

# **Monitoring Urban Conditions through Computer Vision – New Possibilities for Data-driven Urban FM?**

*Hanna Bonekämper, M.A.*

*Karlsruher Institut für Technologie (KIT); Institut für Technologie und Management im Baubetrieb (TMB)*

## **Abstract:**

Rapid urbanization demands new approaches for the continuous monitoring of urban conditions to ensure smooth operation and liveable cities – key goals of successful Urban Facility Management (FM). Traditional inspection methods offer limited scalability and low temporal resolution. This paper explores how computer vision can transform urban condition monitoring through the analysis of street-level imagery, and positions these methods as a technical foundation for dynamic digital twins that continuously capture and analyze urban conditions rather than serving merely as static, formal scale models. Based on a systematic review of recent studies, it highlights data-driven methods – particularly image segmentation and deep learning models such as DeepLabv3+ – used to assess indicators such as greenery, cleanliness, or accessibility. The findings demonstrate computer vision's potential to generate high-frequency data that can feed dynamic digital twins, enabling continuous monitoring and immediate response strategies. Integrating these techniques into Urban FM supports more data-driven decision-making and adaptive management of the built environment.

## **Key words:**

*Urban Facility Management, Dynamic Digital Twins, Computer Vision, Image Segmentation*

## 1. Introduction

With accelerating global urbanization increasing the demand for effective urban monitoring (Hurbean et al., 2021; Amer et al., 2025), the established link between urban environment quality and people's well-being (Vallebuena et al., 2023) underscores this urgency.

Traditional methodologies for assessing urban conditions, ranging from infrastructure quality and walkability to socio-economic vitality, have historically relied on labor-intensive field observations, surveys, and census data (Amer et al., 2025). However, these approaches are often constrained by low update frequency, which hinders the ability to capture the dynamic nature of cities (Amer et al., 2025). To overcome these limitations, the field of urban studies has increasingly integrated big data and machine learning (ML), transforming the way researchers quantify both the tangible and intangible aspects of the built environment. A pivotal advancement in this domain is the utilization of Street View Imagery (SVI), provided by platforms such as Google Street View, Baidu Total View, and Tencent Street View (Cinnamon and Jahiu, 2021). Unlike sensing from an aerial perspective, SVI offers a high-resolution, pedestrian-level point of view that reflects how people experience the city (Cinnamon and Jahiu, 2021). Researchers utilize these 360-degree panoramas as an economical alternative for "virtual audits," enabling the large-scale extraction of detailed features such as greenery, building facades, and sidewalk accessibility (Chowdhury et al., 2024).

Beyond their academic relevance, these methods carry direct operational value for Urban FM. Automated detection of urban decay indicators such as graffiti, potholes, and broken windows (Vallebuena, 2023) can trigger targeted maintenance dispatch instead of relying on citizen complaints or periodic manual inspection rounds. Continuous assessment of street greenery through indices such as the Green View Index or the Street Greenery Comprehensive Evaluation Indicator (Wang et al., 2024) can inform the prioritization of green maintenance and irrigation schedules, while sanitation-focused models that detect garbage and road surface conditions (Wang, 2022) support demand-responsive cleaning routes rather than fixed schedules. Perceived safety and walkability indices (Costa et al., 2023; Fantous and Bennedjai, 2026) can likewise feed into infrastructure investment decisions on lighting, cycling infrastructure, or pedestrian zones. Once refreshed continuously, such indicators can be embedded in dynamic digital twins of urban infrastructure, moving Urban FM

practice from periodic, reactive inspection toward continuous, condition-based operations.

Despite this rapid progress, the fast-evolving field lacks up-to-date, comprehensive overviews of recent computer vision applications for urban monitoring. This paper addresses this gap by synthesizing recent advancements to provide a current state-of-the-art overview for data-driven urban assessment.

This paper addresses this by asking the following research questions:

1. Where and when have studies on urban qualities using computer vision been published?
2. Which models and training datasets have been used in studies so far?
3. Which urban conditions and qualities have already been studied using computer vision methods?

## **2. Methodology**

In Phase 1 of the review process, studies were selected based on specific inclusion and exclusion criteria: publications from 2020 to 2026, written in English, and focused on urban environments, while excluding those with a primary emphasis on traffic (as an infrastructure domain outside the scope of the built environment). The search was conducted using Google Scholar as the primary database. A targeted search string was developed from predefined components: "urban quality" AND "computer vision" AND (observation OR monitoring). For data extraction, the software Publish or Perish queried the Google Scholar API with this string and the specified time frame, exporting the results to Microsoft Excel for processing. This yielded an initial result of 311 sources. Subsequent manual title screening based on relevance reduced the pool to 103 sources, followed by abstract screening, which further narrowed it to 54 studies. These were then differentiated by study type, identifying 46 case studies and 8

literature reviews. Finally, the case studies were categorized according to their geographical origin, year of publication, and associated keywords.

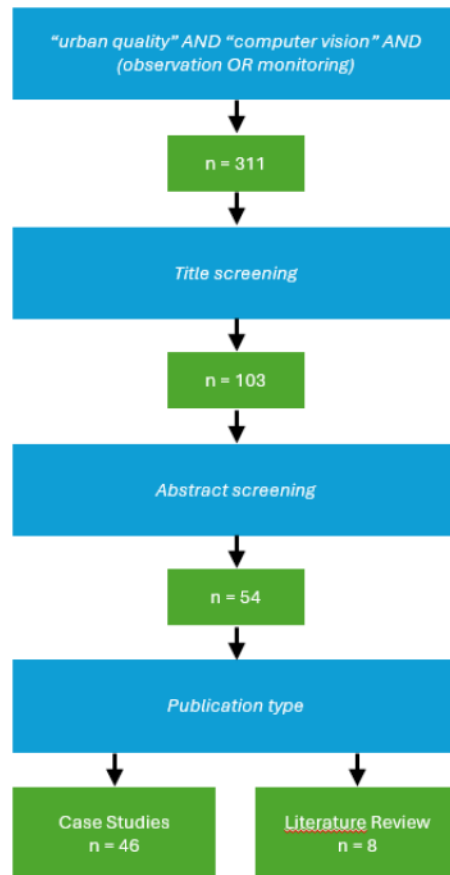


Figure 1: Flowchart of the Literature Search

Phase 2 involved an in-depth analysis of the 46 case studies identified in Phase 1. These were systematically categorized according to key dimensions: image sources, model training approaches, urban characteristics and examined urban qualities, as well as study design. This structured approach enabled a comprehensive mapping of methods and focus areas in computer vision applications for urban studies.

Phase 3 entailed a focused reading of the previously categorized and most relevant sources from Phase 2. This step enabled the extraction of detailed insights, arguments, and methods. Through this targeted analysis, key patterns and contributions were synthesized for the review's objectives.

### 3. Results

The final set of sources from Phase 1 includes 46 case studies and 8 literature reviews. The following analysis focuses on the case studies.

3.1 Geographical and Temporal Distribution of Publications

Approximately 89% of sources were published after 2022, with a peak in 2025, accounting for around 48% of the total publications. This distribution indicates a strong upward trend in research activity in recent years.

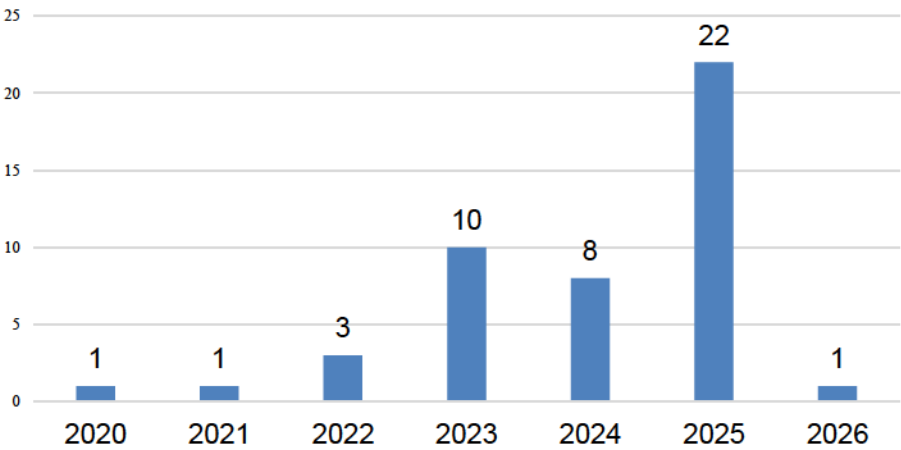


Table 1: Publication years of the examined case studies

The geographical distribution reveals the strong dominance of research conducted in China, accounting for approximately 52% of the publications.

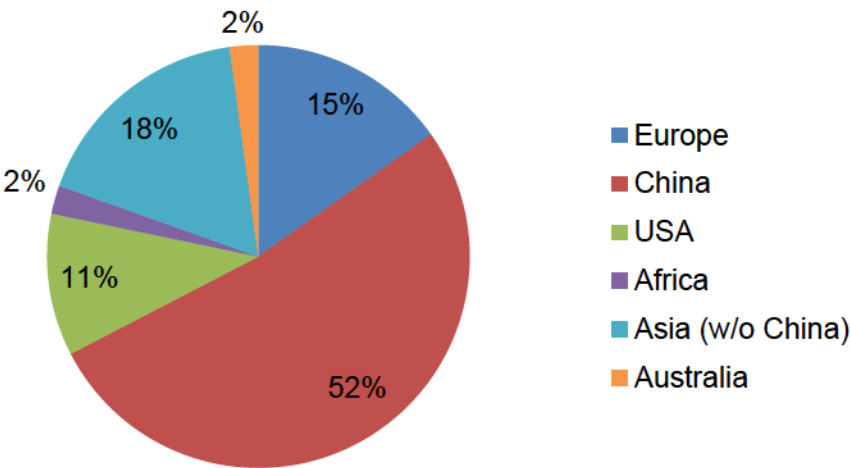


Table 2: Publication origins of the examined case studies

3.2 Keyword Clusters

Based on the identified keyword clusters, the literature review reveals a strong dominance of Google Street View–based approaches. This highlights the central role of street-level imagery for analyzing urban environments. Perception is the second

most prominent cluster, underlining the interest in subjective, human-centered dimensions such as perceived attractiveness and quality of urban space. In contrast, deep learning and semantic segmentation appear less frequently, suggesting that advanced AI-based methods are increasingly used but are not yet the standard across studies. Topics such as urban planning, visual quality, green space, and pedestrian-related aspects are represented to a moderate extent.

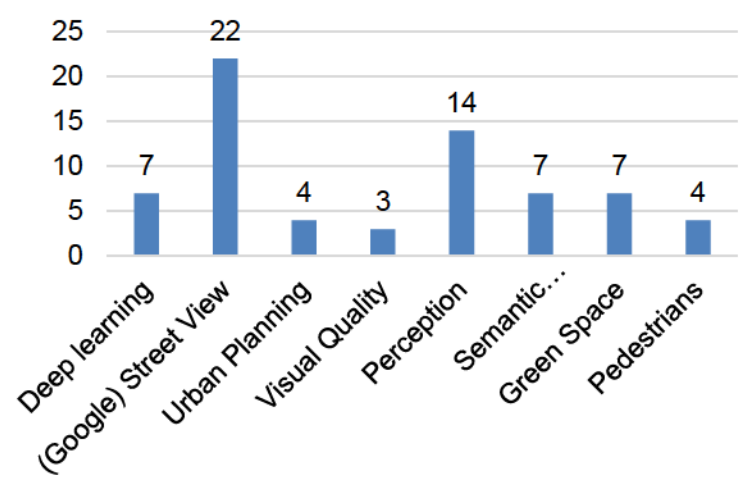


Table 3: Keyword clusters of the examined case studies

### 3.3 Data Sources for Testing the Models

By far the predominant data source was Baidu Street View Images, highlighting the platform's substantial relevance. Baidu Maps, the navigation and mapping service from the Chinese internet conglomerate Baidu and the primary equivalent to Google Maps, covers over 11 million kilometers of roadways nationwide (Xia et al., 2022). These findings align with the high number of Chinese sources identified in Phase 1.

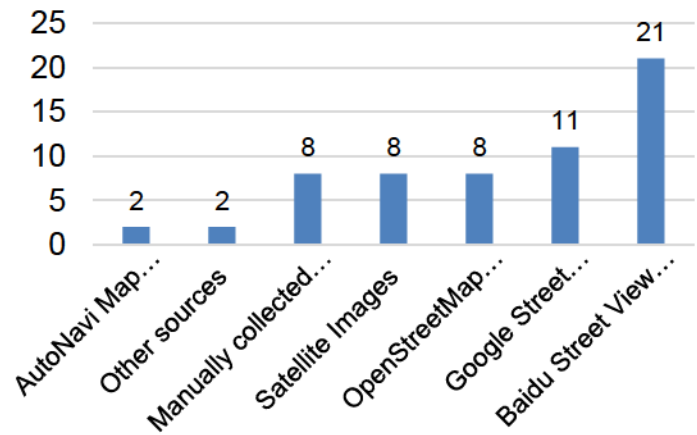


Table 4: Data sources for testing the models

### 3.4 Datasets for Training the Models

The analysis focused on established datasets used for training the computer vision models (excluding case studies without dataset mentions for a substantive overview). It reveals the distribution and significance of these datasets, with two primary ones standing out: ADE20K and MIT Place Pulse 2.0.

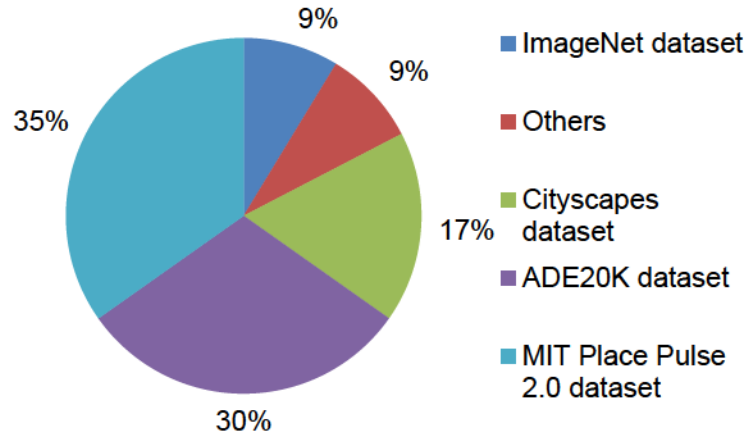


Table 5: Datasets for training the models

ADE20K (Zhou et al., 2019) originates from the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) and was released in 2016. This dataset comprises 25,000 images of complex everyday scenes, annotated with 3,169 distinct labels, with its primary objective being semantic segmentation.

MIT Place Pulse 2.0 (Dubey et al., 2016), from the MIT Media Lab and also released in 2016, is a crowdsourced dataset featuring 110,988 images from 56 cities, alongside 1,170,000 pairwise comparisons from 81,630 online volunteers across six perceptual attributes: safe, lively, boring, wealthy, depressing, and beautiful.

### 3.5 Pre-trained Models

While datasets provide the raw training data for computer vision models, pre-trained models represent ready-to-use neural network architectures already trained on such data. The analysis examined the usage of pre-trained models (excluding case studies without concrete mentions for a substantive overview). DeepLabv3+ (Chen et al., 2018), released in 2018 as the fourth version of the DeepLab series from Google Inc., emerged as the most frequently adopted architecture, accounting for 31% of applications. The "Others" category includes various smaller models, each used in only a single instance across the reviewed literature.

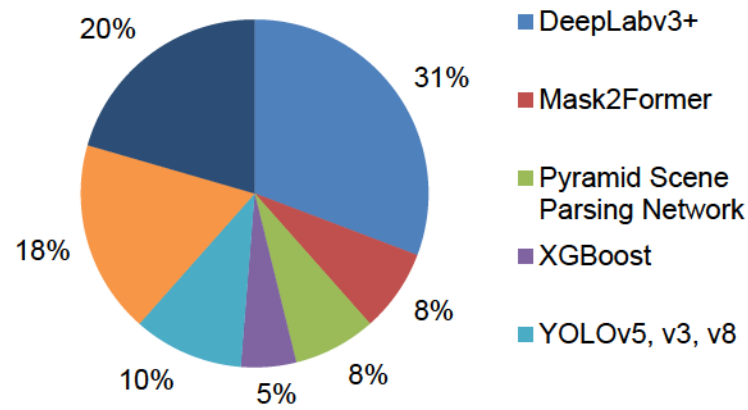


Table 6: Pre-trained models

### 3.6 Methods to Analyze Urban Conditions

The methodology and study design of the case studies revealed diverse approaches, with some models employing multiple methods to analyze urban conditions (categorized across all used techniques). Classification with implicit feature extraction was notable in certain models, which bypass explicit definitions of urban conditions and instead autonomously extract relevant features based on perceptual evaluations (labels). Image segmentation emerged as the most frequently utilized method, primarily because it assigns specific classes to every individual pixel. This granular approach enables a comprehensive and detailed quantitative analysis of represented classes and thus the relevant urban conditions.

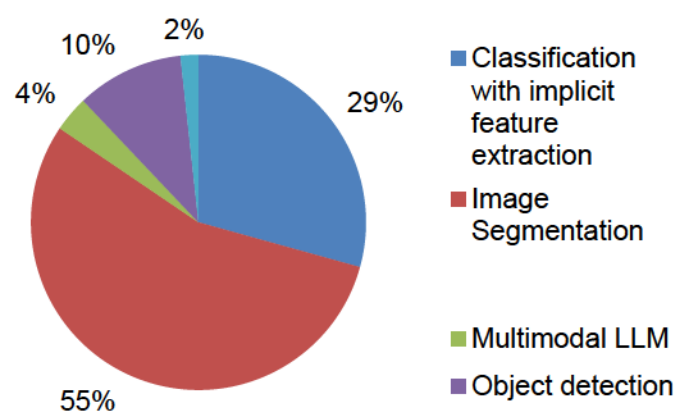


Table 7: Methods to analyze urban qualities

### 3.7 Urban Qualities and Indicators

The following analysis integrates the bibliographic data from the literature overview with the specific findings regarding urban qualities and indicators identified in the research datasets.

**Urban Vitality and Social Perception:** A significant cluster of research focuses on urban vitality and how it is perceived by citizens. Ma (2023) investigates urban vitality in Qingdao, utilizing perceptual emotion indicators such as "Safe," "Beautiful," "Depressing," "Lively," "Wealthy," and "Boring" to identify urban features. Similarly, Li et al. (2025) conduct a multi-level analysis in Hohhot, focusing on street vitality and the built environment. Their research stands out for its comprehensive use of specific indices, including the Population Density Index (PDI), Green View Index (GVI), and Street Cleanliness Index (SCI) to quantify urban conditions. In contrast, Chen et al. (2024) narrow their focus to "urban nature vitality" in San Francisco, primarily using basic physical indicators like buildings, sky, trees, and roads to understand how nature integrates into the urban fabric.

**Visual Quality and Human Perception:** Sottini et al. (2021) focus on how Flickr photo data can be used to assess the perceived quality of urban spaces, analyzing geotagged outdoor images. Based on the classification by Ewing and Handy (2009), their research also considers five key visual quality indicators such as visual enclosure, imageability, human scale, transparency, and complexity to interpret how urban form contributes to the public's perception of city spaces. Gönül and Durak (2025) apply these perceptual methods to a specific historical context, measuring change in historical city centers in Bursa City.

**Environmental Qualities and Specialized Indicators:** Several authors focus on specific environmental or functional urban qualities: Wang et al. (2024) developed the Street Greenery Comprehensive Evaluation Indicator (SGCEI) to identify priority greening areas in Beijing. This differs from the broader Green View Index used by Luo (2022). Fantous and Bennedjai (2026) examine urban walkability in Batna using specialized metrics like the Subjective Walking Preferences Index (SPI). Conversely, Costa et al. (2023) focus strictly on perceived safety for cycling in Berlin, using binary classifications of "safe" versus "unsafe." Vallebuena (2023) identifies urban decay by detecting specific objects like graffiti, potholes, and broken windows, while Wang

(2022) focuses on road sanitation quality, specifically detecting garbage and road surface conditions.

### **3.8 Similarities and Differences in Methodology**

The primary similarity across these studies is the heavy reliance on Deep Learning and Computer Vision (e.g., SegNet, DeepLabv3+, Mask2Former) applied to Street View Imagery (Google, Baidu, or Mapillary). Many authors share the use of the MIT Place Pulse 2.0 dataset for training models to recognize human emotions related to urban space. The main difference lies in the granularity of indicators. While one group of authors (e.g., Ma) uses broad emotional labels to describe urban qualities, other authors (e.g., Fantous and Bennedjai) develop highly specific quantitative indices to provide actionable data for urban planning and maintenance.

## **4. Conclusion**

This review synthesized 54 studies on computer vision applications for urban monitoring, differentiated into 46 case studies and 8 literature reviews. Approximately 89% of sources were published after 2022, peaking in 2025, and China dominates the field with 52% of publications. Keyword clusters highlight Google Street View–based approaches, while Baidu Street View led as the primary data source for model testing; training relied heavily on ADE20K and MIT Place Pulse 2.0, with image segmentation as the leading method.

Beyond these technical patterns, the review shows that computer vision has primarily been used to assess urban conditions and qualities, to measure specific environmental qualities such as greenery, walkability, and cycling safety, and to detect signs of urban decay and sanitation problems such as graffiti, potholes, broken windows, and litter. These studies range from broad, perception-based labels to highly specialized indices like the Street Cleanliness Index, the Street Greenery Comprehensive Evaluation Indicator, and the Subjective Walking Preferences Index, illustrating a shift from generic emotional ratings toward more actionable, planning-oriented indicators.

## **5. Discussion**

For Urban FM, this shift is significant because it moves digital twins beyond their traditional role as static, formal-scale representations of the built environment toward dynamic digital twins that are continuously updated through image-based monitoring

and that can support immediate, condition-based response strategies. Indicators such as street cleanliness, greenery, and accessibility, once computed at scale and refreshed at high frequency, can feed directly into KPI-based FM dashboards, enabling prioritization of maintenance, cleaning, and infrastructure investment based on current rather than historic conditions. This reframes computer vision-based street-level analysis not merely as a research method, but as a candidate data layer for the continuous monitoring components of dynamic digital twins in Urban FM.

Despite these strengths in scalability, the heavy reliance on Chinese platforms and MIT-based perception datasets limits generalizability, particularly in data-scarce regions where urban form, infrastructure standards, and cultural perceptions of quality differ. The exclusion of traffic-focused studies from this review narrows its scope and may overlook integrated urban dynamics relevant to FM practice, such as the interaction between mobility infrastructure and pedestrian-level conditions. Future work should prioritize diverse, multi-scale data fusion, broader geographical representation, and richer indicators of urban change to capture both static urban conditions and dynamic transformations over time — a prerequisite for computer vision to mature into a reliable data source for dynamic digital twins in Urban FM.

## 6. References

- Amer, A. A., Abdel-Moneim, N. M., & Khalil, H. A. E. E. (2025). What deep learning can (and can't) see: A systematic review assessing applicability for quantifying qualities of urban vitality. *Computational Urban Science*, 5(1), 67.
- Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder–decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 801–818).
- Chen, M., Cai, Y., Guo, S., Sun, R., Song, Y., & Shen, X. (2024). Evaluating implied urban nature vitality in San Francisco: An interdisciplinary approach combining census data, street view images, and social media analysis. *Urban Forestry & Urban Greening*, 95, 128289.
- Chowdhury, T., Mandel, I., Ortiz, J., & Ju, W. (2024). Designing a user-centric framework for information quality ranking of large-scale street view images. *arXiv preprint arXiv:2404.00392*.
- Cinnamon, J., & Jahiu, L. (2021). Panoramic street-level imagery in data-driven urban research: A comprehensive global review of applications, techniques, and practical considerations. *ISPRS International Journal of Geo-Information*, 10(7), 471.
- Costa, M., Marques, M., Siebert, F. W., Azevedo, C. L., & Moura, F. (2023). Scoring cycling environments' perceived safety using pairwise image comparisons. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 4046–4051). IEEE.

- Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. In *Proceedings of the European Conference on Computer Vision* (pp. 196–212). Springer.
- Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design*, 14(1), 65–84.
- Fantous, A., & Bennedjai, R. (2026). Urban connectivity and walkability assessment: Spatial analysis and predictive modeling—An integrated approach using space syntax, GIS, and machine learning for pedestrian flow analysis in Batna. *GeoJournal*, 91(1), 8.
- Gönül, A., & Durak, S. (2025). On measuring the change in historical city centres: An attempt at comparing human perception and deep learning through visual quality of street space. *Multimedia Tools and Applications*, 1–23.
- Hurbean, L., Danaiaata, D., Militaru, F., Dodea, A. M., & Negovan, A. M. (2021). Open data-based machine learning applications in smart cities: A systematic literature review. *Electronics*, 10(23), 2997.
- Li, X., Li, B., & Su, Y. (2025). Temporal evolution of multidimensional built environment perceptions and street vitality: A longitudinal analysis in rapidly urbanizing cities. *Sustainability*, 17(18), 8428.
- Luo, J., Zhai, S., Song, G., He, X., Song, H., Chen, J., ... Feng, Y. (2022). Assessing inequity in green space exposure toward a “15-minute city” in Zhengzhou, China: Using deep learning and urban big data. *International Journal of Environmental Research and Public Health*, 19(10), 5798.
- Ma, Z. (2023). Deep exploration of street view features for identifying urban vitality: A case study of Qingdao City. *International Journal of Applied Earth Observation and Geoinformation*, 123, 103476.
- Sottini, V. A., Barbierato, E., Capecchi, I., Borghini, T., & Saragosa, C. (2021). Assessing the perception of urban visual quality: An approach integrating big data and geostatistical techniques. *Aestimum*, 79, 75–102.
- Vallebuena, A., & Lee, Y. S. (2023). Measuring urban quality and change through the detection of physical attributes of decay. *Scientific Reports*, 13(1), 17316.
- Wang, H., Zhang, T., Ning, X., Wei, Y., Li, G., & Yang, X. (2024). The development of comprehensive evaluation indicators of street greenery and the identification of priority greening areas: Providing more accurate greening planning. *Ecological Indicators*, 167, 112630.
- Wang, Z., Feng, J., Feng, X., Liu, X., & Zhao, X. (2022). Evaluation algorithm of urban road sanitation quality based on machine vision. In *2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)* (pp. 1007–1010). IEEE.
- Xia, D., Huang, J., Yang, J., Liu, X., & Wang, H. (2022). DuARUS: Automatic geo-object change detection with street-view imagery for updating road databases at Baidu Maps. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 3565–3574).
- Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, 127(3), 302–321.