaining only 25% of street-level features still yields better performance than the fully masked setting. These findings demonstrate that the proposed feature propagation mechanism is effective in mitigating incomplete SVI coverage by leveraging relational context, and that street-level imagery provides complementary semantic cues beyond urban attributes and satellite imagery alone.

6.4. Configurations analysis

Hyperparameter tuning is conducted across multiple dimensions, including image feature embedding sizes (32, 64, 128, 256, 512), hidden dimensions (32, 64, 128, 256, 512), learning rates (1×10^{-2}, 5×10^{-3}, 1×10^{-3}, 5×10^{-4}, 1×10^{-4}), and the number of nearest neighbors for building nodes (3, 5, 8, 10, 15). Fig. 12 presents the sensitivity of our model to different configurations, assessed on the validation set during training. Overall, we observe consistent performance trends across accuracy and macro-F1, with relatively stable outcomes at medium-to-large parameter settings.

For image dimensions, both metrics show the median accuracy and macro-F1 generally increase with larger embedding sizes up to 256, after which improvements plateau, suggesting that 256 dimensions offer a good balance between accuracy and computational cost. Hidden unit size exhibits a similar trend. Performance steadily improves from 32 to 128 units, after which gains become negligible. While 512 units achieve slightly higher accuracies for Amsterdam, they also exhibit larger variance of mean F1 in Berlin, indicating reduced stability. This suggests that 256 hidden units are optimal for balancing performance and model robustness. Learning rate emerges as the most sensitive hyperparameter. Small rate (1×10^{-4}) lead to underfitting, while overly large rates (1×10^{-2}) destabilize training and reduce performance. The best trade-off is consistently observed around 5×10^{-3}, which achieves high accuracy with relatively low variance. For the number of nearest neighbors, overall performance is only weakly sensitive to this parameter, with a slight improvement observed when increasing the neighborhood size from 3 to 5, while further increases yield negligible gains. Considering the increased graph density and computational cost associated with larger neighborhoods, a value of 5 is selected. Based on this analysis, we set the default configuration to an image embedding size of 256, hidden unit size of 256, learning rate of 5×10^{-3}, and 5 nearest neighbors for building nodes.

7. Discussion

7.1. Inferring other building attributes

To further evaluate the capability of our method in predicting additional attributes, we conduct an experiment on the task of inferring building age. Construction year data from Amsterdam are used, focusing on buildings constructed after 1800. In total, 121,974 buildings with age labels are obtained (90.5% of all buildings in Amsterdam), and their distribution is summarized in Fig. 13. To simulate a data-limited scenario, the dataset is split into training, validation, and test sets in a 2:2:6 ratio. Using the same parameter settings, we compare performance against the same benchmark models as in the building type task.

Table 8 reports the performance of different models on building age prediction in Amsterdam. Among the baselines, tree-based methods (RF and XGBoost) achieve lower errors than both the vision-based

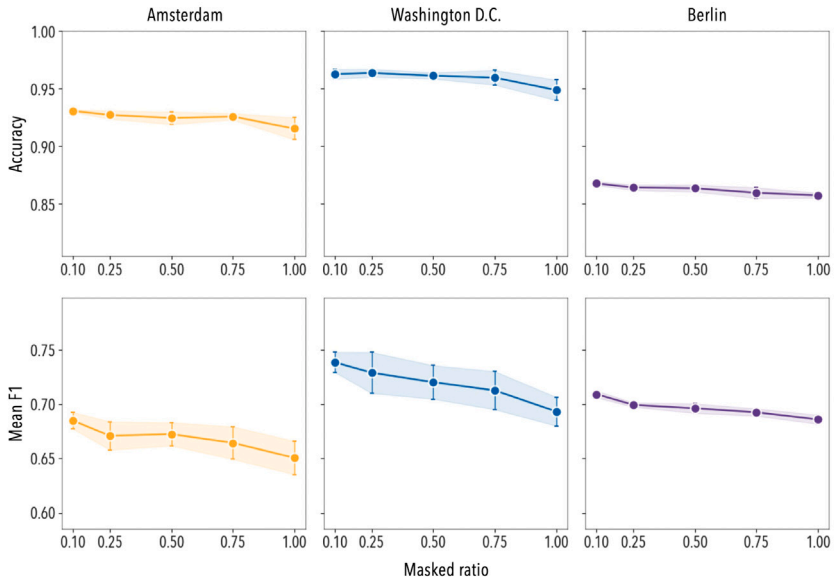

Fig. 11. Model performance under different levels of masked street view imagery across Amsterdam, Washington D.C., and Berlin. The x-axis indicates the ratio of SVI features randomly masked, and the y-axes show Accuracy (top row) and macro-F1 (bottom row). Each point represents the average of 5 runs, with shaded areas and error bars denoting the standard deviation.

Swin Transformer and the GraphSAGE models, reflecting the effectiveness of urban features for this task. When enriched with both SVI and VHR satellite data, our proposed HeteroGraphSAGE consistently outperforms the benchmarks across all metrics. It achieves the best performance with an RMSE of 16.86, MAE of 6.86, and R^2 of 0.83. The breakdown by age group further highlights the benefits of multi-modal integration for XGBoost and HeteroGraphSAGE models, with our approach achieves the lowest RMSE across all age categories. In particular, incorporating SVI markedly reduces errors for the youngest buildings ($age \leq 50$), while the full integration yields substantial improvements for the older buildings ($50 < age \leq 100$ and $age > 100$), dropping the RMSE to 10.23 and 21.28, respectively, underscoring its advantage in capturing both morphological and visual signals for predicting construction periods.

7.2. Implications and future works

This study demonstrates the value of combining heterogeneous graphs and cross-view imagery for large-scale building attribute prediction. By integrating hierarchical urban features with satellite and street-level visual cues, our framework moves beyond building-centric aggregation and provides a more holistic representation of the urban environment.

First, by explicitly modeling the city as a heterogeneous graph, our approach captures the hierarchical relations among buildings, urban plots, streets, and intersections. This design goes beyond homogeneous graph frameworks that treat buildings in isolation and allows feature propagation across multiple urban layers. The findings suggest that incorporating such hierarchy not only improves building attribute classification but also opens avenues for extending prediction to other multi-factor attributes, such as building energy performance (Hu et al., 2022), building-level Local Climate Zones (Li et al., 2025b), or urban heat exposure (Liu et al., 2025b).

Second, the systematic integration of urban features with satellite and street-level imagery provides a balanced and robust predictive framework. Prior research has demonstrated the complementary strengths of morphological descriptors and top–down imagery (Lei et al., 2024; Kong et al., 2024; Wang et al., 2024; Yap et al., 2025), but facade-level cues remain underutilized at scale. Our results show that cross-modal fusion enhances recognition of minority and visually distinctive classes while reducing ambiguity in categories with similar morphology or top–down appearance. This step forward indicates broader potential for applying fine-grained, multi-modal observations

Table 7
Performance and efficiency of different vision backbones on Amsterdam, Washington D.C., and Berlin.

Backbone	Pretrained weight	Params (M)	Amsterdam			Washington D.C.			Berlin		
			Time (s)	Acc	mF1	Time (s)	Acc	mF1	Time (s)	Acc	mF1
ResNet18	IMAGENET1K_V1	11	70.49	92.61%	0.62	66.14	95.84%	0.69	154.62	85.93%	0.68
ResNet50		24	111.50	92.55%	0.66	96.45	95.81%	0.71	258.35	86.04%	0.70
ViT-B/16		86	257.17	92.24%	0.63	237.60	95.45%	0.70	572.51	85.82%	0.69
Swin-T		28	160.85	92.65%	0.64	144.41	96.27%	0.70	342.26	85.79%	0.69
DINOv2-S	vit_small_patch14_dinov2.lvd142m	22	142.48	92.51%	0.62	136.95	96.16%	0.73	285.19	85.96%	0.70
DINOv2-B	vit_base_patch14_dinov2.lvd142m	86	348.84	92.98%	0.65	321.35	96.08%	0.73	799.30	85.79%	0.70
DINOv3-B	dinov3-vitb16-pretrain-lvd1689m	86	306.42	93.10%	**0.68**	282.24	**96.38%**	**0.74**	716.51	**86.56%**	0.70
DINOv3-L	dinov3-vitl16-pretrain-lvd1689m	303	992.72	**93.20%**	0.66	916.22	96.13%	**0.74**	2264.00	86.50%	**0.71**

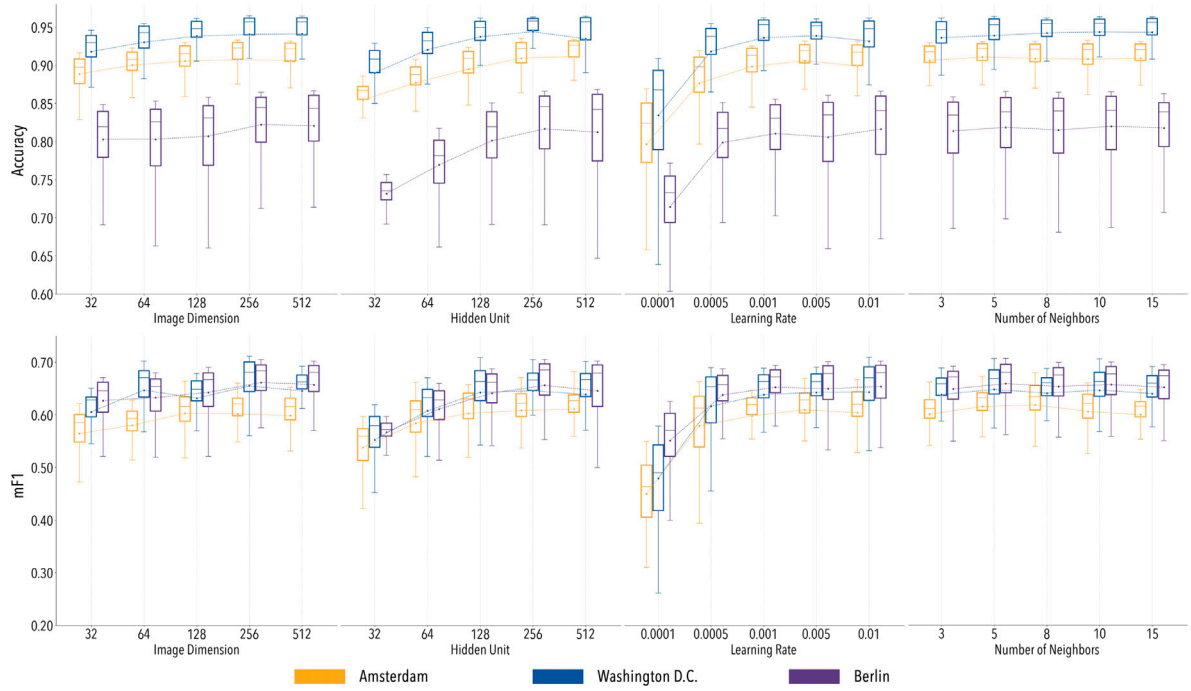

Fig. 12. Sensitivity analysis of model performance with respect to hyperparameter configurations. Boxplots show distributions of validation accuracy and macro-F1 across data from Amsterdam, Washington D.C. and Berlin. Lines connect mean values to highlight performance trends.

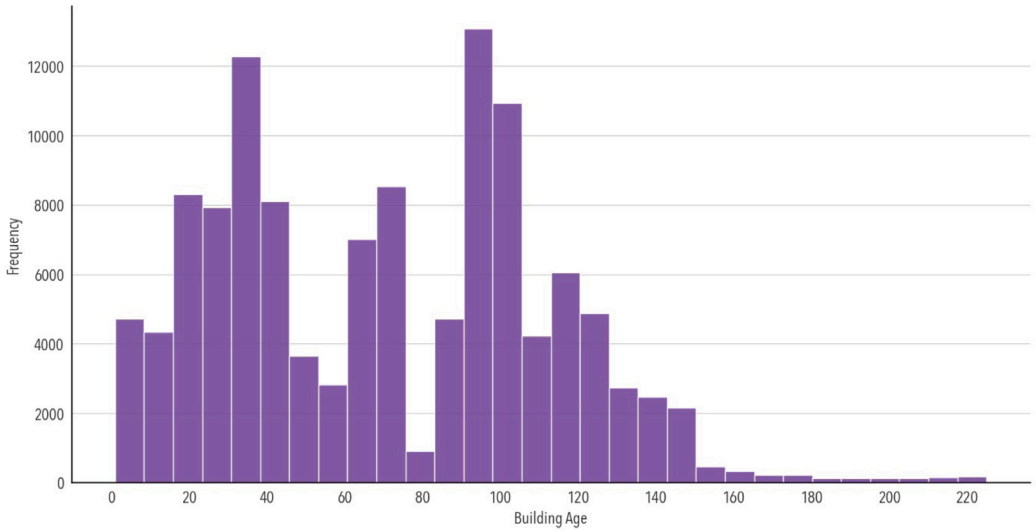

Fig. 13. Distribution of building age data in Amsterdam from 1800 to the present.

to tasks such as material characterization (Lei et al., 2024), and disaster vulnerability assessment (Li et al., 2025a).

Third, validation across multiple global cities and attributes underscores the robustness of the proposed method. We show that visual cues substantially enhance building attribute prediction across contexts, even in data-scarce environments. This highlights opportunities for future global-scale research, where similar frameworks could be used to infer building-level indices relevant for sustainability and equity, such as solar irradiance potential (Yu et al., 2025), or environmental exposure (Yu et al., 2016; Li et al., 2022b). The ability to generalize across diverse urban contexts is important for addressing data inequalities among regions.

While the proposed framework advances the integration of heterogeneous graphs and cross-view imagery for building attribute prediction, limitations remain that offer avenues for future research. A first limitation concerns the coverage and quality of SVI. Although SVI provides unique vertical and facade-level cues that are indispensable for identifying building types, materials, and architectural periods, its availability, particular for crowdsourced platform, is highly uneven across areas and countries (Hou and Biljecki, 2022). These issues introduce noise into the learning process and can bias predictions toward well-documented urban contexts. In this work, we mitigate these challenges by incorporating preprocessing of street-level building images through the OpenFACADES toolkit,[5] which automates quality assessment by detecting occlusions, distortions, and incomplete facades. Also, feature propagation of street-level feature is introduced

[5] https://github.com/seshing/OpenFACADES

Table 8
Performance comparison of different models for building age prediction in Amsterdam.

Model	Input data	Overall			RMSE by age group		
		RMSE	MAE	R^2	$age \leq 50$	$50 < age \leq 100$	$age > 100$
ResNet50	VHR	21.84	13.10	0.72	21.48	15.94	28.28
Swin-T		21.99	12.74	0.72	21.03	15.43	29.58
Random Forest	urban features (75 dim.)	19.70	9.95	0.77	21.57	11.17	25.01
XGBoost	urban features (75 dim.)	19.51	10.48	0.78	21.08	11.75	24.70
	(+) VHR	20.40	12.17	0.76	21.89	12.49	25.88
	(+) SVI	19.10	11.15	0.79	20.24	11.64	24.59
	(+) SVI (+) VHR	19.34	11.49	0.78	20.40	11.88	24.95
GraphSAGE	urban features (60 dim.)	20.48	11.96	0.76	22.03	13.68	25.13
HeteroGraphSAGE	urban features (60 dim.)	19.81	11.21	0.77	20.89	12.61	25.28
	(+) VHR	18.81	8.28	0.79	21.52	11.95	21.72
	(+) SVI	17.35	7.26	0.82	18.23	11.07	22.19
	(+) SVI (+) VHR	**16.86**	**6.86**	**0.83**	**18.22**	**10.23**	**21.28**

on semantic similarity graphs, while this remains a mitigation strategy rather than a full solution. In future work, it will be valuable to test the extent to which incorporating high-quality imagery, for example, from commercial platforms, where higher-resolution, multi-temporal, and more systematically curated SVI datasets are increasingly available. This have potential to further stabilize building-level representations.

A second area for improvement involves the integration of temporal information. Our current framework focuses on static representations of buildings and their urban context. Yet, many attributes, such as building use, occupancy, and even facade conditions, are dynamic over time. Incorporating temporal layers into heterogeneous graphs, or linking with longitudinal imagery archives, could allow the prediction of changes in building attributes and enrich studies on urban dynamics. This is especially relevant for monitoring processes such as urban evolution, gentrification, or the degradation of building stock, which are critical for long-term planning.

A third limitation concerns cross-city generalization. While the proposed framework performs well within individual cities, our exploratory tests and empirical experience suggest that models trained on one city do not transfer reliably to others. This indicates that the learned representations may still encode substantial city-specific characteristics. Addressing this limitation, particularly in the context of data-scarce cities, remains an important direction for future work and may require domain adaptation strategies, multi-city pretraining, or explicit modeling of cross-city heterogeneity.

8. Conclusion

Reliable information on buildings is fundamental for understanding cities and supporting various urban applications. Yet many attributes, including use type, construction year, and number of floors, remain scarce, fragmented, or inconsistently recorded across regions. These gaps limit large-scale urban analytics and restrict the development of open and comprehensive geospatial resources. To address this challenge, we developed a hierarchical, multi-modal graph neural network framework that integrates urban features with cross-view imagery. Our method models the city as a heterogeneous graph consisting of buildings, urban plots, streets, and intersections, enriched with morphological, topological, functional, socio-demographic, and environmental indicators. Visual features from very high-resolution satellite imagery and street view imagery are further incorporated, with a semantic similarity graph used to propagate facade information and mitigate incomplete street-level coverage. This design enables a holistic representation of buildings that captures both their local appearance and broader urban context. The contributions of this work are: (1) introducing a heterogeneous graph framework that explicitly captures hierarchical urban relations; (2) systematically integrating urban features with cross-view imagery; and (3) validating the framework across multiple global cities and attributes.

Our experiments show that the proposed method enhances recognition of minority classes, achieves more balanced classification across diverse building categories, and improves robustness in cases where categories share similar morphology or top–down appearance. At the same time, results highlight potential challenges, including misclassification of ambiguous or hybrid-use buildings and the impact of uneven SVI coverage. These findings imply that combining heterogeneous graphs with cross-view imagery provides a scalable and robust pathway to enrich building-level information. Beyond predicting building attributes, the approach also provides opportunities to infer other urban indicators shaped by diverse and complex factors, such as energy usage, solar access, or environmental exposure, supporting more equitable and sustainable urban analytics on a global scale.

CRediT authorship contribution statement

Xiucheng Liang: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Winston Yap:** Writing – review & editing, Software, Methodology, Conceptualization. **Filip Biljecki:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is part of the project Large-scale 3D Geospatial Data for Urban Analytics, which is supported by the National University of Singapore, Singapore under the Start Up Grant. The first and second authors acknowledge the NUS Graduate Research Scholarship granted by the National University of Singapore (NUS). We express our gratitude to the members of the NUS Urban Analytics Lab for the valuable discussions. We also acknowledge the contributors of OpenStreetMap, Mapillary and other platforms for providing valuable open data resources and code that support urban research and applications.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.isprsjprs.2026.02.016.

References

Basaraner, M., Cetinkaya, S., 2017. Performance of shape indices and classification schemes for characterising perceptual shape complexity of building footprints in GIS. Int. J. Geogr. Inf. Sci. 31, 1952–1977. http://dx.doi.org/10.1080/13658816.2017.1346257.

Batty, M., 2009. Cities as complex systems: Scaling, interaction, networks, dynamics and urban morphologies. In: Encyclopedia of Complexity and Systems Science. Springer, pp. 1041–1071.

Biljecki, F., Chew, L.Z.X., Milojevic-Dupont, N., Creutzig, F., 2021. Open government geospatial data on buildings for planning sustainable and resilient cities. http://dx.doi.org/10.48550/arXiv.2107.04023, URL: http://arxiv.org/abs/2107.04023, arXiv:2107.04023.

Biljecki, F., Chow, Y.S., Lee, K., 2023. Quality of crowdsourced geospatial building information: A global assessment of OpenStreetMap attributes. Build. Environ. 237, 110295.

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: A review. Landsc. Urban Plan. 215, 104217.

Biljecki, F., Ledoux, H., Stoter, J., 2017. Generating 3D city models without elevation data. Comput. Environ. Urban Syst. 64, 1–18. http://dx.doi.org/10.1016/j.compenvurbsys.2017.01.001.

Biljecki, F., Sindram, M., 2017. Estimating building age with 3D GIS. ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci. IV-4-W5, 17–24. http://dx.doi.org/10.5194/isprs-annals-IV-4-W5-17-2017.

Boeing, G., 2022. Street network models and indicators for every urban area in the world. Geogr. Anal. 54, 519–535.

Creutzig, F., Lohrey, S., Bai, X., Baklanov, A., Dawson, R., Dhakal, S., Lamb, W.F., McPhearson, T., Minx, J., Munoz, E., et al., 2019. Upscaling urban data science for global climate solutions. Glob. Sustain. 2, e2.

De Sabbata, S., Liu, P., 2023. A graph neural network framework for spatial geodemographic classification. Int. J. Geogr. Inf. Sci. 37, 2464–2486. http://dx.doi.org/10.1080/13658816.2023.2254382.

Dibble, J., Prelorendjos, A., Romice, O., Zanella, M., Strano, E., Pagel, M., Porta, S., 2019. On the origin of spaces: Morphometric foundations of urban form evolution. Environ. Plan. B: Urban Anal. City Sci. 46, 707–730. http://dx.doi.org/10.1177/2399808317725075.

Du, S., Zhang, F., Zhang, X., 2015. Semantic classification of urban buildings combining VHR image and GIS data: An improved random forest approach. ISPRS J. Photogramm. Remote Sens. 105, 107–119. http://dx.doi.org/10.1016/j.isprsjprs.2015.03.011, URL: https://www.sciencedirect.com/science/article/pii/S092427161500091X.

Elmqvist, T., Andersson, E., Frantzeskaki, N., McPhearson, T., Olsson, P., Gaffney, O., Takeuchi, K., Folke, C., 2019. Sustainability and resilience for transformation in the urban century. Nat. Sustain. 2, 267–273. http://dx.doi.org/10.1038/s41893-019-0250-1, URL: https://www.nature.com/articles/s41893-019-0250-1. Publisher: Nature Publishing Group.

Fan, Z., Feng, C.-C., Biljecki, F., 2025. Coverage and bias of street view imagery in mapping the urban environment. Comput. Environ. Urban Syst. 117, 102253.

Feldmeyer, D., Meisch, C., Sauter, H., Birkmann, J., 2020. Using OpenStreetMap data and machine learning to generate socio-economic indicators. ISPRS Int. J. Geo-Information 9, 498.

Fleischmann, M., Arribas-Bel, D., 2022. Geographical characterisation of British urban form and function using the spatial signatures framework. Sci. Data 9, 546. http://dx.doi.org/10.1038/s41597-022-01640-8.

Florio, P., Politis, P., Krasnodebska, K., Uhl, J.H., Melchiorri, M., Martinez, A.M., Kakoulaki, G., Pesaresi, M., Kemper, T., 2025. GHS-OBAT: Global, open building attribute data reporting age, function, height and compactness at footprint level. Data Brief 111751.

Frantz, D., Schug, F., Okujeni, A., Navacchi, C., Wagner, W., van der Linden, S., Hostert, P., 2021. National-scale mapping of building height using sentinel-1 and sentinel-2 time series. Remote Sens. Environ. 252, 112128.

Ghione, F., Mæland, S., Meslem, A., Oye, V., 2022. Building stock classification using machine learning: A case study for Oslo, Norway. Front. Earth Sci. 10, 886145.

He, Z., Yao, W., Shao, J., Wang, P., 2024. UB-FineNet: Urban building fine-grained classification network for open-access satellite images. ISPRS J. Photogramm. Remote Sens. 217, 76–90. http://dx.doi.org/10.1016/j.isprsjprs.2024.08.008, URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271624003186.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Herfort, B., Lautenbach, S., Porto De Albuquerque, J., Anderson, J., Zipf, A., 2023. A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap. Nat. Commun. 14, 3985. http://dx.doi.org/10.1038/s41467-023-39698-6, URL: https://www.nature.com/articles/s41467-023-39698-6.

Hou, Y., Biljecki, F., 2022. A comprehensive framework for evaluating the quality of street view imagery. Int. J. Appl. Earth Obs. Geoinf. 115, 103094. http://dx.doi.org/10.1016/j.jag.2022.103094, URL: https://linkinghub.elsevier.com/retrieve/pii/S1569843222002825.

Hu, Y., Cheng, X., Wang, S., Chen, J., Zhao, T., Dai, E., 2022. Times series forecasting for urban building energy consumption based on graph convolutional network. Appl. Energy 307, 118231. http://dx.doi.org/10.1016/j.apenergy.2021.118231.

Jia, C., Du, Y., Wang, S., Bai, T., Fei, T., 2019. Measuring the vibrancy of urban neighborhoods using mobile phone data with an improved PageRank algorithm. Trans. GIS 23, 241–258. http://dx.doi.org/10.1111/tgis.12515.

Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X.X., 2018. Building instance classification using street view images. ISPRS J. Photogramm. Remote Sens. 145, 44–59. http://dx.doi.org/10.1016/j.isprsjprs.2018.02.006, URL: https://www.sciencedirect.com/science/article/pii/S0924271618300352.

Kirkley, A., Barbosa, H., Barthelemy, M., Ghoshal, G., 2018. From the betweenness centrality in street networks to structural invariants in random planar graphs. Nat. Commun. 9, 2501. http://dx.doi.org/10.1038/s41467-018-04978-z.

Kong, B., Ai, T., Zou, X., Yan, X., Yang, M., 2024. A graph-based neural network approach to integrate multi-source data for urban building function classification. Comput. Environ. Urban Syst. 110, 102094. http://dx.doi.org/10.1016/j.compenvurbsys.2024.102094.

Kumar, S., Pal, S.K., Singh, R.P., 2018. A novel method based on extreme learning machine to predict heating and cooling load through design and structural attributes. Energy Build. 176, 275–286.

Lei, B., Liu, P., Milojevic-Dupont, N., Biljecki, F., 2024. Predicting building characteristics at urban scale using graph neural networks and street-level context. Comput. Environ. Urban Syst. 111, 102129. http://dx.doi.org/10.1016/j.compenvurbsys.2024.102129, URL: https://linkinghub.elsevier.com/retrieve/pii/S0198971524000589.

Lei, B., Stouffs, R., Biljecki, F., 2023. Assessing and benchmarking 3D city models. Int. J. Geogr. Inf. Sci. 37, 788–809.

Li, H., Deuser, F., Yin, W., Luo, X., Walther, P., Mai, G., Huang, W., Werner, M., 2025a. Cross-view geolocalization and disaster mapping with street-view and VHR satellite imagery: A case study of Hurricane IAN. ISPRS J. Photogramm. Remote Sens. 220, 841–854. http://dx.doi.org/10.1016/j.isprsjprs.2025.01.003, URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271625000036.

Li, J., Huang, X., Tu, L., Zhang, T., Wang, L., 2022a. A review of building detection from very high resolution optical remote sensing images. GIScience Remote. Sens. 59, 1199–1225. http://dx.doi.org/10.1080/15481603.2022.2101727.

Li, S., Liu, P., Stouffs, R., 2025b. Fine-grained local climate zone classification using graph networks: A building-centric approach. Build. Environ. 278, 112928.

Li, M., Xue, F., Wu, Y., Yeh, A.G., 2022b. A room with a view: Automatic assessment of window views for high-rise high-density areas using city information models and deep transfer learning. Landsc. Urban Plan. 226, 104505.

Li, W., Yu, J., Chen, D., Lin, Y., Dong, R., Zhang, X., He, C., Fu, H., 2025c. Fine-grained building function recognition with street-view images and GIS map data via geometry-aware semi-supervised learning. Int. J. Appl. Earth Obs. Geoinf. 137, 104386. http://dx.doi.org/10.1016/j.jag.2025.104386, URL: https://linkinghub.elsevier.com/retrieve/pii/S1569843225000330.

Liang, X., Chang, J.H., Gao, S., Zhao, T., Biljecki, F., 2024. Evaluating human perception of building exteriors using street view imagery. Build. Environ. 263, 111875.

Liang, X., Xie, J., Zhao, T., Stouffs, R., Biljecki, F., 2025. OpenFACADES: An open framework for architectural caption and attribute data enrichment via street view imagery. ISPRS J. Photogramm. Remote Sens. 230, 918–942.

Lindenthal, T., Johnson, E.B., 2025. Machine learning, architectural styles and property values. J. Real Estate Financ. Econ. 71, 353–384.

Liu, P., Biljecki, F., 2022. A review of spatially-explicit GeoAI applications in urban geography. Int. J. Appl. Earth Obs. Geoinf. 112, 102936. http://dx.doi.org/10.1016/j.jag.2022.102936.

Liu, P., Chen, Y., Liang, X., Li, H., Biljecki, F., Stouffs, R., 2025a. A graph neural network for small-area estimation: Integrating spatial regularisation, heterogeneous spatial units, and Bayesian inference. Int. J. Geogr. Inf. Sci. 1–39.

Liu, P., Lei, B., Huang, W., Biljecki, F., Wang, Y., Li, S., Stouffs, R., 2025b. Sensing climate justice: A multi-hyper graph approach for classifying urban heat and flood vulnerability through street view imagery. Sustain. Cities Soc. 118, 106016.

Lu, Z., Im, J., Rhee, J., Hodgson, M., 2014. Building type classification using spatial and landscape attributes derived from LiDAR remote sensing data. Landsc. Urban Plan. 130, 134–148. http://dx.doi.org/10.1016/j.landurbplan.2014.07.005.

Milojevic-Dupont, N., Hans, N., Kaack, L.H., Zumwald, M., Andrieux, F., de Barros Soares, D., Lohrey, S., Pichler, P.-P., Creutzig, F., 2020. Learning from urban form to predict building heights. PLoS One 15, e0242010.

Nachtigall, F., Milojevic-Dupont, N., Wagner, F., Creutzig, F., 2023. Predicting building age from urban form at large scale. Comput. Environ. Urban Syst. 105, 102010. http://dx.doi.org/10.1016/j.compenvurbsys.2023.102010, URL: https://linkinghub.elsevier.com/retrieve/pii/S019897152300073X.

Ogawa, Y., Zhao, C., Oki, T., Chen, S., Sekimoto, Y., 2023. Deep learning approach for classifying the built year and structure of individual buildings by automatically linking street view images and GIS building data. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16, 1740–1755.

Ozuduru, B.H., Webster, C.J., Chiaradia, A.J.F., Yucesoy, E., 2021. Associating street-network centrality with spontaneous and planned subcentres. Urban Stud. 58, 2059–2078. http://dx.doi.org/10.1177/0042098020931302.

Prieto-Curiel, R., Schumann, A., Heo, I., Heinrigs, P., 2022. Detecting cities with high intermediacy in the African urban network. Comput. Environ. Urban Syst. 98, 101869. http://dx.doi.org/10.1016/j.compenvurbsys.2022.101869.

Raghu, D., Bucher, M.J.J., De Wolf, C., 2023. Towards a 'resource cadastre' for a circular economy–urban-scale building material detection using street view imagery and computer vision. Resour. Conserv. Recycl. 198, 107140.

Ramalingam, S.P., Kumar, V., 2023. Automatizing the generation of building usage maps from geotagged street view images using deep learning. Build. Environ. 235, 110215. http://dx.doi.org/10.1016/j.buildenv.2023.110215, URL: https://www.sciencedirect.com/science/article/pii/S0360132323002421.

Rosser, J.F., Boyd, D.S., Long, G., Zakhary, S., Mao, Y., Robinson, D., 2019. Predicting residential building age from map data. Comput. Environ. Urban Syst. 73, 56–67. http://dx.doi.org/10.1016/j.compenvurbsys.2018.08.004.

Rossi, E., Kenlay, H., Gorinova, M.I., Chamberlain, B.P., Dong, X., Bronstein, M., 2022. On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features. http://dx.doi.org/10.48550/arXiv.2111.12128, arXiv:2111.12128.

Roth, J., Martin, A., Miller, C., Jain, R.K., 2020. SynCity: Using open data to create a synthetic city of hourly building energy estimates by integrating data-driven and physics-based methods. Appl. Energy 280, 115981.

Schug, F., Frantz, D., van der Linden, S., Hostert, P., 2021. Gridded population mapping for Germany based on building density, height and type from Earth observation data using census disaggregation and bottom-up estimates. PLoS One 16, e0249044.

Sun, M., Zhang, F., Duarte, F., Ratti, C., 2022. Understanding architecture age and style through deep learning. Cities 128, 103787.

Tang, J., Xu, L., Yu, H., Jiang, H., He, D., Li, T., Xiao, W., Zheng, X., Liu, K., Li, Y., et al., 2025. A dataset of multi-level street-block divisions of 985 cities worldwide. Sci. Data 12, 456.

Tooke, T.R., Coops, N.C., Webster, J., 2014. Predicting building ages from LiDAR data with random forests for building energy modeling. Energy Build. 68, 603–610.

Wang, X., Guan, X., Cao, J., Zhang, N., Wu, H., 2020. Forecast network-wide traffic states for multiple steps ahead: A deep learning approach considering dynamic non-local spatial correlation and non-stationary temporal dependency. Transp. Res. Part C: Emerg. Technol. 119, 102763.

Wang, Y., Zhang, Y., Dong, Q., Guo, H., Tao, Y., Zhang, F., 2024. A multi-view graph neural network for building age prediction. ISPRS J. Photogramm. Remote Sens. 218, 294–311. http://dx.doi.org/10.1016/j.isprsjprs.2024.10.011, URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271624003885.

Wang, S., Zhao, C., Jiang, Q., Zhu, D., Ma, J., Sun, Y., 2025. Application of graph convolutional neural networks and multi-sources data on urban functional zones identification, A case study of Changchun, China. Sustain. Cities Soc. 119, 106116.

Wang, Y., Zhu, D., 2024. A hypergraph-based hybrid graph convolutional network for intracity human activity intensity prediction and geographic relationship interpretation. Inf. Fusion 104, 102149.

Westrope, C., Banick, R., Levine, M., 2014. Groundtruthing OpenStreetMap building damage assessment. Procedia Eng. 78, 29–39.

Wu, W.-B., Ma, J., Banzhaf, E., Meadows, M.E., Yu, Z.-W., Guo, F.-X., Sengupta, D., Cai, X.-X., Zhao, B., 2023. A first Chinese building height estimate at 10 m resolution (CNBH-10 m) using multi-source earth observations and machine learning. Remote Sens. Environ. 291, 113578.

Wurm, M., Schmitt, A., Taubenböck, H., 2016. Building types' classification using shape-based features and linear discriminant functions. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 9, 1901–1912. http://dx.doi.org/10.1109/JSTARS.2015.2465131.

Xu, Y., He, Z., Xie, X., Xie, Z., Luo, J., Xie, H., 2022. Building function classification in Nanjing, China, using deep learning. Trans. GIS 26, 2145–2165. http://dx.doi.org/10.1111/tgis.12934.

Xu, X., Wang, W., Hong, T., Chen, J., 2019. Incorporating machine learning with building network analysis to predict multi-building energy use. Energy Build. 186, 80–97. http://dx.doi.org/10.1016/j.enbuild.2019.01.002.

Xue, J., Jiang, N., Liang, S., Pang, Q., Yabe, T., Ukkusuri, S.V., Ma, J., 2022. Quantifying the spatial homogeneity of urban road networks via graph neural networks. Nat. Mach. Intell. 4, 246–257. http://dx.doi.org/10.1038/s42256-022-00462-y.

Yan, X., Ai, T., Yang, M., Yin, H., 2019. A graph convolutional neural network for classification of building patterns using spatial vector data. ISPRS J. Photogramm. Remote Sens. 150, 259–273. http://dx.doi.org/10.1016/j.isprsjprs.2019.02.010.

Yap, W., Biljecki, F., 2023. A global feature-rich network dataset of cities and dashboard for comprehensive urban analyses. Sci. Data 10, 667.

Yap, W., Stouffs, R., Biljecki, F., 2023. Urbanity: Automated modelling and analysis of multidimensional networks in cities. NPJ Urban Sustain. 3, 45. http://dx.doi.org/10.1038/s42949-023-00125-w.

Yap, W., Wu, A.N., Miller, C., Biljecki, F., 2025. Revealing building operating carbon dynamics for multiple cities. Nat. Sustain. 8, 1199–1210.

Yu, Q., Dong, K., Guo, Z., Xu, J., Li, J., Tan, H., Jin, Y., Yuan, J., Zhang, H., Liu, J., Chen, Q., Yan, J., 2025. Global estimation of building-integrated facade and rooftop photovoltaic potential by integrating 3D building footprint and spatio-temporal datasets. Nexus 2, http://dx.doi.org/10.1016/j.ynexs.2025.100060, URL: https://www.cell.com/nexus/abstract/S2950-1601(25)00007-5. Publisher: Elsevier.

Yu, S., Yu, B., Song, W., Wu, B., Zhou, J., Huang, Y., Wu, J., Zhao, F., Mao, W., 2016. View-based greenery: A three-dimensional assessment of city buildings' green visibility using floor Green view index. Landsc. Urban Plan. 152, 13–26.

Zhang, Y., Liu, P., Biljecki, F., 2023. Knowledge and topology: A two layer spatially dependent graph neural networks to identify urban functions with time-series street view image. ISPRS J. Photogramm. Remote Sens. 198, 153–168. http://dx.doi.org/10.1016/j.isprsjprs.2023.03.008, URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271623000680.

Zhao, W., Bo, Y., Chen, J., Tiede, D., Blaschke, T., Emery, W.J., 2019. Exploring semantic elements for urban scene recognition: Deep integration of high-resolution imagery and OpenStreetMap (OSM). ISPRS J. Photogramm. Remote Sens. 151, 237–250. http://dx.doi.org/10.1016/j.isprsjprs.2019.03.019, URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271619300887.

Zhao, K., Liu, Y., Hao, S., Lu, S., Liu, H., Zhou, L., 2021. Bounding boxes are all we need: Street view image classification via context encoding of detected buildings. IEEE Trans. Geosci. Remote Sens. 60, 1–17.

Zhao, W., Persello, C., Stein, A., 2022. Extracting planar roof structures from very high resolution images using graph neural networks. ISPRS J. Photogramm. Remote Sens. 187, 34–45.

Zhu, X.X., Chen, S., Zhang, F., Shi, Y., Wang, Y., 2025. GlobalBuildingAtlas: An open global and complete dataset of building polygons, heights and LoD1 3D models. Earth Syst. Sci. Data Discuss. 2025, 1–31.

Zhu, D., Ma, Z., 2025. Gravity-informed deep flow inference for spatial evolution modeling in panel data. Int. J. Geogr. Inf. Sci. 1–29.

Zhu, D., Zhang, F., Wang, S., Wang, Y., Cheng, X., Huang, Z., Liu, Y., 2020. Understanding place characteristics in geographic contexts through graph convolutional neural networks. Ann. Am. Assoc. Geogr. 110, 408–420. http://dx.doi.org/10.1080/24694452.2019.1694403.