OpenFACADES: An Open Framework for Architectural Caption and Attribute Data Enrichment via Street View Imagery

Xiucheng Liang[a], Jinheng Xie[b], Tianhong Zhao[c], Rudi Stouffs[a], Filip Biljecki[a,d,*]

[a]*Department of Architecture, National University of Singapore, Singapore*
[b]*Department of Electrical and Computer Engineering, National University of Singapore, Singapore*
[c]*School of Artificial Intelligence, Shenzhen Technology University, Shenzhen, China*
[d]*Department of Real Estate, National University of Singapore, Singapore*

Abstract

Building properties, such as height, usage, and material, play a crucial role in spatial data infrastructures, supporting various urban applications. Despite their importance, comprehensive building attribute data remain scarce in many urban areas. Recent advances have enabled the extraction of objective building attributes using remote sensing and street-level imagery. However, establishing a pipeline that integrates diverse open datasets, acquires holistic building imagery, and infers comprehensive building attributes at scale remains a significant challenge. Among the first, this study bridges the gaps by introducing OpenFACADES, an open framework that leverages multimodal crowdsourced data to enrich building profiles with both objective attributes and semantic descriptors through multimodal large language models. First, we integrate street-level image metadata from Mapillary with OpenStreetMap geometries via isovist analysis, identifying images that provide suitable vantage points for observing target buildings. Second, we automate the detection of building facades in panoramic imagery and tailor a reprojection approach to convert objects into holistic perspective views that approximate real-world observation. Third, we introduce an innovative approach that harnesses and investigates the capabilities of open-source large vision-language models (VLMs) for multi-attribute prediction and open-vocabulary captioning in building-level analytics, leveraging a globally sourced dataset of 31,180 labeled images from seven cities. Evaluation shows that fine-tuned VLM excel in multi-attribute inference, outperforming single-attribute computer vision models

1

and zero-shot ChatGPT-4o. Further experiments confirm its superior generalization and robustness across culturally distinct region and varying image conditions. Finally, the model is applied for large-scale building annotation, generating a dataset of 1.2 million images for half a million buildings. This open-source framework enhances the scope, adaptability, and granularity of building-level assessments, enabling more fine-grained and interpretable insights into the built environment. Our dataset and code are available openly at: https://github.com/seshing/OpenFACADES.

Keywords: Building exteriors, Street-level, Volunteered geographic information, ChatGPT, Multi-task learning, SDI

1. Introduction

Buildings, as prominent artifacts within urban settings, serve as vital indicators of the management, transformation, and overall dynamism of the built environment. Their physical characteristics, including geometry, height, function, material, condition, and style, are the key parameters that not only support sustainable urban development but also reflect economic progress and cultural evolution over time (Biljecki et al., 2021). Such rich building-level data has been instrumental in a range of applications, such as urban climate simulations for improved environmental planning (Creutzig et al., 2019), building energy modeling for resource optimization (Kumar et al., 2018; Roth et al., 2020), estimation of urban material stocks for the circular economy (Raghu et al., 2023), and disaster impact assessments to inform effective response and recovery efforts (Westrope et al., 2014). Moreover, these data support more nuanced analyses of population distributions (Schug et al., 2021), socio-economic conditions (Feldmeyer et al., 2020), as well as deeper understanding of the impact on human behaviors (Wang et al., 2016) and public perception (Liang et al., 2024). Hence, more comprehensive and openly accessible geospatial data on building can enable the formulation of nuanced urban planning policies, fostering locally informed and globally connected approaches to efficiently support urban resilience and sustainability (Elmqvist et al., 2019).

---

*Corresponding author

Traditionally, obtaining building attributes has involved expert evaluation, government records, or crowdsourced labeling, which often require field studies. This approach limits coverage and efficiency, leaving many buildings without detailed information. Although platforms like OpenStreetMap (OSM) and government databases now contain diverse urban information, the incompleteness and uneven geographical distribution of global building features hinder their usability across larger regions (Biljecki et al., 2023; Milojevic-Dupont et al., 2023; Lei et al., 2023; Herfort et al., 2023; Florio et al., 2025). With their rapid development, remote sensing-based methods have become a standard approach for extracting building information from aerial and satellite imagery, including attributes such as building height (Wu et al., 2023b; Frantz et al., 2021), and types (Du et al., 2015; Zhao et al., 2019). Remote sensing provides broad coverage, reduces reliance on ground surveys, and enables high-resolution tracking of urban changes over time. In parallel, machine learning methods that leverage geometric and built environment information have been widely applied to enhance the coverage and accuracy of building data (Roy et al., 2023; Nachtigall et al., 2023; Lei et al., 2024; Wang et al., 2024c). Despite the advancements, the top-down perspective poses inherent challenges, as critical vertical details of structures remain difficult to capture from overhead imagery.

The emergence of easily accessible Street View Imagery (SVI) has transformed the way buildings are analyzed, providing a ground-level, bottom-up perspective that captures architectural details often obscured in aerial or satellite imagery (Biljecki and Ito, 2021; Gaw et al., 2022; Zhang et al., 2024a). Leveraging this capability, numerous studies have integrated deep learning with SVI to extract and profile various building attributes, including height (Yan and Huang, 2022; Fan et al., 2024), type and usage (Kang et al., 2018; Zhao et al., 2021; Ramalingam and Kumar, 2023), architectural style (Lindenthal and Johnson, 2021; Sun et al., 2022b), and facade materials (Xu et al., 2023; Raghu et al., 2023; Chen et al., 2024a). Beyond building profiling, these integrations also support a range of practical applications, including risk assessment (Pelizari et al., 2021; Wang et al., 2021), refinement of 3D building models (Zhang et al., 2021), and building energy efficiency estimation (Sun et al., 2022a; Mayer et al., 2023). These advancements have significantly contributed to SVI-based urban studies, enabling fine-grained, large-scale geospatial analyses.

Despite advancements in SVI-based methods for inferring building attributes, various

challenges limit scalability and adaptability: (1) existing datasets struggle with uncertainty due to limited angular coverage in perspective views or distortions in panoramic images, hindering comprehensive observations; (2) reliance on proprietary data restricts accessibility, transparency, and adaptability, with ambiguous licensing further limiting research utility and inclusivity (Helbich et al., 2024); (3) while some efforts align visual data with geolocation, annotations often focus on isolated attributes, requiring separate models. Multi-task learning has been explored (Chen et al., 2022), but class diversity remain constrained, limiting the ability to capture architectural complexity for more inclusive and interpretable analyses. Consequently, SVI-based building datasets offering holistic structural perspectives, fully open data, and comprehensive architectural insights remain scarce.

Vision-language models (VLMs), uniting computer vision (CV) and natural language processing, have demonstrated the ability to interpret complex visual relationships, reason about scenes, and generate semantically rich descriptions (Li et al., 2024a). In the remote sensing domain, vision-language tasks have demonstrated promise for multi-scale feature understanding, multi-task learning, and applications such as visual question answering, image captioning, and semantic segmentation (Zia et al., 2022; Hu et al., 2025; Dong et al., 2024; Wang et al., 2024a). More recently, multimodal large language models (MLLMs) have advanced these capabilities by integrating deep contextual and semantic representations learned from massive, multimodal datasets, thereby enabling more nuanced and precise interpretations of visual data. This versatility highlights their potential to serve as foundational instruments in SVI-based building research, by enhancing the characterization of building properties, streamlining multi-task learning, and transcending predefined label sets in the analysis of facade features.

To advance fine-grained, bottom-up observations of buildings, we propose an open framework, OpenFACADES, that enriches a variety of building properties from a street-level perspective by leveraging multimodal crowdsourced inputs and open-source MLLMs. First, we utilize open-source building footprints and SVI to perform visibility simulations that geospatially align building geometries with corresponding SVI shooting locations. Second, we introduce an innovative pipeline that detects individual buildings based on their visible angles and acquires holistic building images using a custom image reprojection method. Third, we assemble one of the largest global, multi-attribute building image

datasets by combining crowdsourced building attributes with high-quality text descriptions generated by state-of-the-art MLLMs. Leveraging this dataset, we are among the first to introduce tailored MLLMs for building profiling through multi-task learning, encompassing both single- and multi-attribute prediction tasks as well as open-vocabulary captioning. Furthermore, we present an in-depth comparative analysis of model performance across various hyperparameter settings, cross-city generalization scenarios, and image quality variations.

The primary contributions of this work are threefold:

- Developed a reproducible methodology that (1) geolocates, detects, and acquires holistic building images from crowdsourced SVI; (2) integrates these images with crowdsourced building data to create an open and structured building image dataset; and (3) enables future scalability by dynamically retrieving the latest available data from these sources.

- Compiled an open global building dataset, consisting of (1) 31,180 individual building images from seven cities across three continents, annotated with attribute labels from OSM and text descriptions generated by ChatGPT-4o; and (2) large-scale automated annotations on 1.2 million images covering over half a million buildings. Each image is linked to its geospatial location and enriched with diverse attributes (e.g., building type, number of floors, age, and surface material) along with detailed textual descriptions. This forms the OpenFACADES dataset, one of the largest such resources, spanning multiple urban morphologies.

- Introduced the first benchmark open-source MLLMs that (1) perform multi-attribute prediction on buildings, achieving robust and more accurate image labeling performance than zero-shot ChatGPT-4o; (2) generate descriptive captions on architectural features, providing comprehensive information beyond standard building attributes; and (3) demonstrate enhanced robustness and generalizability relative to prior CV models.

In summary, this work presents a comprehensive and reproducible framework that leverages multimodal crowdsourced data to develop a global street-level building dataset for training multimodal models. This approach enhances the scope, adaptability, and ac-

curacy of urban analysis, enabling more detailed and interpretable assessments of the built environment.

2. Related work

2.1. Existing street-level building datasets

With advances in geospatial artificial intelligence technologies, research in recent years has increasingly leveraged remote sensing datasets such as high-resolution satellite and aerial imagery, and LiDAR to enhance urban applications. These datasets enable object-based image analysis, pixel-based classification, and semantic segmentation of urban structures, providing critical insights for land use mapping, urban morphology analysis, and spatiotemporal change detection. As key urban components, buildings have spurred the creation of domain-specific datasets and methodologies to support applications such as urban sustainability evaluation through rooftop attributes extraction (Wu and Biljecki, 2021), infrastructure management via automated land cover classification (Boguszewski et al., 2021), and disaster management through assessing damage (Gupta et al., 2019; Li et al., 2025a).

SVI, rapidly emerging as a prominent proximal remote sensing data source, has been leveraged to generate spatially enriched urban datasets that facilitate fine-grained semantic understanding of complex urban scenes (Biljecki and Ito, 2021). Among these, building-centric SVI datasets enable facade-level feature extraction, offering images that capture textural, material, and architectural features of building exteriors for environmental modeling. For example, building age and architectural style have long been studied for their links to building thermal performance (Tooke et al., 2014; Aksoezen et al., 2015; Nouvel et al., 2017) and real estate pricing (Zietz et al., 2008; Lindenthal and Johnson, 2021). Recent advances include the work of Sun et al. (2022b), which applies deep convolutional neural networks (CNNs) to classify buildings in Amsterdam, the Netherlands, into architectural periodization categories (e.g., revival, postwar). Material characterization (Xu et al., 2023; Chen et al., 2024a), another aspect critical for building energy simulation (Nouvel et al., 2017), also supports circular economy objectives by enabling lifecycle material tracking (Raghu et al., 2023) and risk assessment (Wang et al., 2021). Among these

6

efforts, Raghu et al. (2023) employ a multi-city material categories (brick, stucco, etc.) using geotagged SVI perspective views, aligning visual patterns with ground-truth material information for scalable building classification. Combining the aspects of building age and material, Ogawa et al. (2023) introduced a method to detect and geolocate buildings from panoramic images, automatically annotating them with objective building data in Kobe, Japan.

Furthermore, building type or usage, a critical attribute in urban remote sensing and land use classification, is also central to street-level research (Kang et al., 2018; Zhao et al., 2021; Lindenthal and Johnson, 2021; Ramalingam and Kumar, 2023; Li et al., 2025b). A seminal work by Kang et al. (2018) introduces the BIC_GSV dataset, a multi-city geospatial database of 19,658 SVI-derived building facades categorized into eight classes (e.g., apartment, church, garage, etc.) across North America. These ground truth labels are generated through view-direction-aligned spatial joins with OSM building footprints, enabling parcel-scale urban pattern analysis. Advancing this, Zhao et al. (2021) developed the BEAUTY dataset, which extends BIC_GSV by incorporating both SVI-based land use classification (e.g., residential, commercial, etc.) and multi-class building detection. Other similar research frameworks have also been applied to large-scale urban studies, integrating additional building attributes such as floor number estimation, abandoned house detection, and seismic risk assessment (Iannelli and Dell'Acqua, 2017; Zou and Wang, 2021; Rosenfelder et al., 2021; Pelizari et al., 2021; Ghione et al., 2022). These workflows not only enable location-based building retrieval but also demonstrate cross-modal alignment of SVI with open geospatial building footprints.

However, several challenges still remain in street-level building research, limiting the scalability and adaptability of current approaches. First, although many efforts have aligned visual information with building geolocation (Kang et al., 2018; Sun et al., 2022b; Ogawa et al., 2023), they are often either reliant on perspective views with restricted angular coverage, limiting visibility of upper building elements, or on panoramic images prone to severe distortions, misaligning with actual observations. Second, while various SVI-based building datasets have been established, their dependence on data derived from proprietary platforms introduces limitations related to accessibility, transparency, and adaptability. The ambiguous licensing terms of such datasets further constrain their utility

for diverse research applications and compromise the integrity of work built upon them, thereby hindering inclusivity within the research community (Helbich et al., 2024). In a recent trend, crowdsourced SVI platforms have garnered attention in urban studies by producing diverse, publicly accessible imagery. Examples include annotating points of interest (Zarbakhsh and McArdle, 2023), image status (Hou et al., 2024), human perception (Yang et al., 2025), and road surface type (Kapp et al., 2025). Among these, Hou et al. (2024) curate a manually labeled dataset to assess 10 million crowdsourced SVIs from 688 cities, enriched with metadata such as platform, weather, and lighting conditions, while Kapp et al. (2025) utilize OSM tags and ChatGPT-4o to label and amplify underrepresented road surface classes, resulting in 9,122 labeled images. These initiatives illustrate the potential of crowdsourced data for broad, inclusive urban analyses.

2.2. *Vision models in urban analytics*

With the rapid development of deep learning techniques over the past decade, diverse methods have been developed to extract urban cues from visual information. In terms of building facade research, in particular, CNNs have been widely employed due to their strong feature representation capabilities. Among them, VGG, DenseNet, and ResNet have been extensively applied to achieve, or serve as benchmarks for, the accurate classification and evaluation of building functions (Kang et al., 2018), materials (Ghione et al., 2022; Raghu et al., 2023), architectural styles (Lindenthal and Johnson, 2021; Sun et al., 2022b; Ogawa et al., 2023), and human perceptions (Liang et al., 2024). Additionally, Vision Transformers (ViTs) have emerged as powerful alternatives, leveraging self-attention mechanisms to capture long-range dependencies in building images. Recent studies have demonstrated the effectiveness of ViTs in urban analytics, achieving state-of-the-art performance in material recognition, and construction period prediction (Raghu et al., 2023; Ogawa et al., 2023). Beyond that, hybrid models combining various model backbones have been further developed to consider multi-dimensional features as input, improving comprehensiveness and generalizability in multi-scale urban analysis (Huang et al., 2023; Jia et al., 2024; Fujiwara et al., 2024).

However, the annotation of building attributes remains a fundamental limitation in these approaches. Labels are often restricted to isolated attributes, such as building type or mate-

rial, necessitating the training of separate models for different objectives. While multi-task learning frameworks have been explored (Chen et al., 2022), class diversity and model scalability remain constrained. Moreover, annotation schemes are typically predefined and rigid due to the availability of data, preventing adaptation to unannotated or emergent building characteristics, such as mixed-use functions or hybrid architectural materials. This lack of multi-dimensional, context-aware labels significantly limits the ability to capture architectural complexity, hindering the development of comprehensive, inclusive, and interpretable approaches for building analysis.

Rapid advancements in LLMs offer new avenues for extracting nuanced insights about complex urban environments. Notably, VLMs combine visual and linguistic modalities, leveraging deep semantic reasoning to establish rich connections between visual concepts and textual descriptions (Wu et al., 2023a; Li et al., 2024a). Building on these capabilities, recent work in remote sensing demonstrates how VLMs can exceed traditional CV methods by producing more context-aware and human-like interpretations (Al Rahhal et al., 2022; Zia et al., 2022; Hu et al., 2025), thereby providing not only precise visual recognition but also a semantic understanding of objects and their relationships within complex environments. In terms of street-level building research, recent studies have explored the state-of-the-art models for automated building annotation. For example, Li et al. (2024b) employed ChatGPT-4o to generate structured multi-label annotations for buildings using SVIs across multiple cities. Similarly, Zeng et al. (2024) assessed the model's performance in zero-shot building age prediction, finding that ChatGPT-4 effectively estimates the construction period of buildings. However, deploying proprietary LLMs such as ChatGPT-4o at scale presents limitations. Model inference relies on API-based access, which incurs high computational costs, making large-scale applications financially and computationally restrictive, which also constrains the efficiency for fine-tuning, limiting their adaptability for domain-specific urban studies. To address these challenges, recent open-source initiatives have produced diverse series of LLMs, including Qwen-VL (Wang et al., 2024b), Llama (Dubey et al., 2024), and InternVL (Chen et al., 2024b), enabling greater customization and efficiency in downstream tasks. These models exhibit unified capabilities to process multi-dimensional inputs, generating context-aware descriptions informed by their pretraining on large-scale, diverse datasets. This capability holds significant potential for

advancing street-level urban analysis, as their ability to interpret human-centric observations closely aligns with how individuals perceive and contextualize the built environment.

Hence, we propose a reproducible methodology for integrating open-source multimodal building data from global cities into a comprehensive dataset, incorporating objective attributes and detailed captions. Table 1 provides an overview of existing SVI datasets related to building attributes, highlighting how our contribution addresses current limitations while significantly expanding the scale, scope, and dimensionality of SVI-based datasets for building-related research. This advancement not only enhances the accessibility and adaptability of building datasets but also paves the way for broader, more inclusive, and scalable applications in urban analytics.

Table 1: Characteristics of existing SVI-based datasets constructed for building-oriented CV and urban research applications, and the features of the dataset we established in this research (GSV: Google Street View).

Studies	Purpose		Lineage			Coverage		Category
	Task	Building attribute	Image source	Image type	No. of labeled images	No. of cities	Continent(s)	
BIC_GSV (Kang et al., 2018)	image classification	type	GSV	perspective	19,658	More than 30	North America	apartment, church, garage, house, industrial, office building, retail, roof (8 categories)
BEAUTY (Zhao et al., 2021)	image classification and object detection	type	GSV	perspective	19,070	More than 30	North America	*Image classification*: residential, commercial, public, industrial (4 categories); *Multi-class detection*: apartment, church, garage, house, industrial, office building, retail, roof (8 categories).
Lindenthal and Johnson (2021)	image classification	age	GSV	perspective	29,177	1	Europe	Georgian, early Victorian, late Victorian/Edwardian, interwar, postwar, contemporary, revival (7 categories).
Raghu et al. (2023)	image classification	surface material	GSV	perspective	985	3	Asia, North America, Europe	brick, stucco, rustication, siding, wood, metal, other (7 categories)
SVI4BuildingFunc (Li et al., 2025b)	object detection	type	GSV	panoramic	15,400	4	North America, Europe	varies by city (e.g., high residential, low residential, commercial, office, walk-up buildings, mixed-up buildings; 5 to 6 categories per city)
OpenFACADES	Image labeling and captioning	type, age, floor, surface material, feature description	Mapillary	individual building images	31,180	7	North America, Europe, Asia	*Type*: apartments, house, retail, office, hotel, industrial, religious, education, public, garage (10 categories); *Surface material*: metal, glass, brick, stone, concrete, wood, plaster (7 categories); *Age*: numeric value; *Floor*: numeric value.

3. Methodology

In this study, we introduce OpenFACADES, a comprehensive framework for acquiring building images from SVI and automatically annotating them with crowdsourced data. This framework facilitates the development of large multimodal models tailored for architectural attributes question-answering and captioning. The framework is structured into three main steps, as illustrated in Figure 1:

(1) Integrating multimodal crowdsourced data. Initially, crowdsourced SVI metadata and building data are collected for research areas. Then, isovist analysis is performed to simulate the theoretical angles of view (AOV) from each camera location to the target structures. SVIs with optimal visibility are then retrieved and filtered by image features to ensure that only high-quality candidates are retained for subsequent analysis.

(2) Retrieving building image data. Based on the geospatial AOVs simulated, we map the relative viewing angles and detect target buildings within the image space. This process enables us to precisely identify associate building information with their visual representations. Then, based on the coordinates of bounding boxes, building images are reprojected from panoramic to perspective view, generating holistic building images. These images further undergo a filtering process to identify high-quality and suitable building views.

(3) Establishing dataset and multimodal models. Building images with available crowdsourced data form a dataset with four label types: single-attribute label, single-attribute Q&A, multi-attribute Q&A, and captioning. Single-attribute labels are derived from building information, while single-attribute Q&A append those attribute labels to targeted questions, generating concise question-to-label pairs. Multi-attribute Q&A and captioning labels are generated using ChatGPT-4o, enabling detailed textual descriptions and structured annotations for comprehensive building attribute analysis. The last three label types are utilized to fine-tune vision-language models, enabling a versatile model for multi-attribute building labeling and captioning with enhanced contextual understanding.

3.1. Integrating multimodal crowdsourced data

The workflow of integrating multimodal crowdsourced data for building analysis is illustrated in Figure 2. The process includes:

01. Integrating multimodal crowdsourced data

Image data preparation (Mapillary)

Building data preparation (OpenStreetMap)

building	apartments
start_date	1905
......	

Isovist analysis + image selection

02. Retrieving building image data

Building detection

Image reprojection

Image filtering

03. Establishing dataset and multimodal models

Street-level building dataset

Single-word Q&A

Multi-class Q&A

Captioning

Open FACADES

Vision language models

Input image

Input text

InternViT-300M

Pixel Unshuffle

MLP Projector

Text Tokenizer

Dynamic High Resolution

InternLM2.5

An **Open F**ramework for **A**rchitectural **C**aption and
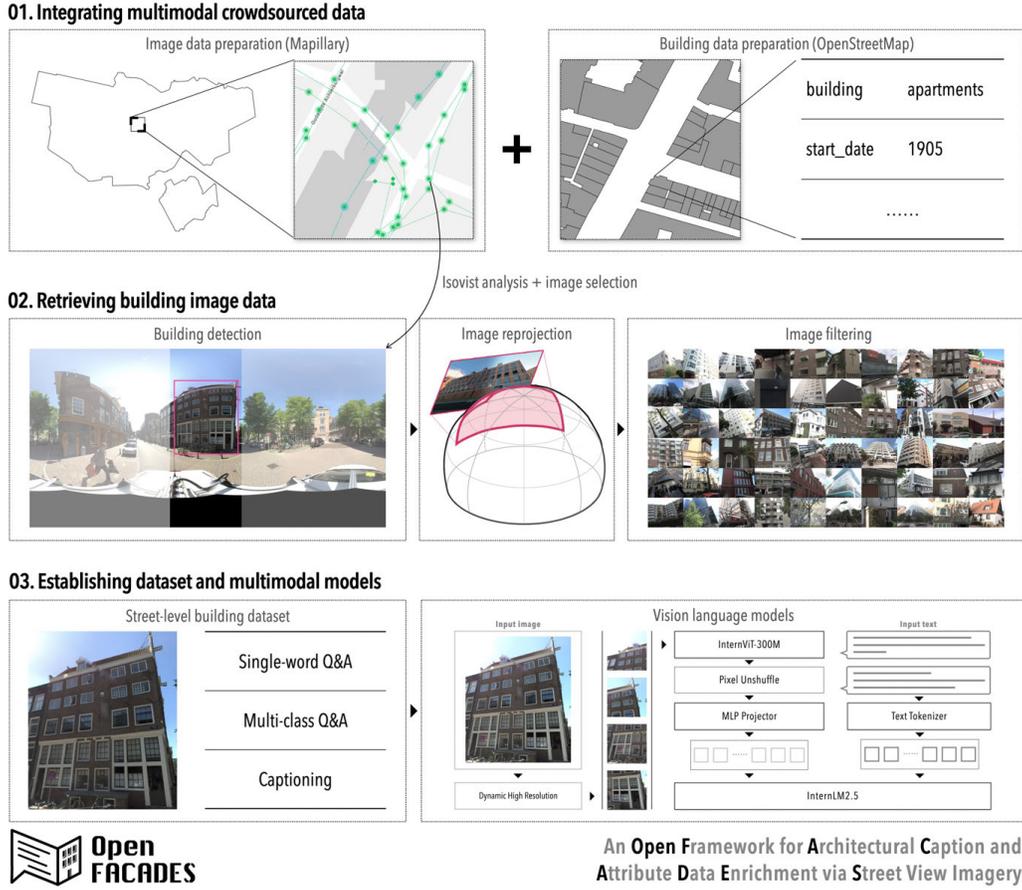**A**ttribute **D**ata **E**nrichment via **S**treet View Imagery

Figure 1: General framework for integrating multimodal crowdsourced data to establish a street-level building dataset and develop multimodal models. Data: (c) Mapillary and OpenStreetMap contributors.

Image data preparation. At the first stage, the raw metadata of street-level image data from crowdsourced platform is obtained within study areas before requesting the images. Here, Mapillary is chosen for its extensive global coverage, high-quality user-generated content, and open-access policies that enable reproducible and scalable urban research (Hou and Biljecki, 2022; Kapp et al., 2025; Danish et al., 2025). Specifically, the metadata, comprising image type (is_pano), location coordinates (computed_geometry), compass angle
(computed_compass_angle), capture time (captured_at), and quality indicator (quality_score), is utilized to structure sorting and quality assessments. Filtering op-

erations select panoramic imagery (`is_pano=True`), remove images captured outside the defined study area, exclude multiple images from the same spatial point to prevent redundant viewpoints, and discard those with poor resolution or quality defects. The output of this phase is a curated set of image metadata, with their corresponding unique image IDs, coordinates, and compass angles, prepared for subsequent spatial analyses.
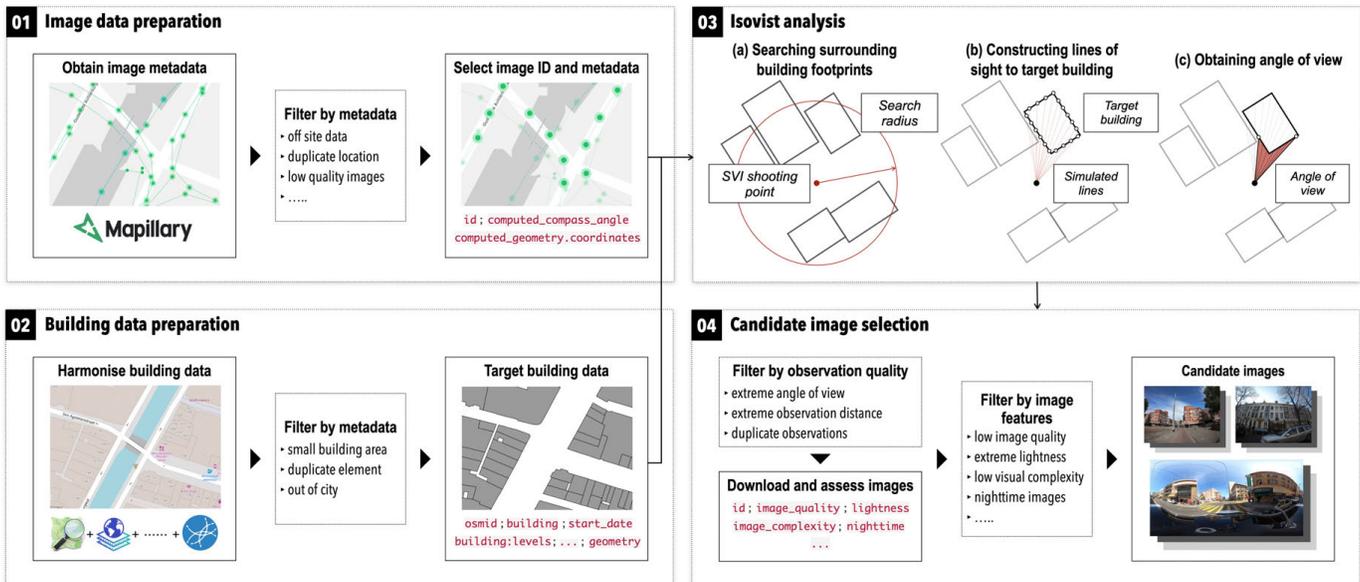

Figure 2: Workflow for obtaining and integrating suitable multimodal crowdsourced data, combining street-level imagery from Mapillary and building information from OpenStreetMap, along with external sources such as Overture Maps and government data, to harmonize building dataset. Data: (c) Mapillary and Open-StreetMap contributors.

Building data preparation.   In parallel, building geometries and associated metadata are retrieved from OpenStreetMap (OSM)[1]. Data harmonization is then conducted to supplement missing building footprints and insufficient building attributes from other data sources, such as Overture Maps[2] and government datasets. Attributes include unique identifiers, building type, facade material, number of floors, construction dates, and polygon

---

[1]https://openstreetmap.org/
[2]https://overturemaps.org/

13

geometries. Inconsistencies and outliers, such as footprints representing insignificant or extraneous structures (e.g., roof and underground structures), duplicates introduced by overlapping contributions, or buildings located outside the target region, are systematically removed. The remaining dataset delivers a precise, consistent, and high-quality representation of the built environment, ready for geometric calculations and alignment with the image data.

Isovist analysis. With both image and building datasets prepared, isovist analysis is applied to compute theoretical AOV from each camera location to the target structures, building on previous studies (Lindenthal and Johnson, 2021; Ogawa et al., 2023; Fan et al., 2025). This analysis identifies each building's perimeter segments that fall within the camera's potential field of view and evaluates the observation efficiency of buildings from specific vantage points. First, a search radius (e.g., 50 meter) is established to identify surrounding buildings from the SVI capture points. Second, sampling points are generated along the polygonal geometries of buildings within the distance threshold, and lines of sight are constructed towards all sampled points of the target buildings. Third, lines of sight intersecting with surrounding building footprints are filtered out, leaving only the largest angular span between the unobstructed lines, which represents the AOV to a building from a given image shooting point. Additionally, the left and right boundaries of the AOV are recorded as azimuth angles relative to the true north, providing detailed spatial orientation for subsequent tasks. This process identifies which buildings are potentially visible from each image capture point, thereby aligning the building information with the corresponding imagery metadata.

Candidate image selection. Based on the theoretical visibility of buildings, the final stage identifies candidate images most likely to provide reliable observations. Criteria derived from the absolute AOV values are used to eliminate images taken from excessive distances, extreme viewing angles, or redundant perspectives. The remained imagery IDs are then used to retrieve image data from Mapillary. Given that crowdsourced SVI can vary in quality and may contain various errors (Hou and Biljecki, 2022), these images are further assessed using quality metrics such as brightness, sharpness, and visual complexity

14

to determine their suitability for inclusion. Images captured under unsuitable conditions (e.g., nighttime, severe overexposure) or containing excessive visual clutter are removed based on the CV models released in NUS Global Streetscapes (Hou et al., 2024). The result is a curated set of candidate street-level images, optimized for integration with building data in subsequent object detection workflows.

3.2. Retrieving building image data

Figure 3 demonstrates the pipeline for extracting and selecting building images from street-level imagery. The process consists of three main steps: Building detection, image reprojection and image filtering.

Building detection. Azimuth angles derived from isovist analysis are first used to map the relative viewing angles of a building within the image space. This conversion defines a focused AOV for the target building before applying object detection. To determine the position of buildings within panoramic imagery, their relative horizontal coordinate ratios are computed as follows:

$$P_{\{l,r\}}^{n,i} = \frac{(A_{\{l,r\}}^{n,i} - H^i + C) \mod 360}{360} \tag{1}$$

where P, which ranges from 0 to 1, represents the left (l) or right (r) horizontal coordinates ratio of building n in the panoramic image i. The term H denotes the yaw angle when the SVI image token, and C is an adjustable calibration constant that ranges from (0-360), depending on the part of the image the view is oriented towards. Typically, C is set to 180 in Mapillary, indicating that the center of the image is the focal direction.

After determining the relative position of buildings in the SVI, images are cropped using the calculated horizontal coordinate ratios to isolate the AOV focused on the target buildings. Within the focused view, object detection is performed to identify buildings. To accomplish this, we employ GroundingDINO, a model equipped with pre-trained weights capable of detecting various objects using human inputs such as category names or referring expressions (Liu et al., 2023). Specifically, we use the "GroundingDINO-B" model checkpoint, which is trained on several widely-recognized object detection datasets, in-
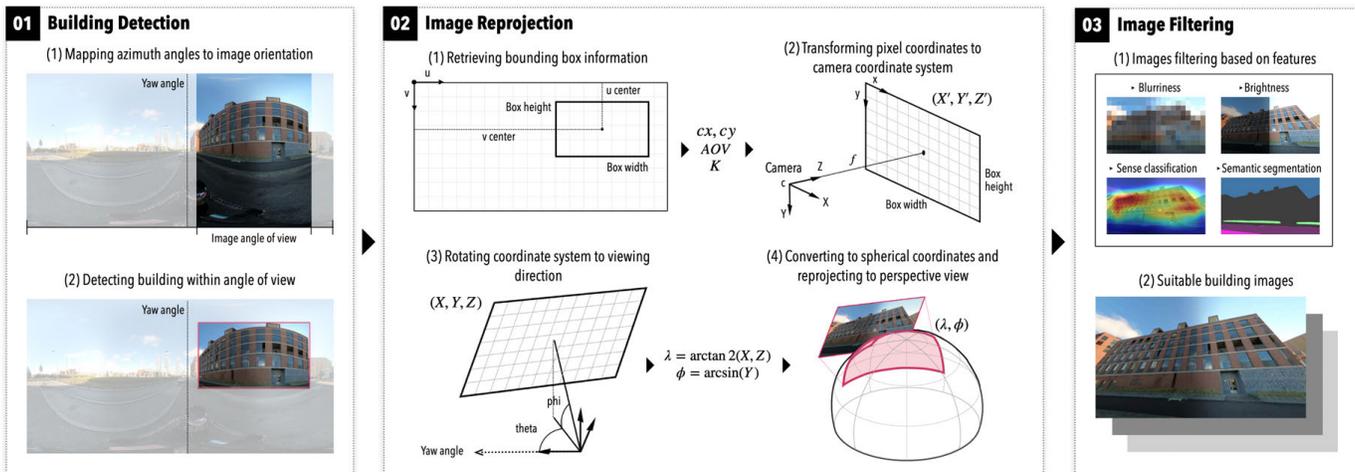
Figure 3: Pipeline demonstrating the extraction and selection of building images from street-level imagery, involving object detection, pixel coordinate transformation and reprojection, and feature-based filtering. Data: (c) Mapillary contributors.

cluding COCO, O365, and OpenImage. By assigning the category name "building" to this open-set detector, we generate bounding boxes around the buildings in each cropped image area. This process constrains the observation area to focus on each building footprint, enabling the association of visual observations with 2D building geometries. Additionally, it facilitates the object detection model in isolating target buildings from surrounding elements, such as adjacent structures and environmental noise.

Image reprojection. Panoramic images are formed by mapping the 3D environment onto a 2D sphere, which causes straight lines and familiar shapes to appear curved or distorted. After retrieving the bounding box information from the object detection, the reprojection process is designed to correct these inherent distortions. The objective of the reprojection is to take the portion of the panoramic image identified by the bounding box and present it as if it were photographed by a standard pinhole camera, providing a more intuitive and distortion-free representation of the detected object.

First, we interpret the bounding box region in terms of pixel coordinates within the panoramic imagery, obtaining the box center as (c_u, c_v), along with its *width* and *height*, which are essential for subsequent tasks. Second, a virtual pinhole camera model is con-

structed based on the specified AOV to a target building and the bounding box *width*. The focal length f and principal point (c_x, c_y) in camera coordinate are computed as:

$$f = \frac{\frac{width}{2}}{\tan\left(\frac{AOV}{2} \cdot \frac{\pi}{180}\right)} \tag{2}$$

$$c_x = \frac{width - 1}{2}, \quad c_y = \frac{height - 1}{2} \tag{3}$$

These values are used to construct the intrinsic camera matrix K, which encapsulates the intrinsic parameters of the virtual pinhole camera. For each pixel (x,y) in the virtual panel, the transformation from the 2D pixel location to a 3D direction in the camera's coordinate system is achieved by applying the inverse of K:

$$K = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} X\prime \\ Y\prime \\ Z\prime \end{bmatrix} = K^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{4}$$

where resulting vector $\mathbf{v}_{cam} = (X\prime, Y\prime, Z\prime)^T$ represents the direction of a ray emanating from the camera center through the corresponding pixel on the virtual image plane.

Third, to determine the approximate view direction of the bounding box, we use the center coordinates (c_u, c_v) of the bounding box in panoramic coordinate system and combined rotation matrix R to align the camera's direction to the rotated direction in 3D space:

$$\theta = (c_u - 0.5) \cdot 360, \quad \phi = (0.5 - c_v) \cdot 180 \tag{5}$$

$$R = R_x(\phi)R_y(\theta) \tag{6}$$

$$\mathbf{v}_{rot} = R\mathbf{v}_{cam} \tag{7}$$

where c_u and c_v are normalized to a range of $[0, 1]$, with c_u as the horizontal center and c_v as the vertical center of the bounding box region. The yaw angle θ defines the horizontal

rotation of the camera and spans from $-180°$ to $180°$. The pitch angle ϕ defines the vertical rotation of the camera and spans from $-90°$ to $90°$. The combined rotation matrix R is formed as the product of two individual rotation matrices based on Rodrigues' formula: $R_y(\theta)$, which rotates the coordinate system around the y-axis (yaw), and $R_x(\phi)$, which rotates the coordinate system around the x-axis (pitch). $\mathbf{v}_{cam}$ is the original direction vector in the camera's coordinate system, while $\mathbf{v}_{rot}$ is the new direction vector after applying the rotations, pointing toward the desired region of the spherical panorama.

Lastly, the rotated 3D direction vector $\mathbf{v}_{rot} = (X, Y, Z)$ is normalized and converted into spherical coordinates, where longitude λ and latitude φ are calculated based on:

$$\lambda = \arctan 2(X, Z), \quad \varphi = \arcsin(Y) \tag{8}$$

The corresponding pixel coordinates $(X_{\text{img}}, Y_{\text{img}})$ in the original panoramic image (equirectangular format) are then derived as:

$$X_{img} = \left(\frac{\lambda}{2\pi} + 0.5\right)(W_{\text{pano}} - 1), \quad Y_{img} = \left(\frac{\varphi}{\pi} + 0.5\right)(H_{\text{pano}} - 1). \tag{9}$$

At these coordinates, pixel values are sampled from the original panoramic image, and reprojected to generate the rectified perspective view using the `remap` function from OpenCV library. This transformation eliminates the spherical distortions inherent in panoramic imagery, producing a visually intuitive and geometrically corrected view aligned with the detected object. As examples demonstrated in Figure 4, this correction is crucial not only for preserving essential structural details for model interpretation but also for mitigating distortions that could otherwise misalign architectural features. This preprocessing step enhances the model's ability to accurately analyze building attributes.

Image filtering. The features of the detected individual building images are further analyzed to refine and enhance the image dataset. ZenSVI (Ito et al., 2025), an open-source library for street-level imagery analysis, is integrated into the framework to facilitate the extraction of image features. We analyze image features across four key dimensions: blurriness, brightness, semantic segmentation, and scene classification. These dimensions are utilized to identify high-quality and suitable building images for inclusion in the dataset.

Q: Analyze the building shown in the image and provide a detailed description of its architectural features. Then, describe the building type, the building's age, the primary facade material, and the total number of floors.

Perspective view

The building shown in the image is a multi-story structure with a classic architectural design. Let's break down its features in detail:
1. Facade and Structure:
  - The building has a symmetrical facade with a central section flanked by two wings.
  - The facade is composed of a combination of brick and stone materials, giving it a robust and timeless appearance.

2. Windows:
  - The windows are uniformly distributed across the facade, with a consistent grid pattern.
  - The windows are framed with stone or brick, maintaining the building's cohesive design.

3. Total Number of Floors:
-   **Number of Floors**: The building has a total of 6 floors.
......

Panoramic view

The building shown in the image is a multi-story structure with a distinctive architectural design. Let's break down its features in detail:
1. Shape and Structure:
  - The building is cylindrical in shape, which is quite unique compared to more common rectangular or square shapes.
  - It has a series of evenly spaced, arched windows that wrap around the entire circumference of the building.
......

4. Height and Layout:
  - The building has a total of five visible floors, with the top floor being the tallest.
  - The ground floor includes large commercial spaces, as evidenced by the large glass windows and doors.

Our method

The building shown in the image is a multi-story structure with a classic architectural design. Let's break down its features in detail:
1. Facade and Structure:
  - The building has a symmetrical facade with a central section that is slightly protruding.
  - The structure is composed of multiple floors, with a total of 8 visible floors.
  - The building is rectangular in shape.

2. Windows:
  - The windows are uniformly distributed across the facade, with a consistent window height.
  - The windows are rectangular and feature a grid pattern, with some having arched tops.
  - The windows on the ground floor are large and have a glass facade, while those on the upper floors are smaller and have a more traditional design.
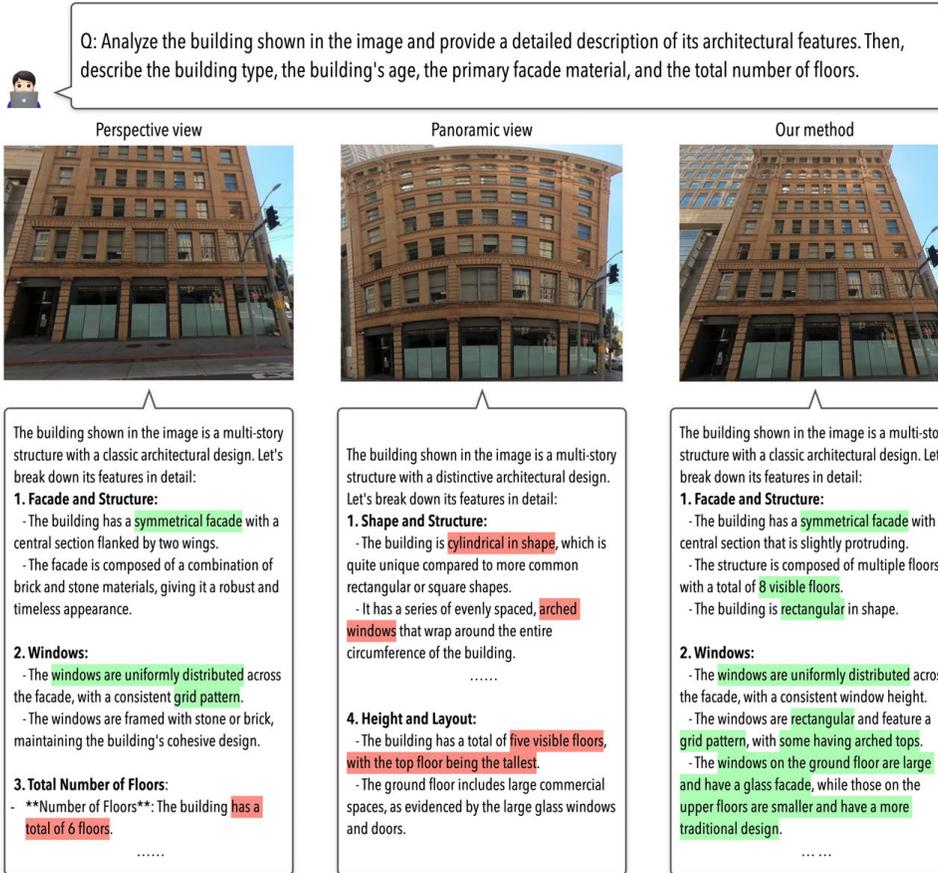... ...

Figure 4: Examples of different types of building images used as input to the vision-language model, resulting in varied responses. By generating a holistic view of individual buildings, our method facilitates a more authentic analysis and interpretation. Data: (c) Mapillary contributors.

Blurriness is evaluated using the OpenCV Laplacian operator to filter out images with motion blur or poor focus, while brightness assessment removes those with suboptimal illumination. A pre-trained Place365 model (Zhou et al., 2017) excludes indoor scenes, and semantic segmentation is applied to detect and minimize occlusions (e.g., trees, vehicles, walls), ensuring that selected images predominantly showcase building facades and maintain high visual quality.

3.3. Establishing dataset and multimodal models

Street-level building dataset. Following the previous process, building information is assigned to the detected buildings in the imagery. In this study, we focus on building type, facade material, construction year (age), and number of floors, which have been identified as essential attributes in prior studies and are supported by relatively sufficient data for model training and evaluation. Specifically, we utilize labels from building data corresponding to the categories: `building`, `building:material`, `start_date`, and `building:levels` in OSM alongside supplementary datasets, as mentioned in Section 3.1. Here, building type and facade material are treated as categorical variables, while construction year and number of floors are represented as numerical values.

From the full set of building data, we sample buildings with available category labels to construct the dataset for subsequent model development. The dataset is assembled and divided into training, validation, and test sets based on the following three principles: (1) ensuring sufficient labels across all classes to avoid biased predictive accuracy; (2) maintaining a balanced geospatial distribution across cities to represent the diversity of architectural designs; and (3) preventing the same building from appearing in both different sets to minimize data leakage.

The dataset contains four types of labels: single-attribute label, single-attribute Q&A, multi-attribute Q&A, and captioning labels. Single-attribute labels are used to fine-tune CV models, serving as the baseline for evaluating the performance of common practices. Single-attribute Q&A labels are derived from those single-attribute labels by appending the label to specific questions about the four building attributes, thereby generating concise question-to-label pairs based on building information. Multi-attribute Q&A and captioning labels are generated using the state-of-the-art multimodal large language model, ChatGPT-4o , through the OpenAI API[3]. This task involves prompting the model to annotate or describe the building features visible in the images, thereby creating an image-text dataset. Figure 5 provides detailed indication of data sources and examples of these labels.

Among these labels, the single- and multi-attribute Q&A outputs share a consistent, structured format constrained by predefined vocabulary, whereas the captioning task pro-

---

[3]https://www.openai.com/

vides a free-form textual description that often embeds the same attributes in more expressive language, forming a hierarchical relationship with the Q&A formats. For example, a building labeled as "brick" (material), "1920" (age), "3" (floors), and "house" (type) in the Q&A tasks might be described in the captioning output as "a three-story house with visible brickwork, built in the early 20th century", reflecting increased richness in narrative form. Inconsistencies can arise between OSM ground truth labels and ChatGPT-generated labels that slight discrepancies (e.g., differing building floor or age estimates) occur when ChatGPT assigns different labels to buildings with ambiguous features.

Image label

Single-attribute label *Source: OSM, government data*

building: apartments   building:levels: 4
start_date: 2006   building:material: NULL

Image-text pairs

Single-attribute Q&A   *Source: OSM, government data*

Q: Assign a brief label for the building type in the photo.
A: apartments
Q: Assign a brief label to indicate the construction year of the building in the image.
A: 2006

Multi-attribute Q&A   *Source: ChatGPT-4o*

Q: Provide concise labels for each category using the following JSON format.
Select appropriate values from the provided options for each category:
{"building_type": "(choose one option from: 'apartments', 'house', 'retail', 'office',
                'hotel', 'industrial', 'religious', 'education', 'public', 'garage')",
 "building_age": "(a 4-digit year indicating the construction date of the building)",
 "floors": "(a numeric value representing the total number of floors)",
 "surface_material": "(choose one option from: 'brick', 'wood', 'concrete', 'metal',
                'stone', 'glass', 'plaster')"}

A: {"building_type":"apartments",
    "building age":"2010",
    "floors":"5",
    "surface_material":"plaster"}

Captioning   *Source: ChatGPT-4o*

Q: Analyze the building shown in the image and provide a detailed description of its architectural features. Then, describe the building type, the building's age, the primary facade material, and the total number of floors in the building.
A: The image depicts a modern mid-rise residential apartment building located in a suburban or urban setting. The building is well-maintained and designed with contemporary architectural elements such as clean lines, neutral tones, and large glass surfaces.
1. Detailed Architectural Features:
Modern Style: The building exhibits a contemporary, minimalist design, characterized by clean lines, large windows, and the use of mixed facade materials.
                        ……
Window Alignment: Windows are symmetrically arranged in vertical lines, emphasizing verticality and a streamlined appearance.
2. Building Type: This is a residential apartment building, likely comprising multiple flats per floor, commonly found in urban or suburban areas.
3. Estimated Building Age: The structure appears to be relatively new, likely built within the last 10–20 years. The materials and style suggest post-2000 construction.
4. Primary Facade Material: The facade is primarily composed of rendered concrete or brick with painted finishes and glass for the balconies.
5. Total Number of Floors: The building has 5 floors: 4 full upper stories above the ground floor, which is partly visible at the base.

Figure 5: Different label types and data collection approaches for developing a street-level building dataset. Data: (c) OSM and Mapillary contributors.

Vision-language models. To address the limitations of traditional CV models in building attribute analysis, we leverage InternVL3 (Zhu et al., 2025), an open-source MLLM designed for unified visual-language reasoning. As depicted in Figure 6, InternVL3 is built on the "ViT-MLP-LLM" paradigm by integrating a scalable vision encoder (Intern-ViT) (Chen et al., 2024c), a multi-layer perceptron (MLP) projector, and a large language model (LLM). The vision encoder is InternViT-300M-448px-V2.5, a distilled variant of the 6B-parameter model optimized via dynamic high-resolution training and next token prediction (NTP) loss (Chen et al., 2024b). This architecture processes 448×448 pixel image tiles through a pixel unshuffle operation, reducing 1024 visual tokens to 256 for efficient cross-modal alignment.

The model is selected for its general-purpose captioning and open-vocabulary classification capabilities, critical for capturing the multifaceted attributes of buildings (e.g., material, style, type) within a unified framework. Unlike conventional models restricted to predefined labels, InternVL's contrastive vision-language pretraining enables semantic reasoning over diverse facade characteristics, aligning with our goal of holistic building profiling. Full-model tuning is conducted through optimizing three components (Figure 6): (1) InternViT-300M Vision Encoder: Retrained on street-level building images to enhance facade feature extraction, leveraging dynamic high-resolution (448px) inputs; (2) MLP Projector: Adjusted to align building-specific visual tokens with textual embeddings in the LLM space; (3) LLM Head: Fine-tuned using the corpus of building characteristic descriptions to generate structured captions. Here, we recast façade profiling as a multi-attribute text prediction task: all attributes and captions are encoded in a fixed-template prompt, and the MLLM is fine-tuned to generate that structured output. A single token-level cross-entropy loss implicitly handles both attribute prediction and free-form captioning, without requiring custom loss functions.

4. Experimental settings

4.1. Implementations

To implement methods mentioned in Section 3.1, we manually select cities that have a sufficient number of panoramic images available through the Mapillary online inter-
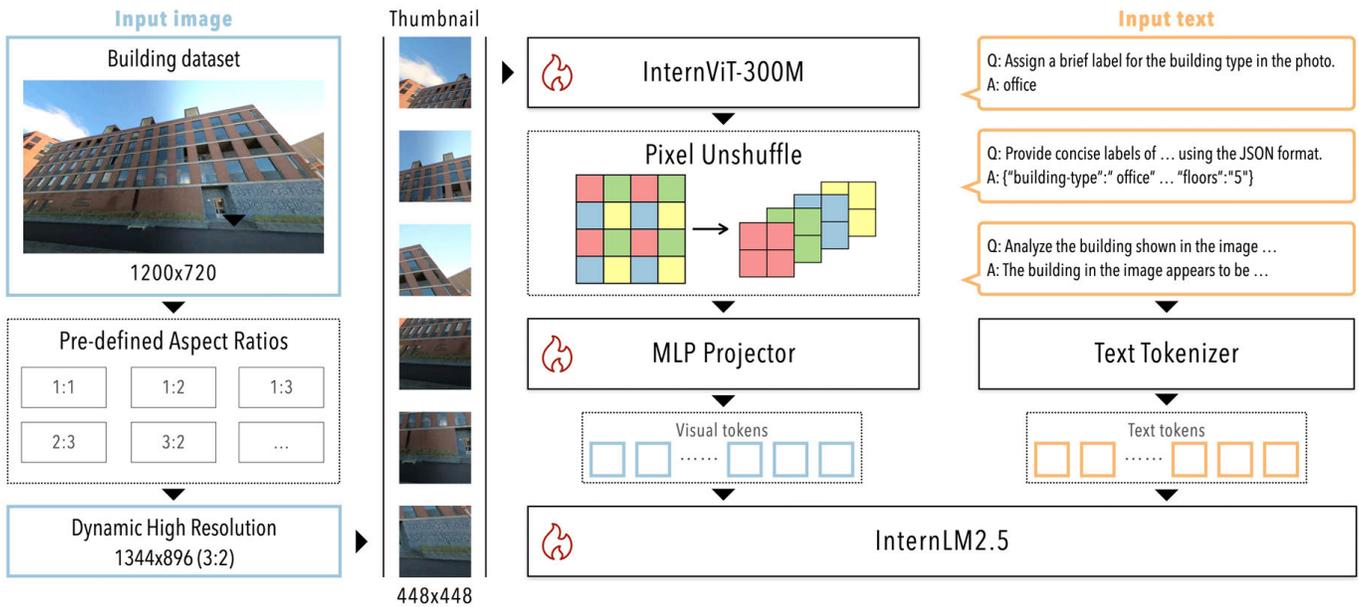
Figure 6: The overall framework of the InternVL series model architecture for building-centric tasks. Data: (c) Mapillary contributors.

face, and that also have a considerable amount of objective building attributes openly. Ultimately, seven cities from three continents are chosen, including Amsterdam, Berlin, Helsinki, San Francisco, Washington D.C., Houston and Manila, balancing the dataset across both selection aspects. Among them, Helsinki is selected due to its rich availability of building material data from the Buildings in Helsinki data[4], while Amsterdam provides diverse data on building age, to add sufficient data on according aspects.

The metadata for panoramic SVIs is first downloaded within the defined city boundaries using the Mapillary Python Software Development Kit[5], while building data is retrieved using OSMnx (Boeing, 2017) and Overture Maps. Subsequently, the data undergoes the process described in Section 3.1 to calculate the angle of view, evaluate observation quality, and identify candidate images. These selected images are then utilized for building detection, image reprojection, and filtering, as detailed in Section 3.2, resulting in a collection

---

[4]https://hri.fi/data/en_GB/dataset/helsingin-rakennukset
[5]https://github.com/mapillary/mapillary-python-sdk

23

Table 2: Image filtering criteria used for selecting building images in this study.

Feature	Description	Selection Rule
Mapillary metadata		
Image Type	Select only panoramic images	`is_pano = True`
Location (Duplicate)	Remove images taken at the same location	Filter duplicates based on coordinates
Image Quality	Use Mapillary quality score $\in$ [0, 1] to retain images with valid or no score	`quality_score ≥ 0.2` or `quality_score = 0`
Building metadata		
Building Area	Retain buildings with sufficient spatial footprint	`area > 20 m`^{2}
Underground	Exclude buildings fully underground	`building:levels ≥ 0` (OSM) or `is_underground = False` (Overture)
Isovist analysis		
Angle of View (AOV)	Exclude extreme observation angles	`10° ≤ AOV ≤ 120°`
Observation Coverage	Ensure valid building visibility within search radius	At least one building within 30m
Max. Images per Building	Cap the number of images per building	$\leq$ 5 images/building
Image-based filters		
File Size	Remove corrupted or very small images	`size ≥ 20 KB`
Semantic Segmentation	Check for sufficient visible building surface	`building_ratio ≥ 0.2`, `wall_ratio ≤ 0.3`, and `vegetation_ratio ≤ 0.75`
Scene Classification	Exclude indoor or irrelevant content	`environment_type = outdoor`
Blurriness	Estimate clarity via Laplacian variance	`blur_score ≤ 30`
Brightness	Filter overly dark or overexposed images	`20 ≤ brightness ≤ 200`

of individual building images for each city.

To ensure the collection of high-quality building images, the filtering criteria we applied are detailed in Table 2. We acknowledge that stringent filtering may inadvertently exclude images from certain regions, e.g., areas where dense greenery obstructs building facades. To address this and support broader applicability, we have made our code flexible, allowing users to adjust filtering thresholds to accommodate different urban contexts and mitigate potential geographic bias. Additionally, other image features, such as distance between the building and the vantage point, visibility coverage of the building perimeter, and other indicators derived from semantic segmentation and scene classification, are generated during the processing pipeline. These features can be further integrated into filtering criteria in future implementation.

4.2. Baselines

Models. To evaluate the effectiveness of large VLM, we compare it against a set of established CV architectures. These models span both classical CNNs and Transformer-based models to serve as baselines across various tasks. All models are implemented using the PyTorch library and initialized with pretrained weights from ImageNet-1K.

- **VGG**: A deep convolutional network characterized by its straightforward architecture of stacked convolutional layers with small receptive fields (Simonyan and Zisserman, 2014). We employ the VGG16 variant in this study, which has previously been applied to building classification tasks (Kang et al., 2018).

- **DenseNet**: Featuring dense connectivity between layers, DenseNet facilitates feature reuse and improves gradient flow while reducing the number of parameters compared to traditional CNNs (Huang et al., 2017). We adopt DenseNet201 as a benchmark model, which has demonstrated effectiveness in building material prediction (Ghione et al., 2022).

- **ResNet**: We include three variants, ResNet18, ResNet50, and ResNet101, which utilize residual learning through skip connections to enable deeper architectures (He et al., 2016). These models are widely used in building-related recognition tasks (Gouveia et al., 2024; Liang et al., 2024).

- **ViT**: Vision Transformers (ViT) divide images into fixed-size patches and use Transformer encoders to capture global context through self-attention mechanisms. Here, we evaluate ViT16, ViT32 and Swin Transformer (Swin_b), which have shown strong performance in previous building classification studies (Raghu et al., 2023; Ogawa et al., 2023).

Each CV model is fine-tuned separately for each building attribute. A systematic grid search is conducted over a range of learning rates (from 1e-7 to 1e-3), and training is performed for up to 64 epochs with early stopping based on validation performance. The checkpoint achieving the best validation metric (accuracy for classification, R-squared for regression) is selected for final evaluation on the test set and used for comparison with the VLM.

Evaluation Metrics. To comprehensively evaluate model performance across different tasks, we adopt a set of widely used metrics, tailored to the respective output types: categorical labels, numerical values, and text.

For tasks such as predicting building type and surface material, we report four standard classification metrics: Accuracy (Acc), macro Precision (Pre), macro Recall (Rec), and macro F1-score (F1). These metrics have been commonly used in prior works on building classification (He et al., 2024; Sun et al., 2022b; Liang et al., 2024) to evaluate model effectiveness. For continuous attributes such as the number of floors and building age, we use R^2, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE). These metrics are standard in evaluating building property predictions (Lei et al., 2024; Wang et al., 2024c). To enable learning from numerical labels, we convert construction year into building age by subtracting the year from 2025 (i.e., *age = 2025 – year*).

Furthermore, the text generated by ChatGPT-4o on the test set serves as a baseline to evaluate performance improvements in image captioning. In this study, we adopt three commonly used metrics in natural language generation: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE_L (Lin, 2004). BLEU evaluates n-gram precision, while METEOR considers precision, recall, synonym matching, and paraphrase alignment to capture semantic relevance more effectively. ROUGE_L measures the longest common subsequence between generated and reference texts, highlighting fluency and textual overlap. These metrics are widely adopted in remote sensing vision-language tasks (Li et al., 2024a), providing a comprehensive assessment of the accuracy and quality of generated descriptions.

4.3. Generalizability and robustness

To align more closely with real-world building profiling practices, we further conduct comparative assessments involving both VLMs and CV models of their generalizability and robustness. These comparisons target two main objectives: (1) assessing their generalizability to unseen data, and (2) evaluating robustness to heterogeneous noises and degradation.

Generalization to unseen city. Generalizing CV models to unseen cities remains a significant challenge due to the diverse and unique architectural features across cities (Sun et al., 2022b). VLMs, pretrained on extensive, high-quality image-text datasets, demonstrate promising potential to overcome these limitations by leveraging their pre-acquired semantic reasoning and contextual understanding capabilities. To investigate this potential, we conducted an experiment on building imagery collected from Brussels, comparing the performance of established CNN and ViT architectures against our fine-tuned VLM. For this evaluation, we retrieve a dataset of 3,687 labeled building images by integrating OSM attributes with buildings detected from Mapillary SVI. The dataset is composed of 3,348 images for building type, 186 for surface material, 1,234 for the number of floors, and 106 for building age. The supplementary material provides the detailed class distribution of this dataset. Due to the limited number of surface material labels across classes, we focus our analysis on the remaining three attributes.

Robustness to varying image quality. Crowdsourced SVI, unlike standardized remote sensing imagery, frequently exhibit diverse quality issues (Hou and Biljecki, 2022). To systematically evaluate the robustness and stability of the model under such conditions, we adopt the methodology proposed by Hendrycks and Dietterich (2019), which assesses model performance in the presence of common image corruptions and perturbations. Following the image quality criteria defined by Hou and Biljecki (2022), we algorithmically generate four types of corruption to the test set: occlusion, motion blur, Gaussian noise, and brightness alteration, as illustrated in Figure 7. We acknowledge that these distortions are simulated rather than naturally occurring, but they provide a controlled and reproducible way to benchmark model behavior under common visual degradation scenarios.

We then evaluate each model's performance under these degraded conditions using Relative Corruption Errors (*Relative* CE) (Hendrycks and Dietterich, 2019). First, the baseline error rate E_{clean}^m is determined for model m on the uncorrupted data. Next, we compute the error rate $E_{c,s}^m$ for each corruption type c at severity level s ($1 \le s \le 3$). In classification tasks (building type and surface material), the error rate is defined as $1 -$ *Accuracy*, whereas for numerical values (predicting number of floors and building age), it is defined as $1 - R^2$. Finally, to account for the varying difficulties introduced by each

corruption, we normalize these error rates by dividing by the ResNet50 baseline error. *Relative* CE is calculated as:

$$RelativeCE_c^f = (\sum_{s=1}^{3} E_{s,c}^f - E_{clean}^f)/(\sum_{s=1}^{3} E_{s,c}^{ResNet50} - E_{clean}^{ResNet50}) \tag{10}$$

This normalization provides a clearer measure of how much each model's performance declines under different corruptions. Averaging these *Relative* CE from four types of corruptions results in the *Relative m*CE, which represents the overall relative performance degradation when the models encountering corruptions.
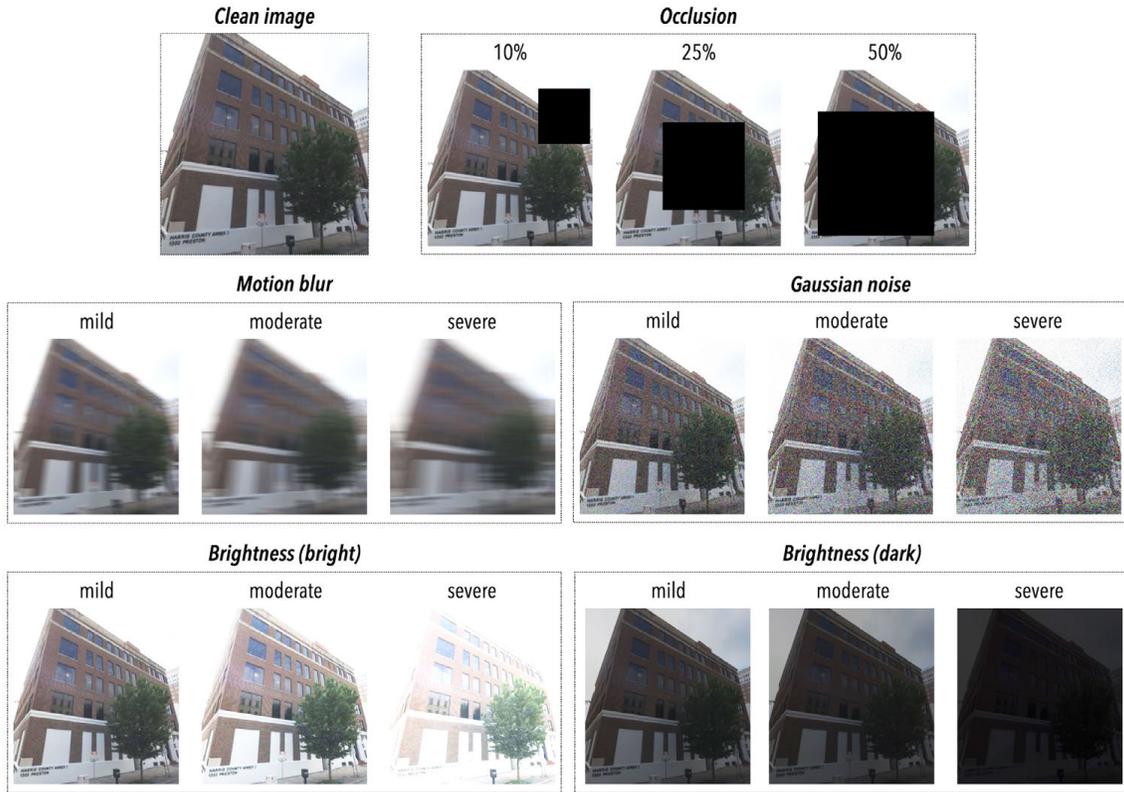
Figure 7: Examples of image corruption and perturbation for robustness experiments, consisting of four categories of algorithmically generated images based on common quality issues in crowdsourced imagery. Each type of corruption has 3 levels of severity (except for brightness which has twice 3 levels of severity), resulting in a total of 15 corruption levels. Data: (c) Mapillary contributors.

4.4. Ablation experiments

As one of the very first studies to apply fine-tuned open-source VLMs to the task of building profiling, we conduct a series of ablation experiments to explore and identify effective fine-tuning strategies.

Fine-tuning settings. To determine optimal training configurations, we first compared two fine-tuning strategies: full fine-tuning, where all model parameters are updated, and parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA). LoRA introduces trainable rank-decomposition matrices into the model while freezing the original weights, enabling more efficient training. This comparison allowed us to evaluate the trade-offs between flexibility, performance, and training cost under different optimization schemes.

We then performed a systematic grid search over learning rates and the number of training epochs. The default learning rate of 4e-5 is used as a baseline, and we evaluat additional values (e.g., 8e-6, 4e-6, and 4e-4) within a predefined range to observe their effects on model stability and performance. We also vary the number of training epochs from 1 to 5 while keeping all other hyperparameters fixed. For each setting, the model is evaluated on a held-out validation set.

Model size. We compare multiple model variants to assess how model capacity influences performance in the context of building profiling. Specifically, we evaluate InternVL3 models with 1B and 2B parameters, as well as InternVL2.5 model which has 4B parameters avaliable. These experiments are conducted to understand the scalability of different VLM sizes and to determine whether increase model complexity leads to substantial gains across all building profiling tasks.

Data size. To assess the impact of training data volume, we conduct an ablation study by fine-tuning InternVL3-2B on varying proportions of the complete image-text dataset. We evaluate performance on both attribute prediction and image captioning tasks. This analysis quantifies how model accuracy scales with additional data and offered insights into the marginal benefit of increased data volume in multimodal learning.

5. Results

5.1. Street-level building dataset

Table 3 provides a detailed breakdown of the number of buildings and SVI images retrieved, individual buildings detected with associated images, and the ratio of completeness for each city. While completeness varies among cities due to differences in the availability and quality of Mapillary images uploaded for specific locations, around 50% of buildings in city centers can be observed and analyzed.

Table 3: Summary of building footprints, image retrieval, and detection completeness across cities and regions in the building dataset.

City	Total building footprints	Total images retrieved	Total individual building images	Buildings with images	Percentage detected	City center completeness (2.5km×2.5km)
Europe						
Amsterdam	195,188	203,570	330,235	120,154	61.6%	83.6%
Helsinki	63,972	20,035	20,479	8,930	14.0%	42.5%
Berlin	497,703	408,166	287,065	137,930	27.7%	46.7%
North America						
San Francisco	160,659	62,521	91,874	34,510	21.5%	39.4%
Houston	399,883	304,030	238,934	91,774	23.0%	53.8%
Washington D.C.	161,190	269,420	201,955	86,144	53.4%	57.4%
Asia						
Manila	105,904	68,706	48,951	23,911	22.6%	22.7%
Total	1,617,019	1,414,288	1,219,493	503,353	31.1%	49.4%

As discussed in Section 3.3, building images with available attributes are sampled to construct a class-sufficient dataset for model development, resulting in a total of 31,180 images. Figure 8 illustrates the distribution of images across categories for each attribute, comprising 17,530 images for building type, 2,871 for surface material, 7,228 for floors, and 5,927 for age.

The dataset is split into training, validation, and test sets in a 6:1:3 ratio. As detailed in Table 4, the training set comprises 20,056 OSM-sourced labeled samples from 19,443
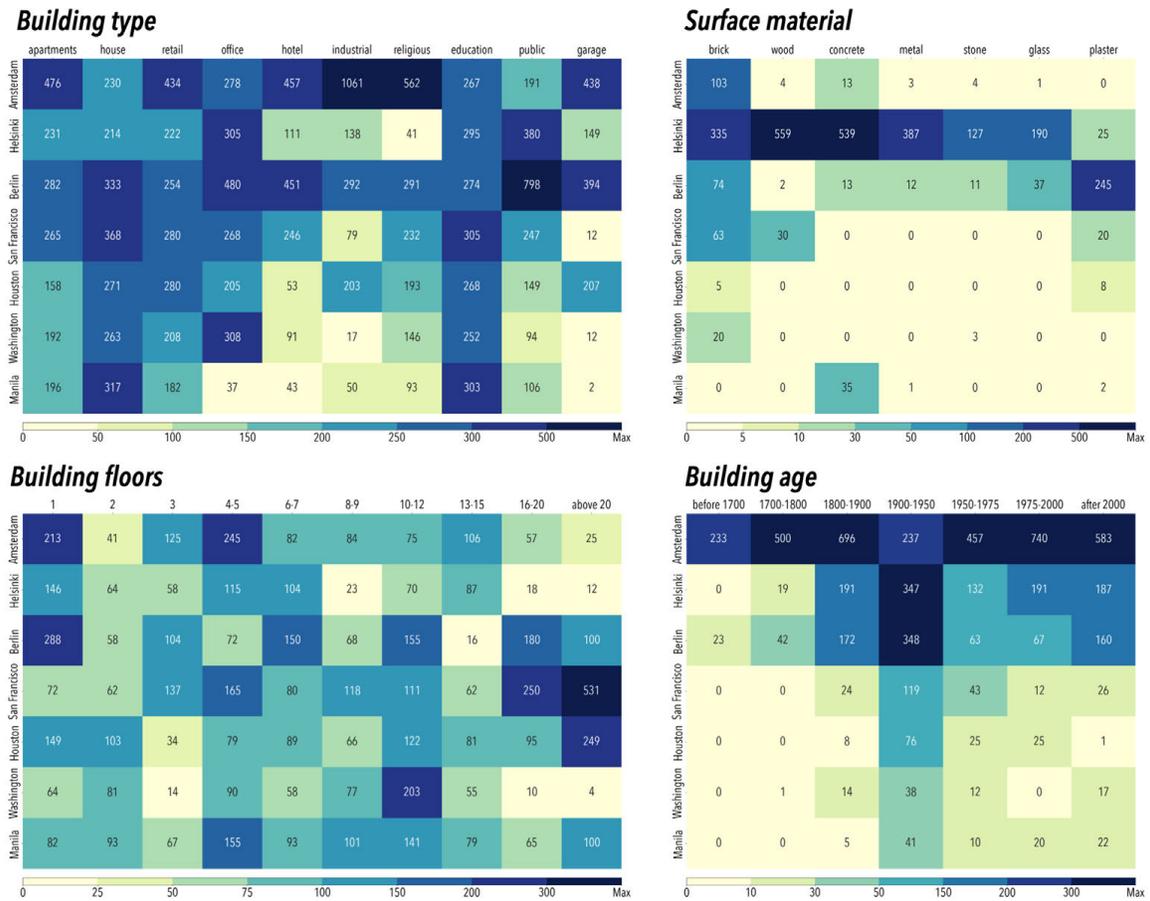
Building type

	apartments	house	retail	office	hotel	industrial	religious	education	public	garage
Amsterdam	476	230	434	278	457	1061	562	267	191	438
Helsinki	231	214	222	305	111	138	41	295	380	149
Berlin	282	333	254	480	451	292	291	274	798	394
San Francisco	265	368	280	268	246	79	232	305	247	12
Houston	158	271	280	205	53	203	193	268	149	207
Washington	192	263	208	308	91	17	146	252	94	12
Manila	196	317	182	37	43	50	93	303	106	2

0 50 100 150 200 250 300 500 Max

Surface material

	brick	wood	concrete	metal	stone	glass	plaster
Amsterdam	103	4	13	3	4	1	0
Helsinki	335	559	539	387	127	190	25
Berlin	74	2	13	12	11	37	245
San Francisco	63	30	0	0	0	0	20
Houston	5	0	0	0	0	0	8
Washington	20	0	0	0	3	0	0
Manila	0	0	35	1	0	0	2

0 5 10 30 50 100 200 500 Max

Building floors

	1	2	3	4-5	6-7	8-9	10-12	13-15	16-20	above 20
Amsterdam	213	41	125	245	82	84	75	106	57	25
Helsinki	146	64	58	115	104	23	70	87	18	12
Berlin	288	58	104	72	150	68	155	16	180	100
San Francisco	72	62	137	165	80	118	111	62	250	531
Houston	149	103	34	79	89	66	122	81	95	249
Washington	64	81	14	90	58	77	203	55	10	4
Manila	82	93	67	155	93	101	141	79	65	100

0 25 50 75 100 150 200 300 Max

Building age

	before 1700	1700-1800	1800-1900	1900-1950	1950-1975	1975-2000	after 2000
Amsterdam	233	500	696	237	457	740	583
Helsinki	0	19	191	347	132	191	187
Berlin	23	42	172	348	63	67	160
San Francisco	0	0	24	119	43	12	26
Houston	0	0	8	76	25	25	1
Washington	0	1	14	38	12	0	17
Manila	0	0	5	41	10	20	22

0 10 30 50 150 200 300 Max

Figure 8: The distribution of the building images categorized by objective building attributes—type, age, floor, and surface material—selected for each city in dataset.

Table 4: Summary of training, validation, and test data used in this study.

Data Type	Source	Train	Validation	Test
Image label				
Single-attribute label	OSM	20,056	2,737	9,190
Image-text pairs				
Single-attribute Q&A	OSM	20,056	–	–
Multi-attribute Q&A	ChatGPT-4o	19,443	–	9,016
Captioning	ChatGPT-4o	19,443	–	9,016
Total pairs		58,942	–	18,032

unique images (some buildings carry multiple attribute labels). Those same 19,443 images are used to generate 19,443 multi-attribute Q&A pairs and 19,443 descriptive captions via ChatGPT-4o (2024-08-06 release), yielding a total of 58,942 image–text pairs for VLM training. The held-out validation set guides checkpoint selection and hyperparameter tuning for both CV and VLM models based on their performance on multi-attribute prediction tasks. The test set, consisting of 9,016 images with ChatGPT generated Q&A and captions, is reserved for final model evaluation and comparison. To better capture ambiguous architectural features, we also include an auxiliary classification setup by introducing prompts such as "alternate_building_type" and "alternate_surface_material", which elicit the top-two predictions from the MLLMs.

We acknowledge that the current dataset has limitations, particularly in terms of geographic diversity across continents and the availability of data for certain attributes, such as surface material and building age. Nevertheless, to the best of our knowledge, this dataset is both large and comprehensive compared to previous efforts highlighted in Section 2. Additionally, the reproducible framework established in this study enables future expansion of the dataset as more building images and their associated attributes become available through crowdsourced platforms. This iterative refinement could progressively enhance the dataset's scope and utility for broader applications.

5.2. Model performance

5.2.1. General performance

After we fine-tuned the baseline models mentioned in Section 4.2, we compare their performance with that of the zero-shot ChatGPT and fine-tuned InternVL3-2B on four building characteristics, as summarized in Tables 5a and 5b. To facilitate a fair comparison, we integrate GPT-generated data to supplement the missing OSM data in the training set for the CV models.

The fine-tuned InternVL3-2B achieves the highest overall performance, particularly in predicting building type and number of floors. Although CV models slightly outperform the VLM in surface material classification and building age estimation, they require attribute-specific architectures and domain-specific tuning. In contrast, VLMs provide a unified and adaptable framework, delivering comparable or superior results across multiple prediction tasks. Notably, zero-shot ChatGPT also demonstrates strong capability across all tasks, achieving performance close to that of some fine-tuned CV models. This suggests its potential as a practical tool for supplementing building data when labeled samples are limited. This advantage of VLM can be attributed primarily to the semantic reasoning and contextual understanding inherent in pretrained VLMs, which significantly enhance their generalization across diverse tasks. Moreover, VLMs can simultaneously infer multiple target variables and implicitly model correlations among tokens associated with different tasks, whereas the CV baselines considered here are independently fine-tuned for each attribute. The combination of robust predictive performance and enhanced functionality makes VLMs a compelling alternative to traditional approaches in future applications.

5.2.2. Performance by cities

Figure 9 presents the performance of three model variants: (1) the InternVL3-2B model before fine-tuning and (2) after fine-tuning, as well as (3) a ChatGPT-4o reference baseline, on seven cities and four building attributes. In general, the fine-tuned InternVL3-2B outperforms its non-fine-tuned counterpart, showing consistent gains in predicting all four attributes. These improvements are particularly notable in Berlin and San Francisco, where building type, material, and floor performance improve substantially. Amsterdam and Helsinki also exhibit moderate but still positive gains for different tasks.

Table 5: Validation performance comparison among zero-shot ChatGPT, fine-tuned InternVL3-2B and CV models.

(a) Performance on classification tasks of building type and surface material

Attribute	Model	LR	Epoch	Acc (%)	mPre	mRec	mF1	Acc@2 (%)
Building type	ChatGPT (zero-shot)	-	-	<u>57.69</u>	<u>0.649</u>	<u>0.566</u>	<u>0.571</u>	<u>75.06</u>
	DenseNet	1e-5	14	53.42	0.540	0.529	0.529	69.39
	VGG	5e-4	9	47.66	0.492	0.470	0.462	62.03
	ResNet18	1e-4	19	50.59	0.509	0.499	0.493	66.28
	ResNet50	1e-4	24	52.20	0.526	0.512	0.504	65.98
	ResNet101	5e-5	13	53.58	0.543	0.527	0.524	68.52
	ViT16	5e-6	7	53.60	0.549	0.532	0.530	69.35
	ViT32	1e-5	5	52.24	0.519	0.515	0.507	67.93
	Swin_b	5e-6	17	56.11	0.562	0.558	0.554	73.14
	InternVL3-2B	8e-6	3	**61.27**	**0.661**	**0.602**	**0.609**	**77.31**
Surface material	ChatGPT (zero-shot)	-	-	65.41	0.581	0.614	0.553	79.07
	DenseNet	1e-4	5	65.92	0.575	0.610	0.566	81.45
	VGG	1e-5	12	57.02	0.504	0.529	0.494	74.06
	ResNet18	1e-5	23	63.78	0.560	0.587	0.554	79.57
	ResNet50	5e-5	30	65.79	0.601	0.610	0.578	81.08
	ResNet101	5e-5	25	66.67	<u>0.608</u>	0.627	0.582	79.70
	ViT16	5e-5	18	64.79	0.571	0.587	0.564	81.20
	ViT32	1e-5	13	60.15	0.538	0.554	0.511	77.07
	Swin_b	5e-6	33	**69.17**	**0.623**	**0.655**	**0.612**	**85.84**
	InternVL3-2B	8e-6	3	<u>68.05</u>	0.598	<u>0.634</u>	<u>0.588</u>	<u>81.70</u>

(b) Performance on prediction tasks of number of floors and building age.

Attribute	Model	LR	Epoch	R2 (↑)	MAE (↓)	MAPE (%) (↓)	RMSE (↓)
Number of floors	ChatGPT (zero-shot)	-	-	0.721	<u>2.36</u>	<u>38.66</u>	5.01
	DenseNet	5e-5	5	0.774	2.53	41.91	4.51
	VGG	5e-5	12	0.689	3.10	42.43	5.28
	ResNet18	5e-5	16	0.741	2.82	38.81	4.82
	ResNet50	5e-5	23	0.768	2.48	41.74	4.56
	ResNet101	5e-5	14	<u>0.777</u>	2.48	41.97	<u>4.48</u>
	ViT16	5e-6	6	0.765	2.63	46.47	4.59
	ViT32	1e-6	22	0.727	2.83	45.08	4.95
	Swin_b	1e-5	10	0.773	2.41	41.48	4.52
	InternVL3-2B	8e-6	3	**0.789**	**2.13**	**36.93**	**4.35**
Building age	ChatGPT (zero-shot)	-	-	0.645	31.63	<u>58.79</u>	57.07
	DenseNet	1e-5	21	0.720	31.31	78.36	50.68
	VGG	1e-4	7	0.559	42.55	121.31	63.59
	ResNet18	5e-5	23	0.714	32.63	74.44	51.23
	ResNet50	1e-4	18	0.721	32.26	85.92	50.55
	ResNet101	5e-5	18	**0.738**	<u>30.61</u>	78.77	**49.01**
	ViT16	5e-6	29	0.719	30.89	80.02	50.77
	ViT32	1e-5	23	340.675	33.82	71.89	54.61
	Swin_b	5e-6	27	<u>0.723</u>	32.72	93.77	<u>50.40</u>
	InternVL3-2B	8e-6	3	0.710	**28.06**	**58.27**	51.50

Figure 9: Model performance on building attributes across different cities before and after fine-tuning the InternVL3-2B model, compared to the baseline performance of ChatGPT-4o. Building type and surface material are evaluated using classification accuracy (Acc.), while number of floors and building age are assessed using R-squared (R2). "NA" indicates cities with insufficient data for model evaluation (ground-truth instances fewer than 20 in test set).

Despite the overall upward trend, improvement magnitude varies across cities and attributes, which may due to several reasons. First, the availability of diverse and distinct samples plays a crucial role: cities with a richer variety of building facades (e.g., Amsterdam, Berlin) yield more pronounced performance boosts. Conversely, location with more homogeneous or ambiguous building styles (e.g., Manila) shows relatively smaller gains. Second, crowdsourced labels in certain cities may be incorrect or insufficient, which can adversely affect the model's ability to learn reliable city-specific patterns, restraining potential performance gains.

Nevertheless, when benchmarked against the zero-shot ChatGPT-4o baseline, the fine-tuned InternVL3-2B model demonstrates generally competitive or superior performance. These results confirm that open-access VLMs can achieve near-state-of-the-art performance at no additional licensing cost once adequately fine-tuned on relevant datasets. This highlights the effectiveness of VLMs in predicting multiple building attributes across global cities, providing a cost-effective solution for a wide range of urban remote sensing applications.

35

5.2.3. Performance by categories

Figure 10 presents the confusion matrices illustrating the performance of our VLM on building type and surface material on different categorical labels. Overall, the model demonstrates robust performance for most classes. In terms of well-predicted labels, visually distinctive building types such as apartments and houses show consistently high accuracies. These categories often have defining features (e.g., apartment blocks characterized by uniform facades and repetitive windows) that the model effectively captures. Similarly, for surface material, high-frequency and visually salient classes like brick, wood, and glass yield strong performances. Conversely, certain labels are harder to classify, yielding relatively lower accuracies. For building type, hotel or public categories are frequently misidentified as office, suggesting significant overlap in their institutional architectural appearance (e.g., multi-stories, ordered elements). Likewise, plaster and concrete exhibit misclassifications due to shared grayscale tones and blank textures.
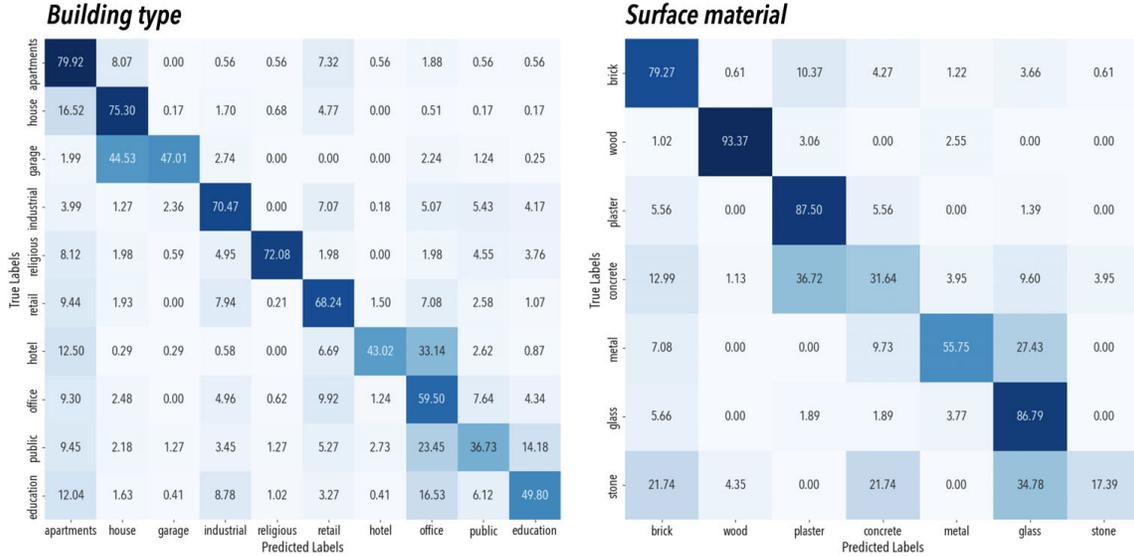
Figure 10: Confusion matrices illustrating the performance of the InternVL3-2B model on classifying different categories of building type and surface material. Darker cells indicate higher prediction accuracy (%).

Figure 11 illustrates the model's ability to predict the number of floors and building age. The heatmaps provide a detailed visualization of prediction accuracy relative to ground-truth values, with darker colors indicating higher frequencies of accurate pre-

dictions within each bin. For the number of floors (left), the heatmap demonstrates strong prediction performance for lower-rise buildings. As the number of floors increases beyond five, prediction performance start to decline. High-rise structures (stories beyond 15) exhibit more frequent errors, highlighting the challenge in distinguishing floor count for taller buildings from single viewpoint images.

For building age (right), the heatmap reveals stronger predictive performance for more recent buildings (post-1900), reflecting their typically more distinctive and recognizable architectural characteristics. Conversely, older buildings (pre-1900) exhibit larger errors, with significant overlaps between adjacent historical periods. The subtle external distinctions, combined with modifications such as renovations and retrofits, contribute to the difficulty in accurately classifying older structures based solely on visual appearance (Sun et al., 2022b). Additionally, the model's performance is likely influenced by the limited representation of older building periods within the original training dataset. Nevertheless, the overall R-squared above 0.7 confirms the model's capability in capturing broad temporal patterns of floor count and building age from images.

Taken together, these results suggest that the model is capable of inferring attributes across various ranges. However, inherent visual ambiguities, particularly among structurally or stylistically similar categories, contribute to overlaps in predictions. Enhancing the quality, diversity, and coverage of crowdsourced data would be a valuable step toward improving the dataset and the model's performance.

5.3. Generalizability and robustness

Generalization to unseen city. CV models with good performance are selected for experiments in this section. Tables 6a and 6b show that the VLM model exhibits superior generalizability compared to commonly used computer vision models, particularly in the tasks of building type and age prediction. This improved performance can be attributed to the VLM's ability to leverage semantic reasoning and contextual understanding from large-scale image-text pretraining, allowing it to adapt more effectively to diverse architectural features. While the performance on number of floors prediction is comparable across models, the lower accuracy of Swin_b on other attributes suggests a reliance on localized visual features, which limits its generalization capability. Despite these strengths, the VLM's low
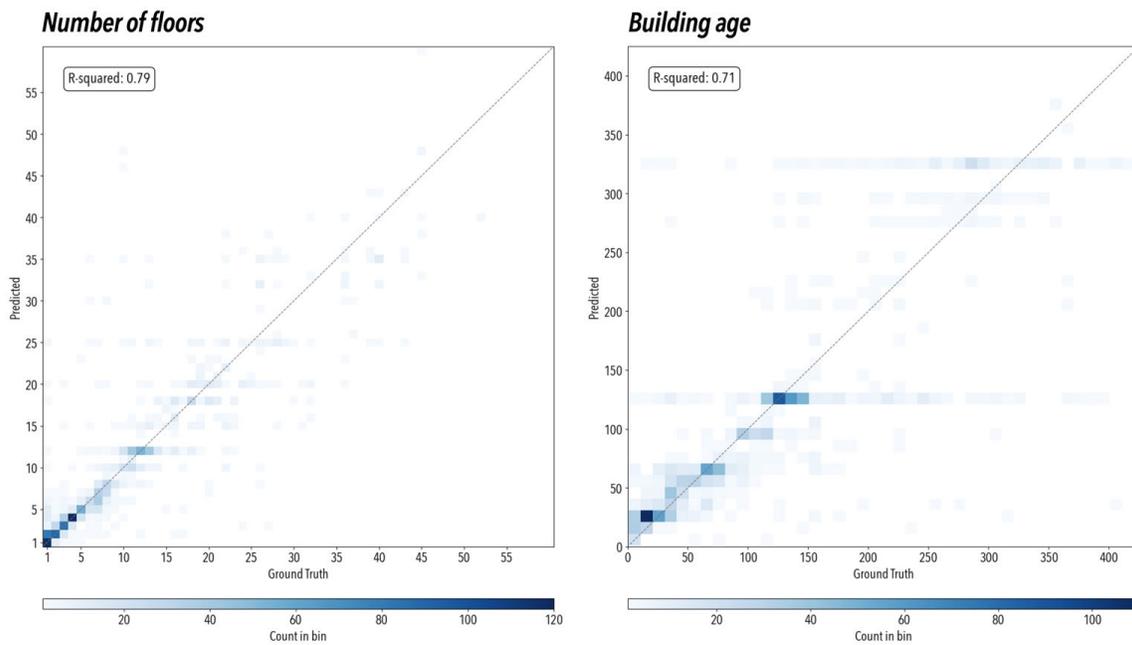
Figure 11: Heatmaps illustrating the performance of the InternVL3-2B model in predicting the number of floors (left) and building age (right). Darker colors indicate a higher frequency of predictions within each ground-truth bin.

macro-F1 score for building-type classification highlights persistent challenges, notably the underrepresentation of minority classes and ambiguity in building features. To address these issues, assembling a more balanced and diverse building imagery dataset and incorporating domain-specific fine-tuning are recommended. Collectively, these findings underscore the potential of VLMs to enhance robustness and generalizability across varied urban contexts.

Table 6: Validation performance comparison of different models on building images in Brussels.

(a) Performance on classification task of building type

Attribute	Model	Acc (%)	mPre	mRec	mF1
	DenseNet	31.12	0.25	0.33	0.21
	ResNet101	31.20	0.24	0.34	0.21
Building type	ViT16	32.54	0.26	0.32	0.21
	Swin_b	24.58	0.26	0.33	0.20
	InternVL3-2B	**63.80**	**0.44**	**0.50**	**0.42**

(b) Performance on prediction tasks of number of floors and building age

Attribute	Model	RMSE ($\downarrow$)	MAE ($\downarrow$)	MAPE (%) ($\downarrow$)	R^2 ($\uparrow$)
	DenseNet	1.82	1.13	30.20	0.610
	ResNet101	1.90	1.15	33.46	0.576
Number of floors	ViT16	1.79	1.15	33.10	0.624
	Swin_b	1.74	1.08	29.39	**0.645**
	InternVL3-2B	**1.73**	**0.90**	**26.20**	**0.645**
	DenseNet	55.50	43.22	74.89	0.200
	ResNet101	51.55	39.83	172.65	0.166
Building age	ViT16	51.44	38.85	168.38	0.157
	Swin_b	54.92	42.27	195.16	0.170
	InternVL3-2B	**44.67**	**30.76**	**66.66**	**0.600**

Robustness to varying image quality. As described in Section 4.3, we evaluate the robustness of the VLM against image corruptions by testing it on a perturbed dataset derived from the test set. Figure 12 illustrates the model's performance under varying severity levels of occlusion, motion blur, Gaussian noise, and brightness distortions. In general, the
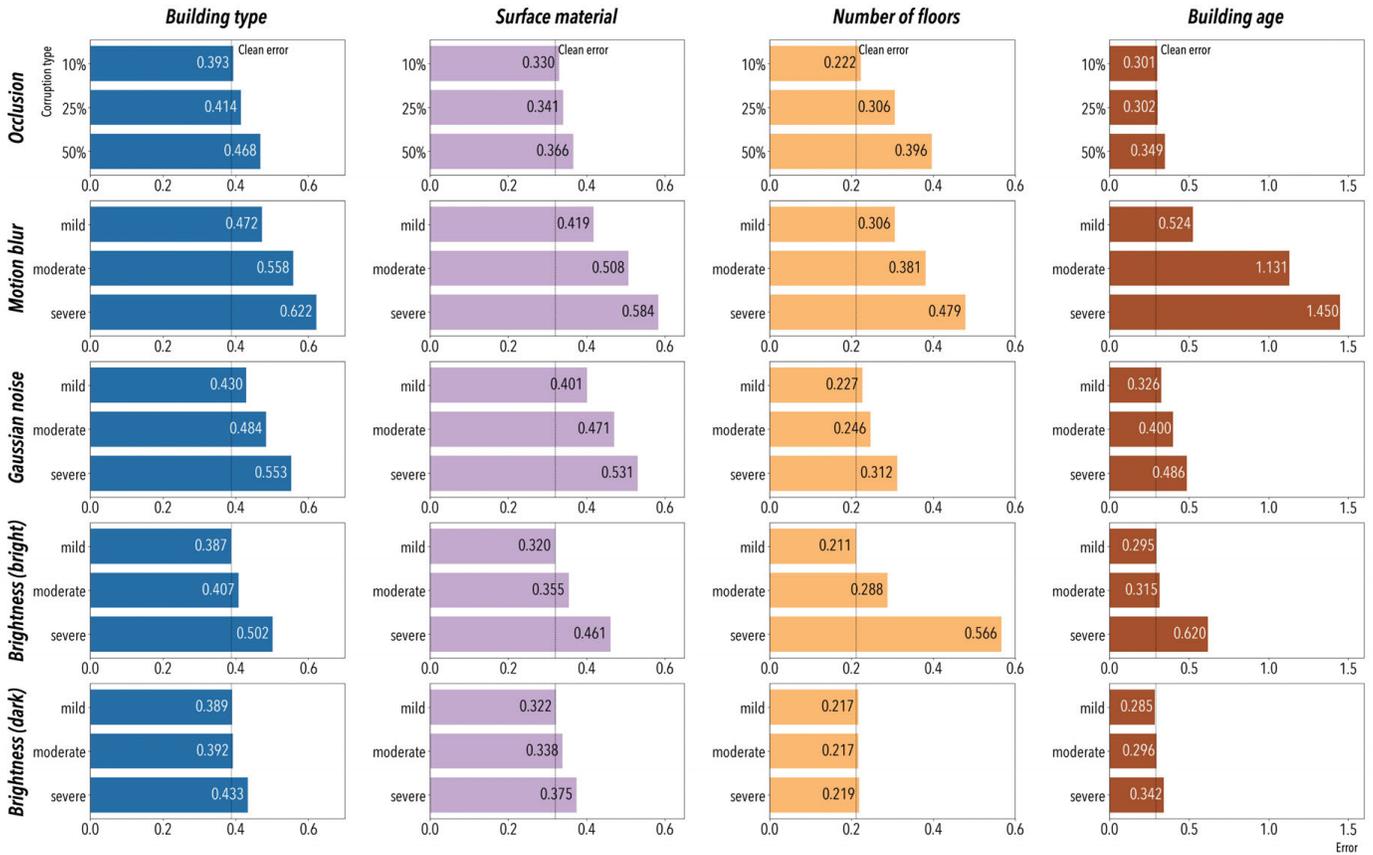
Figure 12: Error rates of the VLM under different severity levels of image corruption. The dotted line represents the clean error obtained from the original test set, serving as a baseline for comparison.

model demonstrates resilience, with performance dropping by less than 10% under most mild and moderate image corruptions. In particular, the model remains significantly stable in handling lighting variations and occlusion, both of which are common challenges in crowdsourced image datasets. However, the model experiences a significant performance drop when confronted with moderate to severe noise and blurriness. In particular, the model is most affected by motion blur, where the error rate for the number of floors prediction increases from 0.21 (clean error) to 0.48, and the error rate for building age prediction rises from 0.29 (clean error) to 1.45. These findings emphasize the necessity of image preprocessing techniques to filter out degraded images during the image selection stage.

Furthermore, the *Relative m*CE is computed for different models using ResNet50 as the baseline. Table 7 presents the relative error rates across different building attribute prediction tasks, indicating each model's stability compared to the baseline. In general, different models demonstrate various capability in handing corruption data. Multi-attribute prediction VLM (InternVL3-2B) demonstrates superior stability compared to single-attribute CV models in most cases, having lowest mean *Relative m*CE among all attributes. Especially, when distinguishing building type and building age when encountering occlusion, motion and brightness variations, VLM demostrate well stability. CNN models' stability performs comparably to more advanced models in the tasks of number of floors prediction, while ViT model performs superior in handling data with Gaussian noise. This outcome implies that additional domain-specific constraints or specialized training strategies might be required to enhance performance on crowdsourced image data.

In conclusion, building on insights from CV baselines, VLMs not only demonstrate robust and generalizable features for tackling diverse tasks based on crowdsourced data, but they also represent a promising framework for large-scale or cross-regional implementations that demand multi-feature prediction and flexible adaptation.

5.4. Ablation experiments

Fine-tuning settings. As shown in Table 8, we compare model performance across different settings, including LoRA, full fine-tuning, number of epochs, and learning rate. We summarize overall performance using two aggregate metrics: Overall Accuracy (OA), computed as the average of accuracies, and Overall R^2 (OR), computed as the average R^2. In general, full fine-tuning consistently outperforms LoRA across most tasks. Although LoRA offers parameter efficiency and retains the original checkpoint's general capabilities, it typically yields lower performance. Therefore, we adopt full fine-tuning for further investigation.

We further investigated the effect of training epochs under the default learning rate of 4e-5. Performance generally improves from epoch 1 to 3, with diminishing or unstable gains beyond that point. For example, at epoch 3, full fine-tuning achieves strong and balanced results, including a peak OA of 64.02% and competitive OR. Although epoch 4 yields a slightly higher OR (0.725), the improvements are marginal and come at the cost of

Table 7: Relative prediction error compared to ResNet50 under different corruptions and perturbations for different objective attributes.

Attribute	Model	Occlusion	Motion	Noise	Brightness	Relative mCE
Building type	ResNet50	1.00	1.00	1.00	1.00	1.00
	DenseNet	2.10	0.98	0.86	0.78	1.18
	ResNet101	0.89	0.95	0.61	0.54	0.75
	ViT16	0.76	0.83	**0.29**	1.22	0.78
	Swin_b	0.88	1.12	0.30	0.47	0.69
	InternVL3-2B	**0.51**	**0.42**	0.39	**0.46**	**0.44**
Surface material	ResNet50	1.00	1.00	1.00	1.00	1.00
	DenseNet	1.26	1.09	1.20	1.16	1.18
	ResNet101	1.49	0.98	1.25	1.02	1.19
	ViT16	**0.48**	0.72	**0.50**	1.48	0.80
	Swin_b	0.99	1.11	1.03	0.70	0.96
	InternVL3-2B	0.59	**0.66**	0.64	**0.47**	**0.59**
Number of floors	ResNet50	1.00	1.00	1.00	1.00	1.00
	DenseNet	2.01	1.62	0.86	1.24	1.43
	ResNet101	0.80	1.19	**0.31**	0.69	0.75
	ViT16	1.07	1.41	0.46	1.61	1.14
	Swin_b	0.66	1.87	0.49	0.84	0.97
	InternVL3-2B	**0.63**	**0.78**	0.42	**0.54**	**0.59**
Building age	ResNet50	1.00	1.00	1.00	1.00	1.00
	DenseNet	2.65	0.94	0.99	1.43	1.50
	ResNet101	1.04	0.78	0.48	1.10	0.85
	ViT16	0.50	0.78	**0.27**	1.80	0.84
	Swin_b	0.67	1.11	2.90	1.43	1.53
	InternVL3-2B	**0.30**	**0.69**	0.48	**0.54**	**0.50**

Table 8: Validation performance across epochs and learning rates for LoRA and full-tuning schemes.

Epoch	LR	Acc (%)		mF1		OA	R^2		MAE		OR
		Type	Material	Type	Material		Floor	Age	Floor	Age	
LoRA											
1	4e-5	60.29	60.29	0.583	0.513	60.29	0.739	0.524	2.64	40.44	0.632
2	4e-5	60.66	62.50	0.593	0.532	61.58	0.729	0.560	2.60	38.01	0.645
3	4e-5	60.52	63.24	0.592	0.537	61.88	0.745	0.595	2.52	35.79	0.670
4	4e-5	60.11	64.34	0.583	0.548	62.22	0.741	0.593	2.52	35.73	0.667
5	4e-5	60.34	64.86	0.586	0.553	62.60	0.757	0.571	2.51	35.97	0.664
Full-tuning											
1	4e-5	62.64	63.97	0.608	0.538	63.31	0.719	0.637	2.50	33.25	0.678
2	4e-5	62.02	64.71	0.599	0.542	63.37	0.735	0.672	2.40	30.95	0.704
3	4e-5	63.32	64.71	0.618	0.542	**64.02**	0.734	0.680	2.38	30.44	0.707
4	4e-5	62.09	65.07	0.605	0.545	<u>63.58</u>	0.748	0.701	2.31	29.03	**0.725**
5	4e-5	62.41	64.34	0.608	0.541	63.38	0.741	0.696	2.39	29.14	<u>0.719</u>
3	4e-4	10.31	16.54	0.030	0.085	13.43	-0.291	-0.060	7.88	79.28	-0.176
3	8e-5	62.36	63.24	0.609	0.546	62.80	0.736	0.646	2.43	31.73	0.691
3	4e-5	63.32	64.71	0.618	0.542	**64.02**	0.734	0.680	2.38	30.44	<u>0.707</u>
3	8e-6	63.43	63.97	0.619	0.540	<u>63.70</u>	0.762	0.665	2.19	31.63	**0.714**
3	4e-6	63.39	62.50	0.620	0.532	62.94	0.748	0.653	2.33	31.91	0.701

additional computation. Thus, epoch 3 is selected as the most efficient convergence point.

We also experimented with several learning rates, including the default 4e-5 and smaller values such as 4e-6. A learning rate of 8e-6 offers improved stability and slightly better OR. In addition, our observations suggest that smaller learning rates are especially beneficial for fine-tuning on relatively small datasets, as they help retain general-purpose capabilities such as open-ended reasoning and semantic alignment. Based on these comparisons, we adopt full fine-tuning with 3 training epochs and a learning rate of 8e-6 as our final configuration. This setting provides the best trade-off between predictive performance, training efficiency, and the preservation of the pretrained strengths of large VLMs.

Model size.   We further compare the performance of VLM variants with different model sizes under both zero-shot and fine-tuned settings, as shown in Tables 9a and 9b. In the zero-shot setting, larger model sizes or more recent pretrained weights generally yield better predictions.

The results confirm that fine-tuning is essential for domain adaptation, consistently improving performance across all building profiling tasks. For example, classification accuracy improves by 7–15 percentage points over zero-shot baselines, while R^2 values for floor and age prediction also increase significantly. After fine-tuning, model performance becomes more comparable across sizes, though task-specific variations remain. InternVL2.5–4B achieves the highest accuracy in classifying building type and surface material, whereas InternVL3 variants perform better on floor count and age prediction. These findings suggest that models in the 1–2B parameter range are generally sufficient for building profiling, particularly when training on relatively small datasets.

Table 10 presents captioning evaluation metrics for various models. As with the labeling task, pretrained models already exhibit reasonable performance relative to the ChatGPT-4o reference captions, while fine-tuned models incorporate domain-specific knowledge and produce more coherent, better-structured captions. We also observe that performance generally improves with model size, but doubling the parameter count yields diminishing marginal gains, which is likely a consequence of the dataset's limited scale and the inherent noise in OSM derived ground truth labels. To balance computational cost and accuracy, we therefore select the InternVL3 model with 2 billion parameters for this study.

Data size.   To assess the impact of training set size on performance, we trained InternVL3-2B on varying fractions of the 58,942 image–text pairs listed in Table 4. The left panel of Figure 13 reports accuracy for material and type predictions, while the right panel shows R^2 values for floor and age predictions at each dataset proportion. Table 11 presents BLEU, METEOR, and ROUGE-L scores relative to GPT-generated captions under each condition.

These ablation experiments reveal several important insights. First, in multi-attribute prediction tasks (Figure 13), performance peaks early in both scenarios of adjusting ei-

Table 9: Validation performance of VLM variants with different model sizes under both zero-shot and fine-tuning conditions.

(a) Building type and surface material classification in zero-shot and fine-tuned settings.

Attribute	Model	Size	Acc (%)	mPre	mRec	mF1	Acc@2 (%)
Building type	*Zero-shot*						
	ChatGPT-4o	–	57.69	0.64	0.58	0.57	75.06
	InternVL3	1B	44.31	0.60	0.41	0.41	50.60
		2B	46.99	0.58	0.46	0.46	61.38
	InternVL2.5	4B	47.82	0.61	0.46	0.46	63.06
	Fine-tuned						
	InternVL3	1B	60.87	0.65	0.60	0.60	77.37
		2B	61.27	0.66	0.60	0.61	77.31
	InternVL2.5	4B	62.41	0.67	0.61	0.62	76.69
Surface material	*Zero-shot*						
	ChatGPT-4o	–	65.41	0.58	0.61	0.55	79.07
	InternVL3	1B	56.64	0.62	0.48	0.43	62.16
		2B	61.65	0.55	0.55	0.52	69.55
	InternVL2.5	4B	60.65	0.51	0.53	0.50	76.15
	Fine-tuned						
	InternVL3	1B	67.79	0.61	0.64	0.59	81.20
		2B	68.05	0.60	0.63	0.58	81.33
	InternVL2.5	4B	68.55	0.61	0.64	0.59	81.95

(b) Building floors and building age prediction in zero-shot and fine-tuned settings.

| Attribute	Model	Size	R2 (↑)	MAE (↓)	MAPE (↓)	RMSE (↓)
Building floors	*Zero-shot*					
	ChatGPT-4o	-	0.721	2.36	38.66	5.01
	InternVL3	1B	-3.071	8.82	105.22	19.04
		2B	0.624	3.38	50.41	5.81
	InternVL2.5	4B	0.548	3.68	44.30	6.53
	Fine-tuned					
	InternVL3	1B	0.778	2.16	37.69	4.46
		2B	0.789	2.13	36.92	4.35
	InternVL2.5	4B	0.771	2.32	35.83	4.53
Building age	*Zero-shot*					
	ChatGPT-4o	-	0.645	31.63	58.79	57.07
	InternVL3	1B	0.091	57.93	77.44	91.19
		2B	0.560	41.05	77.10	63.30
	InternVL2.5	4B	0.242	51.93	85.25	84.12
	Fine-tuned					
	InternVL3	1B	0.713	28.09	58.15	51.36
		2B	0.710	28.05	58.27	51.50
	InternVL2.5	4B	0.707	29.11	63.73	51.85
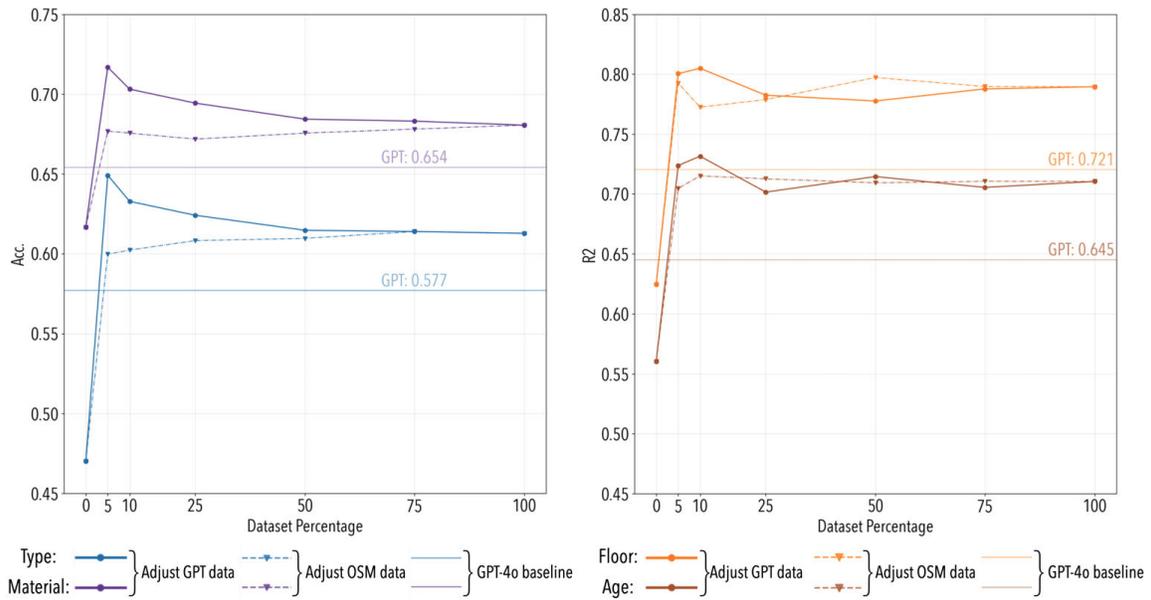
Figure 13: Model performance across varying dataset sizes by adjusting full training data (solid line) and GPT-generated data (dashed lines). The left plot shows accuracy for building type and surface material classification, while the right plot presents R-squared values for floor and age predictions, benchmarked against ChatGPT baselines.

Table 10: Captioning performance metrics (%) in zero-shot and fine-tuned settings, evaluated against ChatGPT-4o-generated reference captions.

Model	Size	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Zero-shot							
InternVL3	1B	46.99	29.55	19.64	13.09	36.56	29.64
InternVL3	2B	36.47	22.61	15.16	10.21	28.28	27.13
InternVL2.5	4B	44.14	28.81	20.27	14.39	37.99	31.13
Fine-tuned							
InternVL3	1B	56.60	41.62	32.65	25.99	44.46	43.31
InternVL3	2B	56.91	41.91	32.93	26.25	44.91	43.55
InternVL2.5	4B	52.32	37.68	29.00	22.64	41.02	40.43

ther OSM data (single-attribute Q&A) or GPT-generated data (multi-attribute Q&A and image captions). Even smaller datasets (around 5–10% of the full corpus) yield notable performance gains, highlighting the VLM's ability to learn effectively in data-constrained scenarios. This behavior can be attributed to the pre-trained semantic relationships embedded in the VLM's latent space from its foundational training. Fine-tuning on limited data stabilizes outputs by aligning task-specific features with the model's pre-existing knowledge distribution. Second, performance rises gradually when adding OSM ground truth data for most attribute prediction tasks, while GPT-generated data slightly diminishes performance gains. One possible explanation is that OSM data encodes structured, human-validated geographic knowledge, whereas GPT-generated samples may introduce inaccuracies or hallucinated features. Mitigating such noise by refining annotation procedures or excluding low-quality samples could improve overall accuracy and robustness (Chen et al., 2024b). Third, integrating single-attribute Q&A data derived from OSM labels appears to constrain descriptive richness across tasks (Table 11). This trade-off likely reflects task interference in a multi-task learning setup, where optimizing for structured attribute prediction can suppress the model's ability to generate diverse captions. To address this limitation, one could increase model capacity, curate high-quality OSM–GPT hybrid datasets, or employ techniques such as knowledge distillation to balance structured output with generative expressiveness.

In summary, our ablations show that (1) full fine-tuning of open-source MLLMs deliv-

Table 11: BLEU, METEOR, and ROUGE-L evaluation across OSM and GPT splits.

| Dataset | | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L |
OSM	GPT						
-	-	36.35	22.54	15.11	10.16	28.32	27.11
5%	100%	57.17	42.09	33.03	26.31	44.97	43.53
10%	100%	57.50	42.40	33.33	26.59	45.29	43.74
25%	100%	57.25	42.17	33.12	26.39	45.07	43.63
50%	100%	56.71	41.73	32.75	26.09	44.62	43.35
75%	100%	56.97	41.92	32.89	26.18	44.88	43.39
100%	5%	53.48	37.78	28.57	21.87	41.12	39.25
100%	10%	54.36	38.62	29.37	22.64	41.67	40.02
100%	25%	54.06	38.75	29.71	23.08	41.78	40.65
100%	50%	55.24	40.13	31.13	24.50	43.02	41.92
100%	75%	55.92	41.26	32.08	25.59	44.09	42.84
100%	100%	56.84	41.82	32.81	26.13	44.81	43.40

ers stronger domain adaptation than adapter-based methods, with fewer epochs and smaller learning rates achieving comparable knowledge gains; (2) small-scale models strike the best efficiency–performance balance, while larger models maintain an edge in free-form captioning; and (3) a modest, well-curated dataset captures most of the benefits of large-scale pretraining, although over-emphasis on structured labels can modestly reduce generative richness. Together, these findings underscore the importance of harmonizing fine-tuning scope, model capacity, data quality, and training objectives when extending vision–language models to building-specific tasks.

6. Discussion

6.1. Image labeling and captioning

Detected buildings across seven global cities, introduced in Section 5.1, are subsequently processed by the fine-tuned VLM to generate objective attributes and captions. Overall, data for half a million buildings are enriched using 1.2 million images, each linked to its geographical location. For buildings with multiple observations, the most frequently assigned categories are retained. Figure 14 compares the availability of building properties

before and after enrichment in Washington, D.C. The proposed approach effectively enhances building-level information, particularly for surface material and building age. The supplementary material presents the distribution of class labels for each attribute across the 1.2 million-building dataset.

Beyond the predefined labels, our dataset includes text annotations for each building image, providing a richer source of information for categorizing architectural features. These captions capture intricate details beyond standard classifications, including facade styles, structural elements, and mixed-use characteristics, offering a more nuanced understanding of urban form. By extracting key descriptors, Figure 15 showcases examples of mixed-use buildings and diverse facade styles identified in Washington, D.C., and San Francisco. This methodology introduces additional dimensionalities for architectural feature analysis, allowing for more detailed characterizations of urban landscapes. Moreover, it facilitates fine-grained comparisons across cities, helping to reveal and interpret regional architectural trends and stylistic variations.

6.2. Application of building image dataset

Despite the centrality of objective building attributes in urban analytics, their scarcity still persists across cities (Biljecki et al., 2023). Our open-source framework OpenFACADES addresses this gap by combining SVI, which captures pedestrian-scale visuals, with building data to train an MLLM for unified attribute extraction and semantic description. The methodology begins by integrating crowdsourced SVI metadata with geometrical building data using isovist analysis to identify relevant images. Buildings are then detected based on their angles of view within image space, followed by an automated process of reprojecting and filtering them into individual building images. Lastly, a subset of this dataset is used to construct an image-text dataset designed for three tasks for VLM fine-tuning: single-attribute Q&A, multi-attribute Q&A, and captioning. Our experiments indicate that the fine-tuned VLM demonstrates strong performance in multi-attribute prediction, surpassing CV models and outperforming zero-shot ChatGPT-4o baselines. Deploying the VLM at scale, we annotate and release data of half a million buildings with both objective attributes and textual descriptions, derived from 1.2 million images across seven global cities, contributing to a scalable and automated approach for building property enrichment.

50

Figure 14: Comparison of OSM building data (left) and the building attributes inferred using our method (right) in Washington D.C., illustrating attributes: building type, surface material, number of floors, and age. Data: (c) OpenStreetMap contributors.
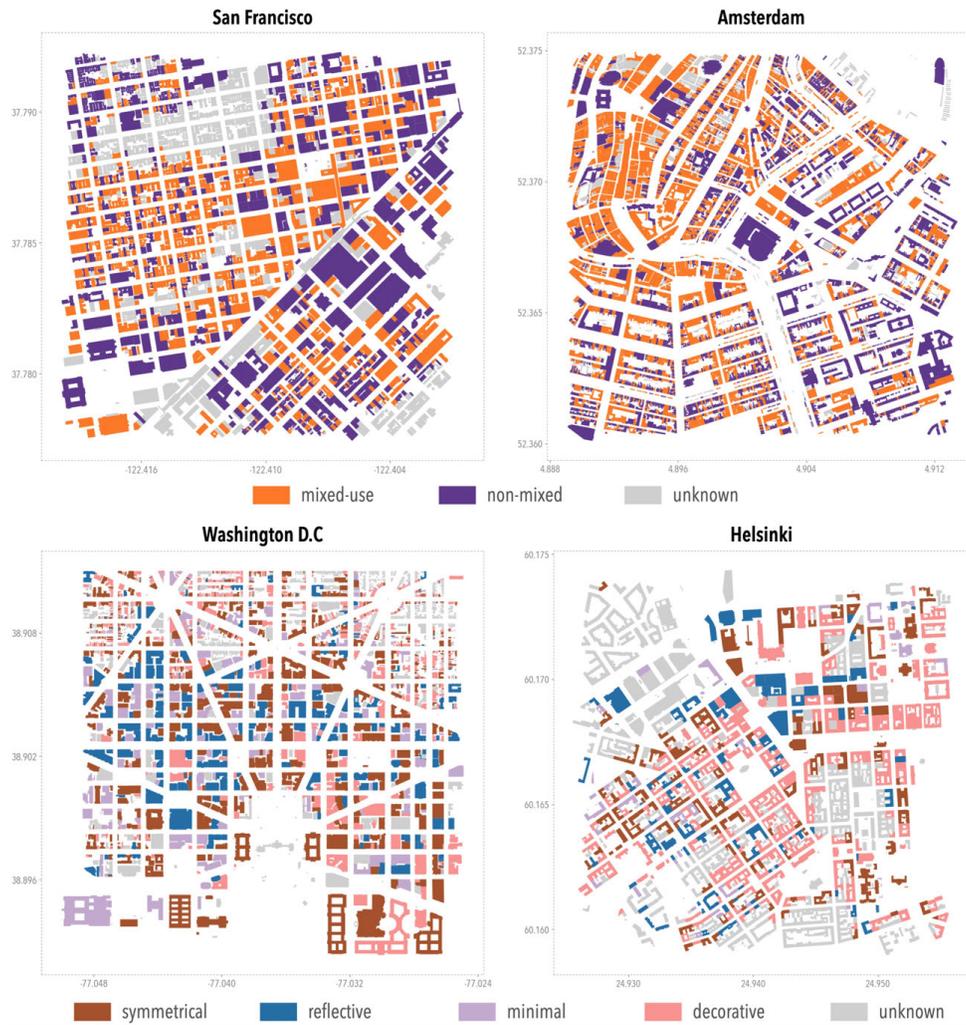
Figure 15: Spatial distribution of mixed-use buildings (top) and facade styles (bottom) in different cities. Data: (c) OpenStreetMap contributors.

Our study directly features three main contributions to building research. First, our methodology detects holistic building facades and reprojects them into undistorted individual images, ensuring comprehensive visual coverage while reducing the uncertainty inherent in panoramic imagery. This pipeline can be integrated with existing methods to detect buildings from diverse viewing angles and associate them with geolocation, enabling nuanced and holistic observation for exterior modeling (Zhang et al., 2021), facade material segmentation (Tarkhan et al., 2025), and window-to-wall ratio calculation (De Simone et al., 2024). Second, this work introduces an inclusive and efficient pipeline to utilize both crowdsourced data and open-sourced LLMs for street-level research. This pipeline not only overcomes the challenge of relying on proprietary datasets, but also circumvents the high costs and limited adaptability associated with proprietary LLM APIs, making advanced analytical techniques more accessible and reproducible to the research community. Future studies might adjust the pipeline to customized tasks to incorporate fine-grained visual information with tailored building data based on their objectives, such as building conditions (Zou and Wang, 2021; Zhang et al., 2024b), human perceptual indicators (Liang et al., 2024) and seismic structural types (Pelizari et al., 2021).

Third, we present unified benchmark VLMs that perform multi-task learning on building facades, generating descriptive captions while maintaining robust multi-class predictions of objective attributes. In particular, we:

- Demonstrate that full fine-tuning of a open-source VLM backbone yields state-of-the-art multi-attribute extraction. It surpass zero-shot ChatGPT and matching or exceeding specialized CNN/ViT baselines on building type, material, floor count and age, simultaneously generating coherent captions within one unified framework. The model delivers consistent performance across categories and cities.

- Reveal generalizability and robustness. In cross-city evaluations and synthetic corruption tests, our fine-tuned VLM surpasses CV models by maintaining high accuracy on unseen urban contexts and showing resilience to occlusion, blur, and lighting distortions, underscoring its suitability for heterogeneous, crowdsourced SVI.

- Investigate efficient training paradigms. Through systematic ablations, we show that

(i) full fine-tuning of open-source VLMs with a low learning rate over a few epochs, (ii) the use of a small-scale backbone, and (iii) a well-curated, balanced dataset together capture most benefits of large-scale pretraining by minimizing data, compute, and architectural complexity while preserving accurate multi-attribute prediction and high-quality captioning.

Finally, we apply the pipeline at scale by generating labels and captions for half a million buildings in seven cities, laying a foundation for future urban analyses. For instance, integrating these labeled data with geospatial information can add new dimensions to urban functional zone classification (Zhang et al., 2023), including potential insights into 3D functional zoning (Lin et al., 2024). The unified model also infers multi-dimensional building properties relevant for applications such as modeling building electricity consumption (Rosenfelder et al., 2021), estimating material stocks (Raghu et al., 2023), and assessing structural risk (Wang et al., 2021). Additionally, captions offer an extra layer of information about building facades, enabling the identification of mixed-use buildings or stylistic variations. This linguistic data holds promise for exploring urban identity, supporting text-image-based generative design, and serving as an additional feature layer in multimodal model training.

6.3. Limitations and future work

Despite the advancements presented in this study, limitations remain. First, while this study incorporates captioning data for fine-tuning VLMs, these captions are generated using commercial state-of-the-art LLMs rather than human-labeled ground truth, leaving their accuracy and reliability unverified. A systematic human evaluation would be valuable for future research to assess captioning quality, consistency, and semantic accuracy. Additionally, leveraging open-access models offers a more sustainable approach for scalable dataset expansion in future studies. Knowledge distillation, in which a smaller model learns from a larger teacher, offers a promising self-supervised approach to improve generalization across urban contexts.

Second, while the fine-tuned model exhibits strong generalizability across cities, the quality of crowdsourced data remains a crucial factor (Biljecki et al., 2023; Hou and Biljecki, 2022). Although multiple strategies were employed in this study to mitigate data

quality issues, several challenges persist. These include incorrect or incomplete building labels, inconsistent geometric information, non-standardized image formats, and misaligned image coordinates, each contributing to various uncertainties. Additionally, as no manual image selection was performed in this study, potential biases in data collection remain unaddressed. Future work should focus on enhancing dataset reliability through improved data filtering mechanisms. Automated repetition detection, heuristic rule-based filtering, and uncertainty-aware sampling could refine image selection and minimize inconsistencies in building attribute annotations (Chen et al., 2024b).

7. Conclusion

This comprehensive study advances spatial data infrastructures and urban data science by introducing a novel framework, OpenFACADES, which leverages volunteered geographic information to enrich building profiles on a global scale using street-level imagery and multimodal large language models. We harvest multimodal crowdsourced data and apply isovist analysis, object detection, and a tailored reprojection method to geolocate and acquire holistic building images, thereby establishing a comprehensive global building image dataset. A selection of this open dataset is then utilized for fine-tuning VLMs, enabling large-scale enrichment of building profiles through multi-attribute prediction and open-vocabulary captioning. This framework provides a scalable solution for capturing multi-dimensional fine-grained architectural details and urban morphological characteristics.

Our findings also demonstrate that VLMs generally outperform conventional CNN-based models and zero-shot GPT-4o baselines in predicting building attributes while generating linguistically grounded descriptions. This methodological advancement has enabled the creation of a large-scale dataset covering half a million buildings across seven global cities. The enriched dataset further facilitates a more nuanced and expansive exploration of urban environments, with potential applications in energy modeling, risk assessment, and sustainable development.

Beyond its immediate applications, we envision this framework as a foundation for comprehensive building profiling, capturing not only physical attributes but also the socio-

economic and cultural narratives embedded within the built environment. This advancement has significant implications for urban research, including large-scale built environment analysis, building simulation, and policy-driven planning strategies.

Acknowledgments

Author contributions

X.L.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization, Project administration. J.X.: Methodology, Software, Resources, Writing - Review & Editing. T.Z.: Methodology, Resources, Writing - Review & Editing. R.S.: Writing - Review & Editing. F.B.: Conceptualization, Resources, Writing - Review & Editing, Supervision, Funding acquisition.

Appendix A. Supplementary material

The following is the Supplementary material related to this article. [Link]

References

Aksoezen, M., Daniel, M., Hassler, U., Kohler, N., 2015. Building age as an indicator for energy consumption. Energy and Buildings 87, 74–86.

Al Rahhal, M.M., Bazi, Y., Alsaleh, S.O., Al-Razgan, M., Mekhalfi, M.L., Al Zuair, M., Alajlan, N., 2022. Open-ended remote sensing visual question answering with transformers. International Journal of Remote Sensing 43, 6809–6823.

Banerjee, S., Lavie, A., 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65–72.

Biljecki, F., Chew, L.Z.X., Milojevic-Dupont, N., Creutzig, F., 2021. Open government geospatial data on buildings for planning sustainable and resilient cities. URL: http://arxiv.org/abs/2107.04023, doi:10.48550/arXiv.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. Building and Environment 237, 110295.

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and gis: A review. Landscape and Urban Planning 215, 104217.

Boeing, G., 2017. Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Computers, environment and urban systems 65, 126–139.

Boguszewski, A., Batorski, D., Ziemba-Jankowska, N., Dziedzic, T., Zambrzycka, A., 2021. Landcover. ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1102–1110.

Chen, F.C., Subedi, A., Jahanshahi, M.R., Johnson, D.R., Delp, E.J., 2022. Deep Learning–Based Building Attribute Estimation from Google Street View Images for Flood Risk Assessment Using Feature Fusion and Task Relation Encoding. Journal of Computing in Civil Engineering 36, 04022031. URL: https://ascelibrary.

org/doi/10.1061/%28ASCE%29CP.1943-5487.0001025, doi:10.1061/(ASCE)CP.1943-5487.0001025. publisher: American Society of Civil Engineers.

Chen, X., Ding, X., Ye, Y., 2024a. Mapping sense of place as a measurable urban identity: Using street view images and machine learning to identify building façade materials. Environment and Planning B: Urban Analytics and City Science , 23998083241279992URL: https://doi.org/10.1177/23998083241279992, doi:10.1177/23998083241279992. publisher: SAGE Publications Ltd STM.

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al., 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 .

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al., 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24185–24198.

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. Global Sustainability 2, e2.

Danish, M., Labib, S., Ricker, B., Helbich, M., 2025. A citizen science toolkit to collect human perceptions of urban environments using open street view images. Computers, Environment and Urban Systems 116, 102207.

De Simone, Z., Biswas, S., Wu, O., 2024. Window to wall ratio detection using segformer. arXiv preprint arXiv:2406.02706 .

Dong, S., Wang, L., Du, B., Meng, X., 2024. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning. ISPRS Journal of Photogrammetry and Remote Sensing 208, 53–69. URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271624000042, doi:10.1016/j.isprsjprs.2024.01.004.

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

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al., 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783 .

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. Nature Sustainability 2, 267–273. URL: https://www.nature.com/articles/s41893-019-0250-1, doi:10.1038/s41893-019-0250-1. publisher: Nature Publishing Group.

Fan, K., Lin, A., Wu, H., Xu, Z., 2024. Pano2Geo: An efficient and robust building height estimation model using street-view panoramas. ISPRS Journal of Photogrammetry and Remote Sensing 215, 177–191. URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271624002727, doi:10.1016/j.isprsjprs.2024.07.005.

Fan, Z., Feng, C.C., Biljecki, F., 2025. Coverage and bias of street view imagery in mapping the urban environment. Computers, Environment and Urban Systems 117, 102253.

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

Florio, P., Politis, P., Krasnodębska, 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 in 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 Sensing of Environment 252, 112128.

Fujiwara, K., Khomiakov, M., Yap, W., Ignatius, M., Biljecki, F., 2024. Microclimate vision: Multimodal prediction of climatic parameters using street-level and satellite imagery. Sustainable Cities and Society 114, 105733.

Gaw, L., Chen, S., Chow, Y., Lee, K., Biljecki, F., 2022. Comparing street view imagery and aerial perspectives in the built environment. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 10, 49–56.

Ghione, F., Mæland, S., Meslem, A., Oye, V., 2022. Building stock classification using machine learning: A case study for oslo, norway. Frontiers in Earth Science 10, 886145.

Gouveia, F., Silva, V., Lopes, J., Moreira, R.S., Torres, J.M., Simas Guerreiro, M., 2024. Automated identification of building features with deep learning for risk analysis. Discover Applied Sciences 6, 466.

Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., Doshi, J., Lucas, K., Choset, H., Gaston, M., 2019. Creating xbd: A dataset for assessing building damage from satellite imagery, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 10–17.

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.

He, Z., Yao, W., Shao, J., Wang, P., 2024. Ub-finenet: Urban building fine-grained classification network for open-access satellite images. ISPRS Journal of Photogrammetry and Remote Sensing 217, 76–90.

Helbich, M., Danish, M., Labib, S.M., Ricker, B., 2024. To use or not to use proprietary street view images in (health and place) research? That is the question. Health

& Place 87, 103244. URL: https://www.sciencedirect.com/science/article/pii/S1353829224000728, doi:10.1016/j.healthplace.2024.103244.

Hendrycks, D., Dietterich, T., 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations .

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. Nature Communications 14, 3985. URL: https://www.nature.com/articles/s41467-023-39698-6, doi:10.1038/s41467-023-39698-6.

Hou, Y., Biljecki, F., 2022. A comprehensive framework for evaluating the quality of street view imagery. International Journal of Applied Earth Observation and Geoinformation 115, 103094. URL: https://linkinghub.elsevier.com/retrieve/pii/S1569843222002825, doi:10.1016/j.jag.2022.103094.

Hou, Y., Quintana, M., Khomiakov, M., Yap, W., Ouyang, J., Ito, K., Wang, Z., Zhao, T., Biljecki, F., 2024. Global Streetscapes — A comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics. ISPRS Journal of Photogrammetry and Remote Sensing 215, 216–238. URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271624002612, doi:10.1016/j.isprsjprs.2024.06.023.

Hu, Y., Yuan, J., Wen, C., Lu, X., Liu, Y., Li, X., 2025. Rsgpt: A remote sensing vision language model and benchmark. ISPRS Journal of Photogrammetry and Remote Sensing 224, 272–286.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Huang, Y., Zhang, F., Gao, Y., Tu, W., Duarte, F., Ratti, C., Guo, D., Liu, Y., 2023. Compre-

hensive urban space representation with varying numbers of street-level images. Computers, Environment and Urban Systems 106, 102043.

Iannelli, G.C., Dell'Acqua, F., 2017. Extensive exposure mapping in urban areas through deep analysis of street-level pictures for floor count determination. Urban Science 1, 16.

Ito, K., Zhu, Y., Abdelrahman, M., Liang, X., Fan, Z., Hou, Y., Zhao, T., Ma, R., Fujiwara, K., Ouyang, J., et al., 2025. Zensvi: An open-source software for the integrated acquisition, processing and analysis of street view imagery towards scalable urban science. Computers, Environment and Urban Systems 119, 102283.

Jia, F., Dong, Q., Huang, Z., Chen, X.J., Wang, Y., Peng, X., Guo, Y., Ma, R., Zhang, F., Liu, Y., 2024. A transformer-based multi-modal model for urban-rural fringe identification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing .

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

Kapp, A., Hoffmann, E., Weigmann, E., Mihaljević, H., 2025. StreetSurfaceVis: a dataset of crowdsourced street-level imagery annotated by road surface type and quality. Scientific Data 12, 92. URL: https://www.nature.com/articles/s41597-024-04295-9, doi:10.1038/s41597-024-04295-9. publisher: Nature Publishing Group.

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 and Buildings 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. Computers, Environment and Urban Systems 111, 102129. URL: https:

//linkinghub.elsevier.com/retrieve/pii/S0198971524000589, doi:10.1016/j.compenvurbsys.2024.102129.

Lei, B., Stouffs, R., Biljecki, F., 2023. Assessing and benchmarking 3d city models. International Journal of Geographical Information Science 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 Journal of Photogrammetry and Remote Sensing 220, 841–854. URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271625000036, doi:10.1016/j.isprsjprs.2025.01.003.

Li, W., Yu, J., Chen, D., Lin, Y., Dong, R., Zhang, X., He, C., Fu, H., 2025b. Fine-grained building function recognition with street-view images and GIS map data via geometry-aware semi-supervised learning. International Journal of Applied Earth Observation and Geoinformation 137, 104386. URL: https://linkinghub.elsevier.com/retrieve/pii/S1569843225000330, doi:10.1016/j.jag.2025.104386.

Li, X., Wen, C., Hu, Y., Yuan, Z., Zhu, X.X., 2024a. Vision-Language Models in Remote Sensing: Current progress and future trends. IEEE Geoscience and Remote Sensing Magazine 12, 32–66. URL: https://ieeexplore.ieee.org/document/10506064/?arnumber=10506064, doi:10.1109/MGRS.2024.3383473. conference Name: IEEE Geoscience and Remote Sensing Magazine.

Li, Z., Su, Y., Zhu, C., Zhao, W., 2024b. Buildingview: Constructing urban building exteriors databases with street view imagery and multimodal large language mode. arXiv preprint arXiv:2409.19527 .

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

Lin, A., Wu, H., Luo, W., Fan, K., Liu, H., 2024. How does urban heat island differ across urban functional zones? insights from 2d/3d urban morphology using geospatial big data. Urban Climate 53, 101787.

Lin, C.Y., 2004. Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, pp. 74–81.

Lindenthal, T., Johnson, E.B., 2021. Machine learning, architectural styles and property values. The journal of real estate finance and economics , 1–32.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al., 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv e-prints , arXiv–2303.

Mayer, K., Haas, L., Huang, T., Bernabé-Moreno, J., Rajagopal, R., Fischer, M., 2023. Estimating building energy efficiency from street view imagery, aerial imagery, and land surface temperature data. Applied Energy 333, 120542.

Milojevic-Dupont, N., Wagner, F., Nachtigall, F., Hu, J., Brüser, G.B., Zumwald, M., Biljecki, F., Heeren, N., Kaack, L.H., Pichler, P.P., et al., 2023. Eubucco v0. 1: European building stock characteristics in a common and open database for 200+ million individual buildings. Scientific Data 10, 147.

Nachtigall, F., Milojevic-Dupont, N., Wagner, F., Creutzig, F., 2023. Predicting building age from urban form at large scale. Computers, Environment and Urban Systems 105, 102010. URL: https://linkinghub.elsevier.com/retrieve/pii/S019897152300073X, doi:10.1016/j.compenvurbsys.2023.102010.

Nouvel, R., Zirak, M., Coors, V., Eicker, U., 2017. The influence of data quality on urban heating demand modeling using 3d city models. Computers, Environment and Urban Systems 64, 68–80.

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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16, 1740–1755.

Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002. Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311–318.

Pelizari, P.A., Geiß, C., Aguirre, P., Santa María, H., Peña, Y.M., Taubenböck, H., 2021. Automated building characterization for seismic risk assessment using street-level imagery and deep learning. ISPRS Journal of Photogrammetry and Remote Sensing 180, 370–386.

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. Resources, Conservation and Recycling 198, 107140. URL: https://www.sciencedirect.com/science/article/pii/S0921344923002768, doi:10.1016/j.resconrec.2023.107140.

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

Rosenfelder, M., Wussow, M., Gust, G., Cremades, R., Neumann, D., 2021. Predicting residential electricity consumption using aerial and street view images. Applied Energy 301, 117407.

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. Applied Energy 280, 115981.

Roy, E., Pronk, M., Agugiaro, G., Ledoux, H., 2023. Inferring the number of floors for residential buildings. International Journal of Geographical Information Science 37, 938–962.

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.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Sun, M., Han, C., Nie, Q., Xu, J., Zhang, F., Zhao, Q., 2022a. Understanding building energy efficiency with administrative and emerging urban big data by deep learning in glasgow. Energy and buildings 273, 112331.

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

Tarkhan, N., Klimenka, M., Fang, K., Duarte, F., Ratti, C., Reinhart, C., 2025. Mapping facade materials utilizing zero-shot segmentation for applications in urban microclimate research. Scientific Reports 15, 5492. URL: https://www.nature.com/articles/s41598-025-86307-1, doi:10.1038/s41598-025-86307-1. publisher: Nature Publishing Group.

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

Wang, C., Antos, S.E., Triveno, L.M., 2021. Automatic detection of unreinforced masonry buildings from street view images using deep learning-based image segmentation. Automation in Construction 132, 103968.

Wang, J., Ma, A., Chen, Z., Zheng, Z., Wan, Y., Zhang, L., Zhong, Y., 2024a. Earth-VQANet: Multi-task visual question answering for remote sensing image understanding. ISPRS Journal of Photogrammetry and Remote Sensing 212, 422–439. URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271624001990, doi:10.1016/j.isprsjprs.2024.05.001.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al., 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 .

Wang, Y., Chau, C.K., Ng, W., Leung, T., 2016. A review on the effects of physical built environment attributes on enhancing walking and cycling activity levels within residential neighborhoods. Cities 50, 1–15.

Wang, Y., Zhang, Y., Dong, Q., Guo, H., Tao, Y., Zhang, F., 2024c. A multi-view graph neural network for building age prediction. ISPRS Journal of Photogrammetry and Remote Sensing 218, 294–311. URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271624003885, doi:10.1016/j.isprsjprs.2024.10.011.

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

Wu, A.N., Biljecki, F., 2021. Roofpedia: Automatic mapping of green and solar roofs for an open roofscape registry and evaluation of urban sustainability. Landscape and Urban Planning 214, 104167.

Wu, M., Huang, Q., Gao, S., Zhang, Z., 2023a. Mixed land use measurement and mapping with street view images and spatial context-aware prompts via zero-shot multimodal learning. International Journal of Applied Earth Observation and Geoinformation 125, 103591. URL: https://linkinghub.elsevier.com/retrieve/pii/S1569843223004156, doi:10.1016/j.jag.2023.103591.

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

Xu, F., Wong, M.S., Zhu, R., Heo, J., Shi, G., 2023. Semantic segmentation of urban building surface materials using multi-scale contextual attention network. ISPRS Journal of Photogrammetry and Remote Sensing 202, 158–168. URL: https://www.sciencedirect.com/science/article/pii/S0924271623001600, doi:10.1016/j.isprsjprs.2023.06.001.

Yan, Y., Huang, B., 2022. Estimation of building height using a single street view image via deep neural networks. ISPRS Journal of Photogrammetry and Remote Sensing 192, 83–98. URL: https://www.sciencedirect.com/science/article/pii/S0924271622002106, doi:10.1016/j.isprsjprs.2022.08.006.

Yang, X., Lindquist, M., Van Berkel, D., 2025. "streetscape" package in r: A reproducible method for analyzing open-source street view datasets and facilitating research for urban analytics. SoftwareX 29, 101981.

Zarbakhsh, N., McArdle, G., 2023. Points-of-Interest from Mapillary Street-level Imagery: A Dataset For Neighborhood Analytics, in: 2023 IEEE 39th International Conference on Data Engineering Workshops (ICDEW), pp. 154–161. URL: https://ieeexplore.ieee.org/document/10148212/authors#authors, doi:10.1109/ICDEW58674.2023.00030. iSSN: 2473-3490.

Zeng, Z., Goo, J.M., Wang, X., Chi, B., Wang, M., Boehm, J., 2024. Zero-shot building age classification from facade image using gpt-4. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 48, 457–464.

Zhang, C., Fan, H., Kong, G., 2021. Vgi3d: an interactive and low-cost solution for 3d building modelling from street-level vgi images. Journal of Geovisualization and Spatial Analysis 5, 18.

Zhang, F., Salazar-Miranda, A., Duarte, F., Vale, L., Hack, G., Chen, M., Liu, Y., Batty, M., Ratti, C., 2024a. Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery. Annals of the American Association of Geographers 114, 876–897.

Zhang, J., Xiang, R., Kuang, Z., Wang, B., Li, Y., 2024b. Archgpt: harnessing large language models for supporting renovation and conservation of traditional architectural heritage. Heritage Science 12, 220.

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 Journal of Photogrammetry and Remote Sensing 198, 153–168. URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271623000680, doi:10.1016/j.isprsjprs.2023.03.008.

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 Transactions on Geoscience and Remote Sensing 60, 1–17.

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 Journal of Photogrammetry and Remote Sensing 151, 237–250. URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271619300887, doi:10.1016/j.isprsjprs.2019.03.019.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2017. Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., Gao, Z., Cui, E., Wang, X., Cao, Y., Liu, Y., Wei, X., Zhang, H., Wang, H., Xu, W., Li, H., Wang, J., Deng, N., Li, S., He, Y., Jiang, T., Luo, J., Wang, Y., He, C., Shi, B., Zhang, X., Shao, W., He, J., Xiong, Y., Qu, W., Sun, P., Jiao, P., Lv, H., Wu, L., Zhang, K., Deng, H., Ge, J., Chen, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W., 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. URL: https://arxiv.org/abs/2504.10479, arXiv:2504.10479.

Zia, U., Riaz, M.M., Ghafoor, A., 2022. Transforming remote sensing images to textual descriptions. International Journal of Applied Earth Observation and Geoinformation 108, 102741.

Zietz, J., Zietz, E.N., Sirmans, G.S., 2008. Determinants of house prices: a quantile regression approach. The Journal of Real Estate Finance and Economics 37, 317–333.

Zou, S., Wang, L., 2021. Detecting individual abandoned houses from google street view: A hierarchical deep learning approach. ISPRS Journal of Photogrammetry and Remote Sensing 175, 298–310. URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271621000915, doi:10.1016/j.isprsjprs.2021.03.020.