# "Does the cafe entrance look accessible? Where is the door?" Towards Geospatial AI Agents for Visual Inquiries

Anonymous ICCV submission

Paper ID *****


Figure 1. We introduce our vision for *Geo-Visual Agents*—multimodal AI agents capable of understanding and responding to nuanced visual-spatial inquiries about the world by analyzing large-scale repositories of geospatial images combined with traditional GIS data sources. For example, *StreetViewAI* [14] (above) makes street view accessible to blind users by combining geographic context, user information, and dynamic street view images into an MLLM, accessed via an AI chat interface and accessible screen reader controls.

## Abstract

*Interactive digital maps have revolutionized how people travel and learn about the world; however, they rely on pre-existing structured data in GIS databases (e.g., road networks, POI indices), limiting their ability to address geo-visual questions related to what the world looks like. We introduce our vision for Geo-Visual Agents—multimodal AI agents capable of understanding and responding to nuanced visual-spatial inquiries about the world by analyzing large-scale repositories of geospatial images, including streetscapes (e.g., Google Street View), place-based photos (e.g., TripAdvisor, Yelp), and aerial imagery (e.g., satellite photos) combined with traditional GIS data sources. We define our vision, describe sensing and interaction approaches, provide three exemplars, and enumerate key challenges and opportunities for future work.*

## 1. Introduction

Over the last two decades, precise location sensing, pervasive internet connectivity, and interactive digital maps have transformed travel planning and in situ navigation, enabling turn-by-turn directions, location-aware search, and dynamic route optimization. Despite these advances, current mapping systems are confined to pre-existing structured geospatial data, leaving a vast repository of visual information—latent within street-level, aerial, and user-contributed imagery—untapped and inaccessible for answering what we term *geo-visual questions*. That is, visually-oriented questions about a location. Imagine, for example, a wheelchair user asking *"Are there stairs leading up to the library on 35th?"* or a blind traveler inquiring *"Where is the door to the cafe and what does it look like?"*

In this workshop paper, we introduce our vision for *Geo-Visual Agents*—multimodal AI agents capable of understanding and responding to nuanced visual-spatial inquiries about the world by analyzing large-scale repositories

of geospatial images (*e.g.,* street-level and aerial imagery) combined with traditional GIS databases (*e.g.,* road networks, POI databases, transit schedules). We envision Geo-Visual Agents acting as *"visual-spatial co-pilots"* across a spectrum of contexts from *a priori* travel planning to *in situ* navigation. Crucially, while we expect many high-value user scenarios where a Geo-Visual Agent is actively sensing and processing visual-spatial data in real-time via AR glasses [12, 33, 62] or smartphone cameras [39, 49, 56], an equally large set of questions can be answered by analyzing existing (and largely untapped) repositories of geo-related imagery—either on-demand (*e.g.,* spinning up an AI agent to query and analyze sources) or via pre-computation.

Our vision moves beyond the current paradigm of geospatial artificial intelligence (GeoAI) [17, 30, 35] such as *CARTO AI* [7] and *SuperMap* [51], which primarily focuses on large-scale data analysis for domain experts. Similarly, our work is related to but distinct from emerging paradigms in GIS research such as "*Autonomous GIS*"—AI-based scientific assistants that help *"reason, derive, innovate, and advance geospatial solutions to pressing global challenges"* [38]. Moreover, because our envisioned agents work primarily via multimodal conversational AI, we draw inspiration from recent work in *Geospatial Visual Question Answering* (GVQA) such as *MQVQA* [61] and *TAMMI* [6], which attempt to imbue multimodal LLMs with domain-specific geographic knowledge; however, again these systems are aimed at analysts and function primarily on remote aerial imagery. While related, our focus is on addressing the personal, interactive, and often immediate needs of an individual planning travel or actively navigating a space.

Below, we expand on our vision including a breakdown of visual-spatial inquiries, modalities of sensing and interaction, and three emerging examples, *StreetViewAI* [14], *Access Scout* [27], and *BikeButler*. Throughout, we highlight key opportunities and open challenges.

## 2. Geo-Visual Queries Across Travel Stages

We envision Geo-Visual Agents providing value across the full mobility cycle from pre-travel planning to *in-situ* navigation. Below, we enumerate four travel stages and opportunities for Geo-Visual Agents therein, focusing on accessibility but also broader user scenarios such as driving and biking. Selecting and fusing data sources will be a function of user task and data availability. For example, pre-travel planning may rely on streetscape images, user-contributed photos, and place-based reviews while in-situ navigation might combine these sources with visual content from a user's real-time camera feed (*e.g.,* from AR glasses) and context sensing (*e.g.,* travel mode inference, location).

**Pre-travel planning.** In this phase, the user is not physically present at a location but planning a future visit. The agent acts as a remote, interactive guide, enabling detailed investigation and reducing uncertainty before travel. For example: (1) a blind parent planning a trip to a park may ask, *"What kind of equipment does the playground have, and does it seem safe?"* (2) A person with a mobility disability virtually investigates a route and inquires *"Are there curb ramps all the way to my doctor's office?"* (3) A potential homebuyer may ask neighborhood-related questions such as *"What do the streets look like?"*, *"Are there tree-lined sidewalks?"*, and *"How much graffiti is there?"*

**While navigating.** During travel itself, the user is under cognitive and physical load, navigating their environment, making route choices, and dynamically avoiding obstacles. Here, the agent provides forward-looking information about the destination or upcoming maneuvers, enhancing situational awareness and facilitating *in situ* travel decisions. For example: (1) A driver approaching a destination asks, *"You said to turn left at the next light. Are there any landmarks?"* (2) A cyclist nearing a decision point queries, *"Is there a protected bike lane at the next intersection, and which side of the road is it on?"* (3) A rail user exiting a train asks, *"Which exit is closest to the library's accessible entrance?"*

**Destination arrival.** When arriving at a destination, the user is faced with a litany of "last 10 meters" problems related to the appearance of their destination, the path to and location of an entrance, and the presence of obstacles or safety issues. For example, (1) approaching their destination, a delivery driver may inquire *"Where is the loading zone for this building?"*; (2) a person meeting a friend in a busy plaza may ask, *"I'm looking for the coffee shop; can you describe its storefront so I can more easily spot it?"*. (3) a blind traveler's ride share arrives for pickup at a busy airport and asks, *"Can you help me find the silver Toyota Camry with license plate KNI667?"*.

**Indoor exploration.** Finally, upon entering a destination, the agent's role can shift to supporting micro-navigation through complex indoor environments like airports, stores, or office buildings. This stage presents a significant data challenge, as comprehensive visual and map datasets for indoor spaces are rare [13]. For example, (1) a customer trying to find the location of a specific item in a hardware store may ask *"Based on the aisle signs, which direction do I go to find the plumbing department?"* (2) A low-vision traveler looking at an airport departure board: *"Can you tell me which gate Delta Flight 850 is leaving from?"*; (3) A wheelchair user in a large convention center: *"Can you guide me to the nearest accessible restroom?"*

Together, these scenarios illustrate how Geo-Visual Agents can transform how we navigate and understand places, enhancing accessibility, offering landmark-based navigation, improving personal safety, and even leading to serendipitous discovery. Below, we describe potential data sources and then outline interaction modalities.

## 3. Sensing and Data Sources

The power of a Geo-Visual Agent lies in its ability to synthesize heterogeneous data sources, fusing visual evidence with structured geospatial data to form a holistic and accurate understanding of a place or route. We focus below on geo-related image sources rather than structured GIS data.

**Streetscape Imagery.** Street view imagery (SVI) [25, 37]—such as *Google Street View* (GSV), *Cyclomedia*, and *Mapillary*—provide a rich, large-scale image archive of the world. GSV alone has over 220 billion images spanning 10 million miles across 100 countries [19]. Such data can be used to analyze road conditions [3], street markings (crosswalks [2, 34], bike lanes [45]), sidewalk infrastructure (sidewalk material [23], curb ramps [21, 41]), bus stops [32], building facades [31], graffiti [52], trees and vegetation [36], neighborhood health indicators [54, 63], and more. Primary limitations include image recency [55], occlusions due to obstructing objects in front of the SVI camera (*e.g.,* buses) [47], and geographic distribution (images are distributed every 10-15 meters along roadways but not foot pathways or inside parks or buildings).

**User-Contributed Photos.** Place-based platforms like *Google Places*, *Yelp*, and *TripAdvisor* contain vast, crowdsourced libraries of photos tied to specific POIs, which provide a useful complement to SVI, including building interiors, curated (business uploaded) shots of storefronts, and pictures of menus, food [16], and social activities (*e.g.,* [60])—all which are often accompanied by user-contributed text (*e.g.,* reviews). We found, however, that analysis of such multimodal data is less common in the literature. The key limitation here is data availability, particularly for unpopular or recently opened places, and social biases in *who* uploads and *why* (*e.g.,* see [4, 58]).

**Aerial Imagery.** Aerial imagery from satellites, airplanes, or drones can provide high-resolution, top-down or oblique (45-degree angle) views of spatial structures, including building footprints, parking lots, vegetation, and pedestrian infrastructure [24]. While remote sensing and photogammetry research has existed for many decades—*e.g.,* for land use classification, agriculture, disaster response, and military analyses [29, 59]—such techniques have not been applied to the Geo-Visual Agent context (*e.g.,* answering queries about parking lot locations, rooftop restaurant patios, or unmapped pedestrian shortcuts). Similar to streetscapes, aerial imagery can suffer from occlusions (from tree cover, clouds), shadows from tall buildings, and lack of availability. In the US, high-resolution aerial imagery is often provided by the federal government such as USGS [53] and NASA [40].

**Robotic scans.** Robots such as autonomous vehicles, ground-based delivery robots, and drones [48, 50] infused with sensor suites (cameras, LiDAR) can generate high-fidelity scans of the environment, producing not just images but 3D reconstructions with mensuration [26]. While a potentially promising future data source, there is currently a lack of open data and APIs.

**Infrastructure-based Cameras.** Infrastructure-based cameras installed for traffic, weather, security, and safety monitoring provide real-time views of cities and uniquely offer dynamic information about pedestrian and car movement, human activity, weather conditions, and transient obstructions [28, 43, 46]; however, while some camera feeds are open—*e.g.,* DOT traffic cameras—most are not and privacy is a key consideration. Moreover, there is a lack of density and availability (*e.g.,* in rural areas).

**First-person Camera Streams.** Finally, first-person camera streams from AR glasses [12, 33, 62], smartphone cameras [5, 39, 49, 56], and dashcams [42, 57] are critical for in-situ travel stages, offering a real-time, egocentric view for navigation, identifying transient obstacles, and reading signs. While primarily used for immediate assistance, these streams could also help update or correct existing geospatial datasets in a continuous feedback loop (*e.g.,* [57]). However, key considerations include high computational and power requirements, robust network connectivity, and privacy concerns for both the user and bystanders.

## 4. Processing and Interpreting with AI

Our vision relies not just on diverse forms of geospatial imagery and pre-existing GIS data but also advances in multimodal AI (*e.g.,* scene understanding [9, 11], object affordances [22, 33], and spatial reasoning [8, 10, 15, 44]) to extract semantic information and object relationships. While some analyses could be pre-computed for known high-value entities (*e.g.,* presence and location of curb ramps [21, 41]), we expect a long-tail of bespoke queries, which will require a Geo-Visual Agent to seek out, analyze, and synthesize image-based sources with pre-existing metadata in GIS databases in real-time.

## 5. Delivering the Answers

Finally, a crucial aspect of our vision is *how* the agent delivers information, which is a function of the user's abilities, their current context, and the complexity and type of data. Regardless of delivery mode, agents need to report uncertainty and data provenance to build trust and mitigate error.

**Audio-First Interfaces:** For hands-free and/or eyes-free operation—essential for drivers, cyclists, and blind and low vision users—audio interfaces are critical (*e.g.,* using earbuds or a smart speaker). The challenge, however, is providing well-structured verbal descriptions to convey complex visual information without overwhelming the user.

**Multimodal Interfaces:** Agents should also select and show relevant imagery. For instance, after describing an entrance, the agent could display a photo of the door

(*e.g.,* drawn from SVI or Yelp). The challenge lies in the AI's ability to select the most appropriate photo(s)—appropriately cropped—from large archives.

**AI-Generated Abstracted Visualizations:** For highly complex spatial information, a raw photo or a long verbal description may be insufficient. An exciting frontier is the agent's ability to generate simplified, abstract diagrams on the fly—akin to a modern *LineDrive* system [1]. Making these abstractions accessible, perhaps tactilely, is also a critical area of open research.

## 6. Case Study Applications

To help showcase and concretize our vision, we highlight three emerging Geo-Visual Agent prototypes.

**StreetViewAI.** Current SVI tools are inaccessible to blind users. Our group is addressing this problem through the design of *StreetViewAI* [14] (Figure 1), which uses context-aware, real-time AI to support virtually exploring routes, inspecting destinations, or even remotely visiting tourist locations such as the Grand Canyon [18]. StreetViewAI provides accessible interactive controls for blind users to pan and move between panoramic images and dynamically converse with a live, multimodal AI agent about the scene and local geography. In a lab study, blind users effectively used StreetViewAI to virtually navigate streetscapes. Key challenges: reconciling users' mental models of SVI, a tendency to over-trust AI, and the difficulty of synthesizing rich visual data into concise audio.

*AI Agent.* StreetViewAI employs three separate AI subsystems. Most relevant is the *AI Chat Agent*, which allows for conversational interactions about the user's current and past street views as well as nearby geography. The agent uses Google's *Multimodal Live API* [20], which supports real-time interaction, function calling, and retains memory of all interactions within a single session. When the user initiates a chat either via typing or speaking, we transmit each GSV interaction along with the user's current view and geographic context (*e.g.,* nearby places, current heading). Thus, they can ask about local geography, current and past views, and object relationships (*e.g., "where is the entrance?"*).

**Accessibility Scout.** Assessing the accessibility of unfamiliar environments is a critical but often laborious job for people with disabilities. While standardized checklists exist, they often fail to account for an individual's unique and evolving needs. *Accessibility Scout* [27] is an LLM-based system designed to address this gap by generating personalized accessibility scans from images—*e.g.,* from *TripAdvisor*, *Yelp*, and *Airbnb*—to identify potential concerns based on self-reported abilities and interests. In user studies, we found that Accessibility Scout's personalized scans were more useful than generic ones and that its collaborative Human-AI approach was effective and built trust.

*AI Agent.* The Accessibility Scout pipeline begins by creating a structured user model in JSON format, initialized from a user's plain text description of their abilities and preferences. To assess an environment, the agent mimics how users assess environmental accessibility by first analyzing an image and the user's intent (*e.g., "going on a date"*) to identify potential tasks a user might perform, such as *"dining"* or *"toileting"*. The agent then decomposes these tasks into primitive motions like *"grabbing"* that are required to complete them. For each task, the agent analyzes the user model, task information, and segmented image to identify and describe environmental concerns. Crucially, the system is designed for Human-AI collaboration; users can provide feedback on identified concerns which the agent uses to update the user model.

**BikeButler.** Existing mapping tools define optimal bike routes using objective data like distance and elevation, but often ignore subjective qualities related to a cyclist's comfort and perceived safety. However, a desirable bike route depends on factors not found in standard GIS databases, such as the presence of tree-lined streets, pavement quality, or bike lane widths. *BikeButler* is an early-stage prototype Geo-Visual Agent that generates personalized cycling routes by fusing structured data from *OpenStreetMap* with visual analysis of SVI. The system creates routes optimized for a user's specific profile (*e.g.,* beginner, expert) and allows them to rate route segments, creating a feedback loop that refines their preferences for future journeys.

## 7. Discussion and Conclusion

In this paper, we introduced our vision for Geo-Visual Agents, dynamic and conversational AI co-pilots that can see and reason about the world in real-time. Our envisioned agents answer nuanced visual questions about the visual world—from a blind user navigating a complex intersection to a cyclist seeking the safest, most pleasant route. Our prototypes offer an initial window into this vision, offering personalized, interactive experiences extending far beyond current mapping services.

Still, significant challenges remain, including: (1) *Dynamic information synthesis*: creating agents that can intelligently select, fuse, and reason over a heterogeneous set of real-time and archived data sources; (2) *Trust and transparency*: communicating uncertainty and data provenance; (3) *Speech UIs*: effectively verbalizing complex visual information concisely via text or speech; (4) *Personalization* learning from a user's unique needs and preferences; (5) *Spatial reasoning* accurately tracking and modeling spatial relationships between objects; (6) *Generative spatial abstractions*: dynamically generating spatial visualizations to help aid understanding.

Addressing these challenges will require a concerted effort across disciplines from computer vision and HCI to accessibility and geospatial science. Join us!

# References

[1] Maneesh Agrawala and Chris Stolte. Rendering effective route maps: improving usability through generalization. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, page 241–249, New York, NY, USA, 2001. Association for Computing Machinery. 4

[2] Dragan Ahmetovic, Roberto Manduchi, James M. Coughlan, and Sergio Mascetti. Mind your crossings: Mining gis imagery for crosswalk localization. *ACM Trans. Access. Comput.*, 9(4), 2017. 3

[3] Shazab Ali, Meng Xu, and Daehan Kwak. Smart roadway monitoring: Pothole detection and mapping via google street. In *Internet Computing and IoT and Embedded Systems, Cyber-physical Systems, and Applications: 25th International Conference, ICOMP 2024, and 22nd International Conference, ESCS 2024, Held as Part of the World Congress in Computer Science, Computer Engineering and Applied Computing, CSCE 2024, Las Vegas, NV, USA, July 22–25, 2024, Revised Selected Papers*, page 151. Springer Nature, 2025. 3

[4] V. Antoniou and A. Skopeliti. Measures and indicators of vgi quality: An overview. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W5: 345–351, 2015. 3

[5] Apple. Detect doors around you using Magnifier on iPhone, 2025. 3

[6] Hichem Boussaid, Lucrezia Tosato, Flora Weissgerber, Camille Kurtz, Laurent Wendling, and Sylvain Lobry. Visual question answering on multiple remote sensing image modalities. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, pages 2319–2328, 2025. 2

[7] CARTO. Genai — ai-powered spatial insights, 2025. 2

[8] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465, 2024. 3

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. 3

[10] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, pages 135062–135093. Curran Associates, Inc., 2024. 3

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[12] Alexander Fiannaca, Ilias Apostolopoulous, and Eelke Folmer. Headlock: a wearable navigation aid that helps blind cane users traverse large open spaces. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*, page 19–26, New York, NY, USA, 2014. Association for Computing Machinery. 2, 3

[13] Jon E. Froehlich, Anke M. Brock, Anat Caspi, João Guerreiro, Kotaro Hara, Reuben Kirkham, Johannes Schöning, and Benjamin Tannert. Grand challenges in accessible maps. *Interactions*, 26(2):78–81, 2019. 2

[14] Jon E. Froehlich, Alex Fiannaca, Nimer Jaber, Victor Tsaran, and Shaun Kane. Streetviewai: Making street view accessible using context-aware multimodal ai. In *The 38th Annual ACM Symposium on User Interface Software and Technology*, page 22, New York, NY, USA, 2025. ACM. 1, 2, 4

[15] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning, 2024. 3

[16] Alessandro Gambetti and Qiwei Han. Aigen-foodreview: A multimodal dataset of machine-generated restaurant reviews and images on social media, 2024. 3

[17] Google. Google earth ai: Our state-of-the-art geospatial ai models. https://blog.google/technology/ai/google-earth-ai/, 2025. Accessed: August 1, 2025. 2

[18] Google. Treks: Grand canyon, 2025. 4

[19] Google. Celebrate 15 years of exploring your world on Street View, 2025. 3

[20] Google. Vertex ai multimodal live api, 2025. 4

[21] Kotaro Hara, Jin Sun, Robert Moore, David Jacobs, and Jon Froehlich. Tohme: detecting curb ramps in google street view using crowdsourcing, computer vision, and machine learning. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, page 189–204, New York, NY, USA, 2014. Association for Computing Machinery. 3

[22] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Comput. Surv.*, 54(3), 2021. 3

[23] Maryam Hosseini, Fabio Miranda, Jianzhe Lin, and Claudio T. Silva. Citysurfaces: City-scale semantic segmentation of sidewalk materials. *Sustainable Cities and Society*, 79: 103630, 2022. 3

[24] Maryam Hosseini, Andres Sevtsuk, Fabio Miranda, Roberto M. Cesar, and Claudio T. Silva. Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery. *Computers, Environment and Urban Systems*, 101:101950, 2023. 3

[25] Yujun Hou and Filip Biljecki. A comprehensive framework for evaluating the quality of street view imagery. *International Journal of Applied Earth Observation and Geoinformation*, 115:103094, 2022. 3

[26] Dingkun Hu and Jennifer Minner. Uavs and 3d city modeling to aid urban planning and historic preservation: A systematic review. *Remote Sensing*, 15(23), 2023. 3

[27] William Huang, Xia Su, Jon E. Froehlich, and Yang Zhang. Accessibility scout: Personalized accessibility scans of built

environments. In *The 38th Annual ACM Symposium on User Interface Software and Technology*, page 18, New York, NY, USA, 2025. ACM. 2, 4

[28] Gaurav Jain, Basel Hindi, Zihao Zhang, Koushik Srinivasula, Mingyu Xie, Mahshid Ghasemi, Daniel Weiner, Sophie Ana Paris, Xin Yi Therese Xu, Michael Malcolm, Mehmet Kerem Turkcan, Javad Ghaderi, Zoran Kostic, Gil Zussman, and Brian A. Smith. Streetnav: Leveraging street cameras to support precise outdoor navigation for blind pedestrians. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA, 2024. Association for Computing Machinery. 3

[29] Bhargavi Janga, Gokul Prathin Asamani, Ziheng Sun, and Nicoleta Cristea. A review of practical ai for remote sensing in earth sciences. *Remote Sensing*, 15(16), 2023. 3

[30] Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. Geoai: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, 34(4):625–636, 2020. 2

[31] Hyejin Kim, Seula Park, and Jiyoung Kim. A study on barrier-free entrance object detection using deep learning in street view imagery. In *2024 IEEE International Conference on Big Data (BigData)*, pages 8716–8718, 2024. 3

[32] Minchu Kulkarni, Chu Li, Jaye Jungmin Ahn, Katrina Oi Yau Ma, Zhihan Zhang, Michael Saugstad, Kevin Wu, Yochai Eisenberg, Valerie Novack, Brent Chamberlain, and Jon E. Froehlich. Busstopcv: A real-time ai assistant for labeling bus stop accessibility features in streetscape imagery. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, New York, NY, USA, 2023. Association for Computing Machinery. 3

[33] Jaewook Lee, Andrew D. Tjahjadi, Jiho Kim, Junpu Yu, Minji Park, Jiawen Zhang, Jon E. Froehlich, Yapeng Tian, and Yuhang Zhao. Cookar: Affordance augmentations in wearable ar to support kitchen tool interactions for people with low vision. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA, 2024. Association for Computing Machinery. 2, 3

[34] Meiqing Li, Hao Sheng, Jeremy Irvin, Heejung Chung, Andrew Ying, Tiger Sun, Andrew Y Ng, and Daniel A Rodriguez. Marked crosswalks in us transit-oriented station areas, 2007–2020: A computer vision approach using street view imagery. *Environment and Planning B: Urban Analytics and City Science*, 50(2):350–369, 2023. 3

[35] Wenwen Li and Chia-Yu Hsu. Geoai for large-scale image analysis and machine vision: Recent progress of artificial intelligence in geography. *ISPRS International Journal of Geo-Information*, 11(7), 2022. 2

[36] Xiaojiang Li, Carlo Ratti, and Ian Seiferling. Quantifying the shade provision of street trees in urban landscape: A case study in boston, usa, using google street view. *Landscape and Urban Planning*, 169:81–91, 2018. 3

[37] Yongchang Li, Li Peng, Chengwei Wu, and Jiazhen Zhang. Street view imagery (svi) in the built environment: A theoretical and systematic review. *Buildings*, 12(8), 2022. 3

[38] Zhenlong Li, Huan Ning, Song Gao, Krzysztof Janowicz, Wenwen Li, Samantha T. Arundel, Chaowei Yang, Budhendra Bhaduri, Shaowen Wang, A-Xing Zhu, Mark Gahegan, Shashi Shekhar, Xinyue Ye, Grant McKenzie, Guido Cervone, and Michael E. Hodgson. Giscience in the era of artificial intelligence: A research agenda towards autonomous gis, 2025. 2

[39] Alice Lo Valvo, Daniele Croce, Domenico Garlisi, Fabrizio Giuliano, Laura Giarré, and Ilenia Tinnirello. A navigation and augmented reality system for visually impaired people. *Sensors*, 21(9), 2021. 2, 3

[40] National Aeronautics and Space Administration and U.S. Geological Survey. Landsat data access. https://landsat.gsfc.nasa.gov/data/data-access/, 2025. Free access to Landsat satellite imagery archive dating back to 1972. Joint NASA-USGS program providing continuous Earth observation data. 3

[41] John S. O'Meara, Jared Hwang, Zeyu Wang, Michael Saugstad, and Jon E. Froehlich. Rampnet: A two-stage pipeline for bootstrapping curb ramp detection in streetscape images from open government metadata. In *Workshop on Vision Foundation Models and Generative AI for Accessibility: Challenges and Opportunities at ICCV 2025*. IEEE, 2025. Workshop Paper. 3

[42] Sangkeun Park, Joohyun Kim, Rabeb Mizouni, and Uichin Lee. Motives and concerns of dashcam video sharing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, page 4758–4769, New York, NY, USA, 2016. Association for Computing Machinery. 3

[43] Yurii Piadyk, Joao Rulff, Ethan Brewer, Maryam Hosseini, Kaan Ozbay, Murugan Sankaradas, Srimat Chakradhar, and Claudio Silva. Streetaware: A high-resolution synchronized multimodal urban scene dataset. *Sensors*, 23(7), 2023. 3

[44] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12977–12987, 2024. 3

[45] Luís Rita, Ricky Nathvani, Miguel Peliteiro, Tudor-Codrin Bostan, Emily Muller, Esra Suel, A. Barbara Metzler, Tiago Tamagusko, and Adelino Ferreira. Using deep learning and google street view imagery to assess and improve cyclist safety in london. *Sustainability*, 15(13), 2023. 3

[46] Joao Rulff, Giancarlo Pereira, Maryam Hosseini, Marcos Lage, and Claudio Silva. Towards data-informed interventions: Opportunities and challenges of street-level multimodal sensing, 2024. 3

[47] Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, and Jon Froehlich. Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery. 3

[48] Hunsoo Song, Joshua Carpenter, Jon E. Froehlich, and Jinha Jung. Accessible area mapper for inclusive and sustainable

urban mobility: A preliminary investigation of airborne point clouds for pathway analysis. In *1st ACM SIGSPATIAL Workshop on Sustainable Mobility (SuMob 2023)*, 2023. 3

[49] Xia Su, Han Zhang, Kaiming Cheng, Jaewook Lee, Qiaochu Liu, Wyatt Olson, and Jon E. Froehlich. Rassar: Room accessibility and safety scanning in augmented reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2024. Association for Computing Machinery. 2, 3

[50] Xia Su, Ruiqi Chen, Jingwei Ma, Chu Li, and Jon E. Froehlich. Flymethrough: Human-ai collaborative 3d indoor mapping with commodity drones. In *The 38th Annual ACM Symposium on User Interface Software and Technology*, page 14, New York, NY, USA, 2025. ACM. 3

[51] SuperMap. Ai gis, 2025. 2

[52] Eric K. Tokuda, Roberto M. Cesar, and Claudio T. Silva. Quantifying the presence of graffiti in urban environments. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–4, 2019. 3

[53] U.S. Geological Survey. Earthexplorer. https://earthexplorer.usgs.gov/, 2025. Query and order satellite images, aerial photographs, and cartographic products. Provides access to over 40 years of Landsat data and various aerial photography collections. 3

[54] Zeyu Wang, Koichi Ito, and Filip Biljecki. Assessing the equity and evolution of urban visual perceptual quality with time series street view imagery. 145:104704. 3

[55] Zeyu Wang, Yingchao Jian, Adam Visokay, Don MacKenzie, and Jon E. Froehlich. Street view for whom? an initial examination of google street view's urban coverage and socioeconomic indicators in the us. Under review, 2025. Submitted for review. 3

[56] Chris Yoon, Ryan Louie, Jeremy Ryan, MinhKhang Vu, Hyegi Bang, William Derksen, and Paul Ruvolo. Leveraging augmented reality to create apps for people with visual disabilities: A case study in indoor navigation. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, page 210–221, New York, NY, USA, 2019. Association for Computing Machinery. 2, 3

[57] Aziza Zhanabatyrova, Clayton Frederick Souza Leite, and Yu Xiao. Automatic map update using dashcam videos. *IEEE Internet of Things Journal*, 10(13):11825–11843, 2023. 3

[58] Guiming Zhang and A-Xing Zhu. The representativeness and spatial bias of volunteered geographic information: a review. *Annals of GIS*, 24(3):151–162, 2018. 3

[59] Lefei Zhang and Liangpei Zhang. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):270–294, 2022. 3

[60] Mengxia Zhang and Lan Luo. Can consumer-posted photos serve as a leading indicator of restaurant survival? evidence from yelp. *Management Science*, 69(1):25–50, 2023. 3

[61] Meimei Zhang, Fang Chen, and Bin Li. Multistep question-driven visual question answering for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023. 2

[62] Yuhang Zhao, Elizabeth Kupferstein, Brenda Veronica Castro, Steven Feiner, and Shiri Azenkot. Designing ar visualizations to facilitate stair navigation for people with low vision. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, page 387–402, New York, NY, USA, 2019. Association for Computing Machinery. 2, 3

[63] Shengyuan Zou and Le Wang. Detecting individual abandoned houses from google street view: A hierarchical deep learning approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:298–310, 2021. 3