This is the Accepted Manuscript version of an article published by Elsevier in the journal *ISPRS Journal of Photogrammetry and Remote Sensing* in 2024, which is available at:
https://doi.org/10.1016/j.isprsjprs.2024.06.023
Cite as: 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.

Global Streetscapes – A comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics

Yujun Hou[a], Matias Quintana[b], Maxim Khomiakov[a,c], Winston Yap[a], Jiani Ouyang[a,d], Koichi Ito[a], Zeyu Wang[a], Tianhong Zhao[e], Filip Biljecki[a,f,*]

[a]*Department of Architecture, National University of Singapore, 4 Architecture Dr, 117566, Singapore*
[b]*Singapore-ETH Centre, Future Cities Lab Global Programme, CREATE campus, 1 Create Way, #06-01 CREATE Tower, 138602, Singapore*
[c]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Bygning 324, Kgs. Lyngby, 2800, Denmark*
[d]*State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Luoyu Road 129, Wuhan, 430079, China*
[e]*College of Big Data and Internet, Shenzhen Technology University, 3002 Lantian Road, Shenzhen, 518118, China*
[f]*Department of Real Estate, National University of Singapore, 15 Kent Ridge Drive, 119245, Singapore*

Abstract

Street view imagery (SVI) is instrumental for sensing urban environments, benefitting numerous domains such as urban morphology, health, greenery, and accessibility. Billions of images worldwide have been made available by commercial services such as Google Street View and crowdsourcing services such as Mapillary and KartaView where anyone from anywhere can upload imagery while moving. However, while the data tend to be plentiful, have high coverage and quality and are used to derive rich insights, they remain simple and limited in metadata as characteristics such as weather, quality, and lighting conditions remain unknown, making it difficult to evaluate the suitability of the images for specific analyses. We introduce Global Streetscapes — a dataset of 10 million crowdsourced and free-to-use SVIs sampled from 688 cities across 210 countries and territories, enriched with more than 300 camera, geographical, temporal, contex-

*Corresponding author (filip@nus.edu.sg)

tual, semantic, and perceptual attributes. The cities included are well balanced and diverse, and are home to about 10% of the world's population. Deep learning models are trained on a subset of manually labelled images for eight visual-contextual attributes pertaining to the usability of SVI — panoramic status, lighting condition, view direction, weather, platform, quality, presence of glare and reflections, achieving accuracy ranging from 68.3% to 99.9%, and used to automatically label the entire dataset. Thanks to its scale and ready-to-use pre-computed standard semantic information, the data can be readily used to benefit existing use cases and to unlock new applications, including multi-city comparative studies and longitudinal analyses, as affirmed by a couple of use cases in the paper. Moreover, the automated processes and open-source code facilitate the expansion of the dataset and new updates. With the rich manually annotated information, some of which are provided for the first time, and diverse conditions present in the images, the dataset also facilitates assessing the properties of crowdsourced SVIs and provides a benchmark for training and evaluating future computer vision models. We make the Global Streetscapes dataset and the code to reproduce and use it publicly available in `https://github.com/ualsg/global-streetscapes`.

Keywords: Urban analytics, volunteered geographic information, data fusion, GeoAI, machine learning, spatial data infrastructure

---

1. Introduction

Street view imagery (SVI) is rapidly emerging as a prominent geospatial data source for sensing, measuring, and understanding our complex and dynamic urban environments, rivalling traditional remote sensing sources such as satellite imagery [1, 2, 3, 4, 5]. Such development is spurred by the increasing coverage and ease of access to data and advancements in computer vision (CV) techniques to automatically extract a vast array of information from it. SVI has been used on its own or in conjunction with other urban data (e.g. street networks, buildings information, demographic and socioeconomic data, questionnaires and surveys, etc.) to drive urban research topics including but not limited to spatial data infrastructures [6, 7, 8], urban mobility (e.g. bikeability [9, 10], walkability [11, 12], traffic speed [13]), urban infrastructure assessment [14], physical disorder [15], 3D building models reconstruction [16], urban greenery [17, 18, 19], urban health [20], urban perception [21, 22, 23, 24, 25, 26, 27], urban density [28], real estate prices [29, 30], socioeconomic and cultural activities and behaviours [31, 32, 5], and urban soundscapes [33]. The availability of historical SVI has

also enabled understanding changes in urban landscapes [34, 35, 26].

Prominent sources include commercial SVI from Google Street View, Baidu Maps, and Tencent Street View, as well as crowdsourced SVI managed by two services—Mapillary and KartaView [36, 37]. Notably, as an emerging form of Volunteered Geographic Information (VGI), crowdsourced SVI captures the urban environment from a more diverse range of viewpoints, locations, and ambient conditions than their commercial counterparts, because they can be collected anywhere, anytime, by anyone, akin to other VGI such as OpenStreetMap [4].

Such diversity enables a variety of applications and may be a cornerstone in benchmarking certain CV innovations that rely on testing on realistic data, which is heterogeneous in nature, in contrast with the controlled and standardised acquisition protocols of commercial providers. Some particular advantages of crowdsourced imagery are that it can cover informal settlements and less developed regions, which do not tend to be captured in commercial data [38, 39, 40, 41], it is released under a liberal license, and in some areas, it may be more dynamic due to temporally finer granularity [42]. While commercial data continue to reign supreme [1], the use of crowdsourced imagery has been picking up momentum in the international scientific literature [43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54].

However, each having pros and cons, both commercial and crowdsourced SVI suffer from a common issue — there is limited information about the characteristics of images, and this limitation prevents selecting only images with the desired quality and properties, potentially adversely affecting the reliability of studies. For example, for bikeability or walkability studies, it would be ideal if only images captured from the right point of view (e.g. by cyclists, pedestrians) or the right platform (e.g. cycleways, sidewalks) are used [55]; studies on perception could benefit by controlling the ambient conditions (e.g. filter SVIs by weather, season, day/night time) to minimise their confounding effects on the study result; urban morphology studies relying on perspective images (in places with limited availability for panoramas) to calculate metrics such as building view index and sky view index could achieve higher accuracy when using images facing the front or back of the road compared to images facing sideways [36]; while night images could be considered noise in some applications, they could be potentially useful in assessing nocturnal urbanscapes (e.g. night-time safety perception) or capturing diurnal variations in urban activities.

Despite the rapid growth of urban research powered by SVI, disproportionately little attention has been given to the above-mentioned issues beyond merely acknowledging them. A variety of images have often been used without researchers being adequately informed of the data quality and fitness for use, largely

due to the absence of automated means to produce such information. This bottleneck could limit the accuracy and effectiveness of studies relying on SVIs (especially crowdsourced SVIs), consequently limiting the usability of such data and its potential as a remarkably versatile data source. Further, obtaining and processing imagery, especially for several cities, can be cumbersome and computationally prohibitive.

We identify two main challenges contributing to this conundrum: 1) the lack of information (metadata) to describe the contextual characteristics of SVIs to aid in the assessment of their suitability for use, and 2) the lack of benchmark datasets to facilitate advancements in algorithms to extract such contextual information from SVIs automatically. While there exist a variety of SVI datasets to benchmark CV prediction tasks related to understanding urban scenes and places (e.g. Cityscapes [56], Mapillary Vistas [57], Mapillary Street-Level Sequences [58], BDD100K [59] etc.), they mostly focus on applications and problems concerned in the CV domains such as autonomous driving, augmented reality, and mobile robotics, and thus do not provide much information for usage in urban applications or lack worldwide coverage. Though the urban science fields have also substantially benefitted from the CV advancements powered by these datasets—mainly in the use of semantic segmentation and object detection techniques to automatically extract information from SVIs, even greater potentials could be achieved from SVI datasets tailored for urban research, especially one that could considerably enhance the usability of SVI by providing comprehensive auxiliary information of the images to assess their fitness for use.

These recent advancements reaffirm the timeliness, novelty, and potential of the work introduced in this paper, which focuses on the development of a comprehensively enriched and labelled SVI dataset, named *Global Streetscapes*, to advance the use of SVI in urban science research. The open data repository we developed consists of metadata about over 10 million entirely crowdsourced SVIs obtained from Mapillary and KartaView, sampled from 688 cities worldwide. All images are enriched with extensive metadata, geographical, temporal, and contextual information (totalling more than 341 attributes) to facilitate their use and to promote their integration with other data. Specifically, images are merged with other global geospatial datasets based on their locations and have further temporal attributes calculated based on their capture timestamps. A subset of the images has been manually tagged with eight contextual properties that we deem helpful to describe the conditions of the images for evaluating their fitness for use but are not available in their metadata, and are commonly used in research [60]. Using these manual tags, we developed models for the associated classification

tasks and ran inference for the remaining images to label them with these attributes automatically. These manual tags can also serve as benchmarks for future approaches and support their continuous development. To further enhance the usability of the dataset for urban science and geospatial applications, we also ran inference from models that have been trained on well-known datasets (e.g. Mapillary Vistas [57], Places [61], Place Pulse 2.0 [62, 63]) to expedite specific tasks commonly associated with SVI-driven urban applications, such as semantic segmentation, instance detection, scene type classification, and human perception classification [64]. Figure 1 shows the distribution of geographical locations and quantity of SVIs in Global Streetscapes. Figure 2 shows a mosaic of thousands of images from Global Streetscapes, demonstrating the diverse locations, scenes, viewpoints, ambient and camera settings included in the dataset.

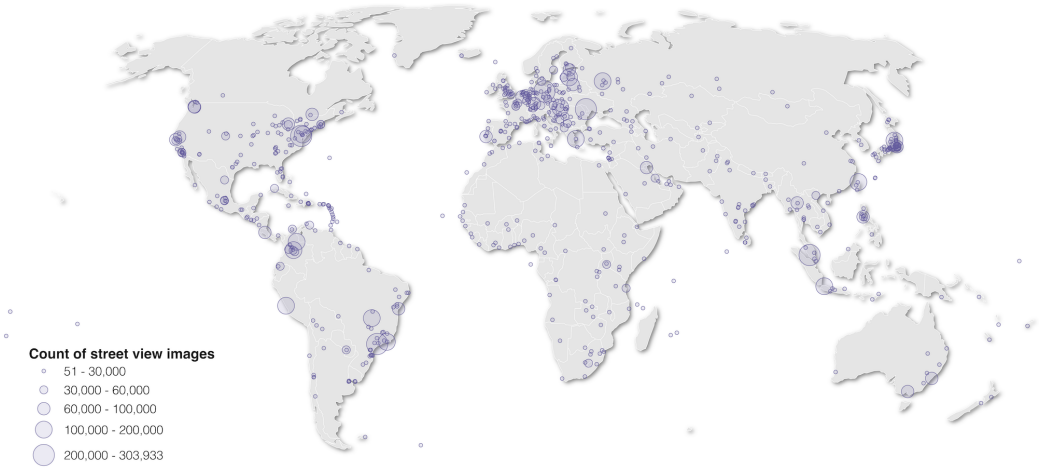

Figure 1: Overview of the geographic coverage of the Global Streetscapes dataset across 688 cities, illustrating the amount of SVIs available for each city.

Our key contributions, both scientific and practical, and spanning acquisition, processing, harmonisation and utilisation of street view image data and analytics, are summarised as follows, and will be further elaborated in the continuation of the paper:

- Constructed Global Streetscapes, a worldwide dataset of 10 million crowdsourced SVIs sampled from Mapillary and KartaView, covering 688 cities around the world, which account for about 10% of the world's population, enriched with more than 300 camera, geographical, temporal, contextual,

perceptual, and semantic attributes and has wide geographical, environmental, and temporal diversity.

- Developed a reproducible framework to 1) sample and synthesise crowdsourced SVIs from two different sources, 2) enrich them with rich auxiliary information to facilitate their usage and integration with external datasets, and 3) enable future updates by fetching the latest available data from the aforementioned sources.
- Created the first dataset with curated manual labels and baseline CV models for benchmarking models for SVIs and urban data science.
- Discussed how Global Streetscapes could potentially answer novel research questions and drive new applications or enhance existing ones.
- Reduced the entry barrier to SVI research by providing a diverse, vast, off-the-shelf solution, increasing equity and participation from researchers who are less experienced with computational methods, and saving time and effort for seasoned researchers.

2. Existing datasets

In recent years, many open SVI datasets have been constructed to aid the development of tools and algorithms to sense the complex urban environments, by providing rich labels at either the element or the scene level, or both [3].

For element-level sensing, Cityscapes [56] and Mapillary Vistas [57] are among the most prominent street-level imagery datasets to provide labels for a large number of street object categories for benchmarking semantic segmentation and instance detection algorithms to extract urban street elements. While both contain 25,000 high-resolution images, the Mapillary Vistas images exhibit a wider range of ambient conditions (e.g. weather, lighting condition) and geographical coverage, hence better approximating real-world scenarios and providing a more robust benchmark. Some datasets, on the other hand, focus on specific elements in the street scene. For example, the Mapillary Traffic Sign Dataset (MTSD) [65] details over 300 traffic sign classes with 100,000 high-resolution images that come with bounding box annotations. Similar to Mapillary Vistas, these images were gathered from locations all around the world under diverse ambient conditions. However, for both Mapillary Vistas and MTSD, though they have contextual

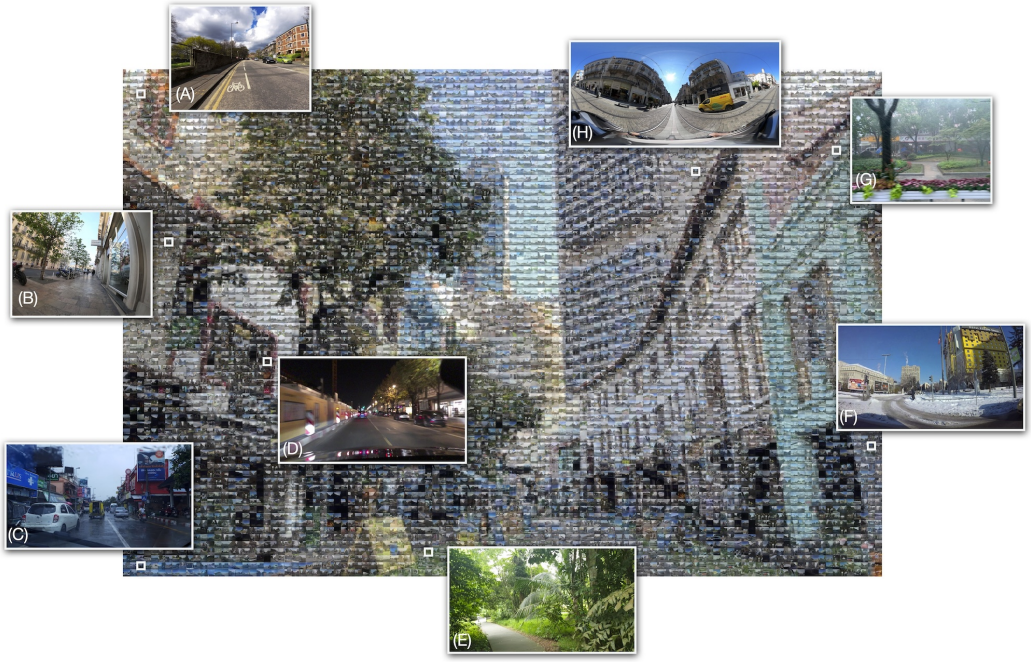

Figure 2: A mosaic of 5,720 SVIs from Global Streetscapes (0.057% of the dataset), demonstrating the diverse scenes, viewpoints, ambient conditions, and camera settings included in the dataset: perspective images facing the front of the road taken on a cycleway in Edinburgh (A), a sidewalk in Toulon (B), a rainy road in Chiang Mai (C), a night road in Berlin (D), a walking trail in a nature park in Singapore (E), or a snowy road in Sarajevo (F); a perspective image facing the side of the road taken in a rainy day in Pudong (G); a panorama taken from a cyclist riding along tram tracks in Orléans (H). The mosaic forms a single SVI featuring the unique streetscape at the Circular Road in Singapore, with lines of iconic shophouses against a backdrop of modern development. Our heterogeneous dataset will increase geographical diversity, scene variety, cultural and landscape mixture, and equity in urban analytics. Source of imagery: Mapillary and KartaView contributors.

Table 1: Characteristics of existing open SVI datasets constructed for computer vision and urban research applications, and our Global Streetscape dataset (MSLS: Mapillary Street-Level Sequences; MTSD: Mapillary Traffic Sign Dataset; MPOINSLI: Mapillary POI-Neighborhood Street-Level Images; OSV-5M: OpenStreetView-5M; CV: computer vision; UR: urban research; '-': information unavailable; '✓': information available; '✓*': information and manual label available). Besides a rich set of metadata, our dataset offers advantages such as multidisciplinary use and multi-source provenance.

Dimension	**Description**	**Cityscapes**	**Mapillary Vistas**	**MTSD**	**MSLS**	**BDD100K**	**Place Pulse 2.0**	**MPOINSLI**	**OSV-5M**	**Global Streetscapes (*Our work*)**
Purpose	Domain: task	CV: semantic urban scene understanding		CV: traffic signs detection and classification	CV: lifelong place recognition	CV: heterogeneous multitask learning	UR: human perception	UR: POI characteristics	CV: geolocation	UR+CV: SVI usability, scene understanding, urban analytics
Lineage	Image source	Self on-site collection	Mapillary	Mapillary	Mapillary	On-site collection by drivers	Google Street View	Mapillary	Mapillary	Mapillary, KartaView
Coverage	No. of images	25,000	25,000	100,000	1.6 million	120 million	110,988	167,743	5.1 million	10 million
	No. of cities	50	N/A	N/A	30	N/A	56	1	70k	688
	No. of continents	1	6	6	6	1	6	1	6	6
	No. of years covered	1	N/A	N/A	9	N/A	6	7	13	> 13
Metadata enrichment	Street network	-	-	-	-	-	-	-	-	✓
	Degree of urbanisation	-	-	-	-	-	-	-	-	✓
	Administrative area	-	-	-	-	-	-	-	✓	✓
	Spatial index	-	-	-	-	-	-	-	-	✓
	POI	-	-	-	-	-	-	✓	-	-
	Land cover	-	-	-	-	-	-	-	✓	-
	Soil type	-	-	-	-	-	-	-	✓	-
	Driving side	-	-	-	-	-	-	-	✓	-
	Distance to sea	-	-	-	-	-	-	-	✓	-
	Climate	-	-	-	-	-	-	-	✓	✓
	Season	-	-	-	-	-	-	-	-	✓
	Weather	-	-	-	-	✓*	-	-	-	✓*
	Lighting condition	-	-	-	✓	✓*	-	-	-	✓*
	Platform	-	-	-	-	-	-	-	-	✓*
	View directions	-	-	-	✓	-	-	-	-	✓*
	Panoramic status	-	-	-	✓	-	-	-	-	✓*
	Quality levels	-	-	-	-	-	-	✓	-	✓*
	Presence of glare	-	-	-	-	-	-	-	-	✓*
	Presence of windshield reflection	-	-	-	-	-	-	-	-	✓*
	Scene type	-	-	-	-	✓*	-	-	-	✓
	Perception	-	-	-	-	-	✓*	-	-	✓
	Semantic segmentation	✓*	✓*	-	-	✓*	-	-	-	✓
	Instance detection	✓*	✓*	✓*	-	✓*	-	-	-	✓

diversity—which is good for catering to diverse use cases—information describing the ambient conditions is not included in the accompanying metadata. The lack of labels to describe the contextual characteristics of SVIs makes it difficult to develop automated means to efficiently extract such information.

The Mapillary Street-Level Sequences (MSLS) [58] and the BDD100K [59] datasets provide scene-level labels that partially describe the diverse contextual characteristics of SVIs. The MSLS is constructed for lifelong place recognition, containing 1.6 million images that spread over 30 major cities in 6 continents captured over nine years. The dataset provides auxiliary information describing the view direction, lighting condition, and panoramic status of the images. This information was derived from the image's metadata instead of manual labels, so its reliability largely depends on the accuracy of the metadata, which could greatly vary among the wide range of devices used to capture the SVI. The BDD100K [59] is a driving dataset constructed for heterogeneous multitask learning, providing labels for weather, lighting conditions, and scene types alongside manual labels for semantic segmentation and instance detection. While BDD100K totals a massive count of 120 million images, the images were all collected in the United States, without a worldwide geographical representation. Further, it does not provide other visual features pertaining to the use of SVIs, such as view direction, platform, and quality. GSV-Cities [66] is another dataset constructed for place recognition, with 560,000 Google Street View images from more than 40 cities over a 14-year period that have diverse appearance variations, though such varying ambient conditions are not annotated.

While not primarily focused on SVIs, other well-known datasets that characterise urban places and objects including Places [61] and ADE20K [67] have been widely utilised in urban analytics as well. Although the above datasets annotate SVIs with a series of rich semantic features, they are not enriched with additional metadata such as street networks and a variety of other features that we will describe later, which could limit their versatility. For example, researchers have used the Places dataset [61] with OSM information to produce accurate classification of rural, urban roads, and highways [68].

Multiple SVI datasets have been curated for urban research purposes as well. One such prominent dataset is the MIT Place Pulse 2.0 dataset [62], which established means to quantitatively assess urban scenes for multiple dimensions of human perception toward the built environments. It contains 110,988 images and 1,170,000 pairwise comparisons provided by 81,630 online volunteers along six perceptual attributes: safe, lively, boring, wealthy, depressing, and beautiful. These images, obtained from Google Street View, cover 56 cities across six conti-

nents and span a time range of six years. As the images are not crowdsourced, they are more standardised in terms of data collection, i.e. image quality and conditions are comparable, and thus do not exhibit much diversity in scene conditions. Information pertaining to the ambient conditions of the images is unavailable. Such perceptual datasets could be used to reveal changes in urban environments as well [69]. In addition, datasets have been curated to classify architectural styles [70] and age [71], identify shop storefronts [72], and detect road damage [73].

There is also an increasing effort to merge SVIs with additional data to support urban research. Merging points of interest (POIs) data with SVI, The Mapillary POI-Neighborhood Street-Level Images (MPOINSLI) dataset [44] curates 167,743 SVIs from Mapillary that have been filtered to have a view of 6,732 unique POIs and their neighbourhoods, potentially supporting further analyses such as POI-scene recognition and fine-grained land use classification. On the other hand, [74] developed a modelling framework, URBAN-i, that detects informal settlements, pedestrians, and vehicle types from aerial imagery and SVI, and combines them with spatiotemporal data (location coordinates, date, and time) to map urban dynamics. These efforts demonstrate the potential and usefulness of enriching SVIs with other data sources to support diverse urban analytics applications.

OpenStreetView-5M (OSV-5M) is, at the time of writing, the largest worldwide, open-access SVI dataset constructed for geolocation using crowdsourced SVIs from Mapillary [75]. The authors randomly sampled one image per cell on a 100 $\times$100m grid across the entire world, to achieve a balanced distribution worldwide. Each image is further enriched with additional metadata including the associated administrative area, land cover, soil type, driving side, and distance to the sea. While such auxiliary information is beneficial for the purpose of geolocation, it may not be sufficient for our purpose of enhancing SVI usability for urban analytics. Nonetheless, it is yet another dataset that demonstrates the value of crowdsourced SVI.

Scene diversity also plays an important role in research on the cross-domain performance of CV algorithms. VALERIE22 is a dataset of synthetic street scenes generated to study domain-specific factors that influence perception performance of deep neural networks [76]. The highly diverse 3D scenes are generated by varying multiple factors including the width of a street or pavement, scene type, materials for roads and sidewalks, placement and density of cars, vegetation, road elements and pedestrians, position of camera, time of the day, etc. The dataset also comes with a rich set of metadata describing the specific scene and semantic features. However, it is not geographically diverse, hence not fully fulfilling our

research objectives.

While all these SVI datasets have considerably contributed to CV and urban research, none could adequately address the two research challenges described in Section 1. It is observed that enriching SVI with other data sources could further enhance its usability and drive novel applications. Additionally, most of the datasets strive to cover broad geographical regions and diverse conditions to approximate real-world scenarios to robustly assure the generalisability of models or approaches derived from the dataset. These findings support our research objectives outlined in Section 1 and underline the importance and need for a dataset that not only exhibits geographic, temporal, environmental, and viewpoint diversity, but also comes with comprehensive labels to inform its usage and rich auxiliary data sources to inspire new applications. Moreover, while most existing datasets focus only on a single source of data (e.g. GSV), our dataset is one of the first efforts to synthesise more than one source of crowdsourced SVI. Such harmonisation can improve data availability as the coverage of one source can complement that of the other. It is not uncommon that some cities only have data from one source but not from the other.

Table 1 summarises the characteristics of various datasets related to our research objectives and how our dataset, Global Streetscapes, differentiates itself and contributes to making the best use of SVI.

3. Methodology

3.1. City selection and data download

Cities were selected from the SimpleMaps World Cities Database [77], a widely used dataset [78], which contains 42,905 cities across 239 countries, using a combination of methods (Figure 3A). As the initiation of a global-scale dataset, we first focused on cities with a population greater than 50,000 for a greater chance of data availability of crowdsourced imagery. This is also following the recommendation by the UN Statistical Commission which delineates a city based on a total population of at least 50,000 as one of their criteria [79], considering that the dataset is constructed primarily for use in urban research. This resulted in 8,729 cities across 189 countries for further sampling: For each of the 189 countries, 6% of the available cities from each country were randomly sampled, yielding 501 candidate cities for data download (Figure 3A). This stratified sampling helps to balance the geographical coverage of the candidate cities and ensure countries are represented proportionally to the number of cities they have, preventing the

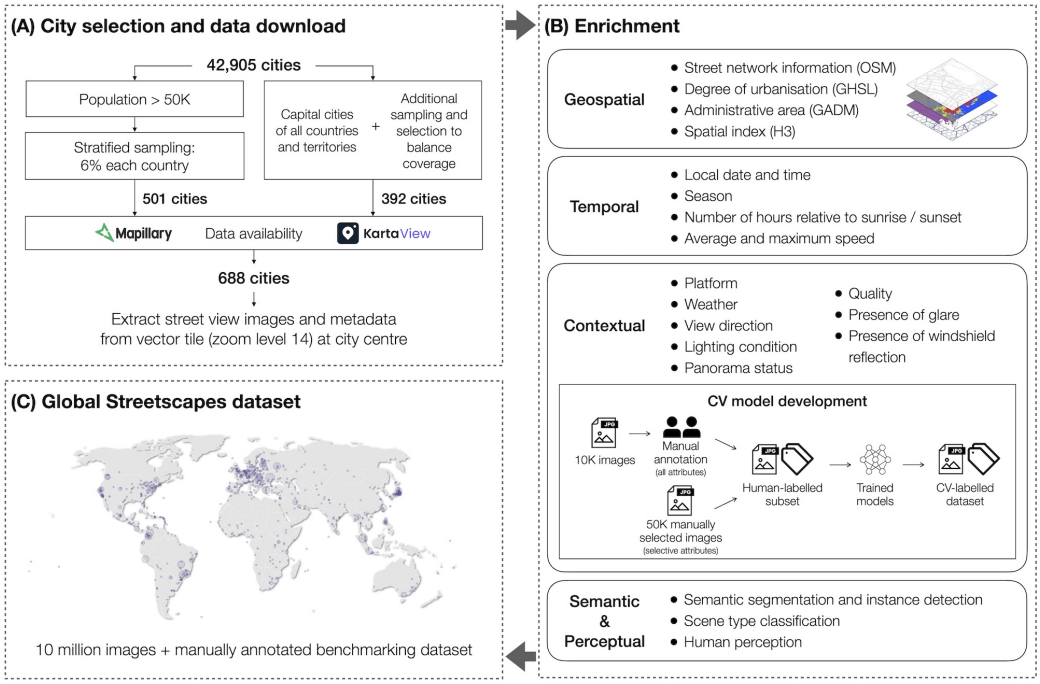

Figure 3: The methodology framework, from city selection and data download (A) to data enrichment (B), to produce the Global Streetscapes dataset (C).

cases where a disproportionately large (or small) number of cities are sampled from a single country.

To ensure all territories are represented in the dataset, we added the capital cities of all countries and territories to the candidate cities (if they were not already included). To further balance the geographic distribution of training data, we used a worldwide 1 × 1 km grid from WorldPop [80] to randomly sample 100 cells around the world and extract a small number of SVIs from these cells. The nearest cities to these images were determined from the World Cities Database using a *k*-d tree method, and were added to the candidate cities. A number of cities were also manually added to the candidate cities, following a visual inspection of the geographical distribution of the candidate cities. The above measures resulted in 392 more cities added to the pool of candidate cities (Figure 3A).

For each of the candidate cities, the city's point coordinates (latitude and longitude) were extracted from the World Cities Database. Mapillary and KartaView SVIs, along with their metadata, were then downloaded from the vector tile (which measures approximately 2.4 km × 2.4 km at the Equator, at zoom level of 14) associated with the city's coordinates, using the Mapillary Python Software Development Kit (SDK) [81] and the KartaView Application Programming Interface (API) [82]. Depending on data availability, only the cities with data available from either Mapillary or KartaView were included in the final dataset.

This entire process resulted in a large collection of SVIs spanning 688 cities from 210 countries and regions (Figure 1). The cities are of varying sizes, from having hundreds or thousands of inhabitants to being home to millions of urban dwellers. Together, these cities cover approximately 10% of the world's population. In total, 10 million SVIs were obtained, with 8.89 million (88.9%) of them from Mapillary and 1.11 million (11.1%) from KartaView.

3.2. Geospatial enrichment

All SVIs in the dataset are merged with multiple geospatial data sources to provide the geographical context for each SVI (e.g. the street and the administrative area it is located in, and the degree of urbanisation of the area it is located in), and to spatially index the image to promote future integration with other data (Figure 3B).

Street network information. The street networks within the 2.4 × 2.4 km sample area in each city were extracted from OpenStreetMap (OSM)[1] using the

[1]https://www.openstreetmap.org

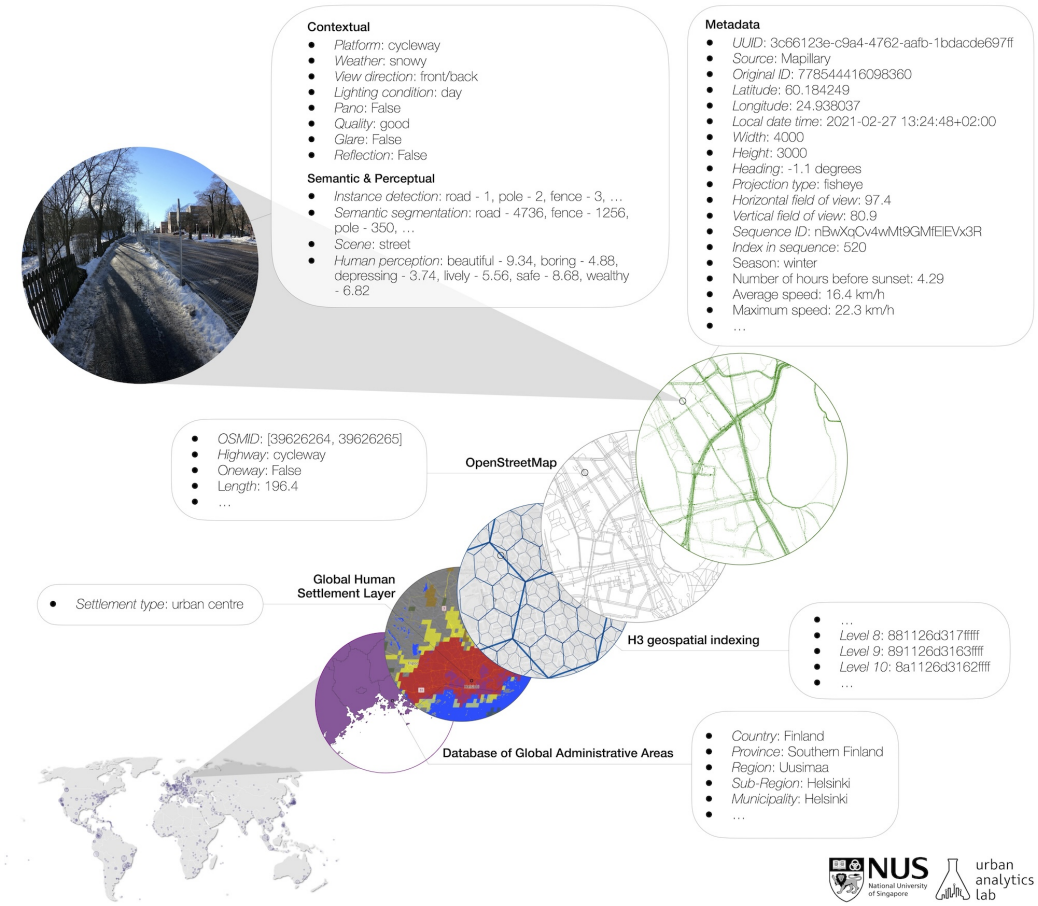

Figure 4: We processed the original temporal and camera metadata of all street view images in our dataset, further enriched them with multiple sources of spatial data based on location coordinates, and labelled them with contextual, semantic, and perceptual attributes using computer vision. The area featured in this illustration is in Helsinki, Finland, and shows the different data sources and the kind of data it provides. Source: Mapillary, KartaView, OpenStreetMap, Uber, European Commission, GADM.

Python package OSMnx [83], and matched with the SVIs based on distance. By associating SVIs with their surrounding street networks, we can learn about the street context of where the SVIs locate, e.g. the name, address, and type of the street, to inform their fitness for use, or analyse the distribution of SVIs at the street level, to inform data coverage [4]. This data fusion also enables the aggregation of SVI data [4, 84] and further, its fusion with other data types (e.g. building data, population data, POI data, social sensing data, satellite imagery), using streets as the spatial unit, to represent urban space more comprehensively and develop expressive, multimodal machine-learning models [45, 84].

Degree of urbanisation. The degree of urbanisation at the location of each image was obtained from the Global Human Settlement Layer (GHSL) published by the European Commission [85], and provides one of the following eight values: urban centre, dense urban cluster, semi-dense urban cluster, suburban or peri-urban, rural cluster, low density rural, very low density rural, or water. This information provides the settlement context of the images and could help researchers identify images from areas of interest.

Administrative area. The Database of Global Administrative Areas (GADM)[2] is a spatial database of administrative areas around the world, available from the country level to further subdivisions such as counties and provinces. We spatially joined our SVIs with the GADM dataset for all administrative levels available (up to six) at the image's location. These attributes could enable the integration of our data with other data types (e.g. demographic and socioeconomic datasets) that are based on administrative levels.

Spatial index. We spatially indexed our SVIs on the H3 hierarchical indexing system[3], a standardised global grid system that partitions the world into regular hexagonal cells [86], at all 16 levels of resolution. This transformation could further support the integration of our data with other H3-indexed global datasets, such as the Kontur Population dataset[4], and facilitate multi-scale H3-based spatial analysis which is common in urban SVI research [87, 36, 35, 33].

Figure 4 shows an overview of the above-mentioned external sources used for the geospatial enrichment and the related attributes. Further technical details of the geospatial enrichment processes are documented in Appendix A.2.

[2]https://gadm.org/

[3]https://h3geo.org/

[4]https://www.kontur.io/portfolio/population-dataset/

3.3. Temporal enrichment

Local date and time. Mapillary and KartaView provide detailed timestamp of the image capture, unlike GSV which only provides the year and month of capture. For Mapillary, the capture date and time of an image was expressed as a Unix timestamp, while for KartaView, it was given as a date-time string in Greenwich Mean Time (GMT). We converted the timestamps from both sources to the local time zone and expressed the local date time in a unified date-time string format. This provides a more localised temporal context for the image, making it easier to identify the time of day at which the image was captured. The original timestamps given by both sources are also kept in the dataset.

Climate and season. For each city, its Köppen climate classification was obtained based on location, using an API[5] based on a database run by the Institute for Veterinary Public Health and the Provincial Government of Carinthia in Austria[6]. Images from all tropical climate regions were then labelled as 'tropical' for season. For the remaining images, we estimated the season based on the month it was taken in and whether it is located in the northern or southern hemisphere. This information is helpful because the result of some SVI-based studies could be sensitive to seasons, a topic that has recently gained more attention [88, 89, 34]. As a prime example, changes in vegetation through growing and non-growing seasons could affect the calculation of the green view index in urban landscape studies [2].

Number of hours relative to sunrise or sunset. Using the Python package PyEphem [90], we calculated how long an image was taken before or after sunrise or sunset. This information could be used to infer the lighting condition in the image.

Average and maximum speed. Crowdsourced SVIs are collected by volunteers using various modes of transport (e.g. car, bicycle, and on foot) [91, 4]. We calculated the average speed of each geotagged sequence as the ratio of the total sequence length to the time duration between the first and last images of the sequence to deduce the mode of transport associated with the trajectory. The average speed across each segment between each pair of consecutive images was also calculated, and the maximum value across the sequence was used as the maximum speed for the sequence. The variance among the segment speeds was also calculated for each sequence. These statistics could help us infer the transportation mode associated with each sequence, potentially facilitating VGI research on

[5]https://github.com/sco-tt/Climate-Zone-API

[6]https://koeppen-geiger.vu-wien.ac.at/

Table 2: Overview of the eight manually labelled attributes and their characteristics. All attributes are at image level and (*) attributes are in most cases consistent at the sequence level.

Attribute	Data Type	No. of classes	Possible values
Platform*	String	6	driving/walking/cycling surface, railway, fields, tunnel
Weather*	String	5	clear, cloudy, rainy, snowy, foggy
View direction*	String	2	front/back, side
Lighting condition*	String	3	day, night, dusk/dawn
Panoramic status*	Boolean	2	true, false
Quality	String	3	good, slightly poor, very poor
Glare	Boolean	2	true, false
Reflection	Boolean	2	true, false

contributor behaviour [92, 93].

Further technical details of the temporal enrichment processes are documented in Appendix A.2.

3.4. Contextual enrichment by computer vision

3.4.1. Contextual attributes

Besides geospatial and temporal enrichment, as the core part of our work, we have considered numerous other attributes that provide essential contextual information about the image's characteristics to facilitate the 'fitness for use' evaluation.

Platform. The type of platform (e.g. road, cycleway, sidewalk, etc.) on which an image is taken could inform the perspective of an image, i.e. whether it is taken from a car-, cyclist-, or pedestrian-oriented environment. This feature is important for studies on topics such as human perception, walkability, and bikeability, which have attracted heightened interest in very recent years, as these are important aspects of evaluating urban liveability and sustainability. There is even an effort to exclusively collect SVIs on walking infrastructure such as footpaths and sidewalks to facilitate pedestrian-related use cases[7], which validates the importance of having pedestrian perspectives featured in the dataset and also the necessary auxiliary information to identify them. In addition, it would be more appropriate to use pedestrian-perspective SVIs to study walkability instead of SVIs taken from cars, which tend to dominate the data (especially in commercial SVI) and could be easily misincluded in analysis [55].

Weather. The weather conditions in an image could influence its fitness for use by affecting the image quality, lighting, and human perception. For example,

[7] https://footpath.ai/

a place could look vibrant and lively in sunny weather but depressing and unsafe in rainy weather, and using a mix of these images without prior selection could lead to biases in SVI-based urban perception studies. Further, certain weather conditions such as rain, snow, and fog could potentially cause obstruction and reduce the overall clarity of the image, and having this label could help studies that rely on image clarity (e.g. to extract urban features) avoid SVIs taken in these less desirable weather conditions.

View direction. Among the non-panoramic (or 'perspective') SVIs, which are commonly found in crowdsourced SVI [36], SVIs could be taken with the camera facing the front, or the side, of the road, as one moves along the road, producing front-viewing and side-viewing SVIs, respectively. As the two types of SVIs capture the street scene from two different angles, it is important to know the view direction of a SVI to determine whether it is suitable for a use case. For example, for studies that focus on building facades [94, 95], side-viewing SVIs would likely be more suitable, as they are more closely oriented toward the building facades located on the side of the road and could capture more details on these surfaces. On the other hand, studies on urban morphology [96, 97] might find front-viewing SVIs more fitting, as they are focused along the road and give a more complete view of the street canyon. It was also found that, when calculating metrics such as the sky view index or green view index from perspective images, the results are more accurate when using front-viewing SVIs compared to side-viewing SVIs [36]. It is thus important to be informed of the view direction of a perspective SVI before including it in the research data. In some less frequently observed cases, SVIs could be taken from a rear-facing camera attached to the back of a car or bicycle. These rear-viewing SVIs are considered equivalent to front-viewing SVIs for the purpose of labelling and model training in this work, as the two types are hardly distinguishable at the level of a single still image.

Lighting condition. The lighting condition attribute indicates whether an image is taken during the day, night, or dawn/dusk.

Panoramic status. Panoramas contain more information than perspective SVIs. It could thus be beneficial to know whether an image is a panorama. Such information is not always available in the metadata.

Quality. The quality of an image can be affected by a combination of factors such as obstruction, noises, blurriness, glare, low resolution, etc. [4, 98]. To simplify the labelling task, we conceived this attribute as the subjective judgement of the annotator about the quality of an image—whether it is acceptable (or 'good'), 'slightly poor', or 'very poor'. Such characterisation can help filter out low-quality imagery that is undesirable for use cases, but it may unlock new in-

sights for the assessment of the quality of crowdsourced SVI and understanding VGI contributor activity and patterns in relation to the quality of data, which are perennial topics in VGI [99, 100, 101, 102, 103, 104, 105, 106].

Presence of glare and windshield reflection. The presence of glare and windshield reflection could indicate the quality conditions of SVIs. In addition, knowledge of the presence of glare could be useful for examining travel safety and comfort for drivers, cyclists, and pedestrians.

3.4.2. Training data preparation

These contextual attributes (Section 3.4.1) are challenging to be accurately computed from the image's metadata. We thus consider computer vision to be a viable means to extract such information from SVIs without having to rely on metadata availability and quality. To facilitate model development, we sampled a subset (more than 10,000) of our SVIs to be manually annotated with the eight contextual attributes (Figure 3B). Figure 4 shows an example of the manual labels gathered. Details of the sampling and annotation processes are documented in Appendix A.3.

Subsequently, all labels were processed to remove 'unclear' or 'unobservable' records. For 'platform', small or undefined classes such as 'indoor' and 'others' were excluded from the training data, while some other classes (e.g. 'sidewalk', 'pedestrian zone', 'walking trails') were merged into bigger and more generic classes (e.g. 'walking surface') to simplify the classification task. The detailed mapping of these classes can be found in Appendix A.3. Additionally, SVIs that have been labelled as 'panoramic' were removed from the training data for 'view direction' because 'view direction' is not applicable to panoramas. Table 2 lists the data type, number of classes, and possible values for each attribute.

Among the processed labels, class imbalance was observed in certain attributes, including weather, platform, and lighting condition. This is because some classes (e.g. 'rainy', 'foggy', 'snowy') occur much less frequently than others (e.g. 'clear', 'cloudy'), or in some cases, one class could dominate (e.g. 'day' for lighting condition, 'driving surface' for platform). As a result, the quantity of training samples could be insufficient for the smaller classes.

To supplement the training samples for the smaller classes, we manually browsed through the web applications of both Mapillary and KartaView to specifically seek additional SVIs that fall under one (or sometimes, two) of the smaller classes. As such, these supplementary images could have one or two contextual tags, and would only be used in the training for the one or two contextual attributes they were tagged for and not used in the training for the other contextual attributes

(further detail is provided in Appendix A.3). Using this method, we found additional 26,046, 22,772, and 1,995 images to augment the smaller training classes for weather (i.e. 'snowy', 'rainy', 'foggy'), platform (i.e. 'walking surface', 'cycling surface', 'railway', 'fields', 'tunnel'), and lighting condition (i.e. 'night', 'dusk/dawn'), respectively (Figure 3B). Table 3 shows the exact number of SVIs used in the training and testing for each contextual attribute. Figure 5 shows the class distribution among the manual labels available for each contextual attribute, after class augmentation for platform, weather, and lighting condition.

Figure 5: The class distribution among the manual labels for the eight contextual attributes, with some arbitrary example images from each class. Note that the classes for platform, weather, and lighting condition have been augmented with labels from additional manually selected images. Sources of imagery: Mapillary and KartaView contributors.

3.4.3. Model development

To provide a common baseline for benchmarking, we implement state-of-the-art computer vision models using the labelled data. The labelled images were divided into train and test sets following a city-wise 80-20 split, where all images of a city are either on the train set or test set but not both. For the classification problem, we produce the baseline results by weighting the loss weights uniformly to account for class imbalances in the training set, and use a stratified hold-out validation set to qualify our model decisions. The primary metric of interest is the performance on the test set, measured by the accuracy and the macro-averaged precision, recall, and F1 score (Table 3), since this presents generalisation performance. The test metrics are computed upon model convergence, using the test data for inference. Each of the 'Training' attributes in Table 3 is treated as the dependent variable in a classification task, resulting in eight classification models. We chose the MaxViT: Multi-Axis Vision Transformer [107] as it achieves competitive performance on image classification, leveraging both local and global spatial information in the model. For all purposes of training the classification model, we use the MaxViT-Tiny model of 33M parameters. We utilise the AdamW [108] optimiser with learning rate of $3 \cdot 10^{-4}$, weight decay of $1 \cdot 10^{-2}$ and $\beta_{1,2} = (0.90, 0.99)$. Early stopping is applied if the validation loss does not decrease in two epochs. Cross-entropy loss is used with uniform loss weighting to optimise the model. Details about the performance of the models are shown in Table 3.

In general, the models achieved moderate to high accuracy for all eight attributes. The models performed the best for panoramic status and lighting condition, producing high accuracy, precision, recall, and F1 scores (Table 3). For some attributes, such as glare, quality, weather, and platform, greater differences between the accuracy and macro-averaged F1 scores were observed, largely due to the classes being imbalanced and difficulty to accurately predict for small classes (Figure 5). Possible reasons for the model results are discussed in detail in Section 5.5.

To the best of our knowledge, there are no existing benchmarks for the classification of panoramic status, view direction, reflection, and platform. While several works have produced various baselines and datasets for classification of weather [59, 109, 110], lighting conditions [109, 110], and glare [109], achieving accuracy levels comparable to or even higher than our baselines, the data involved in their experiment is considerably different from ours, e.g. [109] used web images which are usually more typical and easily characterised compared to crowdsourced SVIs that come from diverse settings, and the data from [59] and [110] was collected

Table 3: Overview of the computer vision models used for labelling contextual, semantic, and perceptual attributes. For the contextual attributes (platform, weather, view direction, lighting condition, panoramic status, quality, glare, and reflection). The models were trained with our manually labelled data; the number of images involved in training and testing, as well as the accuracy, macro-averaged precision, recall, and F1 score on the test set are reported below.

Attribute	Model	Number of images		Accuracy	Precision	Recall	F1 Score	Training / Inference
		train	test					
Panoramic status	MaxViT	8,372	2,172	0.999	0.995	0.995	0.995	Training
Lighting condition	MaxViT	9,380	3,079	0.962	0.916	0.897	0.905	Training
Glare	MaxViT	8,089	2,115	0.941	0.602	0.698	0.631	Training
View direction	MaxViT	7,632	2,012	0.874	0.735	0.912	0.780	Training
Quality	MaxViT	8,199	2,107	0.799	0.398	0.515	0.410	Training
Reflection	MaxViT	8,112	2,119	0.787	0.745	0.788	0.757	Training
Weather	MaxViT	27,771	8,068	0.755	0.664	0.608	0.599	Training
Platform	MaxViT	25,407	7,311	0.683	0.574	0.582	0.567	Training
Instance detection and semantic segmentation	Mask2Former	-	-	-	-	-	-	Inference
Scene type	VGG16	-	-	-	-	-	-	Inference
Human perception	ViT	-	-	-	-	-	-	Inference

from only one country, which may not capture features that could vary across different geographical regions. For quality classification, the work of [111] presented a deep learning approach to classify images into six quality categories. However, some of the images in their training data were artificially blurred, which may not well represent the cases seen in crowdsourced SVIs. Therefore, our work could be considered a representative baseline for the tasks of classifying the eight contextual attributes in a geographically and contextually heterogeneous SVI dataset.

After all training was complete, the resulting best-performing CV models were used to automatically inference the eight contextual attributes for the remaining dataset (Figure 3B).

3.5. Semantic and perceptual enrichment by computer vision

We processed our SVIs with several existing, state-of-the-art CV algorithms to compute semantic segmentation, instance detection, human perception, and scene classification. Figure 4 shows an example of the inferred labels available. SVI-based urban studies heavily rely on semantic information extracted from SVI but the computation process is often resource intensive, or could impose a technical barrier for researchers who are not well-versed in this aspect. By pre-computing these important attributes and making them readily available, we substantially lower the technical and resource threshold to use SVI in urban research, thereby promoting the use of SVI.

Instance detection and semantic segmentation. We implement a unified image segmentation and detection pipeline, simultaneously producing object counts and pixel counts. More specifically, we utilise the 'Mask2Former' approach [112],

a universal and lightweight transformer architecture applicable to various image segmentation tasks. Mask2Former is trained and validated on the Mapillary Vistas version 1.2 dataset [57], which consists of 65 semantic classes, and reports a mean intersection over union (mIoU) performance of 60.8%.

Adopting Mask2Former offers two key advantages for applying a standardised pipeline to diverse urban contexts: Firstly, it significantly enhances accuracy in identifying fine-grained semantic categories within images. Unlike other models that often overlook small regions in images, Mask2Former excels in this regard. Secondly, it is lightweight and computationally scalable, making it an efficient solution. Readers interested in the specific architecture and training of Mask2Former are referred to [113, 112].

Scene classification. To infer the various scenes featured in the SVIs, we ran the released VGG16 model [114] trained on the Places dataset [61] for all SVIs in our dataset. The Places dataset has more than 10 million images labelled with more than 400 unique scene categories and VGG16 has attained the highest top-1 accuracy on both the validation and test sets. The diverse scene categories contain many that are relevant to SVI, such as street, highway, residential neighbourhood, park, etc. The scene label thus provides an overall semantic context of the SVIs and could help researchers remove outliers that are not of their research interest [6].

Human perception. We utilised six pre-trained perceptual models by [115], respectively predicting, for each SVI, the six dimensions of human perception for the urban built environment, which include 'safe', 'lively', 'wealthy', 'beautiful', 'boring', and 'depressing'. Each model outputs for each SVI a numerical score ranging from 0 to 10 that reflects the magnitude of the perception. The models were pre-trained on MIT Place Pulse 2.0 [62, 63]. The backbone of the model implements a vision transformer pre-trained on ImageNet [116], which is known for its high utility as a dataset for transfer-learning in a broad range of vision tasks. The models achieve an accuracy of 76.7% for 'safe', 77.1% for 'lively', 72.9% for 'wealthy', 76.9% for 'beautiful', 61.6% for 'boring', and 67.2% for 'depressing'.

A summary of the models used for inference is shown in Table 3.

4. Global Streetscapes

The Global Streetscapes dataset, which we release openly at `https://github.com/ualsg/global-streetscapes`, consists of:

- The metadata and all enriched geospatial, temporal, contextual, semantic, and perceptual attributes (described in Section 3) for the entire dataset.

- The manual labels used in model training and testing for the eight contextual attributes outlined in Section 3.4.
- The manually labelled images used for model development (the exact number of images used for each attribute is outlined in Table 3).

The contents of each data file are explained in Appendix A.1. To ensure the reproducibility and continuation of the project, Python scripts to download and enrich the images can be found in our GitHub repository, alongside various Jupyter Notebooks with step-by-step instructions for basic queries and visualisations of the dataset. The images, except for those used for model development, are not hosted, as they can be downloaded from their original sources (Mapillary or Kartaview), using the code in our GitHub repository.

Global Streetscapes contains images sampled from over 688 cities around the world, and their geographical distribution is shown in Figure 1. Among the six continents it covers, Europe has the most number of images, accounting for nearly 40% of the dataset, followed by Asia (23.3%), North America (20.17%), South America (9.83%), Africa (5.49%), and Oceania (1.98%) (Figure 6A). Despite our effort to include all countries and ensure geographical balance, Europe still appears overrepresented (by having the highest number of images) in the dataset. This could be attributed to the high number of countries and cities in Europe, as well as greater data availability there.

The dataset also covers regions with varying degrees of urbanisation (Figure 6B). As SVIs were sampled around the city areas, where more SVIs are available, the urban centre group (as classified on GHSL) understandably has the highest share of images (88.61%). This group is followed by suburban or peri-urban areas (5.02%), dense urban clusters (3.64%), low-density rural areas (1.39%), very low-density rural areas (0.65%), etc. At the same time, due to the vast number of images, even categories with such a low percentage have a large number of images that are sufficient for a variety of use cases.

At the street level, the dataset features a diverse range of locations and urban environment settings as indicated by the OSM road type (Figure 6C), including residential roads (20.07%), secondary roads (16.99%), primary roads (14.68%), footways (14.58%), tertiary roads (13.7%), trunk roads (5.47%), service roads (4.95%), motorways (4.91%), and cycleways (2.09%), etc.

The dataset mostly consists of perspective images (71.07%), followed by fisheye (20.34%), equirectangular (4.77%), and spherical (3.75%) images (Figure 6D). This image type breakdown shows that perspective images dominate crowdsourced SVI, and it is thus essential to develop tools that could better facilitate

the use of such data, e.g. by providing information on the view direction of the image [36].

The images are quite evenly distributed across seasons. Apart from 21.94% of images from the tropical regions, among the remaining images, most (21.42%) of them were taken in autumn, followed by summer (20.27%), winter (18.76%), and spring (17.61%) (Figure 6E). While images are mostly taken during the day hours, a considerable amount is taken at nighttime (Figure 6F).

The average speeds exhibit two peaks at 5 km/h and 15 km/h, with a long tail toward higher speeds (Figure 6G). Figure 6H shows the density distribution of our predicted scores for the six dimensions of human perception ('safe', 'beautiful', 'lively', 'wealthy', 'depressing', and 'boring').

To highlight the temporal diversity inherent in the dataset, which may facilitate longitudinal analyses such as [35, 26], nine cities were selected to visualise their SVI availability in each year from 2014 to 2022 (Figure 7). Notably, certain areas exhibit distinct patterns: some possess data for just a single year (e.g. Ait Melloul), while others span all nine years albeit with a relatively limited number of SVIs (as seen in Warsaw). Conversely, there are instances of substantial SVI data a few years ago but a lack of coverage in recent years (like Ottawa). Such temporal variability much depends on VGI contributor activity and patterns.

Our dataset also exhibits a diverse range of semantic and perceptual characteristics. Figure 8 shows the range and distribution of green view index, sky view index, and perception scores for all six dimensions (Section 3.5), across the entire dataset, with example images showing different visual traits found at eight different statistical values (minimum, maximum, mean, and the 10th, 25th, 50th, 75th, and 90th percentiles).

5. Discussion

5.1. Potential benefits to existing urban research topics

Active mobility. Urban active mobility research can benefit from this dataset and address several previously known hindrances when working with SVI. One such potential improvement is enhanced accuracy of perspectives when analysing walkability and bikeability. Studies have utilised SVI (i.e. GSV in most studies) to examine the relationship between visual features of streets with walkability and bikeability, and some of them reported potential biases in the results due to the discrepancies between the perspectives of vehicles in GSV and that of pedestrians and cyclists [10, 117]. The labels of platforms can enable researchers to build an image classification model, with which they can potentially obtain only suitable

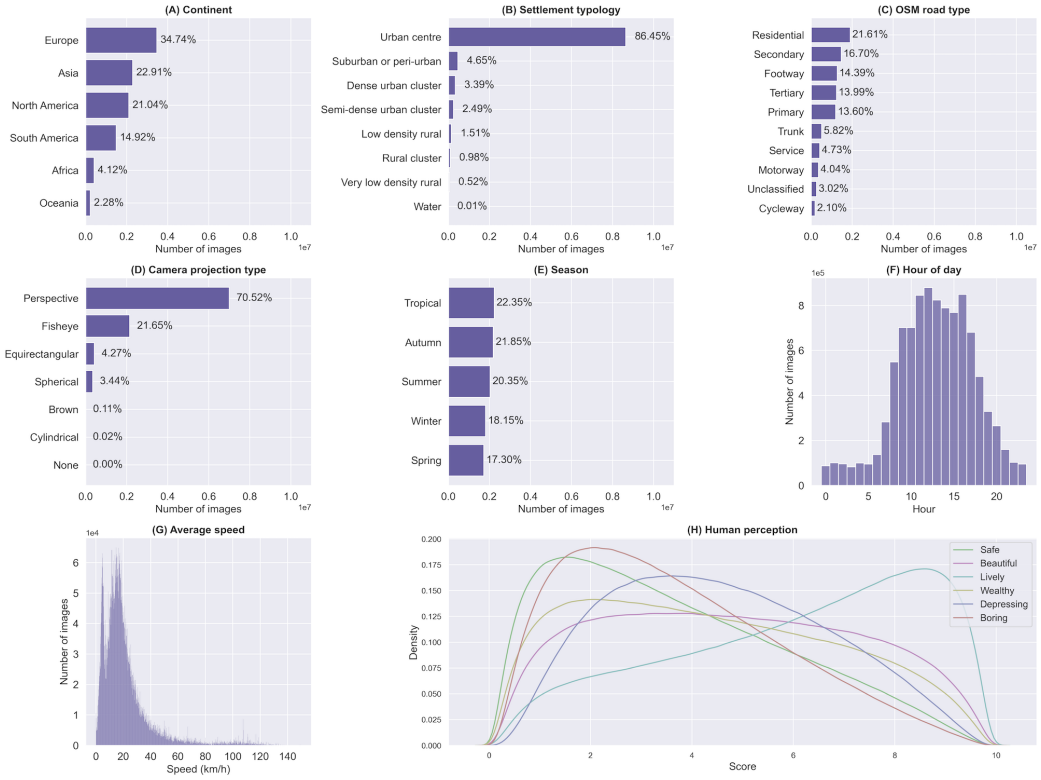

Figure 6: Class or value distribution among the 10 million images for (A) continents covered, (B) settlement typology (degree of urbanisation), (C) OSM road type, (D) camera projection type, (E) season, (F) hour of the day, (G) average speed, and (H) perception scores.

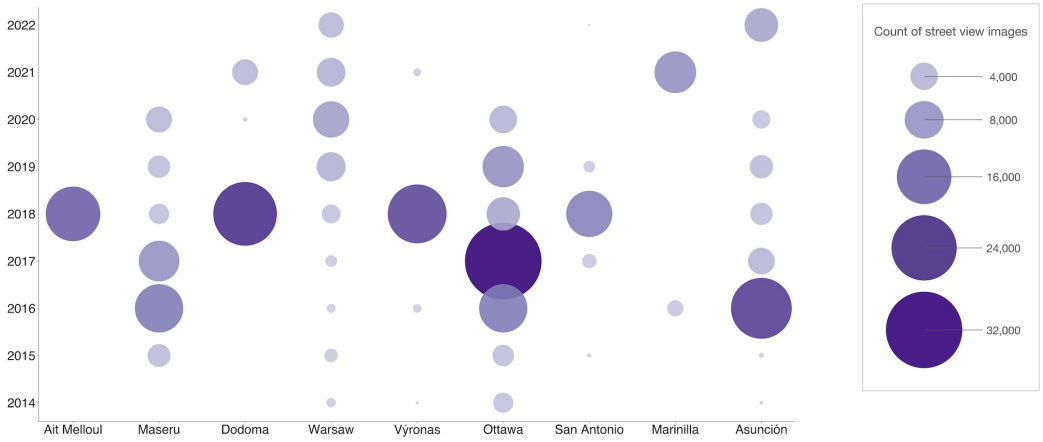

Figure 7: The count of street view images in each year from 2014 to 2022 for nine arbitrary cities, indicating temporal completeness and variation in contributor activity in underlying sources (Mapillary and KartaView).

SVI from crowdsourced services, leading to a more reliable assessment of walkability and bikeability by accurately reflecting target street users' perspectives. Another possible enhancement is a more granular assessment of streets under different conditions. Similar to the issue above, previous studies have not been able to consider various weather and lighting conditions when assessing streets for urban mobility due to the unavailability of such diverse SVI in major proprietary SVI services and the absence of a means to filter SVI from crowdsourced services. Thus, the new dataset can open up research opportunities to analyse streets under specific conditions by classifying SVI into different weather or temporal conditions. For example, Figure 9 shows pairs of SVIs taken from the same location but have contrasting visual characteristics, which could potentially affect downstream analyses if they are directly used without processing. Figure 10 shows an example of selecting images with specific conditions from the dataset.

Perception. SVI remotely senses the urban environment from a uniquely human perspective, complementing the bird-eye view provided by satellite imagery. Existing literature suggests human subjective perception of urban spaces is correlated with built environment features, and SVI provides a convenient means to accurately assess both, substantially enhancing the scalability of studies on quality of urban life [118, 22]. For instance, the 'safe' score predicted from SVIs can be used to quantify the perceived safety in a neighbourhood, which can be further compared with real crime data to reveal the relationship between perceived and

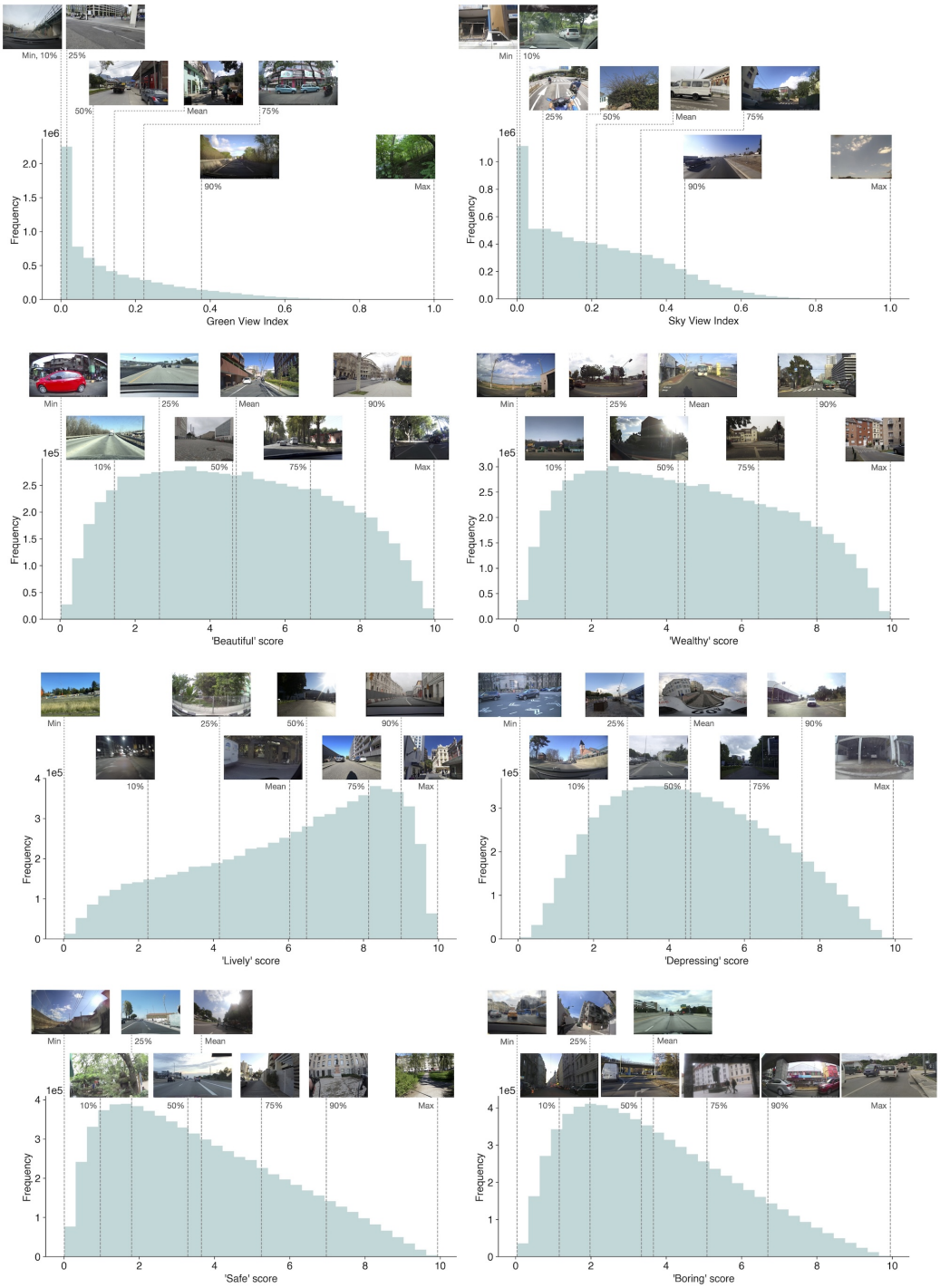

Figure 8: The histograms of green view index values, sky view index values, and perception scores (‘beautiful’, ‘wealthy’, ‘lively’, ‘depressing’, ‘safe’, and ‘boring’) across the entire dataset, with example images at various statistical values (minimum, mean, maximum, and the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles).

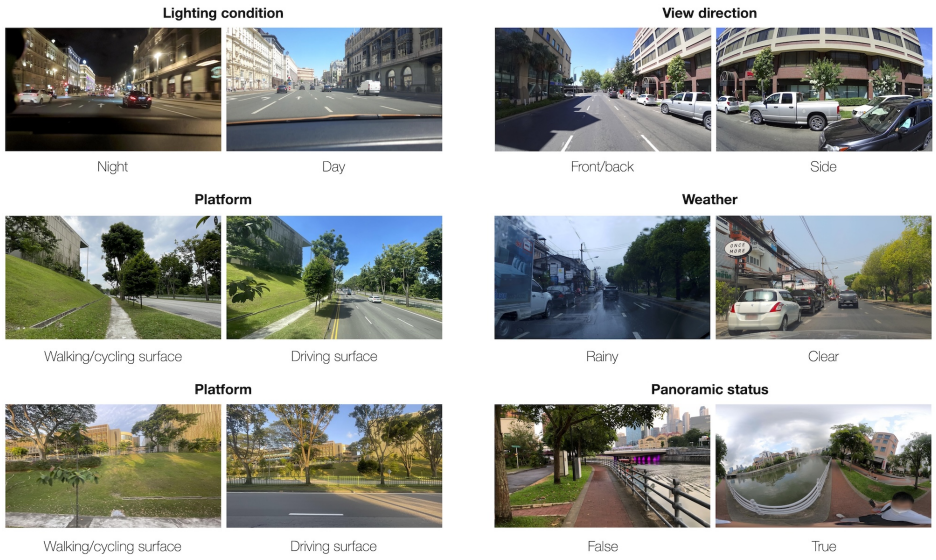

Figure 9: Sets of images taken at the same location but with contrasting visual characteristics, affirming the need for a contextually rich dataset, which will contribute to increasing the fit for purpose and usability of street-level imagery.

actual safety and provide insights into optimising urban management strategies [24]. Moreover, the GeoAI-generated perception scores can be compared with onsite surveys by participants to evaluate the variations in perception bias in different regions worldwide and explore the driving force behind them. Further, Global Streetscapes enhances SVIs by integrating both physical environment features and subjective human perceptions of cities worldwide. This dataset thus offers considerable potential in assessing, at a global scale, various aspects of human-centric urban development, such as social, economic, and cultural developments, housing prices, urban vitality, public mental health, urban crime, etc. These features and use cases could facilitate the development of human-centred GeoAI applications and contribute valuable insights into the planning for liveable cities.

Urban complexity. SVI provides a powerful lens to sense the complexity and dynamics of urban environments at an unprecented scale. Yet, existing SVI datasets often lack crucial semantic and categorical information, such as weather conditions or seasonality, pertaining to the image context. Consequently, most SVI studies tend to treat images statically, neglecting their spatio-temporal context. Global Streetscapes bridges an essential gap by providing comprehensive support for various urban sensing applications. Additionally, it may facilitate the

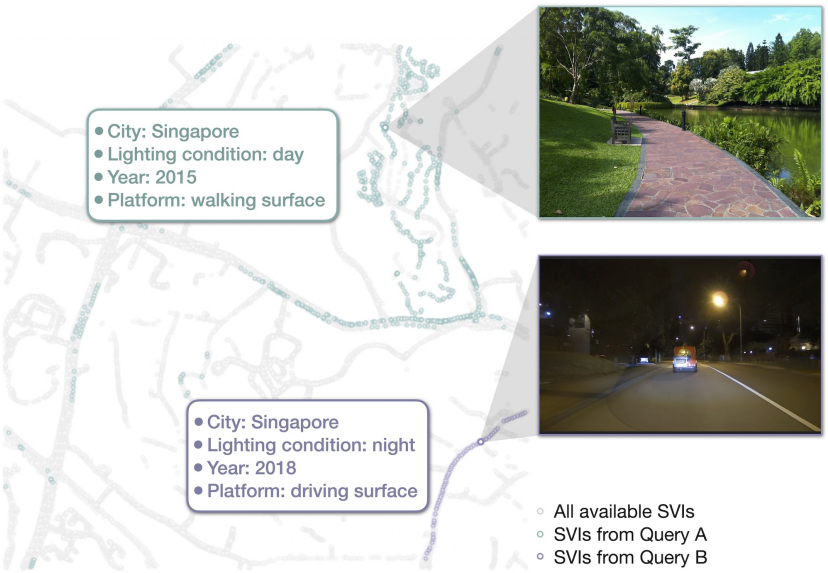

Figure 10: Two example queries from the dataset to select images with desired characteristics. A semantically enriched dataset such as ours may facilitate identifying street-level imagery that is suitable for a particular use case, and might contribute to the development of novel computer vision models and benchmarks.

development of dynamic open-source tools in the field [45].

5.2. Potential research directions

Global-scale applications and comparative studies.

With the expansive coverage and extensive pre-computed attributes, Global Streetscapes could greatly lower the barrier to conducting multi-city studies. Such studies could yield insights into how cities differ or are similar in their morphology, appearance, or how they are perceived. For example, Figure 11 shows how the green view index and sky view index, which was calculated from our pre-computed segmentation values (Section 3.5), aggregated at level-10 H3 grid, vary both within and across cities. Further questions can include: What made cities look different from one another? How are the cities perceived by their residents? How do the residents perceive other cities? As a large-scale, pre-processed dataset, data from Global Streetscapes can also be readily integrated with other analytics frameworks, such as the work by [119] to quantify urban greenness using OSM and SVI data, across multiple scales of analysis.

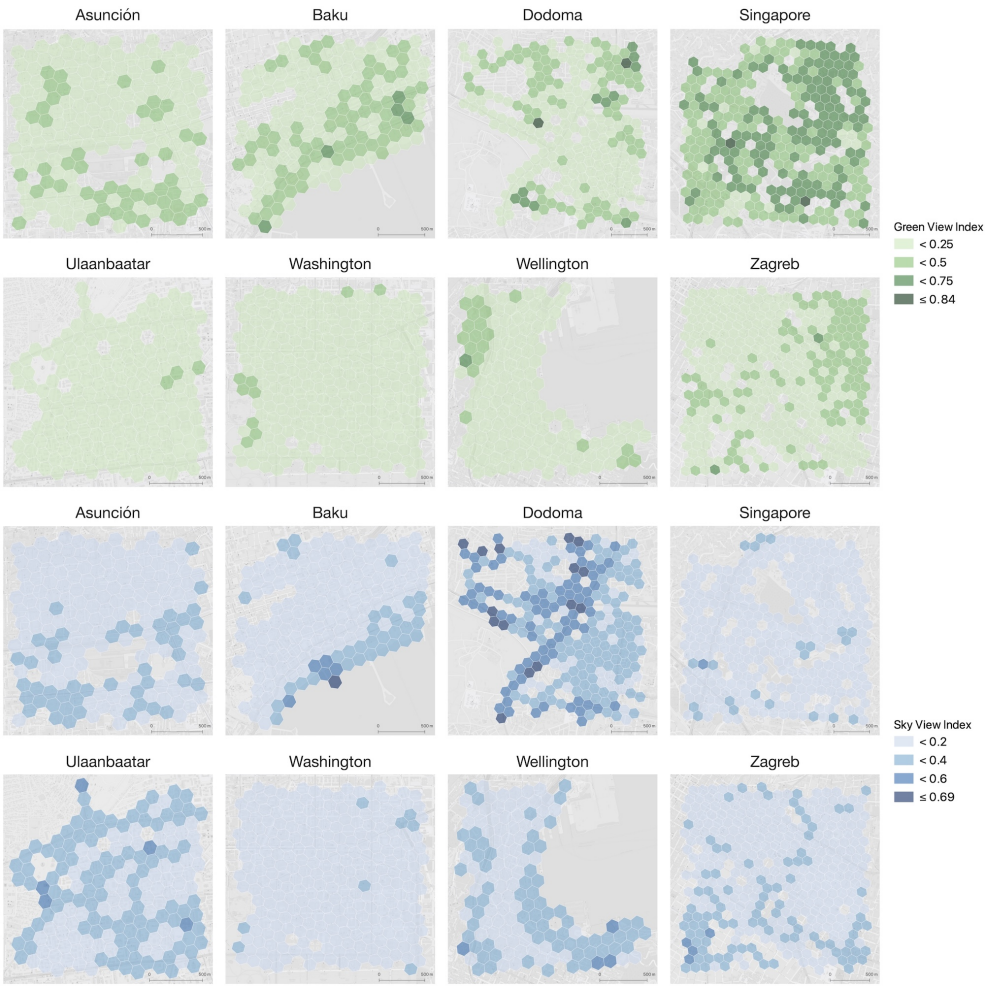

Figure 11: Spatial distribution of green view index and sky view index, aggregated at level-10 H3 grid, across eight cities from different continents. These pre-computed metrics lower the barriers and increase efficiency.

Longitudinal studies As Global Streetscapes also covers a long temporal period, it could potentially support change detections for locations around the world. Long-term changes in urban physical appearance could be detected from changes in widely implemented metrics such as the green, sky, building, and road view indexes [36, 120, 121, 122]. Figure 12 shows an example of tracking changes in the green view index and sky view index for six different cities.

Moreover, numerous contextual parameters provided in Global Streetscapes, such as view direction, platform, and quality, could thus be helpful to standardise conditions and ensure consistency of quality of imagery for a longitudinal analysis [123], where quality and visual conditions inevitably vary, especially for crowdsourced SVI. These conditions could be analysed in conjunction with changes detected from OSM data, another global crowdsourced geospatial dataset, which has wide spatial coverage and rich spatial information but lack visual data, and Global Streetscapes could complement the visual analysis aspect to drive new insights.

Additionally, temporal changes in dynamic aspects of the city (e.g. traffic flow, pedestrian flow) could be investigated as well for cities around the world to study how urban life has evolved. These analyses could be strengthened by incorporating socioeconomic, demographic, and Point-of-interest (POI) data. Optical Character Recognition (OCR) could be run with each image to extract texts on signboards to reflect changes in business activities. The vast spatiotemporal dimensions could give us a fuller picture of how cities have changed over time or when compared to others across the world, identifying urban growth and/or decay.

Data quality. Another promising potential research direction to examine is data quality and contributor analysis of such volunteered SVI data. Some examples of possible dimensions include image quality, spatial coverage, availability of panoramas, update frequency, etc. For example, Figure 13 shows the share of panoramas in 35 cities with more than 10,000 SVIs and more than 5% share of panoramas, which exhibits a substantial variation among cities. As Global Streetscapes contains data for most of the capitals in the world, specific data characteristics could be examined alongside socioeconomic factors such as the Gross Domestic Product (GDP) to assess data equity.

Data generation. Apart from conducting analyses based on existing data, the dataset could also be used to generate new data. As images in Global Streetscapes are spatially indexed at various resolution levels, it also makes it easy to integrate the images with multiple types of external data, such as the Kontur global popula-

Figure 12: Change in green view index and sky view index, aggregated at level-10 H3 grid, across six cities.

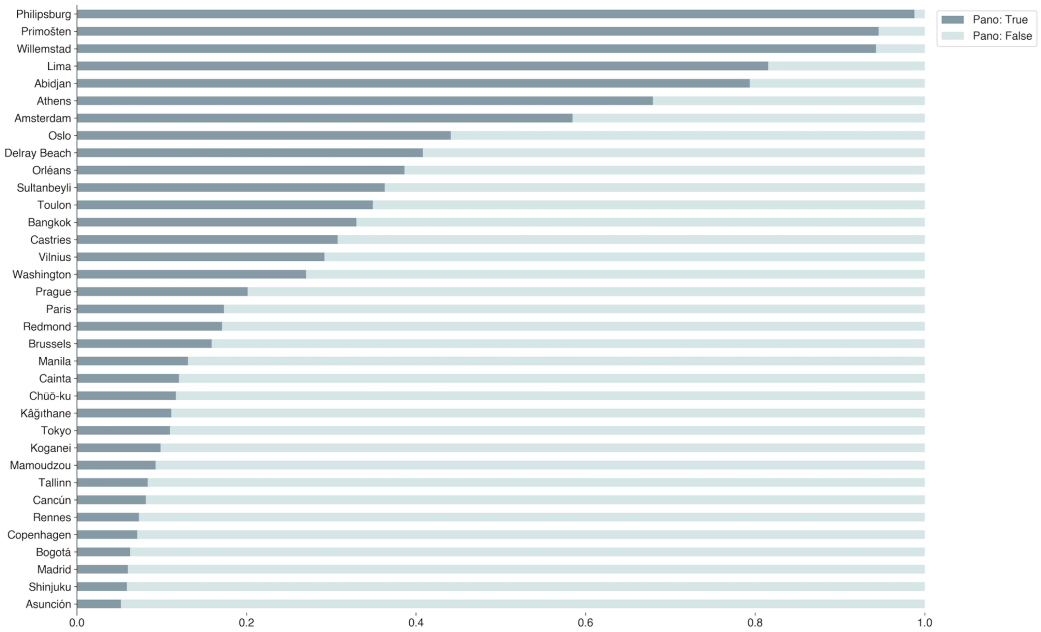

Figure 13: Share of panorama images in 35 cities with more than 10,000 street view images in the dataset.

tion dataset[8]. By associating SVI with population data, machine learning methods could be applied to model the relationship between these two types of data, and it could be potentially possible to predict population density at fine scales by using just SVIs. Population density estimates could be helpful for areas where fine-scale population data is not available. In addition, utilising the technique by [16] to reconstruct 3D building models from single street view images, Global Streetscapes could potentially supply suitable images to construct 3D building models for locations covered in the dataset.

Computer vision. This dataset has captured and labelled diverse scenes and environmental settings for places around the world across multiple years. The same place could have been captured multiple times under different ambient conditions across the years. This dataset could thus potentially be further developed to supply essential data for studying place recognition and geolocalisation. Advancements in such technologies could help build better analytics tools and in turn benefit urban research and analytics, in topics such as tracking changes in the built

[8]https://www.kontur.io/portfolio/population-dataset/

environment and understanding how people interact with different urban spaces. If further supplemented with manual labels for semantic segmentation and object detection, this dataset could potentially help advance existing algorithms by allowing researchers to study the domain discrepancy that arises from different visual conditions, e.g. semantic segmentation in daytime versus nighttime images, as experimented in BDD100K [59].

5.3. Equity

Street view images have propelled urban science and analytics, but many regions such as African countries are largely excluded from the developments relying on street view imagery, with only 20% having partial coverage in mostly large cities [124]. Owing to the presence of Mapillary and KartaView volunteers and their worldwide contributions, our dataset gathers SVIs crowdsourced from not only major and frequently studied cities such as Tokyo, Jakarta, New York City, and Beijing, but also smaller ones and those given less attention such as Kasempa, Caldera, Lobamba, and even small island cities such as Tarawa—some of which are not yet covered by commercial SVI sources such as Google Street View. This dataset's global coverage and ready-to-use methodology thus aim to promote diverse and equitable research. By downloading, managing, cleaning, and processing the data, our effort lowers the entry barriers to researchers and saves efforts, both in time and computational resources, especially for analyses that include multiple cities and vast amounts of data. Further, as it presents an off-the-shelf solution, researchers without experience in SVI can, for the first time, take advantage of such data. Finally, our dataset contains imagery from more than one source (which is uncommon, see Table 1), further facilitating downstream analyses. It increases equity and participation and may bring this novel and emerging urban data source closer to all researchers regardless of their level of expertise with computational methods.

5.4. Dataset longevity and crowd intelligence

To facilitate the continuous growth of the dataset, the entire workflow, from data download to enrichment, has been automated and is reproducible with scripts from our GitHub repository. These scripts can be run on a schedule to continuously refresh data once in a while as new images are continuously being collected and uploaded in Mapillary and KartaView. In case of adding new cities to the dataset, the scripts can be run with the new cities to update the dataset. In addition, when more advanced models for predicting the eight contextual labels are available, we will update labels in the dataset using those models. All version

updates will be documented in the same GitHub repository stated above, and new versions of the dataset will be maintained in the data repository.

Further, dataset users can become dataset creators as well. Interested users can follow our methodology and run our open-access scripts to create and maintain datasets for their own cities. For example, users can collect SVIs for their city and contribute them to Mapillary or KartaView, download the data (which could include contributions from others) with our download scripts, and enrich it with geospatial, temporal, contextual, semantic, and perceptual information by running the data with our enrichment scripts. The processed and enriched SVIs can in turn be used to analyse their own city.

In addition, channels and platforms can be set up for interested volunteers to continuously contribute to Global Streetscapes by adding new labels, verifying existing labels, proposing new attributes to be labelled, or suggesting new open-source data to be integrated. The set of images we have currently manually labelled is only a small fraction of the entire dataset. The model and inference accuracy could thus benefit from having more manual labels. The label accuracy could also be improved by having the same image labelled by more people. This could be potentially implemented in the form of a survey linked in the data and GitHub repositories, where dataset users can volunteer to contribute back to the project by participating in the survey in which they would either add some new labels or verify some existing labels. More accurate labels could also help improve our CV algorithms which would in turn allow us to generate a more accurate and updated dataset for the users. For this reason, interested and savvy users are also encouraged to use the manual labels provided to further advance the related CV models.

By leveraging on crowd intelligence from a large community, we can continuously improve the accuracy and completeness of the dataset, which was itself built from crowdsourced data. With Global Streetscapes, we hope to promote awareness and interest in data sources that comply with open data [125] and the Findability, Accessibility, Interoperability and Reusability (FAIR) principles for scientific data management [126].

5.5. Limitations and future work

While we tried to make geographical coverage as balanced as possible, it is inevitably and ultimately based on the data sources—Mapillary and KartaView. Data availability is subject to user contribution patterns [127, 128, 129], which could depend on various factors, e.g. socioeconomic backgrounds, access to the Internet infrastructure, data policies in different administrations, corporate actors,

etc. Data equality is also an essential topic of discussion, especially in the realm of Volunteered Geographic Information (VGI) [92, 130, 131], and is worth further investigation in the future, as described in Section 5.2. In future dataset updates, settlements with smaller populations (< 50,000) should be systematically included to check for data availability as well for a more inclusive and diverse representation of the world's cities (and settlements) [132].

Figure 14: Example images taken from different walking or driving surfaces to illustrate the difficulties present in platform classification: (A) motorised and non-motorised roads may look similar to each other, and sometimes there is no clear separation between a sidewalk and a road; (B) the design and pavements of sidewalks often look very different even within the same country (the three images are all from Japan); (C) such variation in appearance can become more prominent across different sub-types of platforms (e.g. sidewalk v.s. footpath in a park) and across regions, hindering the development of classification approaches at the global scale.

In addition, to standardise data sampling and simplify processes, we only collected images around the city's coordinates as given by the SimpleMaps World Cities Database [77], in a 2.4 × 2.4 km tile (measured at the Equator), which could be insufficient to represent variations across the entire city and could hinder some applications. In future work, tiles at both the urban core and urban peripherals (at various radii from the centre) could be sampled for image collection so as to give a more representative sample of each city. Further, the city coordinates

provided in the World Cities database could deviate from the real-life downtown areas for some cities, by various extents from hundreds of metres to several kilometres, hence do not necessarily represent the city centre, which on its own is a location that is often subjective.

Another limitation is the class imbalance among the manually labelled data, which is inevitable and part of real-world data. While we augmented the small classes for certain attributes with manually selected images, as described in Section 3.4.2, many of the additional images were from the same sequences. The similarities among images that come from the same sequence could potentially introduce biases to the models. However, as we used a city-wise split for the train-test split, sequences used in training were unlikely to be used again in testing. Hence, any possible biases introduced by the supplementary images should not have overstated the generalisability of the models. In future works, synthetic data, which could be generated with diverse and varying characteristics [76, 133], could potentially be used to supplement the training data instead.

Some ambiguous situations might give rise to ambiguous labels as well. For example, it is difficult to determine how much cloud cover is appropriate for the weather to be considered 'cloudy', which can be difficult even for human judgement. The model could thus potentially confuse between the two classes ('cloudy' and 'clear'). For platform classification, difficulties could arise from the similarities in appearance between different types (Figure 14A) or sub-types (Figure 14C) of platform, as well as the variations in appearance between the same type of platform within (Figure 14B) and across different geographies (Figure 14C).

Notably, our model for quality classification appears to be rather stringent, misclassifying more good-quality images than poor-quality images. This implies that using this model as a filter for image quality could likely result in a pool consisting of mostly good-quality images instead of a pool that is mixed with many poor-quality images.

Some images that are not considered SVI (e.g. aerial photos, imagery from boats) could have been mixed in the sources [4, 134]. However, due to limited computing resources, we were unable to implement checks for all of the 10 million images, for both duplicates and validity. Nevertheless, users who intend to use a smaller subset of the images are encouraged to perform these checks. For image validity checking, users could refer to methods such as computing visual complexity, as implemented in some previous works [135, 45].

Though we have done extensive work to enhance the usability of SVI, especially crowdsrouced SVI, users are advised to exert caution when applying the dataset or the associated CV models. This is because, while exhibiting a range

of advantages, the heterogeneity inherent in the characteristics, quality, and coverage of crowdsourced SVI (which stems from unstandardised and unorganised data collection) could pose challenges to the generalisability and transferability of the models.

6. Conclusion

Benefiting from contributions of myriads of volunteers from all over the world and developing a reproducible framework, we constructed a large open, labelled, processed, and worldwide street-level imagery dataset—Global Streetscapes, the first of its kind.

It consists of more than 10 million images from over 688 cities around the world and is enriched with a comprehensive range of spatial, temporal, semantic, perceptual, and contextual attributes that we believe will be relevant for a variety of downstream analyses and computer vision modelling benchmarking efforts. Because it is derived from crowdsourcing (VGI) services (i.e. Mapillary and KartaView), the dataset has high spatial, temporal, environmental, and viewpoint diversity. These variations are also extensively described in its rich auxiliary information, benefiting an array of use cases that require both a variety of disparate settings and specific scenarios for particular studies, and the production of derivative open datasets [136].

This contribution tackles the main bottleneck of usability of SVI (especially crowdsourced instances), which exhibits heterogeneous conditions and characteristics, hence allowing the development of new global-scale applications. As such, our work provides support for many research lines that could be difficult or erroneous to implement otherwise, including various topics of longitudinal studies and global-scale analysis. To drive the continuous development of this work and beyond, we also provided, for the first time, comprehensive ground-truth contextual labels, and trained state-of-the-art computer vision models for benchmarking. We look forward to following the uses of the dataset and welcome collaborations.

Code availability

The custom code used for the creation, pre-processing, enrichment, and model development of Global Streetscapes is hosted on a public GitHub repository `https://github.com/ualsg/global-streetscapes`. This codebase includes scripts to reproduce the workflow in Figure 3, multiple Python Jupyter notebooks with

data analysis, and thorough documentation in the repository's wiki. The v1.0 release of the Global Streetscapes dataset can be accessed through the above-stated GitHub repository as well.

CRediT author statement

YH: Conceptualisation, Methodology, Software, Formal Analysis, Investigation, Data Curation, Visualisation, Writing - Original Draft; **MQ**: Methodology, Software, Data Curation, Visualisation, Writing - Reviewing & Editing, Project administration; **MK**: Software, Visualisation, Writing - Reviewing & Editing; **WY**: Conceptualisation, Software, Writing - Reviewing & Editing; **JO**: Software, Writing - Reviewing & Editing; **KI**: Conceptualisation, Writing - Reviewing & Editing; **ZW**: Investigation, Writing - Reviewing & Editing; **TZ**: Writing - Reviewing & Editing; **FB**: Conceptualisation, Methodology, Resources, Visualisation, Writing - Reviewing & Editing, Supervision, Funding acquisition.

Funding

This research is part of the projects (i) Large-scale 3D Geospatial Data for Urban Analytics, which is supported by the National University of Singapore under the Start Up Grant R-295-000-171-133 and (ii) Multi-scale Digital Twins for the Urban Environment: From Heartbeats to Cities, which is supported by the Singapore Ministry of Education Academic Research Fund Tier 1. The research was partially conducted at the Future Cities Lab Global at the Singapore-ETH Centre, which was established collaboratively between ETH Zürich and the National Research Foundation Singapore (NRF) under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

Acknowledgements

We thank the colleagues at the Urban Analytics Lab at the National University of Singapore (NUS) for the discussions. We thank the students who annotated the training dataset and Felix Hammer who contributed to the initial stage of the model development. Moreover, we gratefully acknowledge the use of Mapillary, KartaView, OpenStreetMap, SimpleMaps World Cities Database, Global Administrative Area (GADM) database, Global Human Settlement Layer (GHSL) database, Uber H3, and the Climate Zone API[9] which is based on a database by the

[9]https://github.com/sco-tt/Climate-Zone-API

Institute for Veterinary Public Health and the Provincial Government of Carinthia in Austria, and thank the contributors and companies for their continuous efforts to collect data and keep it open and maintained, which enables ours and other research in the community.

Appendix A. Supplementary information

Appendix A.1. Dataset file breakdown

Each SVI record can be identified based on their `uuid, source, orig_id` value, representing the Universally Unique Identifier we assigned to it, the source from which it was obtained (Mapillary or KartaView), and its original identifier given by its source, respectively. These three columns are present in all files of the dataset. The files below, which we host on a public repository accessible through `https://github.com/ualsg/global-streetscapes`, are produced by the workflow depicted in Figure 3A (metadata) and B (geospatial, temporal, contextual, and semantic and perception attributes). Within the dataset, we also include a summary of the contents of each file, in `info.csv`.

City information. The city information associated with each SVI, obtained from the SimpleMaps World Cities Database [77], as well as which continent it belongs to, can be found in `simplemaps.csv`.

Metadata. The metadata obtained from the two image sources is broken down into multiple files named `metadata_kv.csv` for Kartaview data, and `metadata_mly1.csv`, `metadata_mly2.csv`, `metadata_mly3.csv`, and `metadata_mly4.csv` for Mapillary data. The file `metadata_common_attributes.csv` contains attributes found in both Kartaview and Mapillary that we have standardised, including the image's location coordinates, capture date and time (in local timezone), width and height, heading, projection type, horizontal and vertical field of view, identifier of the sequence it belongs to, its order index in the sequence, and the number of available images in its sequence.

Geospatial attributes. Attributes obtained from joining the dataset with data from OSM, GHSL, GADM, and H3 are stored in the files named `osm.csv`, `ghsl.csv`, `gadm.csv`, and `h3.csv`, respectively. The Köppen climate classification associated with each image's location can be found in `climate.csv`.

Temporal attributes. Attributes pertaining to season, the number of hours relative to sunrise / sunset, and the average and maximum speed are stored in the files named `season.csv`, `ephem.csv`, and `speed.csv`, respectively.

Contextual attributes. All contextual attributes of platform, weather, view direction, lighting condition, panorama status, quality, presence of glare, and presence of windshield reflected can be found in `contextual.csv`.

Semantic attributes. The pixel and instance counts from semantic segmentation and instance detection are stored in `segmentation.csv` and `instances.csv`, respectively.

Perception attributes. Attributes related to human perception are stored in `perception.csv`.

Scene classification. The scene recognised in each SVI is stored in `places365.csv`.

Train and test split. The manual labels used for training and testing the contextual models described in Section 3.4.3 can be found in `glare.csv`, `lightning_condition.csv`, `pano_status.csv`, `platform.csv`, `quality.csv`, `reflection.csv`, `view_direction.csv`, and `weather.csv`. The files are located in the folders `train/` and `test/` accordingly.

Appendix A.2. Geospatial and temporal enrichment

Street network information. For each image, the OSM street segment nearest to the image location, within a maximum radius of 10 m, was considered to be associated or matched with that image, and all of its attributes available on OSM were appended to the image. These attributes include the name, ID (on OSM), type, number of lanes, speed limit, length, end nodes of that street, and the distance between it and the image. Even if an image has no associated street found within its 10 m radius, this could indicate a couple of possibilities: the location of the image might involve some positional error, or the image could be taken from off-road locations.

Degree of urbanisation. GHSL is a raster dataset with a resolution of 1 km^2. The settlement typology classification is based on multiple criteria including population density, built-up area size, and contiguity.

Spatial index. Using the Python package h3-py [86], we converted the latitude and longitude coordinates of each image to the index of the hexagonal cell that contains it, for all 16 levels of resolution (with the hexagon edge length ranging from 0.000584 km to 1281 km), supporting spatial analysis across various scales.

Local date and time. To convert the original capture timestamps to their respective local time zone, we first utilised the Python package pytzwhere to identify the timezone corresponding to the image's latitude and longitude coordinates, and then converted the original timestamps from GMT to the identified local time zone using the Python packages pytz and datetime.

Season. For non-tropical images located in the northern hemisphere, those taken from March to May were mapped as Spring, June to August as Summer, September to November as Autumn, and December to February as Winter. For non-tropical images located in the southern hemisphere, those taken from March to May were mapped as Autumn, June to August as Winter, September to November as Spring, and December to February as Summer.

Number of hours relative to sunrise or sunset. Using the Python package PyEphem [90], we calculated the sunrise and sunset times at each image's location on the image's capture date. The image's capture time was then compared against the sunrise and sunset timings to determine how long it was taken before or after sunrise or sunset. If the image was taken within ±12 hours relative to the sunrise (or sunset) time, the number of hours between the image's capture time and the sunrise (or sunset) time was calculated; a positive number indicates 'number of hours after sunrise (or sunset)', whereas a negative number indicates 'number of hours before sunrise (or sunset)'. This gives further information about the lighting condition of the image. If the calculation of the sunrise and sunset times yields `null`, it indicates that it could be either a polar day or polar night, depending on whether it is in spring/summer or autumn/winter, respectively.

Average and maximum speed. We considered the values for average speed invalid, if they were infinite, negative, or beyond 250 km/h (we deemed 250 km/h as a reasonable upper limit for average speed). Sequences with these invalid average speed values were given `null` for all speed metrics. Though some metrics, including the distance of the sequence, time duration of the sequence, sequence average speed (calculated as the ratio of sequence distance to sequence time duration), average segment speed (calculated as the mean of all segment speeds calculated between every two consecutive points in a sequence), maximum segment speed, and segment speed variance, were calculated at the sequence level (i.e. each sequence has one value), they have been appended to each image according to its sequence ID.

Appendix A.3. Contextual enrichment by computer vision: Training data preparation

Greedy sampling. Class imbalance was foreseeable as certain conditions were expected to occur less frequently than others, especially for 'time of day', 'platform', and 'weather' (e.g. night time, cycling path, snowy weather). Thus, we greedily sampled images to make certain smaller classes more available in the training data compared to a completely random sample. Using the calculated time

of day, we grouped all candidate SVIs into 'day' and 'night' categories and sampled 2,700 SVIs from the 'night' category and 8,000 SVIs from the 'day' category. Further, based on the OSM street type associated with each image, we grouped the images according to whether they were likely taken from a 'driving surface', 'walking surface', or 'cycling surface', and sampled similar numbers of SVIs from each category respectively. For the remaining attributes, it was difficult to infer them from the existing non-visual information, so they were not considered in the greedy sampling scheme.

Manual annotation. The 10,000 images used for model development was labelled by two trained annotators independently. To ensure that the annotators obtain a holistic understanding of their task, training was provided by the First Author to the annotators prior to the start of the labelling task. The training included an overview of the research background and objectives, detailed explanation on the meaning of each contextual attribute, and example images illustrating the differences between each class. For instance, for 'quality', example images that have high clarity and little obstruction were shown as 'good quality', while those that suffer from slight obstruction (e.g. by fog or raindrops on the windshield) or slight blurriness, i.e. are clear enough to show the overall scene but not clear enough to show granular details, were shown as 'slightly poor quality', and images that suffer from high blurriness or very poor lighting were shown as 'very poor quality' (note that nighttime images of high or acceptable clarity were not considered as having poor quality, even though they were taken under low-light conditions).

At the start of the labelling process, 50 images were set to be labelled by both annotators as well as the First Author. The labels from both annotators were compared against each other and against the labels from the First Author, and consistent agreement was observed for most of the images, indicating good quality of the labels. Throughout the labelling process, active communication was maintained between the annotators and the First Author to ensure that immediate and careful attention was given to any question raised by the annotators regarding any aspect of the labelling task.

While most of the 10,000 images were labelled only once by either annotator to ensure consistency, 1,600 of them were cross-labelled by both annotators to asses the level of agreement between them. The level of agreement was expressed as the ratio between the number of images with manually agreeable labels and the total number of images, and was evaluated for each contextual attributes. The levels of agreement for 'platform', 'weather', 'view direction', 'time of day', 'presence of glare', 'quality', and 'presence of windshield reflection' were 80.3%,

Table A.4: The mapping between the original and final values for 'platform'.

Mapped	Original
Driving surface	Paved road for vehicular traffic
	Unpaved road for vehicular traffic
	Tunnel
Cycling surface	Cycleway
Walking surface	Sidewalk
	Pedestrian zone/ living street
	Walking trails (e.g. hiking trails/ forest trails/ footpaths in parks etc.)
	Open fields

66.5%, 88.5%, 97.1%, 78.3%, 84.8%, and 80.0%, respectively, which were considered fair and acceptable levels of agreement. Disagreement in the labels for 'weather' largely stemmed from the difficulty to determine the level of cloud cover appropriate for the weather to be considered 'cloudy' or 'clear'. Among these 1,600 cross-labelled images, only images with mutually agreeable labels were used in the subsequent model development for each attribute.

For 'panoramic status', the manual annotation was done solely by the First Author, with the aid of a computer vision model developed at the initial stage of the project based on labels from the Mapillary Street-level Sequences dataset [58].

Class augmentation. For instance, when a sequence of images taken in the nighttime were found on the web application, the entire sequence of images would be downloaded and annotated with one tag, 'night', and if they were also taken from a cycling path, they would be annotated with one more tag, 'cycling surface', and these images would be used to augment both the 'night' images used in the training for lighting condition, and also the 'cycling surface' images used in the training for platform; but if they were taken from a road (which was already a major class) instead, they would be tagged with 'night' only, and used to augment only the 'night' images for lighting condition. We selected a total of more than 50,000 images from 143 different sequences across different continents for the purpose of class augmentation.

Values mapping. The original labels for 'platform', after removing 'unclear', 'indoor', and 'others', consisted of 8 possible values, which were then mapped as 'driving surface', 'cycling surface', or 'walking surface'. The mapping relationship can be found in Table A.4.

References

[1] F. Biljecki, K. Ito, Street view imagery in urban analytics and GIS: A review, Landscape and Urban Planning 215 (2021) 104217. doi:10.1016/j.landurbplan.2021.104217.
URL https://linkinghub.elsevier.com/retrieve/pii/S0169204621001808

[2] X. Li, C. Ratti, I. Seiferling, Mapping Urban Landscapes Along Streets Using Google Street View, Lecture Notes in Geoinformation and Cartography (2017) 341–356doi:10.1007/978-3-319-57336-6_24.

[3] Y. Kang, F. Zhang, S. Gao, H. Lin, Y. Liu, A review of urban physical environment sensing using street view imagery in public health studies, Annals of GIS 26 (3) (2020) 261–275. doi:10.1080/19475683.2020.1791954.
URL https://www.tandfonline.com/doi/full/10.1080/19475683.2020.1791954

[4] Y. Hou, F. Biljecki, A comprehensive framework for evaluating the quality of street view imagery, International Journal of Applied Earth Observation and Geoinformation 115 (2022) 103094. doi:10.1016/j.jag.2022.103094.

[5] F. Zhang, A. Salazar-Miranda, F. Duarte, L. Vale, G. Hack, M. Chen, Y. Liu, M. Batty, C. Ratti, Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery, Annals of the American Association of Geographers (2024) 1–22doi:10.1080/24694452.2024.2313515.
URL http://dx.doi.org/10.1080/24694452.2024.2313515

[6] J. Kang, M. Körner, Y. Wang, H. Taubenböck, X. X. Zhu, Building instance classification using street view images, ISPRS Journal of Photogrammetry and Remote Sensing 145 (2018) 44–59. doi:10.1016/j.isprsjprs.2018.02.006.

[7] M. Ogawa, K. Aizawa, Identification Of Buildings In Street Images Using Map Information, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 984–988. doi:10.1109/ICIP.2019.8803066.
URL https://ieeexplore.ieee.org/document/8803066/

[8] Y. Yan, B. Huang, Estimation of building height using a single street view image via deep neural networks, ISPRS Journal of Photogrammetry and

Remote Sensing 192 (2022) 83–98. doi:10.1016/j.isprsjprs.2022.08.006.
URL http://dx.doi.org/10.1016/j.isprsjprs.2022.08.006

[9] J. B. Cicchino, M. L. McCarthy, C. D. Newgard, S. P. Wall, C. J. DiMaggio, P. E. Kulie, B. N. Arnold, D. S. Zuby, Not all protected bike lanes are the same: Infrastructure and risk of cyclist collisions and falls leading to emergency department visits in three U.S. cities, Accident Analysis & Prevention 141 (2020) 105490. doi:10.1016/j.aap.2020.105490.
URL https://linkinghub.elsevier.com/retrieve/pii/S000145751931098X

[10] K. Ito, F. Biljecki, Assessing bikeability with street view imagery and computer vision, Transportation Research Part C: Emerging Technologies 132 (2021) 103371. doi:10.1016/j.trc.2021.103371.

[11] M. Steinmetz-Wood, K. Velauthapillai, G. O'Brien, N. A. Ross, Assessing the micro-scale environment using Google Street View: The Virtual Systematic Tool for Evaluating Pedestrian Streetscapes (Virtual-STEPS), BMC Public Health 19 (1) (2019) 1246. doi:10.1186/s12889-019-7460-3.
URL https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-7460-3

[12] Y. Li, N. Yabuki, T. Fukuda, Integrating gis, deep learning, and environmental sensors for multicriteria evaluation of urban street walkability, Landscape and Urban Planning 230 (2023) 104603. doi:10.1016/j.landurbplan.2022.104603.
URL http://dx.doi.org/10.1016/j.landurbplan.2022.104603

[13] J. Jiao, H. Wang, Forecasting Traffic Speed during Daytime from Google Street View Images using Deep Learning, Transportation Research Record: Journal of the Transportation Research Board (2023) 036119812311695doi:10.1177/03611981231169531.
URL http://journals.sagepub.com/doi/10.1177/03611981231169531

[14] D. Zünd, L. M. A. Bettencourt, Street View Imaging for Automated Assessments of Urban Infrastructure and Services, The Urban Book Series (2021) 29–40doi:10.1007/978-981-15-8983-6_4.

[15] J. Chen, L. Chen, Y. Li, W. Zhang, Y. Long, Measuring physical disorder in urban street spaces: A large-scale analysis using street view images and deep learning, Annals of the American Association of Geographers 113 (2) (2022) 469–487. doi:10.1080/24694452.2022.2114417.
URL http://dx.doi.org/10.1080/24694452.2022.2114417

[16] H. E. Pang, F. Biljecki, 3D building reconstruction from single street view images using deep learning, International Journal of Applied Earth Observation and Geoinformation 112 (2022) 102859. doi:10.1016/j.jag.2022.102859.
URL https://linkinghub.elsevier.com/retrieve/pii/S1569843222000619

[17] J. Yang, L. Zhao, J. Mcbride, P. Gong, Can you see green? Assessing the visibility of urban forests in cities, Landscape and Urban Planning 91 (2) (2009) 97–104. doi:10.1016/j.landurbplan.2008.12.004.
URL https://linkinghub.elsevier.com/retrieve/pii/S0169204608002314

[18] P. Stubbings, J. Peskett, F. Rowe, D. Arribas-Bel, A Hierarchical Urban Forest Index Using Street-Level Imagery and Deep Learning, Remote Sensing 11 (12) (2019) 1395. doi:10.3390/rs11121395.
URL https://www.mdpi.com/2072-4292/11/12/1395

[19] D. Liu, Y. Jiang, R. Wang, Y. Lu, Establishing a citywide street tree inventory with street view images and computer vision techniques, Computers, Environment and Urban Systems 100 (2023) 101924. doi:10.1016/j.compenvurbsys.2022.101924.
URL http://dx.doi.org/10.1016/j.compenvurbsys.2022.101924

[20] J. M. Keralis, M. Javanmardi, S. Khanna, P. Dwivedi, D. Huang, T. Tasdizen, Q. C. Nguyen, Health and the built environment in United States cities: Measuring associations using Google Street View-derived indicators of the built environment, BMC Public Health 20 (1) (2020) 215. doi:10.1186/s12889-020-8300-1.
URL https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-020-8300-1

[21] F. Zhang, D. Zhang, Y. Liu, H. Lin, Representing place locales using scene elements, Computers, Environment and Urban Systems 71 (2018)

153–164. doi:10.1016/j.compenvurbsys.2018.05.005.
URL https://linkinghub.elsevier.com/retrieve/pii/S0198971517303903

[22] F. Zhang, B. Zhou, L. Liu, Y. Liu, H. H. Fung, H. Lin, C. Ratti, Measuring human perceptions of a large-scale urban region using machine learning, Landscape and Urban Planning 180 (2018) 148–160.

[23] J. Kruse, Y. Kang, Y.-N. Liu, F. Zhang, S. Gao, Places for play: Understanding human perception of playability in cities using street view images and deep learning, Computers, Environment and Urban Systems 90 (2021) 101693. doi:10.1016/j.compenvurbsys.2021.101693.
URL https://linkinghub.elsevier.com/retrieve/pii/S0198971521001009

[24] F. Zhang, Z. Fan, Y. Kang, Y. Hu, C. Ratti, "perception bias": Deciphering a mismatch between urban crime and perception of safety, Landscape and Urban Planning 207 (2021) 104003.

[25] F. Guan, Z. Fang, L. Wang, X. Zhang, H. Zhong, H. Huang, Modelling people's perceived scene complexity of real-world environments using street-view panoramas and open geodata, ISPRS Journal of Photogrammetry and Remote Sensing 186 (2022) 315–331. doi:10.1016/j.isprsjprs.2022.02.012.

[26] Z. Wang, K. Ito, F. Biljecki, Assessing the equity and evolution of urban visual perceptual quality with time series street view imagery, Cities 145 (2024) 104704. doi:10.1016/j.cities.2023.104704.

[27] N. Yang, Z. Deng, F. Hu, Y. Chao, L. Wan, Q. Guan, Z. Wei, Urban perception by using eye movement data on street view images, Transactions in GIS (May 2024). doi:10.1111/tgis.13172.
URL http://dx.doi.org/10.1111/tgis.13172

[28] F. Garrido-Valenzuela, O. Cats, S. Van Cranenburgh, Where are the people? Counting people in millions of street-level images to explore associations between people's urban density and urban characteristics, Computers, Environment and Urban Systems 102 (2023) 101971. doi:10.1016/j.compenvurbsys.2023.101971.
URL https://linkinghub.elsevier.com/retrieve/pii/S0198971523000340

[29] Y. Kang, F. Zhang, S. Gao, W. Peng, C. Ratti, Human settlement value assessment from a place perspective: Considering human dynamics and perceptions in house price modeling, Cities 118 (2021) 103333. doi:10.1016/j.cities.2021.103333.
URL https://linkinghub.elsevier.com/retrieve/pii/S026427512100233X

[30] W. Qiu, W. Li, X. Liu, Z. Zhang, X. Li, X. Huang, Subjective and objective measures of streetscape perceptions: Relationships with property value in shanghai, Cities 132 (2023) 104037. doi:10.1016/j.cities.2022.104037.
URL http://dx.doi.org/10.1016/j.cities.2022.104037

[31] F. Zhang, J. Zu, M. Hu, D. Zhu, Y. Kang, S. Gao, Y. Zhang, Z. Huang, Uncovering inconspicuous places using social media check-ins and street view images, Computers, Environment and Urban Systems 81 (2020) 101478. doi:10.1016/j.compenvurbsys.2020.101478.
URL https://linkinghub.elsevier.com/retrieve/pii/S0198971519306003

[32] Y. Yao, J. Zhang, C. Qian, Y. Wang, S. Ren, Z. Yuan, Q. Guan, Delineating urban job-housing patterns at a parcel scale with street view imagery, International Journal of Geographical Information Science 35 (10) (2021) 1–24. doi:10.1080/13658816.2021.1895170.

[33] T. Zhao, X. Liang, W. Tu, Z. Huang, F. Biljecki, Sensing urban soundscapes from street view imagery, Computers, Environment and Urban Systems 99 (2023) 101915. doi:10.1016/j.compenvurbsys.2022.101915.
URL https://linkinghub.elsevier.com/retrieve/pii/S0198971522001594

[34] Y. Han, T. Zhong, A. G. Yeh, X. Zhong, M. Chen, G. Lü, Mapping seasonal changes of street greenery using multi-temporal street-view images, Sustainable Cities and Society 92 (2023) 104498. doi:10.1016/j.scs.2023.104498.

[35] X. Liang, T. Zhao, F. Biljecki, Revealing spatio-temporal evolution of urban visual environments with street view imagery, Landscape and Urban Planning 237 (2023) 104802. doi:10.1016/j.landurbplan.2023.104802.

[36] F. Biljecki, T. Zhao, X. Liang, Y. Hou, Sensitivity of measuring the urban form and greenery using street-level imagery: A comparative study of approaches and visual perspectives, International Journal of Applied Earth Observation and Geoinformation 122 (2023) 103385. doi:10.1016/j.jag.2023.103385.

[37] L. F. Alvarez Leon, S. Quinn, The value of crowdsourced street-level imagery: examining the shifting property regimes of openstreetcam and mapillary, GeoJournal 84 (2) (2018) 395–414. doi:10.1007/s10708-018-9865-4.
URL http://dx.doi.org/10.1007/s10708-018-9865-4

[38] P. Paar, J. Rekittke, Low-cost mapping and publishing methods for landscape architectural analysis and design in slum-upgrading projects, Future Internet 3 (4) (2011) 228–247. doi:10.3390/fi3040228.
URL http://dx.doi.org/10.3390/fi3040228

[39] S. D. Brunn, M. W. Wilson, Cape town's million plus black township of khayelitsha: Terrae incognitae and the geographies and cartographies of silence, Habitat International 39 (2013) 284–294. doi:10.1016/j.habitatint.2012.10.017.
URL http://dx.doi.org/10.1016/j.habitatint.2012.10.017

[40] M. Bendixen, T. Kanhema, L. L. Iversen, Putting africa on the map, Nature Africa (Aug. 2023). doi:10.1038/d44148-023-00204-1.
URL http://dx.doi.org/10.1038/d44148-023-00204-1

[41] J. Cinnamon, Visual imagery and the informal city: examining 360-degree imaging technologies for informal settlement representation, Information Technology for Development (2024) 1–18doi:10.1080/02681102.2023.2298876.
URL http://dx.doi.org/10.1080/02681102.2023.2298876

[42] M. Martell, N. Terry, R. Sengupta, C. Salazar, N. A. Errett, S. B. Miles, J. Wartman, Y. Choe, Open-source data pipeline for street-view images: A case study on community mobility during covid-19 pandemic, PLOS ONE 19 (5) (2024) e0303180. doi:10.1371/journal.pone.0303180.
URL http://dx.doi.org/10.1371/journal.pone.0303180

[43] W. Yap, J.-H. Chang, F. Biljecki, Incorporating networks in semantic understanding of streetscapes: Contextualising active mobility decisions, Environment and Planning B: Urban Analytics and City Science (2022) 239980832211388doi:10.1177/23998083221138832.
URL http://journals.sagepub.com/doi/10.1177/23998083221138832

[44] N. Zarbakhsh, G. McArdle, Points-of-Interest from Mapillary Street-level Imagery: A Dataset For Neighborhood Analytics, 2023 IEEE 39th International Conference on Data Engineering Workshops (ICDEW) 00 (2023) 154–161. doi:10.1109/icdew58674.2023.00030.

[45] W. Yap, R. Stouffs, F. Biljecki, Urbanity: automated modelling and analysis of multidimensional networks in cities, npj Urban Sustainability 3 (2023). doi:10.1038/s42949-023-00125-w.

[46] T. Inoue, R. Manabe, A. Murayama, H. Koizumi, Landscape value in urban neighborhoods: A pilot analysis using street-level images, Landscape and Urban Planning 221 (2022) 104357. doi:10.1016/j.landurbplan.2022.104357.
URL http://dx.doi.org/10.1016/j.landurbplan.2022.104357

[47] X. Ding, H. Fan, J. Gong, Towards generating network of bikeways from mapillary data, Computers, Environment and Urban Systems 88 (2021) 101632. doi:10.1016/j.compenvurbsys.2021.101632.
URL http://dx.doi.org/10.1016/j.compenvurbsys.2021.101632

[48] S. Lumnitz, T. Devisscher, J. R. Mayaud, V. Radic, N. C. Coops, V. C. Griess, Mapping trees along urban street networks with deep learning and street-level imagery, ISPRS Journal of Photogrammetry and Remote Sensing 175 (2021) 144–157. doi:10.1016/j.isprsjprs.2021.01.016.
URL http://dx.doi.org/10.1016/j.isprsjprs.2021.01.016

[49] G. Palmer, M. Green, E. Boyland, Y. S. R. Vasconcelos, R. Savani, A. Singleton, A deep learning approach to identify unhealthy advertisements in street view images, Scientific Reports 11 (1) (Mar. 2021). doi:10.1038/s41598-021-84572-4.
URL http://dx.doi.org/10.1038/s41598-021-84572-4

[50] A. Middel, N. Nazarian, M. Demuzere, B. Bechtel, Urban climate informatics: An emerging research field, Frontiers in Environmental Science 10 (May 2022). doi:10.3389/fenvs.2022.867434.
URL http://dx.doi.org/10.3389/fenvs.2022.867434

[51] D. Verma, A. Jana, K. Ramamritham, Machine-based understanding of manually collected visual and auditory datasets for urban perception studies, Landscape and Urban Planning 190 (2019) 103604. doi:10.1016/j.landurbplan.2019.103604.
URL http://dx.doi.org/10.1016/j.landurbplan.2019.103604

[52] D. Stowell, J. Kelly, D. Tanner, J. Taylor, E. Jones, J. Geddes, E. Chalstrey, A harmonised, high-coverage, open dataset of solar photovoltaic installations in the uk, Scientific Data 7 (1) (Nov. 2020). doi:10.1038/s41597-020-00739-0.
URL http://dx.doi.org/10.1038/s41597-020-00739-0

[53] M. Helbich, M. Danish, S. Labib, B. Ricker, To use or not to use proprietary street view images in (health and place) research? that is the question, Health & Place 87 (2024) 103244. doi:10.1016/j.healthplace.2024.103244.
URL http://dx.doi.org/10.1016/j.healthplace.2024.103244

[54] I. A. V. Sánchez, S. Labib, Accessing eye-level greenness visibility from open-source street view images: A methodological development and implementation in multi-city and multi-country contexts, Sustainable Cities and Society 103 (2024) 105262. doi:10.1016/j.scs.2024.105262.
URL http://dx.doi.org/10.1016/j.scs.2024.105262

[55] D. Ki, K. Park, Z. Chen, Bridging the gap between pedestrian and street views for human-centric environment measurement: A GIS-based 3D virtual environment, Landscape and Urban Planning 240 (2023) 104873. doi:10.1016/j.landurbplan.2023.104873.

[56] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The Cityscapes Dataset for Semantic Urban Scene Understanding, 2016.

[57] G. Neuhold, T. Ollmann, S. Rota Bulo, P. Kontschieder, The mapillary vistas dataset for semantic understanding of street scenes, in: Proceedings of

the IEEE international conference on computer vision, 2017, pp. 4990–4999.

[58] F. Warburg, S. Hauberg, M. López-Antequera, P. Gargallo, Y. Kuang, J. Civera, Mapillary Street-Level Sequences: A Dataset for Lifelong Place Recognition, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 00 (2020) 2623–2632. doi:10.1109/cvpr42600.2020.00270.

[59] F. Yu, D. Wang, E. Shelhamer, T. Darrell, Deep layer aggregation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2403–2412.

[60] L. Liu, A. Sevtsuk, Clarity or confusion: A review of computer vision street attributes in urban studies and planning, Cities 150 (2024) 105022. doi:10.1016/j.cities.2024.105022.
URL http://dx.doi.org/10.1016/j.cities.2024.105022

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

[62] A. Dubey, N. Naik, D. Parikh, R. Raskar, C. A. Hidalgo, Deep learning the city: Quantifying urban perception at a global scale, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, 2016, pp. 196–212.

[63] P. Salesses, K. Schechtner, C. A. Hidalgo, The collaborative image of the city: mapping the inequality of urban perception, PloS one 8 (7) (2013) e68400.

[64] Y. Ogawa, T. Oki, C. Zhao, Y. Sekimoto, C. Shimizu, Evaluating the subjective perceptions of streetscapes using street-view images, Landscape and Urban Planning 247 (2024) 105073. doi:10.1016/j.landurbplan.2024.105073.
URL http://dx.doi.org/10.1016/j.landurbplan.2024.105073

[65] C. Ertler, J. Mislej, T. Ollmann, L. Porzi, G. Neuhold, Y. Kuang, The Mapillary Traffic Sign Dataset for Detection and Classification on a Global Scale, arXiv (2019). arXiv:1909.04422, doi:10.48550/arxiv.1909.04422.

[66] A. Ali-bey, B. Chaib-draa, P. Giguère, GSV-Cities: Toward appropriate supervised visual place recognition, Neurocomputing 513 (2022) 194–203. doi:10.1016/j.neucom.2022.09.127.

[67] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, A. Torralba, Semantic Understanding of Scenes through the ADE20K Dataset (Oct. 2018). arXiv:1608.05442, doi:10.48550/arXiv.1608.05442.

[68] R. Prykhodchenko, P. Skruch, Road scene classification based on street-level images and spatial data, Array 15 (2022) 100195. doi:10.1016/j.array.2022.100195.

[69] N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser, C. A. Hidalgo, Computer vision uncovers predictors of physical urban change, Proceedings of the National Academy of Sciences 114 (29) (2017) 7571–7576. doi:10.1073/pnas.1619003114.

[70] Z. Xu, D. Tao, Y. Zhang, J. Wu, A. C. Tsoi, Architectural Style Classification Using Multinomial Latent Logistic Regression, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2014, pp. 600–615. doi:10.1007/978-3-319-10590-1_39.

[71] M. Sun, F. Zhang, F. Duarte, C. Ratti, Understanding architecture age and style through deep learning, Cities 128 (2022) 103787. doi:10.1016/j.cities.2022.103787.

[72] S. Moschos, P. Charitidis, S. Doropoulos, A. Avramis, S. Vologiannidis, StreetScouting dataset: A Street-Level Image dataset for finetuning and applying custom object detectors for urban feature detection, Data in Brief 48 (2023) 109042. doi:10.1016/j.dib.2023.109042.

[73] M. Ren, X. Zhang, X. Zhi, Y. Wei, Z. Feng, An annotated street view image dataset for automated road damage detection, Scientific Data 11 (1) (2024) 407. doi:10.1038/s41597-024-03263-7.

[74] M. R. Ibrahim, J. Haworth, T. Cheng, URBAN-i: 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) (2021) 76–93. doi:10.1177/2399808319846517.

[75] G. Astruc, N. Dufour, I. Siglidis, C. Aronssohn, N. Bouia, S. Fu, R. Loiseau, V. N. Nguyen, C. Raude, E. Vincent, L. XU, H. Zhou, L. Landrieu, OpenStreetView-5M: The Many Roads to Global Visual Geolocation (Apr. 2024). arXiv:2404.18873.

[76] O. Grau, K. Hagn, VALERIE22 - A photorealistic, richly metadata annotated dataset of urban environments, in: Proceedings of the 7th ACM Computer Science in Cars Symposium, CSCS '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1–9. doi:10.1145/3631204.3631866.

[77] SimpleMaps.com, World cities database (2023). doi:10.34740/KAGGLE/DSV/5294033.
URL https://www.kaggle.com/dsv/5294033

[78] P. D. Aboagye, A. Sharifi, Post-fifth assessment report urban climate planning: Lessons from 278 urban climate action plans released from 2015 to 2022, Urban Climate 49 (2023) 101550. doi:10.1016/j.uclim.2023.101550.
URL http://dx.doi.org/10.1016/j.uclim.2023.101550

[79] UN Statistical Commission, A recommendation on the method to delineate cities, urban and rural areas for international statistical comparisons technical report (2020).

[80] A. J. Tatem, WorldPop, open data for spatial demography, Scientific Data 4 (1) (2017) 170004. doi:10.1038/sdata.2017.4.

[81] Mapillary, Mapillary Python SDK, https://github.com/mapillary/mapillary-python-sdk (2022).

[82] KartaView, OpenStreetCam API (2.0), http://doc.kartaview.org/.

[83] G. Boeing, OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks, Computers, Environment and Urban Systems 65 (2017) 126–139. doi:10.1016/j.compenvurbsys.2017.05.004.

[84] Y. Huang, F. Zhang, Y. Gao, W. Tu, F. Duarte, C. Ratti, D. Guo, Y. Liu, Comprehensive urban space representation with varying numbers of street-level images, Computers, Environment and Urban Systems 106 (2023) 102043. doi:10.1016/j.compenvurbsys.2023.102043.

[85] M. Schiavina, M. Melchiorri, M. Pesaresi, GHS-SMOD R2023A - GHS settlement layers, application of the Degree of Urbanisation methodology (stage I) to GHS-POP R2023A and GHS-BUILT-S R2023A, multitemporal (1975-2030) (2023). doi:10.2905/A0DF7A6F-49DE-46EA-9BDE-563437A6E2BA.

[86] Uber, h3-py: Uber's H3 Hexagonal Hierarchical Geospatial Indexing System in Python, https://github.com/uber/h3-py (2023).

[87] S. Woźniak, P. Szymański, Hex2vec: Context-Aware Embedding H3 Hexagons with OpenStreetMap Tags, in: Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, GEOAI '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 61–71. doi:10.1145/3486635.3491076.

[88] Y.-C. Chiang, H.-H. Liu, D. Li, L.-C. Ho, Quantification through deep learning of sky view factor and greenery on urban streets during hot and cool seasons, Landscape and Urban Planning 232 (2023) 104679. doi:10.1016/j.landurbplan.2022.104679.
URL http://dx.doi.org/10.1016/j.landurbplan.2022.104679

[89] K. Chen, M. Tian, J. Zhang, X. Xu, L. Yuan, Evaluating the seasonal effects of building form and street view indicators on street-level land surface temperature using random forest regression, Building and Environment 245 (2023) 110884. doi:10.1016/j.buildenv.2023.110884.
URL http://dx.doi.org/10.1016/j.buildenv.2023.110884

[90] B. C. Rhodes, PyEphem: Astronomical Ephemeris for Python, Astrophysics Source Code Library (2011) ascl:1112.014.

[91] S. Chen, F. Biljecki, Automatic assessment of public open spaces using street view imagery, Cities 137 (2023) 104329. doi:10.1016/j.cities.2023.104329.
URL https://linkinghub.elsevier.com/retrieve/pii/S0264275123001415

[92] Y. Yan, C.-C. Feng, W. Huang, H. Fan, Y.-C. Wang, A. Zipf, Volunteered geographic information research in the first decade: a narrative review of selected journal articles in GIScience, International

Journal of Geographical Information Science 34 (9) (2020) 1–27. doi:10.1080/13658816.2020.1730848.

[93] X. Huang, S. Wang, D. Yang, T. Hu, M. Chen, M. Zhang, G. Zhang, F. Biljecki, T. Lu, L. Zou, C. Y. H. Wu, Y. M. Park, X. Li, Y. Liu, H. Fan, J. Mitchell, Z. Li, A. Hohl, Crowdsourcing geospatial data for earth and human observations: A review, Journal of Remote Sensing 4 (0105) (2024). doi:10.34133/remotesensing.0105.

[94] T. Novack, L. Vorbeck, H. Lorei, A. Zipf, Towards Detecting Building Facades with Graffiti Artwork Based on Street View Images, ISPRS International Journal of Geo-Information 9 (2) (2020) 98. doi:10.3390/ijgi9020098.

[95] C. Wang, S. E. Antos, L. M. Triveno, Automatic detection of unreinforced masonry buildings from street view images using deep learning-based image segmentation, Automation in Construction 132 (2021) 103968. doi:10.1016/j.autcon.2021.103968.

[96] Y. Xia, N. Yabuki, T. Fukuda, Sky view factor estimation from street view images based on semantic segmentation, Urban Climate 40 (2021) 100999. doi:10.1016/j.uclim.2021.100999.

[97] M. Ignatius, R. Xu, Y. Hou, X. Liang, T. Zhao, S. Chen, N. Wong, F. Biljecki, Local Climate Zones: Lessons from Singapore and potential improvement with street view imagery, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences X-4/W2-2022 (2022) 121–128. doi:10.5194/isprs-annals-X-4-W2-2022-121-2022.

[98] A. V. Vo, M. Bertolotto, U. Ofterdinger, D. F. Laefer, In search of basement indicators from street view imagery data: An investigation of data sources and analysis strategies, KI - Künstliche Intelligenz 37 (1) (2023) 41–53. doi:10.1007/s13218-022-00792-4.
URL http://dx.doi.org/10.1007/s13218-022-00792-4

[99] F. Biljecki, Y. S. Chow, K. Lee, Quality of crowdsourced geospatial building information: A global assessment of OpenStreetMap attributes, Building and Environment (2023) 110295doi:10.1016/j.buildenv.2023.110295.

[100] H. Senaratne, A. Mobasheri, A. L. Ali, C. Capineri, M. M. Haklay, A review of volunteered geographic information quality assessment methods,

International Journal of Geographical Information Science 31 (1) (2017) 139 – 167. doi:10.1080/13658816.2016.1189556.

[101] A. Y. Grinberger, M. Schott, M. Raifer, A. Zipf, An analysis of the spatial and temporal distribution of large-scale data production events in OpenStreetMap, Transactions in GIS 25 (2) (2021) 622–641. doi:10.1111/tgis.12746.

[102] I. Majic, E. Naghizade, S. Winter, M. Tomko, There is no way! Ternary qualitative spatial reasoning for error detection in map data, Transactions in GIS 25 (4) (2021) 2048–2073. doi:10.1111/tgis.12765.

[103] D. Sarkar, J. T. Anderson, Corporate editors in OpenStreetMap: Investigating co-editing patterns, Transactions in GIS 26 (4) (2022) 1879–1897. doi:10.1111/tgis.12910.

[104] B. Herfort, S. Lautenbach, J. Porto de Albuquerque, J. Anderson, A. Zipf, A spatio-temporal analysis investigating completeness and inequalities of global urban building data in openstreetmap, Nature Communications 14 (1) (Jul. 2023). doi:10.1038/s41467-023-39698-6.
URL http://dx.doi.org/10.1038/s41467-023-39698-6

[105] A. Yang, H. Fan, Q. Jia, M. Ma, Z. Zhong, J. Li, N. Jing, How do contributions of organizations impact data inequality in openstreetmap?, Computers, Environment and Urban Systems 109 (2024) 102077. doi:10.1016/j.compenvurbsys.2024.102077.
URL http://dx.doi.org/10.1016/j.compenvurbsys.2024.102077

[106] X. Zhang, J. An, Y. Zhou, M. Yang, X. Zhao, How sustainable is openstreetmap? tracking individual trajectories of editing behavior, International Journal of Digital Earth 17 (1) (Feb. 2024). doi:10.1080/17538947.2024.2311320.
URL http://dx.doi.org/10.1080/17538947.2024.2311320

[107] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, Y. Li, MaxViT: Multi-Axis Vision Transformer, arXiv (2022). arXiv:2204.01697, doi:10.48550/arxiv.2204.01697.

[108] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

[109] M. R. Ibrahim, J. Haworth, T. Cheng, WeatherNet: Recognising Weather and Visual Conditions from Street-Level Images Using Deep Residual Learning, ISPRS International Journal of Geo-Information 8 (12) (2019) 549. doi:10.3390/ijgi8120549.

[110] M. M. Dhananjaya, V. R. Kumar, S. Yogamani, Weather and Light Level Classification for Autonomous Driving: Dataset, Baseline and Active Learning, in: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), IEEE Press, Indianapolis, IN, USA, 2021, pp. 2816–2821. doi:10.1109/ITSC48978.2021.9564689.

[111] A. Golchubian, O. Marques, M. Nojoumian, Photo quality classification using deep learning, Multimedia Tools and Applications 80 (14) (2021) 22193–22208. doi:10.1007/s11042-021-10766-7.

[112] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, R. Girdhar, Masked-attention mask transformer for universal image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1290–1299.

[113] B. Cheng, A. Schwing, A. Kirillov, Per-pixel classification is not all you need for semantic segmentation, Advances in Neural Information Processing Systems 34 (2021) 17864–17875.

[114] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition (Apr. 2015). arXiv:1409.1556.

[115] J. Ouyang, Code repository for predicting human perception, https://github.com/strawmelon11/human-perception-place-pulse (2023).

[116] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[117] J. Rui, Measuring streetscape perceptions from driveways and sidewalks to inform pedestrian-oriented street renewal in Düsseldorf, Cities 141 (2023) 104472. doi:10.1016/j.cities.2023.104472.

[118] X. Ma, C. Ma, C. Wu, Y. Xi, R. Yang, N. Peng, C. Zhang, F. Ren, Measuring human perceptions of streetscapes to better inform urban renewal: A perspective of scene semantic parsing, Cities 110 (2021) 103086.

[119] S. Mahajan, greenR: An open-source framework for quantifying urban greenness, Ecological Indicators 163 (2024) 112108. doi:10.1016/j.ecolind.2024.112108.

[120] T. Aikoh, R. Homma, Y. Abe, Comparing conventional manual measurement of the green view index with modern automatic methods using google street view and semantic segmentation, Urban Forestry and Urban Greening 80 (2023) 127845. doi:10.1016/j.ufug.2023.127845.
URL http://dx.doi.org/10.1016/j.ufug.2023.127845

[121] H. Zhu, X. Nan, F. Yang, Z. Bao, Utilizing the green view index to improve the urban street greenery index system: A statistical study using road patterns and vegetation structures as entry points, Landscape and Urban Planning 237 (2023) 104780. doi:10.1016/j.landurbplan.2023.104780.
URL http://dx.doi.org/10.1016/j.landurbplan.2023.104780

[122] Y. Lu, E. J. S. Ferranti, L. Chapman, C. Pfrang, Assessing urban greenery by harvesting street view data: A review, Urban Forestry and Urban Greening 83 (2023) 127917. doi:10.1016/j.ufug.2023.127917.
URL http://dx.doi.org/10.1016/j.ufug.2023.127917

[123] M. Li, H. Sheng, J. Irvin, H. Chung, A. Ying, T. Sun, A. Y. Ng, D. A. Rodriguez, Marked crosswalks in US transit-oriented station areas, 2007–2020: A computer vision approach using street view imagery, Environment and Planning B: Urban Analytics and City Science (2022) 239980832211121doi:10.1177/23998083221112157.

[124] M. Bendixen, T. Kanhema, L. L. Iversen, Putting Africa on the map, Nature Africa (8 2023). doi:10.1038/d44148-023-00204-1.

[125] A. D. Singleton, S. Spielman, C. Brunsdon, Establishing a framework for Open Geographic Information science, International Journal of Geographical Information Science 30 (8) (2016) 1507–1521. doi:10.1080/13658816.2015.1137579.

[126] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne,

J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (1) (2016) 160018. doi:10.1038/sdata.2016.18.

[127] L. Juhász, H. H. Hochmair, User Contribution Patterns and Completeness Evaluation of Mapillary, a Crowdsourced Street Level Photo Service, Transactions in GIS 20 (6) (2016) 925–947. doi:10.1111/tgis.12190.

[128] R. Mahabir, R. Schuchard, A. Crooks, A. Croitoru, A. Stefanidis, Crowdsourcing Street View Imagery: A Comparison of Mapillary and OpenStreetCam, ISPRS International Journal of Geo-Information 9 (6) (2020) 341. doi:10.3390/ijgi9060341.

[129] D. Ma, H. Fan, W. Li, X. Ding, The State of Mapillary: An Exploratory Analysis, ISPRS International Journal of Geo-Information 9 (1) (2019) 10. doi:10.3390/ijgi9010010.
URL https://www.mdpi.com/2220-9964/9/1/10

[130] X. Huang, S. Wang, D. Yang, T. Hu, M. Chen, M. Zhang, G. Zhang, F. Biljecki, T. Lu, L. Zou, C. Y. H. Wu, Y. M. Park, X. Li, Y. Liu, H. Fan, J. Mitchell, Z. Li, A. Hohl, Crowdsourcing geospatial data for earth and human observations: A review, Journal of Remote Sensing 4 (Jan. 2024). doi:10.34133/remotesensing.0105.
URL http://dx.doi.org/10.34133/remotesensing.0105

[131] S. Quinn, L. Alvarez León, Every single street? rethinking full coverage across street-level imagery platforms, Transactions in GIS 23 (6) (2019) 1251–1272. doi:10.1111/tgis.12571.
URL http://dx.doi.org/10.1111/tgis.12571

[132] J. Kim, K. M. Jang, An examination of the spatial coverage and temporal variability of google street view (gsv) images in small- and medium-sized

cities: A people-based approach, Computers, Environment and Urban Systems 102 (2023) 101956. doi:10.1016/j.compenvurbsys.2023.101956.
URL http://dx.doi.org/10.1016/j.compenvurbsys.2023.101956

[133] Z. Liu, T. Li, T. Ren, D. Chen, W. Li, W. Qiu, Day-to-Night Street View Image Generation for 24-Hour Urban Scene Auditing Using Generative AI, Journal of Imaging 10 (5) (2024) 112. doi:10.3390/jimaging10050112.

[134] J. Luo, T. Zhao, L. Cao, F. Biljecki, Water View Imagery: Perception and evaluation of urban waterscapes worldwide, Ecological Indicators 145 (2022) 109615. doi:10.1016/j.ecolind.2022.109615.
URL https://linkinghub.elsevier.com/retrieve/pii/S1470160X22010883

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

[136] W. Yap, F. Biljecki, A global feature-rich network dataset of cities and dashboard for comprehensive urban analyses, Scientific Data 10 (1) (Sep. 2023). doi:10.1038/s41597-023-02578-1.
URL http://dx.doi.org/10.1038/s41597-023-02578-1