Article

Urban Analytics and City Science

EPB: Urban Analytics and City Science
2025, Vol. 52(1) 59–75

DOI: 10.1177/23998083241247870
journals.sagepub.com/home/epb
Sage

Aerial-terrestrial data fusion for fine-grained detection of urban clues

Jessica Gosling-Goldsmith
World Bank Group, USA

Sarah Elizabeth Antos
World Bank Group, USA

Luis Miguel Triveno
World Bank Group, USA

Adam R Benjamin
World Bank Group, USA

Chaofeng Wang
University of Florida, USA

Abstract

Those who work in the design, development, and management of cities are often limited by the scarcity of data. Particularly in the Global South, urban databases may be insufficient, out of date, or simply not available. However, digital technology is making it possible to fill gaps and build substantial datasets using "urban clues," or attributes, gathered in high-resolution imagery by sky- and street-based cameras. Aided by machine learning, it is possible to detect specific building characteristics (purpose, condition, size, material, and construction)—yielding an array of geolocated details about the built environment. The resulting composite view can be made available, as we have done, in an open-source portal for use in urban management. The insights gained in this way may help address common urban management challenges, such as locating homes vulnerable to hazards such as flooding or earthquakes, identifying urban sprawl and informal housing, prioritizing infrastructure investments, and guiding public program support. This approach has been applied in Colombia, Guatemala, Indonesia, Mexico, Paraguay, Peru, St Lucia, and St Maarten.

Corresponding authors:
Jessica Gosling-Goldsmith, World Bank Group, 1818 H St NW, Washington, DC, USA.
Email: jssggg@gmail.com

Chaofeng Wang, M.E. Rinker Sr. School of Construction Management, University of Florida, Gainesville, FL, USA.
Email: chaofeng.wang@ufl.edu

Data Availability Statement included at the end of the article

Keywords
Drone imagery, street view imagery, mobile mapping, cities

Introduction

More than half the world's population lives in cities (UNDESA, 2019). As urban populations grow, decision-making is crucial to the sustainability of the urban environment, as well as the infrastructure, education, health, and economic opportunities accessible to its residents (Stokes and Seto, 2019).

Despite advances in smart cities and the internet of things, and growing access to big data from mobile devices and volunteered geographic information, detailed data characterizing the built environment are rarely generated or cohesively and consistently updated in the Global South, and if data are collected, it is both time and resource intensive (Albert et al., 2017; Bennett et al., 2021; Grippa et al., 2018; Vanderhaegen and Canters, 2017). As a result, urban data may be outdated, incomplete, or non-existent (Kuffer et al., 2021). However, there are opportunities to map urban areas more efficiently at the building or unit level. In addition, new tools go beyond traditional surveys, providing a geographic visualization of detailed urban characteristics.

This paper offers a synchronized approach to collecting and processing aerial and street-level imagery to derive and fuse identifiable characteristics at the unit level. We capture detailed building characteristics, such as size, roof condition, and wall material, resulting in a geospatial database that describes thousands of buildings per neighborhood or city. To our knowledge, this is among the first papers demonstrating systematic methods to acquire and fuse building attribute data at this level of granularity and scale.

Urban clues

We refer to the features detected in this approach as *urban clues*, a term proposed by Allan B. Jacobs for quick insights into an area's living conditions that may reveal meaningful patterns about residents' health, security, and well-being (1985). *Looking at Cities* adds to earlier work classifying the composition and visual perception of the urban landscape (Williams, 1954), defining qualities for creating an urban mental map (Lynch, 1960), emphasizing the importance of walking in cities to conduct field observation (J Jacobs, 1961), and searching for clues and patterns in U.S. cities (Clay, 1973).

The urban clues Jacobs outlined in *Looking at Cities* suggest an area can be understood by collecting and interpreting visual characteristics, mainly by walking along streets. This guide remains relevant today when combined with new observation technologies and machine learning. Table 1 shows the urban clues proposed by Jacobs with the corresponding built environment characteristics detected in our approach. In this paper, we demonstrate how these characteristics are derived from drone and street view images.

Recent developments in convolutional neural networks (CNN) and other tools in computer vision have enabled automated systems to extract information from images as successfully as humans (Esteva et al., 2017). When such models are applied at scale, the resulting urban datasets are significant and may be particularly useful in areas lacking data or affected by external forces (Glaeser et al., 2018). The scope of this paper—unlike those primarily focused on machine learning techniques—is to demonstrate the application of machine learning-assisted frameworks to urban science. Our approach relates to research characterizing the urban and built environment using satellite and aerial imagery, street-level imagery, and a combination of the two. The primary focus is on work that has applied machine learning approaches for characterization purposes, particularly those situating the results geographically.

Table 1. Urban characteristics captured by drone and street view imagery, inspired by urban clues in *Looking at Cities* (AB Jacobs, 1985). Apart from building size, most building characteristics are derived through machine learning.

Clue	Characteristic	Description	Drone	Street View
Purpose	Use	Residential, commercial, mixed		x
Size	Area	Estimated roof area in square meters	x	
	Height	Average (mean) height of the building in meters	x	
	Volume	Estimated volume of the building in cubic meters	x	
Material	Roof material	Concrete, metal, mixed, tile, other	x	
	Wall material	brick or concrete block; plaster; wood—polished; wood—crude/plank; adobe; corrugated metal; stone with mud/ashlar with lime or cement; container/trailer; plant material; mix/unclear/other		x
Construction	Masonry	Unreinforced masonry, reinforced masonry, unknown		x
Condition and maintenance	Roof condition	Good, fair, poor, under construction or vacant	x	
	Wall condition	Good, fair, poor		x
	Total condition	Composite estimation using roof and wall condition	x	x

Urban clues in satellite and aerial imagery

Several studies have applied machine learning to satellite or aerial images to classify land parcels (Wu et al., 2009), classify land use (Albert et al., 2017), define urban typologies (Hermosilla et al., 2014), detect urban decay (Deng and Ma, 2015), delineate informal settlements (Owusu et al., 2021; Pratomo et al., 2017), and identify the degree of deprivation in informal settlements (Ajami et al., 2019). CNN models have demonstrated promise in completing visually nuanced and complex tasks, such as differentiating land cover (Elamin and El-Rabbany, 2022) and segmenting buildings from ground (Sariturk et al., 2023). At a unit level, rooftop classification from drone imagery has been demonstrated for eight roof types of varying color and material in traditional villages of Northern China (W Wang et al., 2023) and five roof types based on roof texture and shape in Hunan Province, China (Y Wang et al., 2022). Notably, deep CNNs have been applied to high-resolution drone imagery to detect buildings damaged by earthquakes (Ding et al., 2022) and classify building exteriors for architectural purposes (Maskeliūnas et al., 2022). In a comparison study of property valuation using images collected by satellites, piloted aircraft, or drones, drone imagery held the most promise for its spatial and temporal accuracy and completeness (Koeva et al., 2021). Similarly, higher spatial resolution aerial imagery has largely improved urban land classification accuracy (Cao et al., 2018; Hoffmann et al., 2019). Furthermore, the performance of models classifying buildings from above is improved by adding street-level imagery (Hoffmann et al., 2019).

Urban clues in street-level imagery

New technology has made it possible to classify geolocated, street-level images from crowdsourced snapshots and photographs collected along pathways (e.g., Google Street View, Baidu Maps, Tencent Street View, and Mapillary). Researchers have successfully applied CNNs to such scenes to characterize planned or unplanned areas for analysis, including in the Global South (Ibrahim et al., 2021). For example, such work has distinguished eight building classes (apartment, church, garage, house, industrial, office building, retail, and roof) from street view images (Kang et al., 2018).

A growing body of research has captured fine-grained building characteristics from street-level images. In the real estate sector, building age (Zeppelzauer et al., 2018) and building condition (Koch et al., 2018) have been predicted from photos using machine learning. However, the results were not geographically located for further spatial analysis. Some research has been shown to determine the type of commercial entity (Movshovitz-Attias et al., 2015), provide a soft story screening (Yu et al., 2020), flag the presence of deteriorated components (Zou and Wang, 2021), and note the presence of common elements (e.g., doors, windows, garages) (Dakin et al., 2020).

Other studies have paired a building's predicted characteristics with its footprint to classify architectural epochs in Amsterdam (Sun et al., 2021) and Paris (Lee et al., 2015), assess architectural styles in the United Kingdom (Lindenthal and Johnson, 2021), and examine buildings' vulnerability to natural disasters (Gonzalez et al., 2020). At present, street-level data from Google Street View and Mapillary remain geographically limited (Biljecki and Ito, 2021), and studies using street images are primarily concentrated in North America, China, and Europe (Cinnamon and Jahiu, 2021).

Urban clues from the sky and street

Urban observations that combine data from the sky and street can uncover clues or patterns that may be missed with a single method. For instance, a building's purpose may change over time, which could be undetected in aerial observations (Hoffmann et al., 2019). Street-level information can supplement occluded data in aerial imagery; for example, researchers successfully used street-level imagery to improve land use classification (Cao et al., 2018) and construct a sidewalk network when tree foliage covered sidewalks in the aerial imagery (Ning et al., 2022). Conversely, buildings may be occluded by trees, vehicles, or boundary walls that prevent a clear view of the unit from the street. In some cases, a building's façade may appear sturdy from the road but have a damaged roof identifiable only from the sky. Research has demonstrated that image analysis from above and at the street level improves urban tree species classification (Wegner et al., 2016). Likewise, housing prices have been predicted by considering the surrounding environment from these two vantage points (Chen et al., 2020; Fu et al., 2019).

Data from street-level images can improve aerial image-based machine learning predictions to classify urban land parcels. For instance, ground-level images helped distinguish single-family houses and multi-family residential buildings from mixed residential and commercial land parcels. In this case, shop signs facing the street were detected in ground-level imagery to improve classification accuracy on average by 10 percentage points (Zhang et al., 2017). Similarly, in a comparative study of predicting land use by applying machine learning to aerial images, street-level images, and the two together, Cao et al. (2018) found that combining aerial and street-level images moderately improved the overall accuracy of the predictions. After assessing several aerial image resolutions, the highest spatial resolution (0.3 m per pixel), plus street-level data, provided the best classification results. In another study, housing price predictions were improved with the addition of visual information from the urban street scene (Google Street View) and local neighborhood characteristics (Bing aerial imagery), even though the street view may not contain the house in the analysis (Law et al., 2019).

Our approach differs from this recent work in a few ways. First, our team simultaneously collects the drone and panoramic street view imagery. Acquiring data using a camera mounted to a drone for aerial imagery and a multi-camera sensor mounted to an automobile for street-level imagery guarantees that imagery is collected at a high spatial resolution with positional accuracy and temporal synchronicity to support a current and comprehensive view of the built environment, which resolves common challenges in other urban street view studies that rely on commercial or crowdsourced image repositories (Biljecki and Ito, 2021). For street view imagery, this approach mitigates commercial licensing restrictions and ensures consistent, high-resolution images. Second, our machine learning algorithms predict a breadth of detailed classes (e.g., wall material, wall condition, masonry type, roof material, roof condition). Third, our approach connects the observable characteristics from drone and street view imagery, which is often missing from prior studies. Furthermore, our results are synthesized in a geospatial database and visualized in an open-source portal. This means select building attributes are jointly evaluated for more significant analysis, and additional urban data can be included to assess the local, neighboring context. These results are immediately verifiable by comparing the attributes with the street view and drone imagery in the open-source portal. Ultimately, this method provides informative attributes to characterize buildings, neighborhoods, and cities across several countries in the Global South, including Colombia, Guatemala, Indonesia, Mexico, Paraguay, Peru, St Lucia, and St Maarten (Figure 1).

Materials and methods

For each project area, the following steps were applied: data collection, initial image processing, inference by machine learning algorithms (image segmentation and image classification), and synthesis of the results in a geospatial database for geographic information system (GIS) analysis (Figure 2). In each of the following subsections, the drone image-related tasks are described first, followed by street view image descriptions.

Data collection

Areas of interest were determined through coordination with local governments and national housing authorities. Past projects have typically prioritized areas of critical importance due to natural hazard risks (e.g., riverine flooding, storm surge from tropical cyclones, earthquake effects). This has resulted in project areas focused on specific communities that frequently span 15–25 sq km. Local airspace regulations, such as no-fly zones near airports or over military or national security interests, influenced the finalized extent of a collection area. The topography and terrain of a given city add varying levels of complexity to the data collection. Mild topographic gradients, such as those found in the floodplain areas of Asunción, Paraguay, were opposed by the challenging,

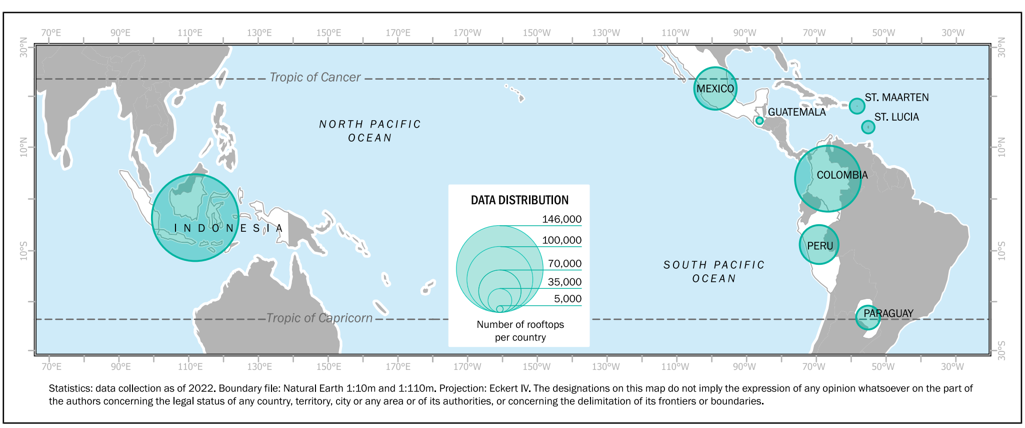

Figure 1. Geographic distribution of projects by country.

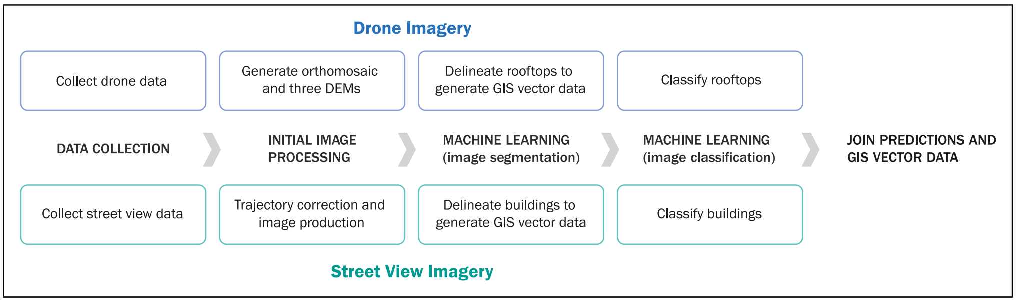

Figure 2. Outline of methodology.

undulating topographic gradients found across the city of Salina Cruz, Mexico, and the island nation of St Maarten.

To fuse drone and street view data, a common geodetic framework is required. Optimally, the same GNSS base station was used for the data collected from the drone and street view sensors. In these instances, the GNSS base station was within 25 km of the project site and part of a continuously operating reference system (CORS) network. If the CORS network was not available, as in the neighborhoods of Asunción, Paraguay, and Padang, Indonesia, a base station was brought to the project site to log static, multi-frequency data at 1 Hz during active drone and street view data collection.

Given the previous findings related to spatial resolution enhancing machine learning accuracy (Cao et al., 2018; Hoffmann et al., 2019), aerial imagery for our projects was collected via sensors mounted to drones (i.e., unmanned aerial systems). Specifically, aerial imagery was collected via a fixed-wing Sensefly eBee X drone equipped with a S.O.D.A. 3D camera. Data collection parameters were set to at least 75% overlap between adjacent images. Also, the S.O.D.A. 3D sensor permitted the collection of both nadir and oblique images along a given flight path, resulting in more aerial data captured on the sides of buildings and the ground under canopies. Flights were conducted to obtain a native spatial resolution of approximately 4 cm, which provided a level of detail roughly eight times the resolution of commercial satellite imagery (e.g., WorldView-3).

Street-level imagery was collected using a car-mounted, street view multi-camera sensor (Figure 3). The sensor suite was a Trimble MX7, which included a Ladybug camera inside a Trimble mobile mapping system. The MX7 took six overlapping pictures every two meters to create a 30-megapixel (MP) panoramic photo. The metadata for each image (JPG) included the timestamp and geographic coordinates (i.e., latitude and longitude in WGS84) of data acquisition and the approximate bearing of each individual camera. To maintain positional accuracy, the Trimble MX7 system used GNSS and an inertial referencing system in addition to a GNSS Azimuth Measurement Subsystem (GAMS) to fine-tune the inertial measurement unit (IMU). Any street blockages, gated communities, one-way streets, or narrow streets were marked in a mobile mapping application. When roads were too narrow for a car, a 360° camera (e.g., GoPro Fusion) could be attached to a backpack to collect data on foot or by motorcycle.

Initial image processing

For drone imagery, the image processing used structure-from-motion (SfM) software that produced a dense point cloud from the imagery. If multiple flights were conducted in a study area, all the images were aligned in the SfM before creating the dense point cloud. Then, the 3D point cloud was cleaned and filtered to classify ground and non-ground points. The digital terrain model (DTM), the

Figure 3. Collecting street-level imagery in Colombia.

digital surface model (DSM), and the digital object model (DOM), which provides the height of non-ground objects (difference between the DSM and DTM), were the three digital elevation models (DEMs) generated from the classified point cloud. The DSM was used to create an orthophoto mosaic. All three DEMs and the orthophoto mosaic were output as GeoTIFFs. The orthophoto mosaic was in natural color with a non-resampled spatial resolution of 4 cm or less, spectral resolution of at least three bands (RGB), and 16-bit radiometric resolution. The DEM layers had a non-resampled spatial resolution of 8 cm or less. If the orthomosaics were between 10 and 30 sq km, image tiles were the preferred input for the machine learning model.

Street-level data came in the form of panoramic and cube images, along with trajectory data (i.e., heading, geographic coordinates, elevation, and timestamp). Some post-processing was required for positional correction due to tall buildings or trees in the urban context. The panoramic images were the six sides of the cube images (capturing the view from the top, bottom, front, back, left, and right of the Trimble MX7 camera), stitched into one rectangular image. These panoramic images were uploaded to Mapillary for others to use (faces and license plates are blurred). The set of cube images was used for image classification. The geolocation data tied the street-level image data to the drone image data through geographic coordinates.

Inference by machine learning algorithms

The collected drone imagery provided optimal geospatial accuracy to delineate rooftops and ideal image resolution to predict rooftop materials and conditions. First, a vector rooftop layer was segmented using a U-Net CNN (Ronneberger et al., 2015). Then, the vector rooftops were manually reviewed for quality control and assurance. In cases where the rooftops were compact, manual edits may be required to separate roofs and finalize the building footprint layer (Figure 4).

Once a footprint layer was prepared, the training dataset was gathered. To standardize training data across countries, annotators referred to a labeling guide specifically developed for this project. The guide included definitions for each class, photographs to illustrate each definition, and examples of edge cases with corresponding instructions. Definitions of roof condition and material labels are listed in Table 2. The roof material was determined by the local building conventions in the study areas, which were mainly concrete or metal. Labels specific to local context were added to the collection of annotations previously used in other cities to detect clues with local understanding. For

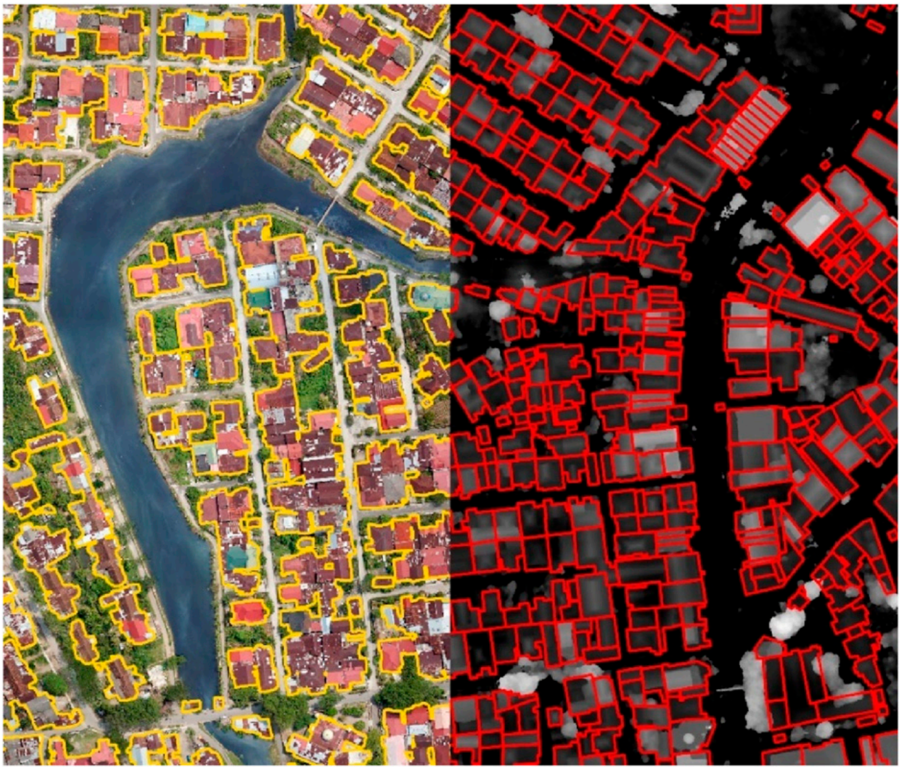

Figure 4. The urban rooftop vector layer, shown in yellow, overlayed on the orthophoto mosaic (left). Clustered rooftops were reviewed and manually divided, if necessary, as shown in red, on top of the digital object model (right) to create a building footprint layer.

Table 2. Algorithms developed to derive building characteristics from drone images.

Building properties observed from the drone	
Roof condition	Roof condition is based on construction appearance, such as patching and coloring (rust). **Good:** new, well-constructed (no holes) and very minimal patching or discoloration. **Fair:** roof is patched or discolored, but still sturdy; may look old or drab but seems livable and not precarious. **Poor:** lots of patching, holes, items to hold it down, or bags to stop leaks. Otherwise, the building may be **under construction** or **vacant.**
Roof material	**Concrete:** more than 50% of the rooftop is visibly concrete. **Metal:** vast majority of the rooftop is covered in metal (70-90%). **Mixed:** multiple materials are used to cover the roof and keep the inhabitants dry. Typically, this is less than 50% concrete. **Tile:** more than 50% of the rooftop is clay tile or metal tile. **Other:** tent or tarp material.

example, labels were added to describe new rooftop materials (e.g., tiles) or labels for wall materials (e.g., adobe) in a new region with a different climate or building construction practices. Rather than using crowdsourced labeling (i.e., from Mechanical Turk), a dedicated team conducted labeling, including local experts, when possible, as the algorithms are only as good as their labels. The condition and material of the rooftops were annotated directly within the GIS rooftop (building footprint) layer.

For drone image classification, a VGG-19 CNN (16 convolution layers and three fully connected layers) (Simonyan and Zisserman, 2015) was trained to predict each rooftop condition and material, using Google Colab (free Tesla k80 GPU). Data augmentation generated greater complexity in the sample and reduced overfitting (horizontal and vertical rotation, flip, shift). Labels were randomly distributed at an 80 to 20 ratio between training and testing. A confidence level represents the certainty of the model in its prediction. An array of predictions was created for each roof with the number of potential classifications. Each prediction had a confidence level assigned, ranging from 0 to 1, with 1 representing 100% probability. The confidence level represented the difference between the highest prediction and the rest of the values in the array.

For street view imagery, we leveraged instance segmentation models, Mask R-CNN (He et al., 2017), to extract building façade attributes from the images, using one NVIDIA Quadro RTX 5000 GPU for the training and model inference. A street view image was first input into a pre-trained backbone CNN (ResNet with 101 layers), which was connected with a Region Proposal Network (RPN) for proposing regions of interest (RoI) from the feature maps generated by the backbone network. Each RoI was then input into two parallel network branches: the mask branch and the bounding box branch. The mask branch was a small, fully convolutional network (FCN) applied to each RoI, predicting a segmentation mask at the pixel level. The box branch is a set of fully connected (fc) layers that yield the class prediction and the bounding box regression. This method can successfully discriminate distinct types of masonry in buildings (Wang et al., 2021). In this study, we expanded the scope to detect and classify more attributes, such as wall material and condition (Table 3).

We collected street-level images of buildings from multiple directions while the vehicle moved along the street. Each building could appear in multiple images; each image usually contained multiple buildings. In the data collection region, we selected a random subset (about 100,000 street view images) for annotation. In each image, building objects were annotated with shapes using the

Table 3. Algorithms developed to derive building characteristics from street view images.

Building properties observed from the street	
Masonry	**Unreinforced masonry:** refers to buildings made from brick, stone or concrete blocks that appear from the outside to be missing concrete columns or beams (or both). Other examples include buildings made from adobe or constructed using timber or wooden frames.
	Reinforced masonry: refers to buildings with confined masonry or concrete frames, which may be called reinforced masonry in some countries. Reinforcement components such as rebar inside of blocks are not always possible to determine from street view analysis but at times, particularly when buildings are "incomplete" rebar is visible.
	Unknown: unknown
Use	**Residential:** used solely for residential purposes
	Commercial: used for commercial purposes (e.g., store)
	Mixed: used for residential and non-residential purposes (e.g., store on the ground floor and residential housing above)
Wall condition	**Good:** new construction and sturdy
	Fair: sturdy but shows signs of aging
	Poor: dilapidated, temporary, self-built, or not well-maintained
Wall material	The predominant classes are **plaster, brick or concrete block, and mix/other/unclear**. Other classes include: **wood—polished; wood—crude/plank; adobe; corrugated metal; stone with mud/ashlar with lime or cement; container/trailer; plant material.**

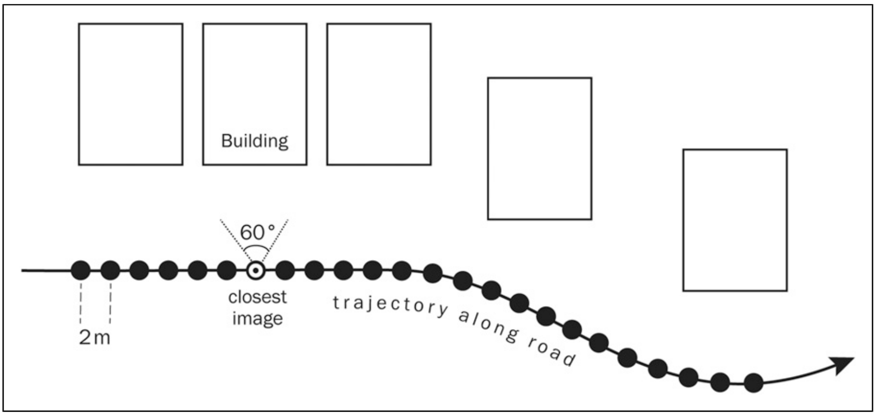

Figure 5. Linking detected building objects to footprints; modified from (Wang et al., 2024).

Computer Vision Annotation Tool (CVAT) labeling platform. Each shape was labeled with its properties, for example, masonry, façade wall material. The model was then trained with the stochastic gradient descent optimizer with a base learning rate of 0.0025. The performance of the trained model is shown in the following section of this paper. We then ran the trained model to infer on all images to cross-validate the prediction of a single building from multiple images. By calculating the direction of the camera, we linked the predicted attributes to the building footprint, as shown in Figure 5.

Synthesize predictions in a geospatial database

Given the geospatial accuracy of the previous steps, rooftop predictions were joined to the building footprint vector layer, and each street view image prediction was joined to the building footprint by its perpendicular location relative to the street view camera location (Wang et al., 2024). This means the predictions and the geospatial data can be analyzed using geographic information system software, such as calculating the building area from the building footprint and the building height using the DOM, which are multiplied to estimate the unit's volume. The resulting dataset can be stored in an open-source portal (https://github.com/GPRH/housing_portal), which makes it accessible for analysis and geographic visualization without specialized geospatial training.

The portal maintains spatial and user data in a PostGIS database. Additional variables can be uploaded to the portal, such as census or hazard data, for further analysis. The frontend is a Single Page Vue.js application. The application is run within docker containers, except for Nginx in production. Street view images that contain faces or personal information are blurred in Mapillary. Building data can be exported in multiple formats (e.g., GeoJSON, GeoPackage, ESRI Shapefile, CSV). Access to each country's portal is managed by an administrator with login permissions exclusive to each user.

Results

Our results demonstrated the ability to synchronously collect drone and street view images, segment and classify buildings and rooftops using machine learning, and fuse the output in a database to access thousands of detailed unit characteristics geographically situated in multiple square kilometers of the urban built environment (Table 4).

Table 4. Drone and street view data produced.

	Colombia	Guatemala	Indonesia	Mexico	Paraguay	Peru	St Lucia	St Maarten
Drone coverage (km^2)	40	4	80	31	17	20	8	28
Car distance (km, inc. overlap)	470	120	1,700	580	230	140	210	320
Panoramic images	411,910	62,036	587,880	337,601	121,368	107,337	109,200	168,691
Cube images	2,792,710	372,216	6,465,264	2,025,606	728,208	644,022	655,200	1,012,146
Street view labels	26,925	–	25,590	8,240	23,417	20,061	–	12,865
Rooftops	109,300	4,900	146,000	67,900	35,100	61,600	9,100	13,000

Table 5. Accuracy and F1-score for roof characteristics derived from drone imagery.

Characteristic	Accuracy (%)	F1-score (%)
Roof condition	77.14	74.67
Roof material	83.64	81.20

Table 6. Accuracy and F1-score from the universal model for select building characteristics derived from street view imagery; modified from (Wang et al., 2024).

Characteristic	Accuracy (%)	F1-score (%)
Masonry	98.97	98.97
Use	94.62	88.31
Wall condition	78.85	69.68
Wall material	97.08	93.30

Table 7. Total condition evaluation derived from roof and wall conditions.

Total condition	Description
Very good	If both the roof and wall conditions are good.
Good	If the roof or wall condition is good and the other is under construction, vacant, or otherwise unknown.
Fair	If both the roof and wall conditions are fair; or if one is fair and the other is good, under construction, vacant, or otherwise unknown.
Poor	If the roof or wall condition is poor and the other is good, fair, under construction, vacant, or otherwise unknown.
Very poor	If both roof and wall condition are poor.

Generalization of machine learning models is commonly difficult: the accuracy of a model trained on a dataset (e.g., drone imagery of a city) can be lower when tested on a new unseen dataset (e.g., drone imagery of another city). To overcome this issue when a new area is analyzed, we added a sample of new labels to capture the local characteristics of that city to the previously trained model to improve the roof material and roof condition predictions (Table 5). For street view images, we

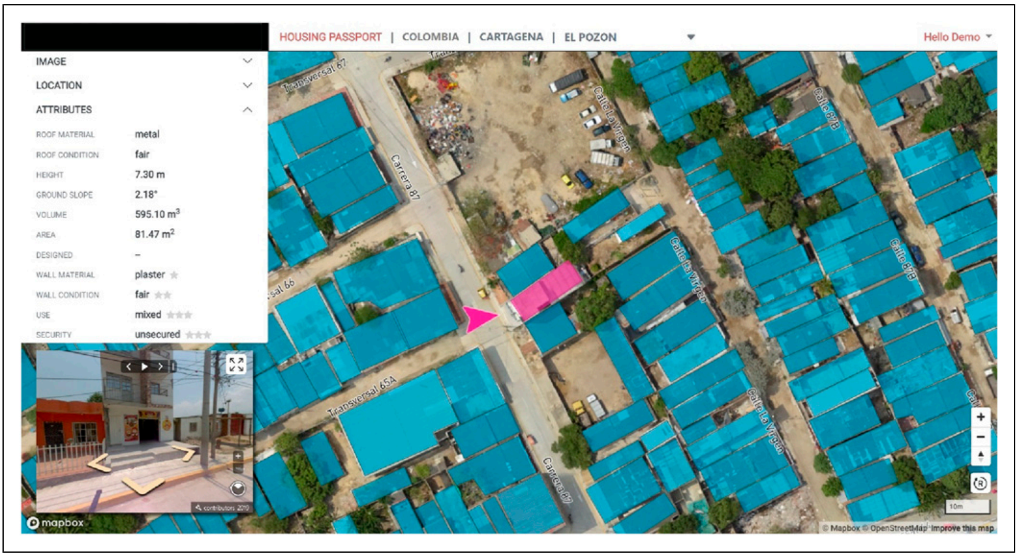

Figure 6. Selected building (pink) seen from the drone and street with related attributes.

experimented with building an extensive dataset containing street-level images taken from six areas of different geographic and demographic features, as shown in Table 4. The experimental dataset contains annotations of four characteristics (Table 6). A universal model was trained on this dataset. Satisfactory generalization in terms of high accuracy was achieved, as shown in Table 6.

Once the rooftop and street view predictions are joined in a geospatial database, they can be combined for further analysis. For example, a total condition screening can be summarized from each building's roof and wall condition predictions. In this case, if the roof and wall conditions are not the same designation of good, fair, or poor, the lower value of the two is taken to summarize the building condition (Table 7).

The portal contains layers for building footprints, drone imagery, and a base map connected to the street view imagery, giving the user an overview of the area of interest and interactive options to identify and compare results. Users can determine a sub-area of analysis, such as a block or neighborhood, and filter by characteristics. Results can be examined against drone imagery and by navigating along roads using imagery dynamically imported from Mapillary Street View service (Figure 6).

Discussion and conclusion

Harmonized, complete, and updated datasets of urban structures are more the exception than the rule due to their traditionally high acquisition cost and fast obsolescence. Yet as areas around the world change to accommodate population growth and adapt to threats from natural hazards, it becomes increasingly important to have recent, granular data on buildings, neighborhoods, and entire cities. Researchers have sought to address this challenge by automating methods of building screenings and land use classifications based on overhead or street-level images, and occasionally fusing data from both perspectives. However, there remains a gap between the research and the datasets available to government authorities and others working in the Global South (Kuffer et al., 2021).

Our work illustrates an opportunity to develop an extensive building stock dataset that spans several square kilometers. The methodology is proven valid for areas as large as 80 square kilometers. It also reaches regions and neighborhoods that have been underserved and overlooked in the past. Very little was found in the literature on collecting and fusing datasets from drone and street-level perspectives at the level of detail and scale offered by our study. Our approach of simultaneously collecting drone and street view imagery ensures a recent, granular, time-synchronized dataset, circumventing reliance on corporate collections that may be temporally limited or volunteer-based collections that may vary in quality (Biljecki and Ito, 2021; Cinnamon and Jahiu, 2021).

The performance of the machine learning classification models was satisfactory, with F1-scores greater than 70% and most well above 80%. Rooftop delineation from drone imagery can be challenging, particularly when buildings are connected or rooftops are otherwise densely grouped. A future goal is to refine the VGG-19 CNN, enabling it to differentiate rooftop tile material (e.g., metal, plastic) for units located in earthquake-hazard areas. In addition, GPT-4V could be used to improve the street view imagery workflow.

The methodology has the potential to be scalable and can be tailored to the needs of different cities. As noted earlier, one technical limitation of the machine learning-based method is the difficulty in generalization. Before a model can be applied with confidence in a new location (e.g., a new geographic area in this study), a validation study is usually needed to ensure the model produces reasonable predictions on the new data. Expanding the training data to include more variations, as we did in this study, is one effective way to overcome this issue.

Flood, storm surge, and earthquake models can be combined with building data in our open-source portal or in other GIS software to identify units vulnerable to hazards. In rapidly changing urban landscapes, new structures are readily detected and measured, locating urban sprawl and informal housing. Detailed characteristics, such as a building's use, size, masonry, and roof and wall conditions, can help to determine where to prioritize infrastructure investments and direct public program support. Governments can apply this approach to support property regularization, valuation, taxation, subsidy allocation, neighborhood improvements, and urban infrastructure development.

One issue emerging from machine learning research is the question of how the results are shared with interested citizens, if at all. Kuffer and colleagues recommend providing the resulting datasets to the local community (2021). Given the sensitivity of our datasets, local governments hold discretionary control. Still, if access is granted, the portal enables users to easily validate the machine learning predictions with drone and street view imagery. In this regard, our portal is valuable for conducting fundamental spatial analysis without requiring additional software or advanced training. Furthermore, the code to create a new portal is freely available, representing an opportunity to share data with those who may not have had access to it prior.

We have extended the urban clues proposed by Jacobs to automate geolocated urban observations from the sky and street. Urban clues detected with this methodology can be used as a guide to generate datasets to inform economic, business, and policy decisions. By applying machine learning to high-resolution imagery recently collected from the sky and street, we identify patterns at the block, neighborhood, and city levels, providing the spatial and temporal accuracy needed to assess detailed building characteristics.

This work could be expanded to identify other clues in the urban environment, for example, to characterize the quality of the urban infrastructure. Potential uses include planning and financing green, resilient, and efficient investments; identifying emergency routes and shelters during disasters; spotting the potential for installing solar energy systems on public buildings; and identifying streets that can be made pedestrian-only. It also could assist ongoing city services like waste, traffic, and infrastructure management.

Our approach is among the first to access drone and street view data to geographically situate detailed observations of the built environment, particularly in locations where data, software, and geospatial expertise may be limited. However, our methodology is applicable worldwide, offering a building screening that empowers decision-makers in urban management and helps to address some of their most important challenges.

Acknowledgments

This methodology, analysis, and data were produced by the World Bank's Global Program for Resilient Housing (GPRH) with the support of the Global Facility for Disaster Reduction and Recovery (GFDRR); Chaofeng Wang with the support from the NVIDIA Applied Research Accelerator Program. We would like to thank Alejandro Cantera López and Brian O'Hare for their contributions to this research.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Jessica Gosling-Goldsmith https://orcid.org/0000-0002-4351-3610

Data Availability Statement

The datasets generated during and/or analyzed during the current study are not publicly available due to privacy concerns. The online portal referenced in the study is available in the GitHub repository: https://github.com/GPRH/housing_portal.

References

Ajami A, Kuffer M, Persello C, et al. (2019) Identifying a slums' degree of deprivation from VHR images using convolutional neural networks. *Remote Sensing* 11(11): 1282. DOI: 10.3390/rs11111282.

Albert A, Kaur J and Gonzalez MC (2017) Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax NS Canada: Association for Computing Machinery, 1357–1366. DOI: 10.1145/3097983.3098070.

Bennett RM, Unger E-M, Lemmen C, et al. (2021) Land administration maintenance: a review of the persistent problem and emerging fit-for-purpose solutions. *Land* 10(5): 509. DOI: 10.3390/land10050509.

Biljecki F and Ito K (2021) Street view imagery in urban analytics and GIS: a review. *Landscape and Urban Planning* 215: 104217. DOI: 10.1016/j.landurbplan.2021.104217.

Cao R, Zhu J, Tu W, et al. (2018) Integrating aerial and street view images for urban land use classification. *Remote Sensing* 10(10): 1553. DOI: 10.3390/rs10101553.

Chen L, Yao X, Liu Y, et al. (2020) Measuring impacts of urban environmental elements on housing prices based on multisource data—a case study of Shanghai, China. *ISPRS International Journal of Geo-Information* 9(2): 106. DOI: 10.3390/ijgi9020106.

Cinnamon J and Jahiu L (2021) Panoramic street-level imagery in data-driven urban research: a comprehensive global review of applications, techniques, and practical considerations. *ISPRS International Journal of Geo-Information* 10(7): 471. DOI: 10.3390/ijgi10070471.

Clay G (1973) *Close-up: How to Read the American City*. Chicago: The University of Chicago Press.

Dakin K, Xie W, Parkinson S, et al. (2020) Built environment attributes and crime: an automated machine learning approach. *Crime Science* 9(1): 12. DOI: 10.1186/s40163-020-00122-9.

Deng C and Ma J (2015) Viewing urban decay from the sky: a multi-scale analysis of residential vacancy in a shrinking U.S. city. *Landscape and Urban Planning* 141: 88–99. DOI: 10.1016/j.landurbplan.2015.05.002.

Ding J, Zhang J, Zhan Z, et al. (2022) A precision efficient method for collapsed building detection in post-earthquake UAV images based on the improved NMS algorithm and Faster R-CNN. *Remote Sensing* 14: 663, MDPI AG. DOI: 10.3390/rs14030663.

Elamin A and El-Rabbany A (2022) UAV-based multi-sensor data fusion for urban land cover mapping using a Deep Convolutional Neural Network. *Remote Sensing* 14: 4298, MDPI AG. DOI: 10.3390/rs14174298.

Esteva A, Kuprel B, Novoa RA, et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639): 115–118. DOI: 10.1038/nature21056.

Fu X, Jia T, Zhang X, et al. (2019) Do street-level scene perceptions affect housing prices in Chinese megacities? An analysis using open access datasets and deep learning. *PLoS One* 14(5): e0217505. DOI: 10.1371/journal.pone.0217505.

Glaeser EL, Kominers SD, Luca M, et al. (2018) Big data and big cities: the promises and limitations of improved measures of urban life. *Economic Inquiry* 56(1): 114–137. DOI: 10.1111/ecin.12364.

Gonzalez D, Rueda-Plata D, Acevedo AB, et al. (2020) Automatic detection of building typology using deep learning methods on street level images. *Building and Environment* 177: 106805. DOI: 10.1016/j.buildenv.2020.106805.

Grippa T, Georganos S, Zarougui S, et al. (2018) Mapping urban land use at street block level using OpenStreetMap, remote sensing data, and spatial metrics. *ISPRS International Journal of Geo-Information* 7(7): 246. DOI: 10.3390/ijgi7070246.

He K, Gkioxari G, Dollár P, et al. (2017) Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)* 2980–2988. DOI: 10.1109/ICCV.2017.322.

Hermosilla T, Palomar-Vázquez J, Balaguer-Beser Á, et al. (2014) Using street based metrics to characterize urban typologies. *Computers, Environment and Urban Systems* 44: 68–79. DOI: 10.1016/j.compenvurbsys.2013.12.002.

Hoffmann EJ, Wang Y, Werner M, et al. (2019) Model fusion for building type classification from aerial and street view images. *Remote Sensing* 11(11): 1259. DOI: 10.3390/rs11111259.

Ibrahim MR, Haworth J and Cheng T (2021) From urban scenes to mapping slums, transport modes, and pedestrians in cities using deep learning and computer vision. *Environment and Planning B: Urban Analytics and City Science* 48(1): 76–93. DOI: 10.1177/2399808319846517.

Jacobs J (1961) *The Death and Life of Great American Cities*. New York: Random House.

Jacobs AB (1985) *Looking at Cities*. Cambridge: Harvard University Press.

Kang J, Körner M, Wang Y, et al. (2018) Building instance classification using street view images. *ISPRS Journal of Photogrammetry and Remote Sensing* 145: 44–59. DOI: 10.1016/j.isprsjprs.2018.02.006.

Koch D, Despotovic M, Sakeena M, et al. (2018) Visual estimation of building condition with Patch-level ConvNets. In: *Proceedings of the 2018 ACM Workshop on Multimedia for Real Estate Tech*. New York: Association for Computing Machinery, 12–17. DOI: 10.1145/3210499.3210526.

Koeva M, Gasuku O, Lengoiboni M, et al. (2021) Remote sensing for property valuation: a data source comparison in support of fair land taxation in Rwanda. *Remote Sensing* 13(18): 3563. DOI: 10.3390/rs13183563.

Kuffer M, Wang J, Thomson DR, et al. (2021) Spatial information gaps on deprived urban areas (slums) in low-and-middle-income countries: a user-centered approach. *Urban Science* 5: 72. DOI: 10.3390/urbansci5040072.

Law S, Paige B and Russell C (2019) Take a look around: using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology* 10(5): 1–19. DOI: 10.1145/3342240.

Lee S, Maisonneuve N, Crandall D, et al. (2015) Linking past to present: discovering style in two centuries of architecture. In: *2015 IEEE International Conference on Computational Photography (ICCP)*. Piscataway: IEEE, 1–10. DOI: 10.1109/ICCPHOT.2015.7168368.

Lindenthal T and Johnson EB (2021) Machine learning, architectural styles and property values. *The Journal of Real Estate Finance and Economics*. DOI: 10.1007/s11146-021-09845-1.

Lynch K (1960) *The Image of the City*. Cambridge: The MIT Press.

Maskeliūnas R, Katkevičius A, Plonis D, et al. (2022) Building façade style classification from UAV imagery using a pareto-optimized deep learning network. *Electronics* 11: 3450, MDPI AG. DOI: 10.3390/electronics11213450.

Movshovitz-Attias Y, Yu Q, Stumpe MC, et al. (2015) Ontological supervision for fine grained classification of Street View storefronts. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE, 1693–1702. DOI: 10.1109/CVPR.2015.7298778.

Ning H, Ye X, Chen Z, et al. (2022) Sidewalk extraction using aerial and street view images. *Environment and Planning B: Urban Analytics and City Science* 49(1): 7–22. DOI: 10.1177/2399808321995817.

Owusu M, Kuffer M, Belgiu M, et al. (2021) Towards user-driven earth observation-based slum mapping. *Computers, Environment and Urban Systems* 89: 101681. DOI: 10.1016/j.compenvurbsys.2021.101681.

Pratomo J, Kuffer M, Martinez J, et al. (2017) Coupling uncertainties with accuracy assessment in object-based slum detections, case study: Jakarta, Indonesia. *Remote Sensing* 9(11): 1164. DOI: 10.3390/rs9111164.

Ronneberger O, Fischer P and Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, et al. (eds) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 Lecture Notes in Computer Science*. Cham: Springer, Vol. 9351. DOI: 10.1007/978-3-319-24574-4_28.

Sariturk B, Kumbasar D and Seker DZ (2023) Comparative analysis of different CNN models for building segmentation from satellite and UAV images. *American Society for Photogrammetry and Remote Sensing* 89(2): 97–105. DOI: 10.14358/PERS.22-00084R2.

Simonyan K and Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: *3rd International Conference on Learning Representations (ICLR 2015)*. San Diego: Conference Track Proceedings, 1–14.

Stokes EC and Seto KC (2019) Characterizing and measuring urban landscapes for sustainability. *Environmental Research Letters* 14(4): 045002. DOI: 10.1088/1748-9326/aafab8.

Sun M, Zhang F and Duarte F (2021) Automatic building age prediction from street view images. In: *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*. Piscataway: IEEE, 102–106. DOI: 10.1109/IC-NIDC54101.2021.9660554.

UNDESA (United Nations Department of Economic and Social Affairs) (2019) *World Urbanization Prospects: The 2018 Revision*. New York: UNDESA.

Vanderhaegen S and Canters F (2017) Mapping urban form and function at city block level using spatial metrics. *Landscape and Urban Planning* 167: 399–409. DOI: 10.1016/j.landurbplan.2017.05.023.

Wang C, Antos SE and Triveno LM (2021) Automatic detection of unreinforced masonry buildings from street view images using deep learning-based image segmentation. *Automation in Construction* 132: 103968. DOI: 10.1016/j.autcon.2021.103968.

Wang C, Antos SE, Gosling-Goldsmith J, et al. (2024) Assessing climate disaster vulnerability in Peru and Colombia using street view imagery: a pilot study. *Buildings* 14(1): 14. https://www.mdpi.com/2075-5309/14/1/14.

Wang W, Shi Y, Zhang J, et al. (2023) Traditional village building extraction based on improved Mask R-CNN: a case study of Beijing, China. *Remote Sensing* 15(10): 2616. DOI: 10.3390/rs15102616.

Wang Y, Li S, Teng F, et al. (2022) Improved Mask R-CNN for rural building roof type recognition from UAV high-resolution images: a case study in Hunan Province, China. *Remote Sensing* 14: 265, MDPI AG. DOI: 10.3390/rs14020265.d.

Wegner JD, Branson S, Hall D, et al. (2016) Cataloging public objects using aerial and street-level images — urban trees. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE, 6014–6023. DOI: 10.1109/CVPR.2016.647.

Williams SH (1954) Urban aesthetics: an approach to the study of the aesthetic characteristics of cities. *Town Planning Review* 25(2): 95–113.

Wu S-S, Qiu X, Usery EL, et al. (2009) Using geometrical, textural, and contextual information of land parcels for classification of detailed urban land use. *Annals of the Association of American Geographers* 99(1): 76–98. DOI: 10.1080/00045600802459028.

Yu Q, Wang C, McKenna F, et al. (2020) Rapid visual screening of soft-story buildings from street view images using deep learning classification. *Earthquake Engineering and Engineering Vibration* 19(4): 827–838. DOI: 10.1007/s11803-020-0598-2.

Zeppelzauer M, Despotovic M, Sakeena M, et al. (2018) Automatic prediction of building age from photographs. In: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. New York: Association for Computing Machinery, 126–134. DOI: 10.1145/3206025.3206060.

Zhang W, Li W, Zhang C, et al. (2017) Parcel-based urban land use classification in megacity using airborne LiDAR, high resolution orthoimagery, and Google Street View. *Computers, Environment and Urban Systems* 64: 215–228. DOI: 10.1016/j.compenvurbsys.2017.03.001.

Zou S and 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. DOI: 10.1016/j.isprsjprs.2021.03.020.

Jessica Gosling-Goldsmith is a cartographer and geospatial consultant based in the Netherlands. Her work has appeared in numerous publications, including those for international organizations such as the United Nations and the World Bank. She has been a geospatial team member of the World Bank's Global Program for Resilient Housing since 2019. She holds an MSc in Cartography through the Erasmus program jointly offered by the Technical University of Munich, Technical University of Vienna, Technical University of Dresden, and University of Twente.

Sarah Elizabeth Antos is a Senior Land Administration Specialist at the World Bank who has worked extensively with spatial data in the field of international development. In particular, her work has focused on remote sensing and the classification of high-resolution imagery. Before joining the World Bank, she worked at the World Health Organization and US's Office of Foreign Disaster Assistance, where she used satellite imagery and survey data to improve disaster recovery efforts. She holds a bachelor's and master's degree in Geography from The George Washington University.

Luis Miguel Triveno is a Senior Urban Development Specialist at the World Bank. He has led and executed social, urban and rural development reform projects in over 25 countries in Europe, Asia, Africa, and Latin America. Between 2012 and 2015, he was Chief Executive Officer of Pro-expansion, a think-tank and reform-oriented organization based in Lima, Peru. Between 2005 and 2012, he served as Research Economist, Chief Economist, Chief Operating Officer and Chief Executive Officer of Hernando de Soto's Institute for Liberty and Democracy.

Adam R Benjamin, Ph.D., is a senior geospatial consultant and has been a drone specialist team member for the World Bank's Global Program for Resilient Housing since 2019. He has over 15 years of experience in reality capture, remote sensing, and geospatial analysis. He previously worked for over a decade as the Geomatics Specialist and technical lead for unmanned aerial systems mapping operations at the University of Florida in Fort Lauderdale. He is a Professional Surveyor and Mapper (PSM), a certified GIS Professional (GISP), and holds an FAA Remote Pilot certificate.

Chaofeng Wang, Ph.D., is specialized in computational mechanics, uncertainty and risk quantification, AI, and their applications in the natural and built environment. He joined the University of Florida as an Assistant Professor of Artificial Intelligence. Before that, he earned degrees in Engineering Mechanics and Civil Engineering from Central South University and Clemson University and worked as a postdoctoral scholar at the University of California, Berkeley.