This is the Accepted Manuscript version of an article published by Elsevier in the journal *Computers, Environment and Urban Systems* in 2027, which is available at: https://doi.org/10.1016/j.compenvurbsys.2026.102509
Cite as: Gu Y, Liang X, Fan Z, Hou Y, Biljecki F (2027): Street view video for urban sensing: Potential for measuring urban vitality. *Computers, Environment and Urban Systems*, 131: 102509.

Street View Video for Urban Sensing: Potential for Measuring Urban Vitality

Youlong Gu[a], Xiucheng Liang[a], Zicheng Fan[a], Yujun Hou[a,b], Filip Biljecki[a,c,*]

[a]*Department of Architecture, National University of Singapore, 4 Architecture Dr, 117566, Singapore*
[b]*Department of Geography and Planning, University of Liverpool, Liverpool, L69 7ZT, United Kingdom*
[c]*Department of Real Estate, National University of Singapore, 15 Kent Ridge Dr, 119245, Singapore*

Abstract

Urban vitality is a key quality of the livability and attractiveness of a city, with street-level social dynamics recognized as a central representation. However, existing approaches to capturing such dynamics, primarily through pedestrian volume, rely on static imagery, location-based proxies, or custom sensors. Such methods are typically discontinuous, limited in spatio-temporal resolution, costly, and overlook the role of sound in shaping environments. To address these gaps, this study explores the feasibility of using Street View Video (SVV), a novel, promising, yet under-utilized data source, for capturing street-level dynamics. Leveraging widely available crowdsourced city tour videos as a representative form of SVV, we develop an automated and scalable framework for high-precision geo-localization of SVV by extracting textual and depth information. Further, we propose a visual-auditory pipeline that integrates continuous pedestrian activity and sound environments as complementary dimensions that reveal street-level social dynamics. Using Melbourne as a case study, we identified 892 high-quality and geolocatable videos totaling over 541 hours and achieved an accuracy of 77% for geo-localization. We generated spatio-temporal maps that not only reproduced well-studied pedestrian volume patterns, but also previously neglected aspects, including human actions (walking, standing, sitting), sound intensity, and sound types (speech, music, vehicle). Correlation and quadrant analyses further explored the relationships among these vitality dimensions, revealing distinctive street characteristics and highlighting the limitations of conventional road classifications. Overall, these findings demonstrate the potential of SVV as a multimodal and reproducible source of urban vitality indicators, providing a proof-of-concept for future urban sensing applications.

Keywords: urban analytics, computer vision, volunteered geographic information, built environment, urban vibrancy, virtual tour

*Corresponding author. Email: filip@nus.edu.sg

© 2025, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

Urban vitality has long been regarded as a fundamental dimension of urban life, central to evaluating a city's livability, attractiveness, and its capacity for sustainable development (Montgomery, 1998; Gehl, 2011; Hall and Pfeiffer, 2013; Zeng et al., 2018; Huang et al., 2020; Chen et al., 2023; Mouratidis and Delclòs-Alió, 2026). It encompasses not only the intensity of economic and social activities, but more fundamentally reflects the potential of urban space to support human engagement — factors closely linked to residents' well-being, sense of belonging, and spatial identity (Dale et al., 2010; Yin, 2017; Chen et al., 2020; Huang et al., 2020). Among its various dimensions, street-level social dynamics are widely recognized as one of its most direct and visible expressions (Jacobs, 1961; Whyte, 1980; Gehl, 2013; Doan and Zhang, 2025), capturing not only the functional use of street public space but also its attractiveness, accessibility, and capacity to promote everyday life, while reflecting broader patterns of local economic and cultural vibrancy (Jacobs, 1961; Gehl, 2011). However, car-centric urban planning has gradually marginalized public spaces (Appleyard, 1980; Newman and Kenworthy, 1998; Sheller and Urry, 2000), reducing walkability and weakening social interactions. Against this backdrop, the ability to reliably monitor and assess street dynamics has become increasingly crucial for sustainable and human-centered urban development (Whyte, 1980; Yigitcanlar and Teriman, 2015).

Previous studies have measured street-level dynamics using a variety of quantitative methods and data sources, with the most direct approach involving the estimation of pedestrian volume. For example, researchers have applied object detection techniques to extract large-scale pedestrian count from street view imagery (SVI) (Yin et al., 2015; Chen et al., 2020, 2022). Others have utilized location-based services (LBS) data sources, including GPS trajectories, mobile signaling records, and heatmaps derived from digital map services to estimate pedestrian density (Wu et al., 2021; Liu et al., 2021; Wang et al., 2023). Furthermore, video cameras and sensors have been deployed to monitor pedestrian flows along specific street segments (Hou et al., 2020; Tokuda et al., 2020; Li et al., 2022; Salazar-Miranda et al., 2023). Other studies have extended the focus beyond volume alone, examining pedestrian behaviors such as movement patterns, group dynamics, and socio-cultural practices (Zaki and Sayed, 2018; Loo and Fan, 2023; Salazar-Miranda et al., 2025).

Although these approaches have contributed valuable insights for identifying street dynamics with various types of spatial data, none has yet been perfected and each faces notable limitations. First, most existing methods achieve high resolution in either space or time, but rarely both. For example, SVI usually provides broad, spatially continuous coverage across a city but is updated infrequently, often yearly. By contrast, fixed or mobile cameras can capture continuous temporal dynamics but usually cover only limited point-based areas (Doan and Zhang, 2025). Second, data accessibility and cost remain a concern. LBS data, for instance, offers fine-grained spatio-temporal resolution, but their accessibility to researchers is often constrained by reliance on provider agreements or the high costs of acquisition and deployment (Oliver et al., 2020). Third, while much prior work has emphasized pedestrian flows, the urban sound environment remains an important yet underexplored dimension. Street sounds fundamentally shape the atmosphere of public spaces and influence perceptions of vitality (Schafer, 1977; Kang et al., 2016; Aletta and Kang, 2018; Cao and Kang, 2024). Although prior studies have mapped urban sound environments and explored interactions with the built form (Kang et al., 2016; Kang, 2023), few have jointly considered pedestrian activity and sound environment in assessing street-level dynamics for measuring vitality at a large scale.

To address these limitations, we draw inspiration from recent pioneering work that lever-

ages video sources at street level — including city tour videos (Hartmann and Purves, 2023), webcams (He et al., 2025), and dashcam footage (Oda and Yoshimura, 2024) — for urban sensing tasks such as analyzing transportation modes and pedestrian distribution. We argue that such commercial and crowdsourced video data represent a readily available and accessible yet underutilized resource, offering key advantages including continuous spatial observation while capturing previously overlooked aspects of the urban environment such as ambient sound. These characteristics make them particularly promising for addressing the aforementioned challenges in urban vitality research and other urban sensing tasks.

Building on these emerging examples, we posit that a broader class of such street-level video data exists, whose shared properties can provide unique perspectives and meaningfully extend established urban sensing approaches and instruments such as SVI. Accordingly, we propose the recognition and conceptualization of Street View Video (SVV) as a novel data type — an umbrella term encompassing these similar forms of street-level video data. By systematically identifying, classifying, and studying SVV, we can deepen our understanding of this emerging data source and examine how, and to what extent, it can be utilized in urban research. To structure our investigation of this new data type, we pose the following research questions:

- Which data forms can be classified under the concept of SVV?
- How can SVV be processed to retrieve geospatial information?
- How can SVV be utilized to measure urban vitality?

To systematically address these questions, we first reviewed the literature to identify potential forms of SVV and define the concept of it. Using the recently emerging and culturally influential city tour videos as example data, we investigate urban vitality within high-density urban centers, where economic and cultural activities concentrate and which have been a primary focus of research in this field (Delclòs-Alió et al., 2019; Guo et al., 2021; Liu et al., 2021; Li et al., 2022; Salazar-Miranda et al., 2023; Bai et al., 2025; Chen et al., 2025b). We developed an innovative SVV processing framework for high-precision geo-localization based on textual and depth information to infer the location. Building on this foundation, we designed, for the first time, an automated video quality control framework alongside a visual–auditory pipeline to characterize urban vitality across two key dimensions: pedestrian activity, captured through human volume and actions, and the sound environment, captured through sound intensity and types. As a case study, we implemented our method in the central business district (CBD) of Melbourne, Australia. Among 1,602 videos longer than 10 minutes, we successfully traced 892 high-quality videos, including 801 recorded during the day and 91 at night, covering most areas of the CBD. Across these identified trajectories, our approach achieved a street-level geo-localization accuracy of 77.19%. Leveraging these results, we are among the first to reveal patterns of urban vitality based on the joint distribution of human activity and sound. We generated fine-grained spatial and temporal patterns at the 50-meter street-segment level within the city center, and derived initial insights into street use and characteristics. More broadly, this study serves as a proof-of-concept for the use of SVV to support future research on street experience, public-space quality, and other dimensions that are difficult to capture using conventional datasets.

2. Related work

2.1. Measuring pedestrian activity

As urban vitality has gained increasing prominence in urban studies, pedestrian activity, long regarded as the most direct and widely studied indicator of street-level social dynamics, has become a foundational research focus (Whyte, 1980; Gehl, 2013; Yin et al., 2015). This area has evolved from traditional field-based techniques to more recent data-driven approaches.

Early methods for estimating pedestrian activity primarily relied on field observation, audits, and surveys, which, while providing valuable insights, were labor-intensive, time-consuming, and limited in spatial and temporal coverage and scalability (Sisiopiku and Akin, 2003; Diogenes et al., 2007; Islam et al., 2014). Moreover, survey-based data was susceptible to subjective bias, limiting its representativeness and reliability (Ben-Joseph et al., 2013; Kelly et al., 2013; Yin et al., 2015; Salazar-Miranda et al., 2023).

With the advancement of geospatial technologies, computer vision, and internet-based image services, automated and scalable methods for mapping pedestrian patterns have proliferated (Feng et al., 2021). For instance, Yin et al. (2015) and Chen et al. (2020) pioneered the use of image-based estimation methods using Google's SVI, and Nathvani et al. (2025) further leveraged time-lapsed SVI to capture a more refined temporal dimension. Parallel to this, LBS data sourced from platforms such as Baidu, Tencent, or third-party data providers has emerged as a powerful resource, offering high spatial-temporal granularity that enables researchers to infer individual movement patterns (Wu et al., 2021, 2022; Li et al., 2020; Jiang et al., 2022). Urban heatmaps derived from aggregated user location data on mapping platforms have also been used as a location-based proxy for human activity (Wang et al., 2023; Li and Pan, 2023; Liu et al., 2021).

More recently, video data has received increasing attention as a promising form for pedestrian observation. Researchers have deployed custom cameras on streets and squares (Liang et al., 2020; Hou et al., 2020; Ceccarelli et al., 2023), or mounted sensors on vehicles (Tokuda et al., 2020; Salazar-Miranda et al., 2023), to capture movement patterns at relatively high temporal resolution. Compared to static data, these videos offer continuous observation of social behaviors such as walking, talking, and grouping, with potential to reveal socio-cultural phenomena (Salazar-Miranda et al., 2023).

These methods have substantially enhanced the capacity to measure pedestrian activity with distinct analytical strengths. However, several limitations exist. Traditional static SVI lacks the ability to capture dynamic human flows (Doan and Zhang, 2025). LBS data, while offering high spatial and temporal precision, is often costly to access and may suffer from representativeness bias (Oliver et al., 2020; Gravitz-Sela et al., 2025). Emerging video-based and mobile-sensor approaches allow for more detailed and accurate observations. However, their limited coverage and high implementation costs continue to constrain scalability and accessibility (Hu et al., 2023; Lian et al., 2024). These challenges highlight the need for a new source of data for measuring pedestrian activity, one that is more widely available, dynamically captured, and capable of supporting fine-grained observation at scale. Such a resource would offer critical opportunities for deepening our understanding of the spatial-temporal dynamics of pedestrian activity.

2.2. Mapping urban sound environment

Urban sound is an integral component of the built environment, essential for understanding place and space, and it significantly shapes public health and perceptions of the city (Andringa

and Lanser, 2013; Hong and Jeon, 2017; Thulin, 2018; Gasco et al., 2020). Measuring and mapping sound is thus a critical part of modern acoustic research, enabling us to understand how the sound environment affects people, fosters urban vitality, and informs planning for more livable cities (Aletta, 2016; Aletta and Kang, 2018; Van Kamp et al., 2019; Kang, 2023).

Existing approaches can be broadly divided into direct field collection and proxy data methods. Field-based measurements have long been the mainstream, offering reliable acoustic information. Typically, these involve sampling sounds at fixed stations or with mobile units (Job et al., 2016; Hong and Jeon, 2017; Zambon et al., 2017; Aumond et al., 2018b,a), capturing sources such as traffic noise or natural sounds, and interpolating results across larger areas, with additional contributions from citizen science and crowdsourced collection (Can et al., 2023; Othman et al., 2024; Boumchich et al., 2024).

In recent years, proxy data approaches have gained attention as a way to reduce costs while scaling up coverage. For example, Aiello et al. (2015) used geotagged social media photos, deriving sound-related words from tags to map the distribution of sound categories across city streets. Zhao et al. (2023) and Huang et al. (2024a) combined street view imagery with field measurements to produce high-resolution soundscape and noise maps that support urban sensing.

While both direct and proxy methods offer valuable perspectives, each has drawbacks. Field-based approaches are labor-intensive and difficult to scale, hindering comparative analyses across space, time, or cultures (Kogan et al., 2017). Proxy approaches achieve broader coverage but face concerns of representativeness, as they rely on indirect indicators rather than acoustic measurements. These limitations underscore the need for new data sources that have the potential to capture sound directly at a lower cost, with higher resolution, while also serving as a missing dimension in holistic assessments of urban vitality.

3. Definition — Street View Video

With advances in map-based services and computer vision, researchers have increasingly relied on visual representations of cities, with Street View Imagery (SVI) being one of the most widely used and studied sources (Anguelov et al., 2010; Biljecki and Ito, 2021). However, due to the inherently discontinuous nature of image data, researchers have increasingly turned to video sources to capture urban environments and their dynamic changes in greater detail, for example, 360-degree virtual videos (Zhu et al., 2019; Cinnamon and Jahiu, 2023; Yang et al., 2025) and UAV-based aerial videos (Zhu et al., 2019; Wang et al., 2024), with street-level videos emerging as the most common and widely adopted form.

We posit that video data at the street level may share certain distributional and structural characteristics. Identifying and comparing these forms provides a unique opportunity to systematically examine their features and assess their potential for urban research tasks. Based on commonly observed types, we distinguish six forms of street-level video, as illustrated in Figure 1. These can be broadly categorized into two groups: stationary video data and mobile video data, depending on whether the recording device remains stationary or moves during data collection.

Stationary video data are captured by devices installed at specific sites such as buildings, poles, or other urban infrastructure. Public webcams represent one of the most accessible forms, often available on various commercial platforms. These cameras are typically installed at different locations and provide continuous 24-hour live streaming. They have achieved widespread global coverage for over a decade (Jacobs et al., 2009), and their use has surged during the COVID-19 pandemic, particularly for monitoring plazas, landmarks, and other pub-

lic areas (Jarratt, 2020). Recent studies have increasingly drawn on these resources, confirming their value for traffic analysis and the study of urban vitality (He et al., 2025; Galich et al., 2025; Chen et al., 2025b). Another form, surveillance footage (e.g., CCTV systems) from municipal or institutional monitoring systems, is even more widely deployed across cities. Such data have been instrumental in evaluating traffic flows, public safety, environmental health, and pedestrian social activity (Grabowski and Czyżewski, 2020; Zou et al., 2021; Shorfuzzaman et al., 2021; Jin et al., 2024; Dardour et al., 2025; Gravitz-Sela et al., 2025; Ibrahim and Lyons, 2025). However, as they are typically managed by government agencies, accessibility remains limited.

In contrast, mobile video data are collected by devices that move through urban environments, commonly mounted on vehicles or carried by individuals. Dashcams, widely deployed on private and commercial vehicles and often equipped with GPS, offer new opportunities for urban sensing. Researchers have collaborated with companies and institutions to leverage dashcam data to analyze pedestrian volumes and patterns (Oda and Yoshimura, 2024; Lian et al., 2024), with some crowdsourced datasets even extending to over 500 cities worldwide (Alam et al., 2025). Vehicle-mounted mobile sensors are also frequently used in traffic research, enabling specialized analyses such as street vegetation assessment or pedestrian activity detection (Tokuda et al., 2020; Hu et al., 2022; Salazar-Miranda et al., 2023; Zhang et al., 2025b). While these custom setups provide rich analytical potential, they often demand greater resources and are constrained by limited spatial coverage (Hu et al., 2023; Lian et al., 2024). In addition, user-generated street-level videos with geotags provide valuable observational material. Social sharing videos, often created through spontaneous personal uploads from various devices such as mobile phones to social media platforms, offer immediacy and a human perspective, supporting near real-time analysis of events such as natural disasters (Lewis and Park, 2018; Lombardo et al., 2025; Rodríguez et al., 2025), and providing a unique lens for perceiving the city image (Chen et al., 2025a). However, the lack of standardization introduces challenges regarding data quality and coverage. By contrast, city tour videos are relatively more standardized, typically recorded with chest-mounted or dashboard cameras, documenting first-person point of view (POV) walking, cycling, or driving experiences through urban areas. Their growing popularity and broad geographic coverage have increasingly highlighted their potential as a source for urban sensing (Hartmann and Purves, 2023).

Based on the characteristics and applications of the data forms outlined above, and with the aim of enabling a more systematic investigation of this emerging class of data, we summarize their shared features and introduce the concept of Street View Video (SVV):

> Street View Video (SVV) is a geo-locatable video collected at or near street-level, which provides observations of the urban environment, sometimes including sound.

Building on these commonalities, each data type also exhibits distinct advantages and limitations. Table 1 summarizes typical characteristics when aggregated at the scale of a single city. Stationary SVV types generally feature high temporal coverage but low spatial continuity, owing to their fixed, point-based deployment. Surveillance networks broaden spatial coverage by observing many sites at once, yet these sites remain discrete, so continuity stays low. On the other hand, these sources score high on standardization, offering stable quality, fixed perspectives, and precise geolocation. Such properties make them suitable for longitudinal urban studies at specific sites. In contrast, mobile video data are more heterogeneous. Depending on the data availability and research objectives, both dashcam footage and mobile sensors

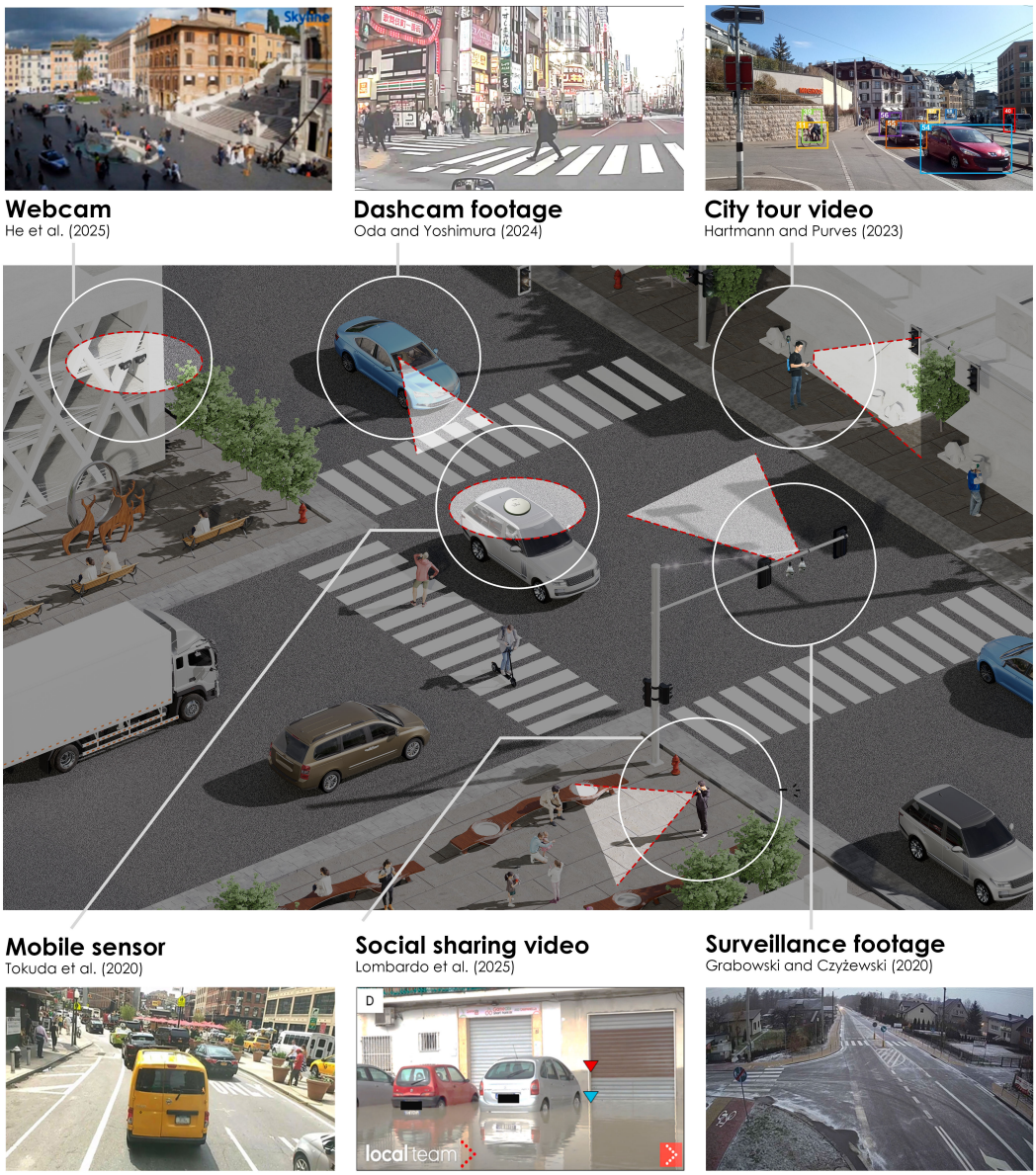


Figure 1: An illustration of six SVV types within an urban environment, accompanied by representative example images.

can yield varying degrees of coverage, although the latter is usually of higher quality, since research-driven collection allows controlled acquisition and calibrated equipment. Both often include GPS information and support fine spatio-temporal scales, enabling research on traffic, pedestrians, and urban environments from a road-level perspective. Social sharing videos and city tour videos stand out for their broad accessibility due to their crowdsourced nature and the frequent inclusion of sound, but typically exhibit the lowest consistency and usually carry coarse geolocation (e.g., city or street level), requiring additional inference for finer positioning. Between the two, city tour videos offer relatively higher coverage, continuity, and quality as they more often originate from video-sharing platforms and are typically produced with better recording equipment as well as a stronger tourism-oriented framing.

Table 1: Comparison of six SVV types across key characteristics when aggregated at the scale of a single city.

Street View Video						
Characteristic	**Stationary**		**Mobile**			
	Webcams	**Surveillance footage**	**Dashcam footage**	**Mobile sensors**	**Social sharing video**	**City tour video**
Spatial coverage	Medium	High	Data-dependent	Data-dependent	Low	Medium
Temporal coverage	High	High	Data-dependent	Data-dependent	Low	Medium
Spatial continuity	Low	Low	High	High	Data-dependent	High
Quality	Medium	High	Medium	High	Low	Medium
Geolocation	Explicit	Explicit	Explicit	Explicit	Implicit	Implicit
Perspective	Overhead	Overhead	Road	Road	Any	Walkway/Road
Sound	No	No	No	Data-dependent	Yes	Yes
Typical source	Commercial/Institutional	Institutional	Institutional	Commercial/Researcher	Crowdsourced	Crowdsourced

For the study of urban vitality, we adopt city tour videos as our exemplar SVV, motivated not only by their capacity to capture urban sound, but more importantly, by their broad availability and widespread popularity across urban centers.

4. Methodology

Based on the city tour video type, we propose an automated and scalable research framework for measuring pedestrian activity and sound environments to represent urban vitality, as illustrated in Figure 2. This framework includes data collection (Step A), geo-localization (Step B), visual-auditory assessments (Step C), and spatial-temporal analysis (Step D).

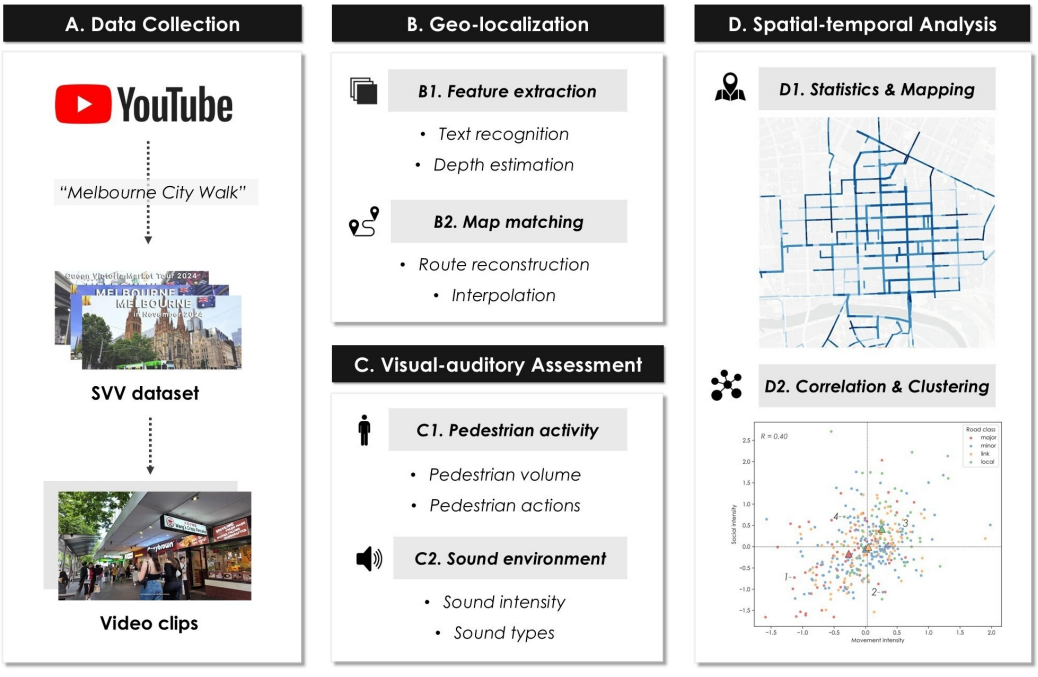


Figure 2: A framework for data downloading, geo-locating, visual-auditory assessing, and spatial-temporal analyzing based on city tour videos.

4.1. Data collection

For the empirical investigation, we choose Melbourne, Australia, as the case study area for two reasons. First, the city provides a rich availability of SVV data suitable for our analysis, based on the preliminary check. Second, for one of the key tasks in this study — pedestrian detection, the city offers official sensor data from the Pedestrian Counting System of the City of Melbourne (https://www.pedestrian.melbourne.vic.gov.au/), which operates 141 sensors across the city. From these, we select 38 outdoor sensors covering the central business district (CBD) for use in our comparative analysis. The specific location of the main study area, along with the corresponding street names and sensor locations, is shown in Figure 3.

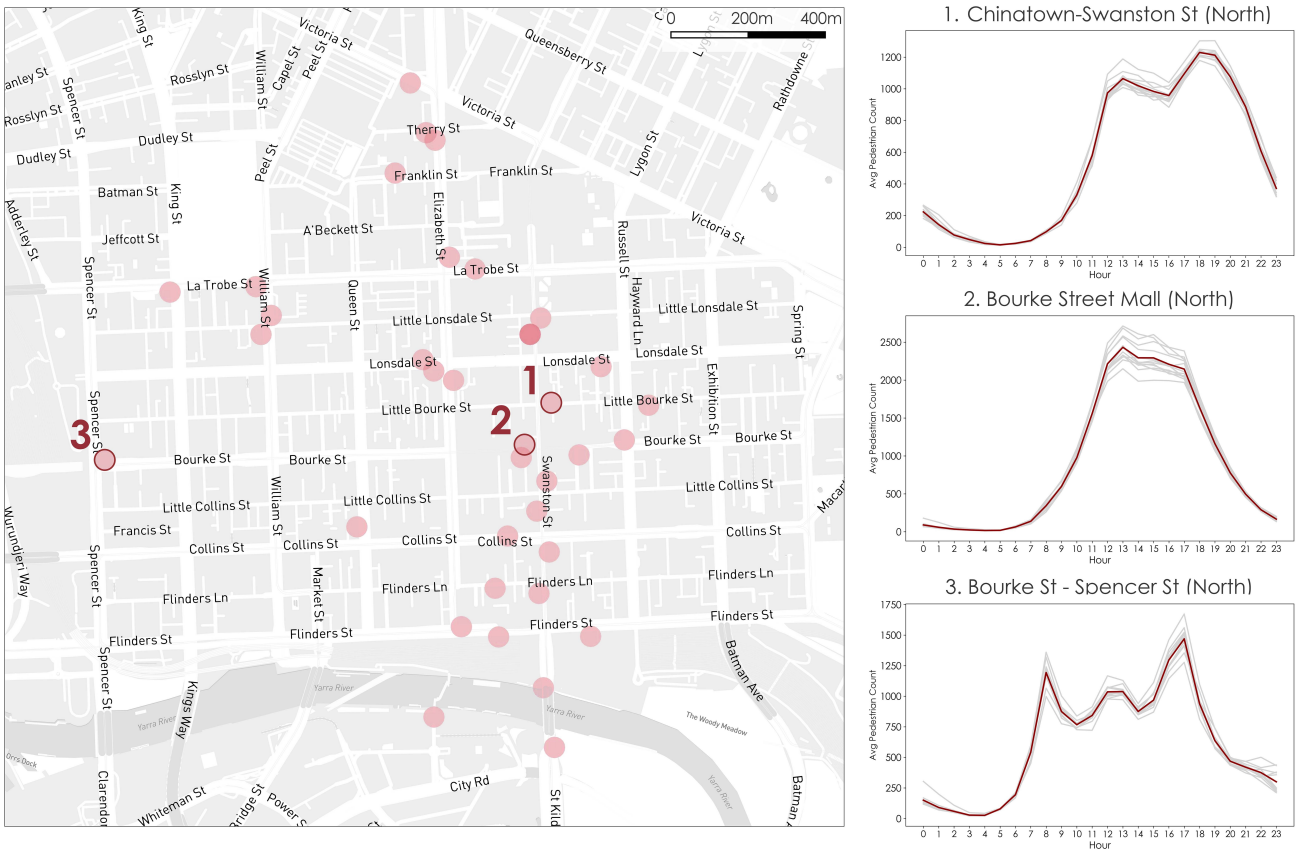


Figure 3: Study area in Melbourne's CBD, including major street names and the locations of the selected pedestrian sensors, along with three example sensor pedestrian count profiles across the entire 2024, with the red line indicating the annual hourly average. Base map © Mapbox and OpenStreetMap contributors.

We implement a systematic data collection protocol utilizing the Python package yt-dlp to retrieve relevant videos from YouTube, one of the largest video platforms with the most available videos worldwide (yt dlp, 2024). The search phrase 'Melbourne city walk' is used because such queries are commonly employed by content creators to document walking routes through city centers. Only videos with a duration exceeding 10 minutes are retained to ensure sufficient spatial coverage along urban streets. Videos are downloaded in 1080p resolution for analysis, and video clips are extracted at 10-second intervals for further assessment. Considering that such videos often include introductory preview segments (e.g., channel introductions

or highlight clips), the first 90 seconds of each video are skipped to ensure continuity in the sampled content.

4.2. Geo-localization

Our geo-localization framework consists of two components: feature extraction (Figure 4 Step (1)) and map matching (Figure 4 Step (2)), which together enable the inference of specific locations for each image frame. We then validate localization accuracy by comparing our results with YouTube timestamps, which include the temporally ordered street names, using the Longest Common Subsequence metric.

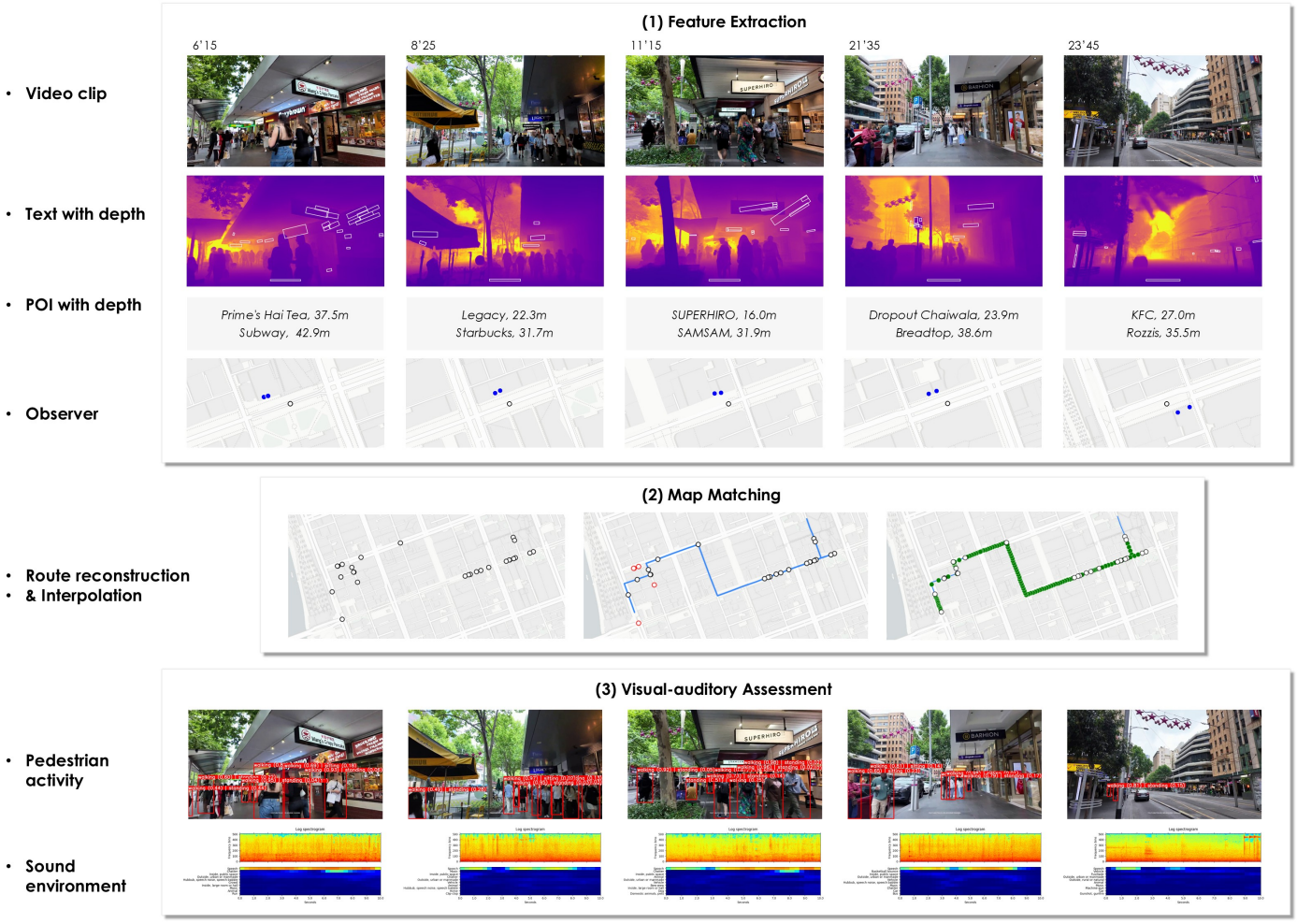


Figure 4: Example workflow from a single street view video, illustrating three stages: (1) feature extraction of textual and depth information for localization and observer position estimation, (2) map matching to align estimated observers with the street network, and (3) visual–auditory assessment integrating pedestrian activity and sound environment analysis. Video frames © respective YouTube video creators.

4.2.1. Feature extraction

We extract image frames at 10-second intervals to match the duration of each video clip. As this study focuses on walking behaviour, we first apply a platform classifier to each image

frame to identify videos that were likely recorded while driving or cycling; such videos are removed entirely from the dataset (Hou et al., 2024).

For the remaining frames, two types of information are derived: textual features and depth features. The former is obtained using the PaddleOCR v5 Python package, which outputs recognized text and the coordinates of its corresponding bounding boxes (Cui et al., 2025). The latter is generated using the Depth Anything v2 package, specifically its *metric depth estimation* function, which estimates the distance of each pixel from the camera in meters (Yang et al., 2024). This model is trained on the Virtual KITTI dataset and further fine-tuned on the pre-trained encoder (Cabon et al., 2020). The dataset contains synthetic replicas of real-world scenes under diverse illumination and weather conditions (e.g., rain, fog, and sunset), which helps the model generalize to a wide range of environmental variations commonly encountered in SVV data.

Because the approach relies on textual cues embedded in the extracted image frame, it remains robust to variations in original recording conditions, including differences in walking speed, camera height, and camera orientation. Provided that geographic text is visible and the video resolution is sufficiently clear (1080p in this study), reliable spatial information can be extracted. The OCR outputs are then combined with depth estimates to compute the average depth of each text bounding box (see Figure 4 Step (1)).

4.2.2. Map matching

In the subsequent stage, we first focus on single-frame analysis. Texts identified in each frame are matched against entries in the local OSM database using *Levenshtein distance*, a string similarity measure that computes the minimum number of character edits required to transform one word into another. To account for variations such as spelling errors, abbreviations, or partial captures, we adopt a fuzzy matching approach. A threshold of 70% similarity is applied, meaning that candidate matches are retained only if their textual similarity score reaches or exceeds this value. Next, to address the challenge of chain stores appearing in multiple locations, we employ a local searching radius of 30 meters. A successful recognition requires at least two points of interest (POIs) to be simultaneously identified within this radius, while frames with only one or no matched POIs are discarded. For frames with multiple recognized POIs, we incorporate the average depth values obtained earlier. By comparing the Euclidean distances between the observer and different POIs, we approximate the observer's position on the road network, using network data processed with the Python package OSMnx (Boeing, 2017). It is important to note that in many cities, bidirectional roads in OSM are represented as two separate segments in opposite directions. For simplicity, we merge these into a single line.

We then extend this single-frame analysis to all frames in a video, generating all retained observation points with their corresponding order of appearance. For trajectory reconstruction, we employ the GoTrackIt Python package, which uses the processed road network and the ordered estimated observer positions to reconstruct the original walking trajectory while filtering outliers (zdsjjtTLG, 2024). Because the reconstruction follows the temporal order of observations, it can reproduce the original walking path even when recording speeds vary across segments, and in rare cases it can capture situations where the creator walks back and forth along the same street.

As this recognition method relies on visible text, missing points naturally occur in areas without explicit textual information. To address this, we treat the recognized points as anchors and interpolate missing points uniformly along the reconstructed trajectory between the

first and last recognized points (see Figure 4 Step (2)). Since the extracted frame interval is consistent with each video clip's duration, we can accurately map each video segment to its corresponding road segment. This interpolation also has the added benefit of adapting the spacing of interpolated points to the distance between anchors, thereby better reflecting speed variations of the video creator along different street segments.

4.2.3. Validation of localization

Finally, we validate the localization results. For some YouTube videos, creators provide additional metadata such as filming date, temperature, and, crucially, timestamps (e.g., *[01:19] Spencer Street*). Given that such videos are typically recorded continuously while walking through the city, connecting the street names in the timestamps yields the ground-truth trajectory. We extract and standardize all such timestamp descriptions and match the contained street names against our localized street names using the aforementioned *Levenshtein distance*. As the order of street appearances is critical for validity, we further compare the two sequences using the *Longest Common Subsequence (LCS)* method. The LCS identifies the longest ordered subsequence of matching street names between the two sequences, where two names are considered identical if their fuzzy similarity exceeds a predefined threshold. This ensures that the comparison captures not only the presence of correct street names but also their correct relative order along the walking trajectory. Formally, we define the localization accuracy as:

$$\text{Accuracy} = \frac{|LCS(\text{GroundTruth}, \text{Recognized})|}{|\text{Recognized}|}$$

where $|LCS(\text{GroundTruth}, \text{Recognized})|$ denotes the length of the longest common subsequence between the YouTube timestamp descriptions and the output of our localization method, and |Recognized| is the total number of recognized street names.

This ratio reflects the proportion of the recognized trajectory that is consistent with the ground truth. As LCS captures only the relative order and not necessarily contiguous order, this measure also tolerates occasional missing or additional timestamp entries(e.g., "video introduction", "street corner", "intersection").

4.3. Visual-auditory assessments

Following video geo-localization, we conduct visual-auditory tasks (Figure 4 Step (3)) to derive indicators of pedestrian activity and sound environments for measuring urban vitality. This study focuses exclusively on pedestrians, as they represent the most prevalent form of street-level human activity and have been widely adopted as a key proxy for vitality (Whyte, 1980; Yin et al., 2015; Chen et al., 2020). Other forms of street use, such as cycling, e-scooter riding, and other micromobility activities, are not considered in the present study but can be incorporated in future work.

4.3.1. Pedestrian activity

Pedestrian activity is represented by the total count of pedestrians and their associated actions identified in the videos. To achieve this, we employ YOLOv11 in combination with an X-CLIP transformer for zero-shot pedestrian action recognition, while BoT-SORT serves as the tracker to enable cross-frame ID reassignment (Jocher and Qiu, 2024; Aharon et al., 2022). Regarding action labels, we follow Gehl, who identifies walking, standing, sitting, and playing as the primary activities in public space (Gehl and Svarre, 2013). Given the diversity and ambiguity

of playing behaviours, we focus on the three most fundamental and observable categories: walking, standing, and sitting. Each pedestrian is assigned to one of these categories, and the sum of all three yields the observed total count. Inference is performed on every 10-second video clip, with key parameter settings including skipping two frames between detections to accelerate inference, a confidence threshold of 0.3, and an IoU threshold of 0.5 to filter low-quality detections.

To assess the reliability of the pipeline, 100 ten-second video clips are randomly sampled from the dataset and manually annotated for the count of pedestrians in each action category. Model predictions are then compared against these ground-truth annotations using bias, MAE, and Pearson's r, with results summarized in Table 2. The model performs well for walking and sitting ($r = 0.88$ and $r = 0.88$, respectively), and the total pedestrian count likewise shows strong agreement ($r = 0.90$, MAE = 2.79). Standing is underpredicted (bias = -1.35, $r = 0.35$), likely owing to visual ambiguity between stationary postures and partial occlusions, which the classifier tends to resolve in favor of walking. Overall, the model can robustly capture pedestrian behaviors, whilst standing behavior may require more cautious interpretation, given its underprediction.

Table 2: Validation of the action recognition pipeline. Bias = mean(pred − GT).

Action	**GT Mean**	**Pred Mean**	**Bias**	**MAE**	**Pearson r**
Walking	20.57	22.60	+2.03	2.31	0.88^{***}
Standing	2.05	0.71	−1.35	1.41	0.35^{**}
Sitting	1.21	1.14	−0.07	0.48	0.88^{***}
Total	23.83	24.45	+0.62	2.79	0.90^{***}

$^{**}p < 0.01$; $^{***}p < 0.001$.

Furthermore, we conduct an exploratory analysis by comparing pedestrian counts derived from three available data sources to understand the extent to which they agree with each other: (i) pedestrian counts recorded by the Melbourne Pedestrian Counting System, (ii) object detection applied to SVIs, and (iii) video-based recognition from SVVs. For the sensor data, we select outdoor sensors located in high-sample SVV areas and aggregate their counts into daytime (06:00 to 18:00) and nighttime (18:00 to 06:00) categories. Each sensor point is assigned a 30 m buffer and spatially joined with the street network. For the SVI data, we download both commercial street imagery from Google Street View (GSV) and crowdsourced imagery from Mapillary across the study area at 50 m intervals (see Appendix Figure A.13 for the spatial distribution of street view data). We then use the Python package ZenSVI to apply the built-in object detection model GroundingDINO to identify persons (Ito et al., 2025; Liu et al., 2024). In addition, considering the multi-temporal nature of crowdsourced imagery, the built-in lighting condition classifier (Hou et al., 2024) is used to categorize images into day, dawn/dusk, or night conditions. Due to the typically limited SVI sample size at night (Fan and Biljecki, 2024), after a preliminary inspection, we restrict comparisons to the daytime results across both SVI datasets. The specific results of these comparisons are discussed in the Section 6.2.

4.3.2. Sound environment

For the sound environment, we use both the overall sound intensity and the distribution of different sound types as indicators. The sound intensity is estimated by computing the root mean square (RMS) amplitude of the waveform and converting it into decibels (dB) as

$$\mathrm{dB} = 20 \cdot \log_{10}(\mathrm{RMS} + \epsilon),$$

where ϵ is a small constant to prevent numerical instability.

For the sound type classification, we adopt the Python package PANNs (Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition). This CNN14 model is pretrained on the AudioSet dataset and achieves a high mean average precision (mAP) of 0.439 on this benchmark (Kong et al., 2020; Gemmeke et al., 2017). As AudioSet contains 527 sound event labels covering diverse environmental and urban sounds (e.g., speech, traffic, and music), the pretrained model can capture major acoustic patterns in street environments. We therefore apply it for inference and retain the three most prominent predicted labels for each audio segment. Following the same procedure as in the previous step, we extract audio in WAV format from the video clips and perform inference on the pretrained model.

Similarly, 100 ten-second audio clips are randomly sampled from the dataset and manually annotated for sound types. Model predictions are evaluated against these ground-truth labels using Top-1 and Top-3 accuracy, and per-class Precision, Recall, and F1, with results summarized in Table 3. The model achieves a Top-1 accuracy of 88.7% and Top-3 accuracy of 98.4%, alongside a macro F1 of 0.83, demonstrating strong overall classification performance. Amongst the three main urban sound categories, Speech attains the highest F1 (= 0.94, recall = 0.97), whilst Music ($F1$ = 0.80) and Vehicle ($F1$ = 0.86) are also reliably identified, showing the model can robustly capture the dominant sound environment of street-level clips.

Table 3: Validation of the sound type classification pipeline.

Class	**Precision**	**Recall**	**F1**	**Support**
Music	0.89	0.73	0.80	11
Speech	0.90	0.97	0.94	37
Vehicle	0.86	0.86	0.86	7
Macro avg	0.88	0.80	0.83	89

Top-1 accuracy = 88.7%; Top-3 accuracy = 98.4%; Metrics are reported for the three dominant classes; macro average is calculated across all classes.

4.3.3. Video quality control

Based on the geo-localization results and subsequent analyses, we conduct video quality control to retain only high-quality videos.

Specifically, using the identified geo-located points, we calculate the distance between consecutive locations. Considering potential positional uncertainties in POI locations, we adopt a relatively tolerant threshold of 50 m per frame (equivalent to 5 m per second). When the distance between two consecutive points exceeds this threshold after accounting for the temporal interval, it likely indicates a failure of the geo-localization method or discontinuities in the video (e.g., jump cuts introduced during editing). Such videos are therefore removed from the dataset.

In addition, we use the sound classification results to detect videos containing commentary or background music. Labels with probabilities greater than 0.3 are considered present. We then examine the overall proportion of the human and music labels within each video. Based on manual inspection, a threshold of 0.8 is adopted, and videos exceeding this threshold are removed. It should be noted that this filtering strategy prioritizes precision over recall, and

thus may also exclude cases where substantial human speech or music is genuinely present in the real-world environment.

All the aforementioned video filtering criteria we apply in this study are summarized in Table 4. We acknowledge that these parameters may vary across regions and video quality conditions. Therefore, we have made our code flexible, allowing users to adjust the thresholds to accommodate different urban contexts and research scopes.

Table 4: Systematic filtering criteria for constructing the SVV dataset.

Feature	Description	Selection rule
Downloading		
Video topic	Ensure contextual relevance	Query: 'Melbourne City Walk'
Video duration	Ensure sufficient coverage	Duration > 10 minutes
Quality control		
Video platform	Exclude non-pedestrian modes	Remove videos labeled as 'driving' or 'cycling'
Spatial continuity	Ensure continuous trajectories	Skip the first 90 seconds; Remove videos with gaps > 50 m between consecutive geo-localized points
Soundtrack	Ensure the ambient audio quality	Remove videos with speech/music proportion > 0.8

4.4. Spatial-temporal analysis

Based on our geo-localization results, we conduct a spatio-temporal analysis to examine the fundamental distribution patterns of the videos, as well as the spatio-temporal trends of visual and auditory dimensions. We further employ correlation and quadrant analyses to assess the relationships across different dimensions of urban vitality as well as the overall characteristics exhibited by the streets.

We first analyze basic video analytics, including the average video length and the length of the walking routes identified. We divide each road into equal 50-meter segments and summarize the frequency with which each road segment appears across all videos in order to assess the coverage of these videos within the city and the overall degree of distribution evenness.

Next, based on the same spatial units, we aggregate the results of each video clip obtained in the visual-auditory assessments. For each metric, we compute the mean to capture the average level. Furthermore, we classify the videos into daytime and nighttime categories based on the original video frames. Specifically, each frame is classified using the same lighting condition classifier described above, which assigns one of three labels: day, dawn/dusk, or night. For each video, we determine the modal label across all frames and use it as the representative lighting condition of the video. In the final categorization, both dawn/dusk and night are grouped into the nighttime category. For simplicity, we only aggregate the overall pedestrian count and sound intensity into daytime and nighttime measures, while other detailed variables, such as different pedestrian actions and sound types, are presented at the full-day level.

Finally, we jointly consider the visual and auditory dimensions. We use Pearson's correlation to assess overall consistency among variables and apply quadrant analysis to capture the integrated characteristics of streets. Following Gehl (2011)'s framework of public life in street spaces, we classify human activities on streets into movement and social interaction dimensions. Visual and auditory variables are then assigned to these two dimensions according to their behavioral relevance. We then measure the behavioral profile of each street segment and examine how different road types are positioned within the two-dimensional activity space.

5. Results

5.1. Geo-localization and quality control

Based on the search results, as of February 2026, we obtained a total of 1,602 valid videos, each exceeding 10 minutes in duration. After removing videos likely recorded from driving or cycling platforms, we applied the geo-localization procedure and successfully located 1,029 videos.

We conducted video quality screening to identify potential issues from the located videos, including geo-localization errors, video discontinuities, and problematic background audio. After removing videos with such issues, 892 videos remained, with a total duration exceeding 32,570 minutes, including 801 recorded during the day and 91 at night. To further validate the utility of the quality assessment framework, we manually inspected the excluded videos, the vast majority of which were confirmed to contain the aforementioned issues. Furthermore, to assess the quality of the final dataset, we randomly sampled 80 videos (approximately 10% of the dataset) for manual inspection, achieving a pass rate of 94% (75/80). The full list of SVVs, the list of videos flagged by the quality assessment pipeline, and the subset used for manual validation are provided in the Code Availability section.

Based on the final SVV list, we then assessed the accuracy of geo-localization by comparing the identified routes with video description data using the Longest Common Subsequence (LCS) method. Using a fuzzy text-matching threshold of 70%, the proportion of successfully matched routes reached 77.19%, indicating a relatively high level of accuracy. This comparison was limited to street segments that were successfully identified, ensuring the accuracy of recognized segments, while those present in the videos but not detected by our method were excluded from the analysis.

We further examined the applicability of the geo-localization method across diverse scenarios within the dataset. Through manual inspection, we identified six challenging scenarios, categorized into three critical failure cases and three moderate environmental variations, as illustrated in Figure 5. Coarse geo-information represents the most prevalent cause of identification failure: since walking routes do not always traverse areas with high-density textual geographic cues, videos predominantly recorded in settings such as urban parks or riverside paths present significant challenges for recognition. Low visual quality issues arise primarily from a small proportion of videos exhibiting severe perspective distortion, blurriness, or excessive camera shake. POI mismatch occurs when discrepancies exist between POI names in the video and those in the reference database, due to temporal gaps or incomplete updates, potentially causing systematic shifts or failures in the results. Moderate environmental variations, by contrast, encompass scenarios that may affect geo-localization but are generally handled with reasonable robustness. As a cultural product, a considerable number of videos capture crowded social events featuring potentially dense human activity; in most cases, successful localization is maintained provided that textual geographic cues are not substantially occluded. Regarding low-light conditions, particularly during nighttime, videos are typically recorded at a sufficient lighting level to remain watchable, thus preserving performance. We also examined the distribution of daytime and nighttime videos in the original list (1,412 and 190, respectively); the proportions remain broadly consistent with those in the final dataset, suggesting that the geo-localization approach is relatively robust to variations between day and night scenes. Adverse weather conditions are rare in the dataset, with heavy rain representing the most notable case, however, effective geo-localization can generally still be achieved under such circumstances.

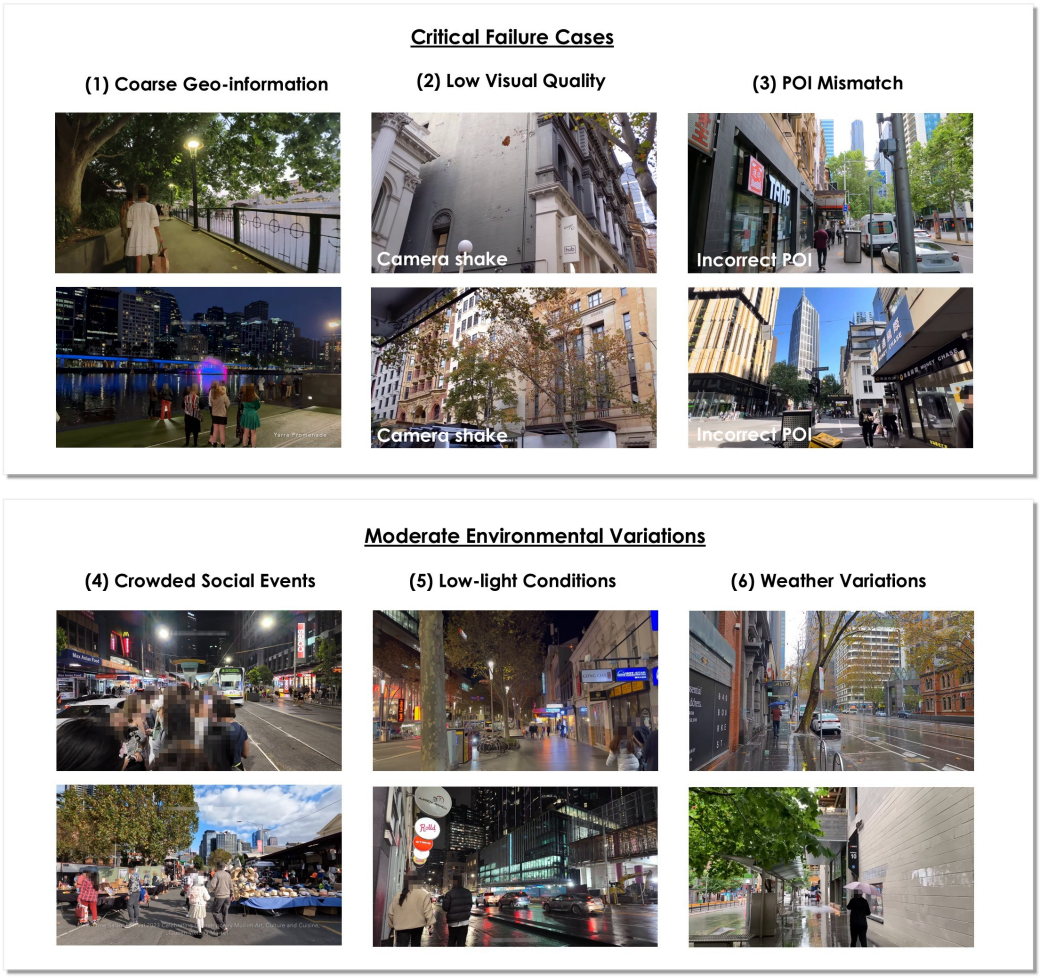


Figure 5: Illustrative examples of the six challenging scenarios identified during geo-localization evaluation, grouped into Critical Failure Cases and Moderate Environmental Variations. Video frames © respective YouTube video creators.

5.2. Video statistics

On the basis of reliable localization, we further compared the identified video duration and route length between daytime and nighttime conditions (Figure 6). Specifically, daytime videos exhibit significantly longer average durations than nighttime videos (907.0 s vs. 664.4 s), which correspond to longer average identified route lengths (823.2 m vs. 497.6 m). These differences are also reflected in higher estimated walking speeds during the daytime (0.90 vs. 0.75 m/s).

To examine spatial distribution evenness, we introduced the indicator *video presence*, defined as the cumulative number of times each street segment appeared across all localized videos (Figure 7). The results revealed highly uneven spatial preferences. For example, segments along Swanston St, Elizabeth St, and Bourke St appeared 125-254 times, accounting for one-fifth of all identified videos. Although nighttime subsets showed a limited sample size compared to their daytime counterparts, both exhibited similar patterns, likely reflecting the popularity of these well-known streets among video creators, which in turn further amplifies

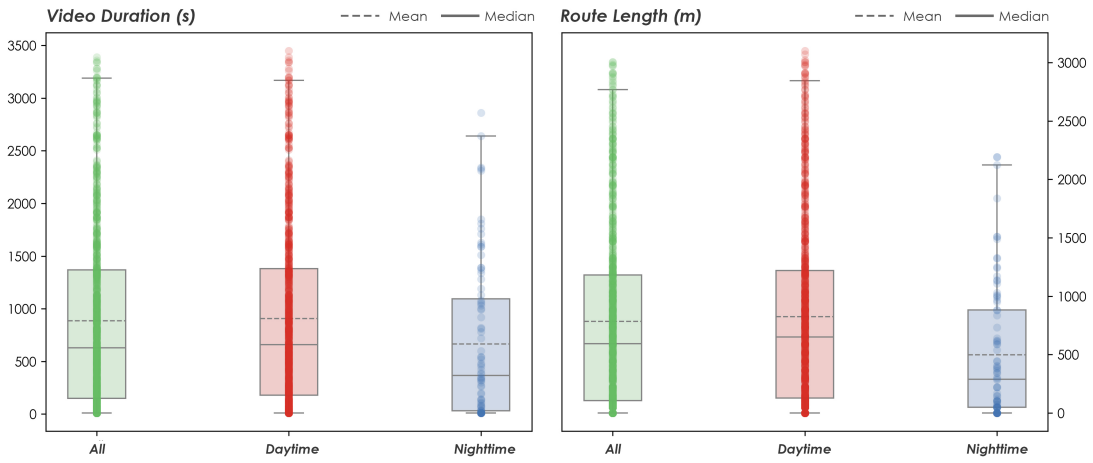


Figure 6: Distribution of video duration and route lengths across all, daytime, and nighttime SVV samples. Outliers in each subset are excluded based on the 1.5×IQR criterion.

their visibility online. By contrast, many street segments had very limited coverage (1–3 appearances), meaning the information they carry is more susceptible to random variation such as time or weather conditions. In subsequent visualizations, such low-sample streets were rendered with thinner widths to distinguish them from well-documented streets, with the legend placed in the upper-left corner to clarify this representation. Interpretive conclusions in this study are drawn primarily from streets with sufficient coverage; low-sample segments are retained for spatial completeness and should be treated as indicative rather than conclusive.

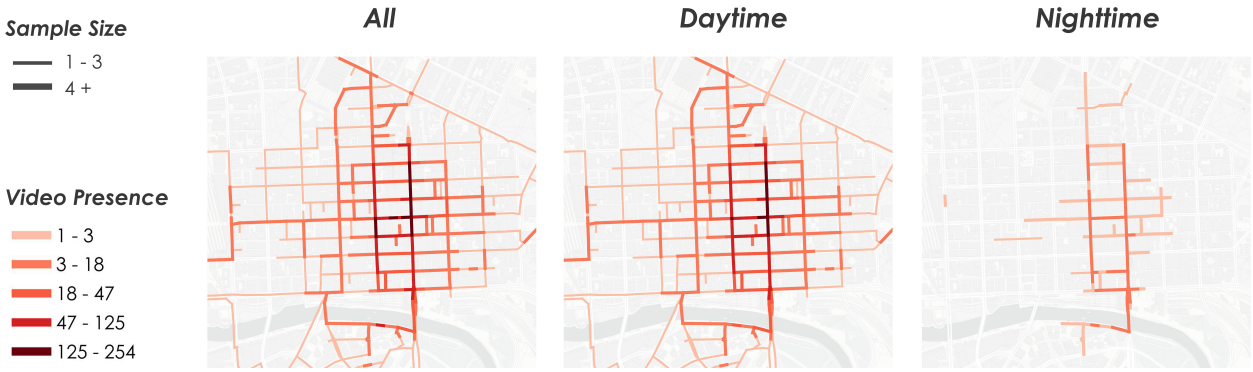


Figure 7: Total count of each street segment identified across all, daytime, and nighttime SVV samples.

5.3. Pedestrian counting and actions

In the first analytic task, we evaluated pedestrian presence across the street segments, further segregated by time of day and specific actions (Figure 8). As expected, popular streets with higher video presence (e.g., Swanston St, Elizabeth St, Bourke St, Little Bourke St, Queen St, Little Collins St, and Southbank Promenade) (Figure 7) also exhibited the highest pedestrian counts, with a maximum of 59 people per 50m segment (averaging over 1 person/m) (Figure 8). Daytime and full-sample trends were similar, while nighttime pedestrians were

particularly concentrated along Southbank Promenade and Bourke St, with levels significantly higher than during the daytime.

We further analyzed three pedestrian actions: walking, standing, and sitting (Figure 8). Walking dominated with over 80% of detected activity, explaining its close alignment with total pedestrian counts. Many high-standing counts clustered at intersections, consistent with waiting at traffic lights. Sitting counts revealed interesting patterns: certain higher values corresponded to segments with ample public seating (e.g., Hardware Ln, parts of Flinders St and Little Bourke St), while high values in areas with few facilities (e.g., Little Collins St) likely reflected errors.



Figure 8: Spatial distribution of pedestrian counts across all, daytime, and nighttime SVV samples, as well as counts for walking, standing, and sitting activities.

5.4. Sound intensity and types

We then analyzed audio intensity and dominant sound types in each clip (Figure 9). Peak sound levels were observed on major vehicular roads (e.g., Bourke St, Spencer St, Flinders St, and Queen St) as well as narrow pedestrian streets (e.g., Little Bourke St, Hardware Ln). In pedestrian-heavy areas, intensity was also elevated. Day–night comparisons revealed distinct patterns: nighttime sound levels were significantly higher along Southbank Promenade and Bourke St, suggesting denser commuting or social activity. The three dominant sound types were speech, vehicle, and music. High speech probability overlapped with some areas of high pedestrian activity (e.g., Bourke St, Queen St, Flinders St, Swanston St, Southbank Promenade). Vehicle noise was concentrated along Collins St, Spencer St, the west part of Bourke

St, and around Southern Cross Station. Interestingly, music intensity was highest not only along known busking areas (Bourke St) but also along Little Bourke St, and Little Lonsdale St, revealing potential new insights into busking hotspots.

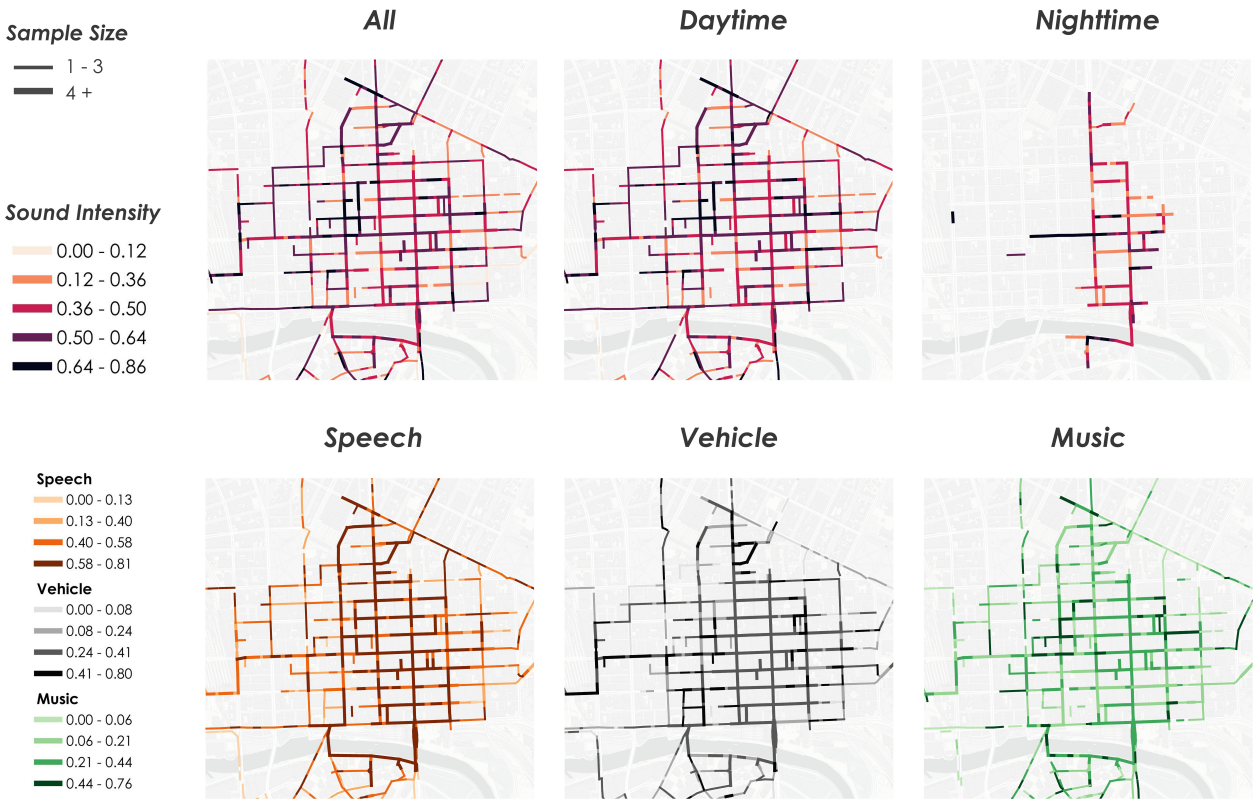


Figure 9: Spatial distribution of sound intensity across all, daytime, and nighttime SVV samples, as well as probability for sound types speech, vehicle, and music.

5.5. Correlation and quadrant analysis

Correlation analysis was conducted across 12 variables, grouped into (i) aggregate measures (pedestrian count, sound intensity) and (ii) type measures (pedestrian actions and sound categories) (Figure 10). For the first group, the daytime pedestrian count correlated strongly with the total pedestrian count (R = 0.84), and the daytime sound intensity correlated strongly with the average sound intensity (R = 0.88). Interestingly, no significant correlation was shown between daytime and nighttime totals. Moreover, a negative relationship was observed between daytime pedestrian counts and nighttime sound intensity, reflecting distinct temporal usage patterns and potential sampling bias in limited nighttime data. Pedestrian counts and sound intensity were weakly correlated ($R = 0.15$), with stronger relationships at night ($R = 0.50$) than during the day ($R = 0.13$), suggesting pedestrian counts might be a better proxy for sound intensity at night. Within the second group, clear positive correlations were observed among pedestrian behaviors (walking, standing, sitting), whereas no strong correlations are found among sound categories (speech, music, vehicle) except for a weak positive correlation between speech and music ($R = 0.16$). Cross-group correlations showed that walking correlated positively with speech ($R = 0.43$), while vehicle noise correlated negatively with all pedestrian

behaviors, reflecting reduced pedestrian presence in vehicle-dominated spaces.

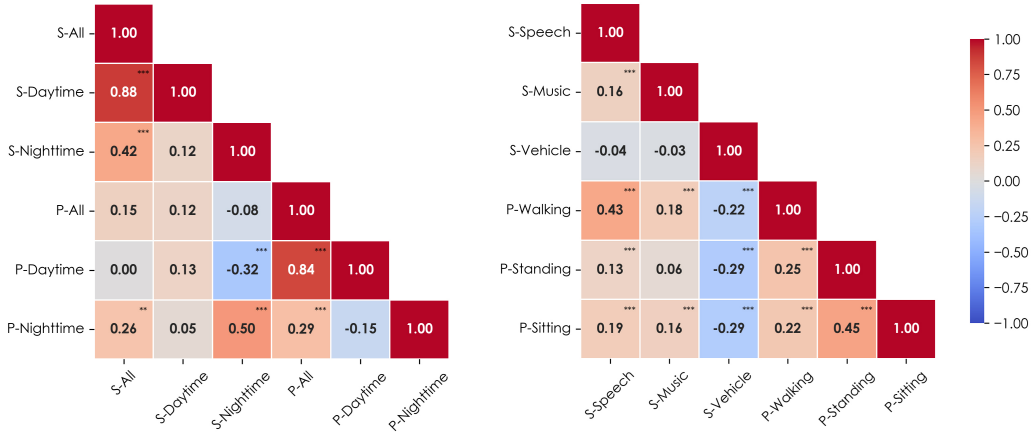


Figure 10: Correlation analysis of 12 variables. *P* in the two groups represents the overall pedestrian count and counts of individual actions, while *S* on the left group denotes the overall sound intensity and on the right group indicates the probability of each sound type. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Finally, to capture integrated street-level characteristics, we performed quadrant analysis. We focused on the three pedestrian action types and three sound categories, and organized them into two behavioral dimensions. The movement dimension represented necessary activities in Gehl (2011)'s framework, while the social interaction dimension captures optional and social activities. Accordingly, walking and standing were assigned to the movement dimension, whereas sitting action, speech, and music sound classes were assigned to the social interaction dimension. As vehicle sounds were negatively associated with pedestrian activities in previous analyses, they were incorporated as an inverse factor within the social interaction dimension. We standardized all variables using Z-scores to capture their deviation from the citywide average. Vehicle sound values were inverted prior to aggregation given its negative influence. Based on these normalized features, we computed the unweighted mean of the corresponding variables to construct the two behavioral scores for each road segment.

The overall scatter distribution showed a moderate positive correlation ($R = 0.40$) (Figure 11), suggesting that movement and social activities tended to co-occur to some extent within street environments. All street segments were then classified into four road hierarchy categories: major, minor, link, and local (see Table 5 for the classification details). For each road class, we calculated the mean center along the two behavioral dimensions and marked its position within the behavioral space (Figure 11); we further computed the mean and standard deviation of distances of the points (each representing a street segment) from their respective mean centers to quantify their dispersion from typical representations (Table 5).

The results in Figure 11 showed that the mean center of all urban streets, represented by the intersection of the dashed lines, was located close to the origin, suggesting an overall balance between the two dimensions. Major streets and local streets were scattered towards opposite directions in the behavioral space, with their mean centers located in the lower-left and upper-right quadrants, respectively. This finding suggests that, on average, major streets are associated with lower levels of pedestrian-related activities and urban vitality, while local streets tend to exhibit higher levels of both movement and social activities. Interestingly, these two categories also displayed the largest behavioral dispersion (higher mean and standard deviation distances; Table 5), indicating the presence of diverse activity patterns within each group. For example, major streets may include both vehicle-dominated corridors (e.g.,

Victoria St) and commercial streets (e.g., Flinders St). Similarly, local streets may range from quiet residential streets to socially active laneways such as Hardware Lane. By contrast, the mean centers of minor and link streets were located close to the origin of the behavioral space and showed relatively small dispersion, suggesting more balanced and consistent behavior patterns across the two dimensions, and comparatively neutral roles in supporting urban activities. Notably, only local streets exhibited positive and above-average values in social intensity spectrum (Figure 11), highlighting their primal role in facilitating public social interactions.

Further examination of within-category correlations between the two dimensions revealed additional differences (Table 5). Major streets showed the strongest correlation ($R = 0.51$), indicating that movement and social activities tended to vary in a more coordinated manner. In contrast, local streets showed the weakest correlation ($R = 0.21$), reflecting greater heterogeneity, where some streets were dominated by movement activities while others supported social interaction.

Overall, these findings suggest that, to a limited extent, the road hierarchy classification could explain the variations among streets in their potential to facilitate pedestrian movement and social interactions, as shown by the distribution pattern of the road class mean centers (Figure 11). Yet, road hierarchy alone is insufficient to clearly distinguish the differences — the majority of the points from different classes scatter across similar spaces (Figure 11). This result highlights the limitations of traditional road classifications in explaining street vitality, suggesting that future research should more explicitly consider functional and contextual characteristics of streets.

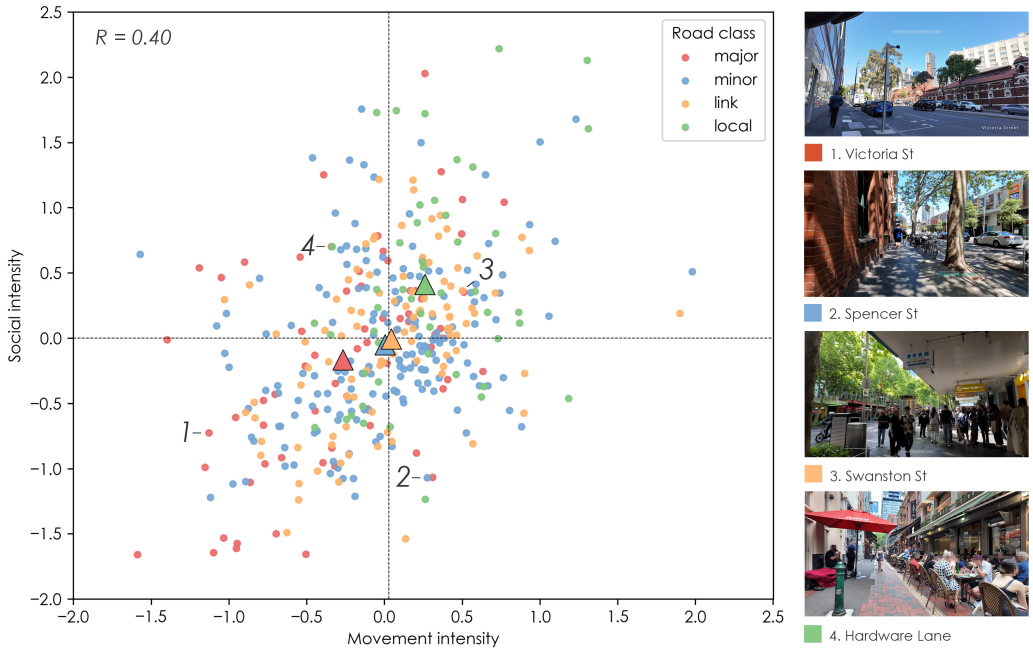


Figure 11: Street segments in the movement–social behavioral space, colored by road hierarchy. Dashed lines indicate citywide means and triangles mark class mean centers. Example streets from the four behavioral quadrants are shown on the right. Video frames © respective YouTube video creators.

Table 5: Behavioral dispersion and correlation between movement and social dimensions across road classes.

Road class	Mean distance from class center	Std. distance from class center	Pearson R	OSM road types
Major	0.851	0.463	0.513	motorway, trunk, primary
Minor	0.678	0.404	0.321	secondary
Link	0.672	0.368	0.404	tertiary
Local	0.810	0.520	0.211	residential, living street

6. Discussion

6.1. An automated multimodal approach for measuring urban vitality

This study investigates the feasibility of SVV for measuring urban vitality in city centers and reveals its spatial-temporal patterns in several important ways.

First, we propose a novel method that leverages city tour video data for high-precision geo-localization by extracting depth and text information, comparing them against the local OSM database, and cross-validating with video metadata. The method achieves a high level of accuracy, demonstrating the strong potential of SVV for downstream urban sensing applications. Its promise lies in its reliance on information already embedded in videos and openly available geospatial data, enabling geo-localization across virtually any urban area. Beyond this form, other types of SVVs, such as crowdsourced social sharing videos, could likewise be geolocated using similar cues. Although in this study we have not yet extended the method to other cities or SVV types, future research could pursue such directions.

Second, based on the geolocated videos, we develop an automated visual-auditory assessment pipeline for measuring urban vitality by extracting pedestrian activity and the sound environment information. Unlike previous studies that have examined these aspects largely in isolation, our study is the first to couple these modalities in a comprehensive framework. Thanks to the high precision of our geo-localization, these patterns can be mapped at fine-grained scales of 50 meters or less, offering micro-level observations — addressing another limitation of sparse granularity in the current vitality research (Doan and Zhang, 2025). The outcomes reveal not only patterns that confirm locally recognized knowledge, such as major pedestrian-friendly streets, but also new data-driven insights that shed light on less visible aspects of urban dynamics, such as pedestrian-sitting hotspots and busking-popular streets.

More specifically, building on the traditional emphasis on pedestrian volumes, our study highlights three complementary dimensions of urban vitality: day-night temporal variations, pedestrian actions, and sound information. These dimensions provide a richer perspective on how streets are used and experienced.

Temporal contrast between daytime and nighttime activities, well established in previous research, is reflected in our findings, which show that many streets exhibit markedly different activity patterns across time (Du et al., 2024; Zhang et al., 2025a; Güller and Varol, 2024). Our mapping (e.g., Figure 8 and Figure 9) identified distinctive street groups (e.g., Southbank Promenade and Bourke St) where pedestrian volume and sound intensity are substantially higher at night than during the day. Recognizing these patterns can inform urban management strategies: streets characterized by strong nighttime activity may benefit from interventions that support the nighttime economy, such as improved lighting and safety measures, whereas in residential areas, these patterns may signal the need for noise mitigation or temporal regulations.

Integrating pedestrian actions with urban sound signals further deepens our understanding of street vitality. Correlation analysis indicates that vehicle sounds are negatively associated with various walking activities, further supporting the long-held urban planning principle that

reducing traffic flow at the street level is critical for promoting walkability (Soto et al., 2022; Appleyard, 1980). Building on classic theories of street life, we organize these signals into two behavioral dimensions: movement and social interaction. Streets dominated by movement signals tend to function primarily as mobility corridors, whereas those characterized by stronger social signals are more likely to support lingering and informal social activities. This distinction provides a useful diagnostic lens for planners: movement-dominated corridors may benefit from traffic calming or widened sidewalks, while socially active streets may require preserving pedestrian priority and minimizing through-traffic.

More broadly, our computational framework suggests that urban vitality cannot be fully captured by static indicators such as road hierarchy alone, echoing long-standing debates in urban design, which highlight how traffic hierarchies tend to prioritize movement over place quality (Marshall, 2004). Although major and local streets differ in their average vitality levels, considerable behavioral diversity exists within each category, indicating that street vitality might also be shaped by micro-scale contextual factors such as building frontage activity and the perceptual qualities of street environments. Integrating visual and auditory signals may therefore better capture how urban streets are experienced in real life, complementing traditional planning indicators and informing more context-sensitive street interventions.

6.2. Comparing data sources for pedestrian activity

For pedestrian activity data derived from different sources, we compared the associations between measurements at both the street segment and sensor point levels (Figure 12).

At the segment level, we aggregated pedestrian observations from SVV and two SVI platforms (GSV and Mapillary) at the street scale. The values shown in Figure 12 are normalized pedestrian counts at each segment from each source. The map of raw pedestrian counts from GSV and Mapillary is available in Appendix A. To ensure consistency, all comparisons involving SVI were restricted to daytime conditions. Video-based observations exhibit moderate positive correlations with both SVI datasets (R = 0.28 for SVV–GSV and R = 0.26 for SVV–Mapillary), whereas the agreement between the two SVI sources is stronger (R = 0.51). Meanwhile, the mean pedestrian counts decreased across the three data sources (19.6, 5.8, and 2.0, respectively). This discrepancy may reflect differences in data acquisition perspective: SVI is typically from vehicle-based viewpoints, whereas SVV is recorded closer to pedestrian space at sidewalk-level perspectives.

At the sensor level, we selected 38 sensor locations that are located outdoors, have high data quality, and spatially overlap well with our study area in the city center. We used the annual average pedestrian counts recorded by the sensors in 2025 as the overall pedestrian activity level at each location. Daytime and nighttime were defined as 6:00–18:00 and 18:00–6:00, respectively. A 30 m buffer was created around each sensor, within which the corresponding SVV and SVI measurements are extracted for comparison. The results indicate that all three street-level datasets show a certain degree of agreement with the sensor measurements. Nighttime SVV and both SVI datasets show strong correlations with the sensors (Sensor–SVV NR = 0.60, Sensor–GSV R = 0.65, Sensor–Mapillary R = 0.56), whereas daytime SVV demonstrates only a moderate correlation (Sensor–SVV DR = 0.30), suggesting greater discrepancies under daytime conditions.

This analysis further illustrates the distinct characteristics of each data type, which we summarize in Table 6. Sensor data provide highly precise spatio-temporal measurements of pedestrian counts at fixed locations, but their continuity remains limited and such systems are not universally available across cities. SVI, by contrast, as the most stable data source, is

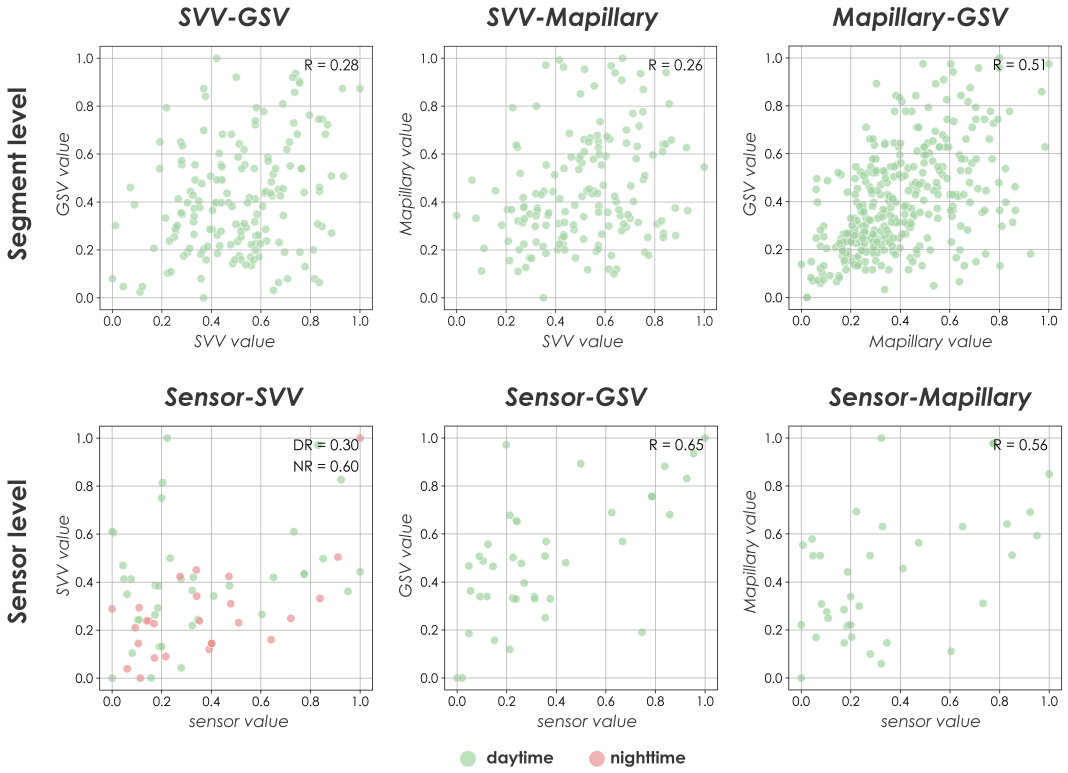


Figure 12: Comparison of normalized pedestrian count across sensors, SVV, and SVI. The SVI–SVV comparison is conducted at the street segment level using all available segments, while the sensor-based comparison is based on 38 sensor locations. Comparisons between sensors and SVV are further distinguished by daytime and nighttime conditions.

easily accessible nowadays and offers the broadest coverage, capturing both pedestrians and surrounding street objects (e.g., vehicles, vegetation). SVV, in comparison, captures pedestrian counts from a sidewalk-level perspective and therefore typically records more pedestrians. Compared to traditional SVI, it also offers broader temporal coverage (day and night). More importantly, it provides access to sound information and human behavioral dimensions, thereby enabling a richer understanding of street-level dynamics. These features also create new opportunities for studying walkability, comfort, and street-level experiences from perspectives that other data sources cannot provide. Nevertheless, SVV data are constrained by their reliance on user-generated content, which often results in uneven spatial and temporal distribution, and the lack of built-in geotagging for certain SVV types necessitates additional geo-localization procedures.

Overall, although pedestrian activity studies rarely have a true ground truth, both the commonly used data sources in previous studies and the data sources proposed in this work show a certain level of consistency in overall patterns, while still exhibiting noticeable discrepancies. These findings suggest that different data sources capture complementary aspects of pedestrian activity. Future research should therefore select data sources based on their specific analytical objectives, or ideally, combine multiple sensing modalities to enable more comprehensive and robust evaluations.

Table 6: Comparison of data sources for pedestrian activity measurement.

Characteristic	Sensor	SVI	SVV
Spatial continuity	Discrete, limited locations	Discrete, citywide	Continuous, popular areas
Temporal coverage	Continuous (24/7/365)	Static snapshot	Broad coverage (day/night)
Update frequency	Real-time	Low (months to years)	Irregular, user uploads
Sound information	–	–	✓
Human behaviour	–	–	✓
Other objects (e.g., cars)	–	✓	✓

6.3. Limitations and future work

In this study, we demonstrate that user-generated SVV presents a vast and promising data source, and an underused instance of user-generated geographic information that warrants more attention and investigation. There are several aspects that future work can build upon.

First, regarding data characteristics and distribution, SVV is inherently shaped by user behavior and preferences. As a form of volunteered geographic information (VGI), it exhibits uneven spatial and temporal coverage, a limitation widely documented in the literature (Wu et al., 2016; Senaratne et al., 2017; Basiri et al., 2019; Huang et al., 2024b). Even within the city center, where this study intentionally focuses, video trajectories tend to concentrate on visually appealing or popular streets. Although such patterns may reflect collective urban attention and street popularity, they can still introduce biases from a data perspective, resulting in a form of survivorship bias at the level of cultural production: more vibrant places of higher touristic appeal are more likely to be captured by city tour videos to begin with, to attract views, while other places are much less visible, making it difficult to understand the less popular places. Outside the city center, suburban and rural areas are much less documented due to lower levels of online content creation and sharing. Consequently, SVV currently provides limited support for analyzing vitality patterns in urban areas that attract less public attention and documentation. Future research should therefore consider integrating complementary data sources such as SVI to calibrate SVV-derived observations and extend the analysis to broader urban areas, thereby improving spatial coverage and analytical robustness.

A second limitation concerns temporal information. Unlike traditional sensing infrastructures that record precise timestamps, SVV rarely provides verifiable recording times, as the upload date of a video may differ significantly from its actual recording date. This discrepancy limits the ability to align observations with specific temporal events such as rush hours or weather conditions. While some creators have begun to include detailed contextual information (e.g., time, date, and weather) in video descriptions, such practices are not yet standardized. In this study, we partially address this issue by applying a lighting-condition classifier to distinguish between daytime and nighttime videos. Nevertheless, future studies could further explore temporal dynamics by focusing on subsets of videos that contain explicit time information or by combining SVV data with other temporally resolved datasets.

Another limitation lies in the variability of data quality. Tour videos often contain various forms of noise, including video editing, background music, unstable camera movement, and diverse filming styles. In this study, we introduced a video quality control framework to filter videos suitable for analysis while maintaining high levels of precision. However, this filtering process inevitably removes a portion of potentially useful data. Future research could develop more comprehensive SVV quality assessment frameworks to better understand how different quality factors influence analytical results and to maximize the effective use of available video data.

Finally, we acknowledge several limitations in the proposed geo-localization framework. Although the method achieved a correctness rate of 77.2% after quality control, its generalizability remains context-dependent. The feature extraction strategy is designed primarily for dense urban centers where geographic textual cues (e.g., street signs and storefront signage) are frequently visible. In suburban or rural environments where such textual information is sparse, the method may perform less effectively. Even within city centers, some environments such as urban parks may lack sufficient textual cues, leading to unmatched videos despite their potential analytical value, as discussed in Section 5.1. In addition, with respect to spatial precision, our current method only resolves matches at the road-segment level, without identifying which specific sidewalk is being traversed. This limitation stems from the inherent uncertainties in the videos, where the frames may capture the near side, the opposite side, or both simultaneously. More precise sidewalk-level recognition would enable richer and more fine-grained urban insights. Therefore, future research could further refine geo-localization techniques by incorporating additional image retrieval-based or classification-based strategies, such as constructing a local reference database of SVIs and encoding video frames as embeddings for direct feature comparison, thereby reducing reliance on text-based matching (Astruc et al., 2024; Clark et al., 2023; Vo et al., 2017). Moreover, stability assessments of video footage combined with segmentation techniques could assist in determining the exact sidewalk being used.

7. Conclusion

For the first time, this study conceptualized Street View Video (SVV) as a novel and promising data source for urban sensing, highlighting its potential to capture the dynamic street-level environments. Building on existing literature, we identified several SVV formats that fall within this conceptualization, providing a foundation for extending and enriching this emerging source of urban data.

Using city tour videos and following the pioneering work of Hartmann and Purves (2023), we proposed a high-precision geolocalization method that combines text and depth information extracted from videos. Given that similar information is also embedded in other SVV formats, the method we developed has the potential to be further extended to broader data sources, supporting various future cross-city and multimodal applications.

We further focused on pedestrian activity and sound environment as key indicators of urban vitality, showing the ability of SVV to capture complementary visual and acoustic dimensions of street environments. To our knowledge, this represents the first attempt to integrate both dimensions at scale through automated processing. While promising, it remains an uneven data source, with limitations that need to be carefully acknowledged and managed.

Overall, this study sets the scene as a proof-of-concept for integrating user-generated SVV into urban analytics workflows. While not a population-representative measure of urban vitality, it demonstrates the feasibility of extracting multimodal indicators and highlights the value of SVV as a complementary data source for future urban sensing and planning research.

Author contributions

Y.G.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration, **X.L.**: Conceptualization, Methodology, Writing - Review & Editing, Visual-

ization, **Z.F.**: Conceptualization, Methodology, Writing - Review & Editing, **Y.H.**: Methodology, Writing - Review & Editing, **F.B.**: Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Data availability

The authors do not have permission to share the data.

Code availability

The code, processing pipeline, and the full video list are available at: https://github.com/youlonggu/Street-View-Video-SVV-Project-Tutorial/

Acknowledgements

We are grateful for the constructive and encouraging feedback from the editors and reviewers, which has significantly improved our paper. We gratefully acknowledge the contributions of video creators on YouTube. We also thank our colleagues at the NUS Urban Analytics Lab for their valuable discussions and insights, and Wu Shuhui for her help in designing the illustration. This research is part of the project Time series Street View Imagery - A new paradigm for urban sensing, which is supported by the Singapore Ministry of Education Academic Research Fund Tier 1. This research is also supported by the NUS Graduate Research Scholarship (X.L.) and the NUS President's Graduate Fellowship (Z.F.).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT for proofreading and assistance for code development. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Appendix A. Supplementary information

Sampled points distribution for Mapillary and GSV, alongside the corresponding street-segment level human object detection counts, are presented in Figure A.13.

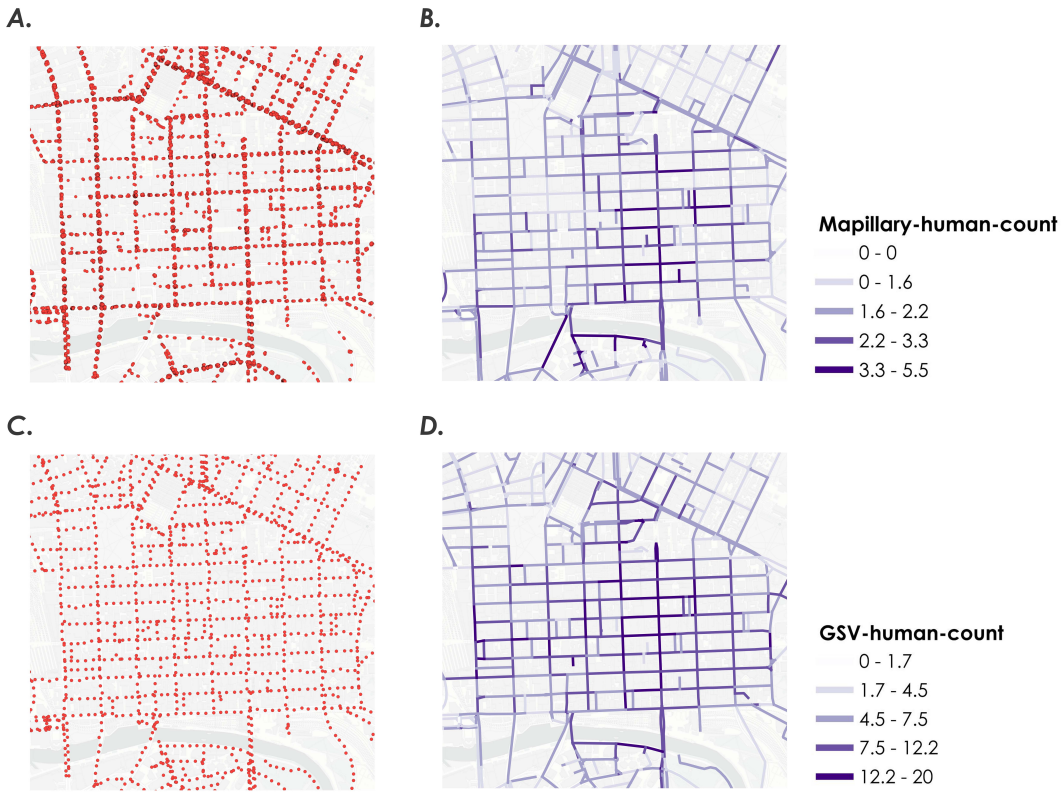


Figure A.13: (A), (C) Sampled SVI points retrieved from Mapillary and GSV at 50 m intervals along the street network, respectively. (B), (D) Mean human counts aggregated at the street-segment level based on object detection results from the corresponding SVI samples.

References

Aharon, N., Orfaig, R., Bobrovsky, B.Z., 2022. BoT-SORT: Robust Associations Multi-Pedestrian Tracking. doi:10.48550/arXiv.2206.14651.

Aiello, L.M., Schifanella, R., Quercia, D., Aletta, F., 2015. Chatty maps: constructing sound maps of urban areas from social media data. Royal Society Open Science doi:10.1098/rsos.150690.

Alam, M.S., Martens, M.H., Bazilinska, O., Bazilinskyy, P., 2025. Understanding global pedestrian behaviour in 565 cities with dashcam videos on youtube .

Aletta, F., 2016. Descriptors and indicators for soundscape design: vibrancy as an example. INTER-NOISE and NOISE-CON Congress and Conference Proceedings 253, 387–392.

Aletta, F., Kang, J., 2018. Towards an Urban Vibrancy Model: A Soundscape Approach. International Journal of Environmental Research and Public Health 15, 1712. doi:10.3390/ijerph15081712.

Andringa, T.C., Lanser, J.J.L., 2013. How Pleasant Sounds Promote and Annoying Sounds Impede Health: A Cognitive Approach. International Journal of Environmental Research and Public Health 10, 1439–1461. doi:10.3390/ijerph10041439.

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., Weaver, J., 2010. Google Street View: Capturing the World at Street Level. IEEE Computer 43, 32–38. doi:10.1109/MC.2010.170.

Appleyard, D., 1980. Livable Streets: Protected Neighborhoods? The ANNALS of the American Academy of Political and Social Science 451, 106–117. doi:10.1177/000271628045100111.

Astruc, G., Dufour, N., Siglidis, I., Aronssohn, C., Bouia, N., Fu, S., Loiseau, R., Nguyen, V.N., Raude, C., Vincent, E., XU, L., Zhou, H., Landrieu, L., 2024. OpenStreetView-5M: The Many Roads to Global Visual Geolocation. doi:10.48550/arXiv.2404.18873.

Aumond, P., Can, A., Mallet, V., De Coensel, B., Ribeiro, C., Botteldooren, D., Lavandier, C., 2018a. Kriging-based spatial interpolation from measurements for sound level mapping in urban areas. The Journal of the Acoustical Society of America 143, 2847–2857. doi:10.1121/1.5034799.

Aumond, P., Jacquesson, L., Can, A., 2018b. Probabilistic modeling framework for multi-source sound mapping. Applied Acoustics 139, 34–43. doi:10.1016/j.apacoust.2018.04.017.

Bai, Z., Huang, S., Wang, L., Qi, J., 2025. Refined understanding of street vitality: comparing the spatial effects of pedestrian vitality and cycling vitality and the spatial heterogeneity of factors. International Journal of Digital Earth 18, 2501769. doi:10.1080/17538947.2025.2501769.

Basiri, A., Haklay, M., Foody, G., Mooney, P., 2019. Crowdsourced geospatial data quality: challenges and future directions. International Journal of Geographical Information Science 33, 1588–1593. doi:10.1080/13658816.2019.1593422.

Ben-Joseph, E., Lee, J.S., Cromley, E.K., Laden, F., Troped, P.J., 2013. Virtual and actual: Relative accuracy of on-site and web-based instruments in auditing the environment for physical activity. Health & Place 19, 138–150. doi:10.1016/j.healthplace.2012.11.001.

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: A review. Landscape and Urban Planning 215, 104217. doi:10.1016/j.landurbplan.2021.104217.

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

Boumchich, A., Picaut, J., Aumond, P., Can, A., Bocher, E., 2024. Blind Calibration of Environmental Acoustics Measurements Using Smartphones. Sensors 24, 1255. doi:10.3390/s24041255.

Cabon, Y., Murray, N., Humenberger, M., 2020. Virtual KITTI 2. doi:10.48550/arXiv.2001.10773.

Can, A., Audubert, P., Aumond, P., Geisler, E., Guiu, C., Lorino, T., Rossa, E., 2023. Framework for urban sound assessment at the city scale based on citizen action, with the smartphone application NoiseCapture as a lever for participation. Noise Mapping 10. doi:10.1515/noise-2022-0166.

Cao, J., Kang, J., 2024. Effect of human sound component on the sociability of urban public spaces—a case study in Sheffield, UK. Humanities and Social Sciences Communications 11, 1523. doi:10.1057/s41599-024-03929-y.

Ceccarelli, G., Messa, F., Gorrini, A., Presicce, D., Choubassi, R., 2023. Deep learning video analytics for the assessment of street experiments: The case of Bologna. Journal of Urban Mobility 4, 100067. doi:10.1016/j.urbmob.2023.100067.

Chen, L., Lu, Y., Sheng, Q., Ye, Y., Wang, R., Liu, Y., 2020. Estimating pedestrian volume using Street View images: A large-scale validation test. Computers, Environment and Urban Systems 81, 101481. doi:10.1016/j.compenvurbsys.2020.101481.

Chen, L., Lu, Y., Ye, Y., Xiao, Y., Yang, L., 2022. Examining the association between the built environment and pedestrian volume using street view images. Cities 127, 103734. doi:10.1016/j.cities.2022.103734.

Chen, M., Wei, Z., Cao, K., 2025a. Multi-perspective perception of city image in Shanghai via massive short videos. International Journal of Applied Earth Observation and Geoinformation 144, 104844. doi:10.1016/j.jag.2025.104844.

Chen, W., Natapov, A., Ali, Y., 2025b. Can urban vitality be seen? Video analytics of social interaction and land use. Land Use Policy 159, 107818. doi:10.1016/j.landusepol.2025.107818.

Chen, Y., Yu, B., Shu, B., Yang, L., Wang, R., 2023. Exploring the spatiotemporal patterns and correlates of urban vitality: Temporal and spatial heterogeneity. Sustainable Cities and Society 91, 104440. doi:10.1016/j.scs.2023.104440.

Cinnamon, J., Jahiu, L., 2023. 360-degree video for virtual place-based research: A review and research agenda. Computers, Environment and Urban Systems 106, 102044. doi:10.1016/j.compenvurbsys.2023.102044.

Clark, B., Kerrigan, A., Kulkarni, P.P., Cepeda, V.V., Shah, M., 2023. Where We Are and What We're Looking At: Query Based Worldwide Image Geo-localization Using Hierarchies and Scenes. doi:10.48550/arXiv.2303.04249.

Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., Zhang, Y., Lv, W., Huang, K., Zhang, Y., Zhang, J., Zhang, J., Liu, Y., Yu, D., Ma, Y., 2025. PaddleOCR 3.0 Technical Report. doi:10.48550/arXiv.2507.05595.

Dale, A., Ling, C., Newman, L., 2010. Community Vitality: The Role of Community-Level Resilience Adaptation and Innovation in Sustainable Development. Sustainability 2, 215–231. doi:10.3390/su2010215.

Dardour, A., El Haji, E., Begdouri, M.A., 2025. Video Surveillance and Artificial Intelligence for Urban Security in Smart Cities: A Review of a Selection of Empirical Studies from 2018 to 2024. Computer Sciences & Mathematics Forum 10, 15. doi:10.3390/cmsf2025010015.

Delclòs-Alió, X., Gutiérrez, A., Miralles-Guasch, C., 2019. The urban vitality conditions of Jane Jacobs in Barcelona: Residential and smartphone-based tracking measurements of the built environment in a Mediterranean metropolis. Cities 86, 220–228. doi:10.1016/j.cities.2018.09.021.

Diogenes, M.C., Greene-Roesel, R., Arnold, L.S., Ragland, D.R., 2007. Pedestrian Counting Methods at Intersections: A Comparative Study. Transportation Research Record 2002, 26–30. doi:10.3141/2002-04.

yt dlp, 2024. yt-dlp: A feature-rich command-line audio/video downloader. https://github.com/yt-dlp/yt-dlp.

Doan, Q.C., Zhang, X., 2025. A systematic review of urban vitality studies: Trends and research opportunities. Land Use Policy 158, 107745. doi:10.1016/j.landusepol.2025.107745.

Du, F., Wang, J., Mao, L., Kang, J., 2024. Daily rhythm of urban space usage: insights from the nexus of urban functions and human mobility. Humanities and Social Sciences Communications 11, 49. doi:10.1057/s41599-023-02577-y.

Fan, Z., Biljecki, F., 2024. Nighttime Street View Imagery: A new perspective for sensing urban lighting landscape. Sustainable Cities and Society 116, 105862. doi:10.1016/j.scs.2024.105862.

Feng, Y., Duives, D., Daamen, W., Hoogendoorn, S., 2021. Data collection methods for studying pedestrian behaviour: A systematic review. Building and Environment 187, 107329. doi:10.1016/j.buildenv.2020.107329.

Galich, A., Stiller, D., Wurm, M., Taubenböck, H., 2025. AI-Based Counting of Traffic Participants: An Explorative Study Using Public Webcams. Future Transportation 5, 87. doi:10.3390/futuretransp5030087.

Gasco, L., Schifanella, R., Aiello, L.M., Quercia, D., Asensio, C., de Arcas, G., 2020. Social Media and Open Data to Quantify the Effects of Noise on Health. Frontiers in Sustainable Cities 2. doi:10.3389/frsc.2020.00041.

Gehl, J., 2011. Life between buildings.

Gehl, J., 2013. Cities for People.

Gehl, J., Svarre, B., 2013. How To Study Public Life. Island Press/Center for Resource Economics, Washington, DC. doi:10.5822/978-1-61091-525-0.

Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M., 2017. Audio set: An ontology and human-labeled dataset for audio events, in: Proc. IEEE ICASSP 2017, New Orleans, LA. doi:10.1109/ICASSP.2017.7952261.

Grabowski, D., Czyżewski, A., 2020. System for monitoring road slippery based on CCTV cameras and convolutional neural networks. Journal of Intelligent Information Systems 55, 521–534. doi:10.1007/s10844-020-00618-5.

Gravitz-Sela, S., Levy, A., Zehavi, S., Bryt, O., Shach-Pinsly, D., Plaut, P., 2025. Analyzing usage patterns from video data through deep learning: The case of an urban park. Computers, Environment and Urban Systems 117, 102229. doi:10.1016/j.compenvurbsys.2024.102229.

Guo, X., Chen, H., Yang, X., 2021. An Evaluation of Street Dynamic Vitality and Its Influential Factors Based on Multi-Source Big Data. ISPRS International Journal of Geo-Information 10, 143. doi:10.3390/ijgi10030143.

Güller, C., Varol, C., 2024. Unveiling the daily rhythm of urban space: Exploring the influence of built environment on spatiotemporal mobility patterns. Applied Geography 170, 103366. doi:10.1016/j.apgeog.2024.103366.

Hall, P., Pfeiffer, U., 2013. Urban Future 21: A Global Agenda for Twenty-First Century Cities. London. doi:10.4324/9781315011523.

Hartmann, M.C., Purves, R.S., 2023. Seeing through a new lens: exploring the potential of city walking tour videos for urban analytics. International Journal of Digital Earth 16, 2555–2573. doi:10.1080/17538947.2023.2230182.

He, K., Li, H., Zhang, H., Hu, Q., Yu, Y., Qiu, W., 2025. Revisiting Gehl's urban design principles with computer vision and webcam data: Associations between public space and public life. Environment and Planning B , 23998083251328771doi:10.1177/23998083251328771.

Hong, J.Y., Jeon, J.Y., 2017. Exploring spatial relationships among soundscape variables in urban areas: A spatial statistical modelling approach. Landscape and Urban Planning 157, 352–364. doi:10.1016/j.landurbplan.2016.08.006.

Hou, J., Chen, L., Zhang, E., Jia, H., Long, Y., 2020. Quantifying the usage of small public spaces using deep convolutional neural network. PLOS ONE 15, e0239390. doi:10.1371/journal.pone.0239390.

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. doi:10.1016/j.isprsjprs.2024.06.023.

Hu, T., Wei, D., Su, Y., Wang, X., Zhang, J., Sun, X., Liu, Y., Guo, Q., 2022. Quantifying the shape of urban street trees and evaluating its influence on their aesthetic functions based on mobile lidar data. ISPRS Journal of Photogrammetry and Remote Sensing 184, 203–214. doi:10.1016/j.isprsjprs.2022.01.002.

Hu, Z., Lam, W.H.K., Wong, S.C., Chow, A.H.F., Ma, W., 2023. Turning traffic surveillance cameras into intelligent sensors for traffic density estimation. Complex & Intelligent Systems 9. doi:10.1007/s40747-023-01117-0.

Huang, B., Zhou, Y., Li, Z., Song, Y., Cai, J., Tu, W., 2020. Evaluating and characterizing urban vibrancy using spatial big data: Shanghai as a case study. Environment and Planning B 47, 1543–1559. doi:10.1177/2399808319828730.

Huang, J., Fei, T., Kang, Y., Li, J., Liu, Z., Wu, G., 2024a. Estimating urban noise along road network from street view imagery. International Journal of Geographical Information Science 38, 128–155. doi:10.1080/13658816.2023.2274475.

Huang, X., Wang, S., Yang, D., Hu, T., Chen, M., Zhang, M., Zhang, G., Biljecki, F., Lu, T., Zou, L., Wu, C.Y.H., Park, Y.M., Li, X., Liu, Y., Fan, H., Mitchell, J., Li, Z., Hohl, A., 2024b. Crowdsourcing Geospatial Data for Earth and Human Observations: A Review. Journal of Remote Sensing 4, 0105. doi:10.34133/remotesensing.0105.

Ibrahim, M.R., Lyons, T., 2025. Transforming CCTV cameras into NO2 sensors at city scale for adaptive policymaking. Scientific Reports 15, 3640. doi:10.1038/s41598-025-86532-8.

Islam, M.S., , Volodymyr, S., , John N., I., , Nalini, R., , Garder, P.E., 2014. Explaining Pedestrian Safety Experience at Urban and Suburban Street Crossings Considering Observed Conflicts and Pedestrian Counts. Journal of Transportation Safety & Security 6, 335–355. doi:10.1080/19439962.2013.877549.

Ito, K., Zhu, Y., Abdelrahman, M., Liang, X., Fan, Z., Hou, Y., Zhao, T., Ma, R., Fujiwara, K., Ouyang, J., Quintana, M., Biljecki, F., 2025. ZenSVI: An open-source software for the integrated acquisition, processing and analysis of street view imagery towards scalable urban science. Computers, Environment and Urban Systems 119, 102283. doi:10.1016/j.compenvurbsys.2025.102283.

Jacobs, J., 1961. The Death and Life of Great American Cities.

Jacobs, N., Burgin, W., Fridrich, N., Abrams, A., Miskell, K., Braswell, B.H., Richardson, A.D., Pless, R., 2009. The global network of outdoor webcams: properties and applications, in: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, New York, NY, USA. pp. 111–120. doi:10.1145/1653771.1653789.

Jarratt, D., 2020. An exploration of webcam-travel: Connecting to place and nature through webcams during the COVID-19 lockdown of 2020. Tourism and Hospitality Research doi:10.1177/1467358420963370.

Jiang, Y., Han, Y., Liu, M., Ye, Y., 2022. Street vitality and built environment features: A data-informed approach from fourteen Chinese cities. Sustainable Cities and Society 79, 103724. doi:10.1016/j.scs.2022.103724.

Jin, A., Li, G., Nie, Q., Chen, N., Li, B., Wen, X., Xue, Y., 2024. Reconstructing urban spatial blind areas in child trafficking: Insights from the internet video surveillance data. Cities 150, 105026. doi:10.1016/j.cities.2024.105026.

Job, J.R., Myers, K., Naghshineh, K., Gill, S.A., 2016. Uncovering Spatial Variation in Acoustic Environments Using Sound Mapping. PLOS ONE 11, e0159883. doi:10.1371/journal.pone.0159883.

Jocher, G., Qiu, J., 2024. Ultralytics yolo11. URL: https://github.com/ultralytics/ultralytics.

Kang, J., 2023. Soundscape in city and built environment: current developments and design potentials. City and Built Environment 1, 1. doi:10.1007/s44213-022-00005-6.

Kang, J., Aletta, F., Gjestland, T.T., Brown, L.A., Botteldooren, D., Schulte-Fortkamp, B., Lercher, P., van Kamp, I., Genuit, K., Fiebig, A., Bento Coelho, J.L., Maffei, L., Lavia, L., 2016. Ten questions on the soundscapes of the built environment. Building and Environment 108, 284–294. doi:10.1016/j.buildenv.2016.08.011.

Kelly, C.M., Wilson, J.S., Baker, E.A., Miller, D.K., Schootman, M., 2013. Using Google Street View to Audit the Built Environment: Inter-rater Reliability Results. Annals of Behavioral Medicine 45, S108–S112. doi:10.1007/s12160-012-9419-9.

Kogan, P., Turra, B., Arenas, J.P., Hinalaf, M., 2017. A comprehensive methodology for the multidimensional and synchronic data collecting in soundscape. Science of The Total Environment 580, 1068–1077. doi:10.1016/j.scitotenv.2016.12.061.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D., 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. doi:10.48550/arXiv.1912.10211.

Lewis, Q.W., Park, E., 2018. Volunteered Geographic Videos in Physical Geography: Data Mining from YouTube. Annals of the American Association of Geographers 108, 52–70. doi:10.1080/24694452.2017.1343658.

Li, J., Zhang, Z., Jing, F., Gao, J., Ma, J., Shao, G., Noel, S., 2020. An evaluation of urban green space in Shanghai, China, using eye tracking. Urban Forestry & Urban Greening 56, 126903. doi:10.1016/j.ufug.2020.126903.

Li, M., Pan, J., 2023. Assessment of Influence Mechanisms of Built Environment on Street Vitality Using Multisource Spatial Data: A Case Study in Qingdao, China. Sustainability 15, 1518. doi:10.3390/su15021518.

Li, Y., Yabuki, N., Fukuda, T., 2022. Exploring the association between street built environment and street vitality using deep learning methods. Sustainable Cities and Society 79, 103656. doi:10.1016/j.scs.2021.103656.

Lian, T., Loo, B.P.Y., Fan, Z., 2024. Advances in estimating pedestrian measures through artificial intelligence: From data sources, computer vision, video analytics to the prediction of crash frequency. Computers, Environment and Urban Systems 107, 102057. doi:10.1016/j.compenvurbsys.2023.102057.

Liang, S., Leng, H., Yuan, Q., Wang, B., Yuan, C., 2020. How does weather and climate affect pedestrian walking speed during cool and cold seasons in severely cold areas? Building and Environment 175, 106811. doi:10.1016/j.buildenv.2020.106811.

Liu, M., Jiang, Y., He, J., 2021. Quantitative Evaluation on Street Vitality: A Case Study of Zhoujiadu Community in Shanghai. Sustainability 13, 3027. doi:10.3390/su13063027.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L., 2024. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. doi:10.48550/arXiv.2303.05499.

Lombardo, M., Totaro, V., Chiaravalloti, F., Petrucci, O., 2025. Street-Scale Hydrodynamic Estimation from Social Media Videos: A Systematic Approach to Urban Floods Data Collection. International Journal of Disaster Risk Reduction , 105419doi:10.1016/j.ijdrr.2025.105419.

Loo, B.P., Fan, Z., 2023. Social interaction in public space: Spatial edges, moveable furniture, and visual landmarks. Environment and Planning B: Urban Analytics and City Science 50, 2510–2526. doi:10.1177/23998083231160549.

Marshall, S., 2004. Streets and Patterns. Routledge, London. doi:10.4324/9780203589397.

Montgomery, J., 1998. Making a city: Urbanity, vitality and urban design. Journal of Urban Design 3, 93–116. doi:10.1080/13574809808724418.

Mouratidis, K., Delclòs-Alió, X., 2026. Urban vitality versus urban livability: Does vibrancy matter for neighborhood satisfaction and neighborhood happiness? Cities 168, 106473. doi:10.1016/j.cities.2025.106473.

Nathvani, R., Cavanaugh, A., Suel, E., Bixby, H., Clark, S.N., Metzler, A.B., Nimo, J., Moses, J.B., Baah, S., Arku, R.E., Robinson, B.E., Baumgartner, J., Bennett, J.E., Arif, A.M., Long, Y., Agyei-Mensah, S., Ezzati, M., 2025. Measurement of urban vitality with time-lapsed street-view images and object-detection for scalable assessment of pedestrian-sidewalk dynamics. ISPRS Journal of Photogrammetry and Remote Sensing 221, 251–264. doi:10.1016/j.isprsjprs.2025.01.038.

Newman, P., Kenworthy, J., 1998. Sustainability and cities: overcoming automobile dependence.

Oda, T., Yoshimura, Y., 2024. Quantifying the vibrancy of streets: Large-scale pedestrian density estimation with dashcam data. Transportation Research Part C: Emerging Technologies 167, 104840. doi:10.1016/j.trc.2024.104840.

Oliver, N., Lepri, B., Sterly, H., Lambiotte, R., Deletaille, S., Nadai, M.D., Letouzé, E., Salah, A.A., Benjamins, R., Cattuto, C., Colizza, V., de Cordes, N., Fraiberger, S.P., Koebe, T., Lehmann, S., Murillo, J., Pentland, A., Pham, P.N., Pivetta, F., Saramäki, J., Scarpino, S.V., Tizzoni, M., Verhulst, S., Vinck, P., 2020. Mobile phone data for informing public

health actions across the covid-19 pandemic life cycle. Science Advances 6, eabc0764. doi:10.1126/sciadv.abc0764.

Othman, E., Cibilić, I., Poslončec-Petrić, V., Saadallah, D., 2024. Investigating Noise Mapping in Cities to Associate Noise Levels with Sources of Noise Using Crowdsourcing Applications. Urban Science 8, 13. doi:10.3390/urbansci8010013.

Rodríguez, R., Robles-Morua, A., Meza-Figueroa, D., Tuxpan-Vargas, J., Cisneros Almazán, R., Perez Ramos, N., Sen-Gupta, B., Martínez-Villegas, N., 2025. Urban flood exposure assessment through citizen science, 2D hydraulic modeling, and artificial intelligence. Journal of Environmental Management 393, 127037. doi:10.1016/j.jenvman.2025.127037.

Salazar-Miranda, A., Fan, Z., Baick, M., Hampton, K.N., Duarte, F., Loo, B.P.Y., Glaeser, E., Ratti, C., 2025. Exploring the social life of urban spaces through AI. Proceedings of the National Academy of Sciences 122, e2424662122. doi:10.1073/pnas.2424662122.

Salazar-Miranda, A., Zhang, F., Sun, M., Leoni, P., Duarte, F., Ratti, C., 2023. Smart curbs: Measuring street activities in real-time using computer vision. Landscape and Urban Planning 234, 104715. doi:10.1016/j.landurbplan.2023.104715.

Schafer, R.M., 1977. The tuning of the world.

Senaratne, H., , Amin, M., , Ahmed Loai, A., , Cristina, C., , Haklay, M.M., 2017. A review of volunteered geographic information quality assessment methods. International Journal of Geographical Information Science 31, 139–167. doi:10.1080/13658816.2016.1189556.

Sheller, M., Urry, J., 2000. The City and the Car. International Journal of Urban and Regional Research 24, 737–757. doi:10.1111/1468-2427.00276.

Shorfuzzaman, M., Hossain, M.S., Alhamid, M.F., 2021. Towards the sustainable development of smart cities through mass video surveillance: A response to the COVID-19 pandemic. Sustainable Cities and Society 64, 102582. doi:10.1016/j.scs.2020.102582.

Sisiopiku, V.P., Akin, D., 2003. Pedestrian behaviors at and perceptions towards various pedestrian facilities: an examination based on observation and survey data. Transportation Research Part F: Traffic Psychology and Behaviour 6, 249–274. doi:10.1016/j.trf.2003.06.001.

Soto, G.W., Whitfield, G.P., Webber, B.J., Omura, J.D., Chen, T.J., Zaganjor, H., Rose, K., 2022. Traffic as a barrier to walking safely in the United States: Perceived reasons and potential mitigation strategies. Preventive Medicine Reports 30, 102003. doi:10.1016/j.pmedr.2022.102003.

Thulin, S., 2018. Sound maps matter: expanding cartophony. Social & Cultural Geography 19, 192–210. doi:10.1080/14649365.2016.1266028.

Tokuda, E.K., Lockerman, Y., Ferreira, G.B.A., Sorrelgreen, E., Boyle, D., Cesar-Jr., R.M., Silva, C.T., 2020. A New Approach for Pedestrian Density Estimation Using Moving Sensors and Computer Vision. ACM Trans. Spatial Algorithms Syst. 6, 26:1–26:20. doi:10.1145/3397575.

Van Kamp, I., Brown, A.L., Schreckenberg, D., 2019. Soundscape approaches in urban planning: implications for an intervention framework. doi:10.18154/RWTH-CONV-238792.

Vo, N., Jacobs, N., Hays, J., 2017. Revisiting im2gps in the Deep Learning Era. doi:10.48550/arXiv.1705.04838.

Wang, Y., Jia, Y., Chen, W., Wang, T., Zhang, A., 2024. Examining safe spaces for pedestrians and e-bicyclists at urban crosswalks: An analysis based on drone-captured video. Accident Analysis & Prevention 194, 107365. doi:10.1016/j.aap.2023.107365.

Wang, Y., Qiu, W., Jiang, Q., Li, W., Ji, T., Dong, L., 2023. Drivers or Pedestrians, Whose Dynamic Perceptions Are More Effective to Explain Street Vitality? A Case Study in Guangzhou. Remote Sensing 15, 568. doi:10.3390/rs15030568.

Whyte, W.H., 1980. The social life of small urban spaces.

Wu, B., Cheng, W.H., Zhang, Y., Mei, T., 2016. Time Matters: Multi-scale Temporalization of Social Media Popularity, in: Proceedings of the 24th ACM international conference on Multimedia, New York, NY, USA. pp. 1336–1344. doi:10.1145/2964284.2964335.

Wu, W., Ma, Z., Guo, J., Niu, X., Zhao, K., 2022. Evaluating the Effects of Built Environment on Street Vitality at the City Level: An Empirical Research Based on Spatial Panel Durbin Model. International Journal of Environmental Research and Public Health 19, 1664. doi:10.3390/ijerph19031664.

Wu, W., Niu, X., Li, M., 2021. Influence of Built Environment on Street Vitality: A Case Study of West Nanjing Road in Shanghai Based on Mobile Location Data. Sustainability 13, 1840. doi:10.3390/su13041840.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth Anything V2. doi:10.48550/arXiv.2406.09414.

Yang, X., Grace, D., Chen, C., Berkel, D.V., Fox, N., Lindquist, M., 2025. Cross-national comparison of the restorative potential of urban forests in different seasons. Landscape and Urban Planning 264, 105491. doi:10.1016/j.landurbplan.2025.105491.

Yigitcanlar, T., Teriman, S., 2015. Rethinking sustainable urban development: towards an integrated planning and development process. International Journal of Environmental Science and Technology 12, 341–352. doi:10.1007/s13762-013-0491-x.

Yin, L., 2017. Street level urban design qualities for walkability: Combining 2D and 3D GIS measures. Computers, Environment and Urban Systems 64, 288–296. doi:10.1016/j.compenvurbsys.2017.04.001.

Yin, L., Cheng, Q., Wang, Z., Shao, Z., 2015. 'Big data' for pedestrian volume: Exploring the use of Google Street View images for pedestrian counts. Applied Geography 63, 337–345. doi:10.1016/j.apgeog.2015.07.010.

Zaki, M.H., Sayed, T., 2018. Automated Analysis of Pedestrian Group Behavior in Urban Settings. IEEE Transactions on Intelligent Transportation Systems 19, 1880–1889. doi:10.1109/TITS.2017.2747516.

Zambon, G., Benocci, R., Bisceglie, A., Roman, H.E., Bellucci, P., 2017. The LIFE DYNAMAP project: Towards a procedure for dynamic noise mapping in urban areas. Applied Acoustics 124, 52–60. doi:10.1016/j.apacoust.2016.10.022.

zdsjjtTLG, 2024. Trackit. https://github.com/zdsjjtTLG/TrackIt.

Zeng, C., Song, Y., He, Q., Shen, F., 2018. Spatially explicit assessment on urban vitality: Case studies in Chicago and Wuhan. Sustainable Cities and Society 40, 296–306. doi:10.1016/j.scs.2018.04.021.

Zhang, J., Li, X., Lian, H., Li, H., Zhang, J., 2025a. Day–Night Synergy Between Built Environment and Thermal Comfort and Its Impact on Pedestrian Street Vitality: Beijing–Chengdu Comparison. Buildings 15, 2118. doi:10.3390/buildings15122118.

Zhang, X., Chen, M., Huang, Y., 2025b. Who gets to use the street? Evaluate the utilization and inclusiveness using crowdsourced videos and vision-language models. Sustainable Cities and Society 134, 106906. doi:10.1016/j.scs.2025.106906.

Zhao, T., Liang, X., Tu, W., Huang, Z., Biljecki, F., 2023. Sensing urban soundscapes from street view imagery. Computers, Environment and Urban Systems 99, 101915. doi:10.1016/j.compenvurbsys.2022.101915.

Zhu, J., Chen, S., Tu, W., Sun, K., 2019. Tracking and Simulating Pedestrian Movements at Intersections Using Unmanned Aerial Vehicles. Remote Sensing 11, 925. doi:10.3390/rs11080925.

Zou, H., Cao, K., Jiang, C., 2021. Spatio-Temporal Visual Analysis for Urban Traffic Characters Based on Video Surveillance Camera Data. ISPRS International Journal of Geo-Information 10, 177. doi:10.3390/ijgi10030177.