Identifying Crucial Objects in Blind and Low-Vision Individuals' Navigation

MD TOUHIDUL ISLAM, Pennsylvania State University, USA
IMRAN KABIR, Pennsylvania State University, USA
ELENA ARIEL PEARCE, Drake University, USA
MD ALIMOOR REZA, Drake University, USA
SYED MASUM BILLAH, Pennsylvania State University, USA

This paper presents a curated list of 90 objects essential for the navigation of blind and low-vision (BLV) individuals, encompassing road, sidewalk, and indoor environments. We develop the initial list by analyzing 21 publicly available videos featuring BLV individuals navigating various settings. Then, we refine the list through feedback from a focus group study involving blind, low-vision, and sighted companions of BLV individuals. A subsequent analysis reveals that most contemporary datasets used to train recent computer vision models contain only a small subset of the objects in our proposed list. Furthermore, we provide detailed object labeling for these 90 objects across 31 video segments derived from the original 21 videos. Finally, we make the object list, the 21 videos, and object labeling in the 31 video segments publicly available. This paper aims to fill the existing gap and foster the development of more inclusive and effective navigation aids for the BLV community.

ACM Reference Format:
Md Touhidul Islam, Imran Kabir, Elena Ariel Pearce, Md Alimoor Reza, and Syed Masum Billah. 2024. Identifying Crucial Objects in Blind and Low-Vision Individuals' Navigation. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24), October 27–30, 2024, St. John's, NL, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3663548.3688538

1 MOTIVATION

Navigating urban environments presents significant challenges for blind and low-vision (BLV) individuals. While advancements in computer vision offer potential solutions through real-time object detection systems, existing datasets for training the underlying models in these systems are often lacking in accessibility-specific annotations. Most contemporary datasets, such as ImageNet [9] and MS-COCO [20], though extensive, do not include critical objects and features that are essential for BLV navigation, such as curb cuts, sidewalk conditions, and specific indoor landmarks. This gap hinders the development of robust navigation aids tailored to the needs of BLV individuals.

To address this issue, we curate a list of 90 objects crucial for BLV navigation by analyzing 21 publicly available videos featuring BLV individuals in various settings. This list is later refined through feedback from focus groups involving BLV individuals and their sighted companions. Our analysis indicates that contemporary datasets cover only a small subset of our proposed objects. We also provide object labeling for the 90 objects across 31 video segments derived from the original videos. By making our object list, the 21 videos, and the labeled video segments publicly available [16][1], we aim to bridge this gap and support the development of more inclusive navigation aids for the BLV community.

[1]https://github.com/Shohan29531/BLV-Road-Nav-Accessibility

2 BACKGROUND AND RELATED WORK

Large datasets contain thousands of images annotated with object/class names, bounding rectangles around objects, textual descriptions, and per-pixel semantic labels. As AI tasks evolve—from object classification and detection to scene understanding, VQA, and semantic segmentation—the effort in annotating data increases, leading to smaller datasets. ImageNet [9] is a classic image classification dataset with over 14M images and 1000 classes. In contrast, MS-COCO [20], popular for scene understanding, has 328K images and 91 classes. Visual Question Answering (VQA) datasets are an exception; for example, VQA v2.0 [12] has 200K images, 614K questions, and 10 answers per question. Semantic segmentation datasets are much smaller due to the laborious nature of pixel-wise annotation [3, 6, 23].

Most large-scale, publicly available datasets are not accessibility-aware, with a few exceptions (e.g.,[14, 29, 34]). They lack annotations crucial for people with disabilities. For instance, snow on sidewalks, areas near curb cuts, and fire hydrants are essential safety information for BLV individuals but such details are missing in mainstream datasets. Theodorou et al. reported that models trained on generic datasets fail to detect objects in pictures taken by blind users, as these images tend to be blurrier and may show out-of-frame objects[34]. The KITTI-15 [3] dataset, used for evaluating AI in autonomous driving, also lacks accessibility-aware annotations like curb cuts, ramps, and crosswalks.

The lack of accessibility awareness in large-scale public datasets is consequential but unsurprising. Most annotators are sighted crowd workers who often lack a nuanced understanding of disability [35] and may ignore individuals with disabilities as "outliers" or label them generically (e.g., "a person with a disability") [8, 21, 22]. Incorporating accessibility awareness in datasets is challenging. Prior attempts [26, 34] revealed several challenges: engaging with target communities, providing accessible data collection tools, ensuring data quality, and balancing data structure with demands on collectors. Consequently, accessibility-focused datasets like VizWiz [14] have significantly fewer annotations (e.g., 30K images in VizWiz vs. over 200K in common VQA datasets [12]).

A possible workaround is to extract data from publicly available video-sharing platforms like YouTube and Vimeo. Previous studies have analyzed YouTube videos to understand people with vision impairments [30–32, 38]. Following this approach, we collected 21 videos from YouTube and Vimeo featuring BLV individuals.

3 IDENTIFICATION OF OBJECTS WITH ACCESSIBILITY IMPACT

Video Collection. We collected free, publicly available videos from YouTube and Vimeo using a systematic keyword-based search. Some keywords used in the search were "blind", "vision impairment", and "visually impaired", along with the terms "white cane", "navigating", "orientation and mobility", and "training". We briefly reviewed the videos from search results and selected videos relevant to blind navigation. In total, we collected 21 suitable videos, 16 from YouTube and 5 from Vimeo (Table 2).

Ethical Considerations. As per the Common Rule of the federal regulations on human subjects research protections (45 CFR 46.104(d)(4)), the collection of existing data is exempt from IRB review when the sources are publicly available. Nonetheless, we contacted the video uploaders on YouTube and Vimeo to give them credit and to solicit their permission.

Identification of Objects with Accessibility Impact. Two researchers watched all videos and noted when relevant objects first appeared in the video in a Google Sheet. An object is deemed relevant (a) if it is on the way of the blind individual featured in the video and affects their course (e.g., they changed direction, they collided with the object); (b) if it provides meaningful feedback (e.g., produces different sounds or tactile feedback); (c) if it affects them physically (e.g., they tripped); (d) if it could be a hindrance in the future (e.g., the electrical box on a powered gate that sticks out into the sidewalk); (e) if it could be mislabeled as something else (e.g., the escalator could be mislabeled as stairs which could lead to someone being hurt from not expecting them to be moving); and (f) if it could be thought of or labeled in the wrong context (e.g., a parked car on the sidewalk could lead a person to believe they were in the street).

This process generated around 80 object categories related to the accessibility of sidewalks and non-visual navigation.

4 REVISING OBJECTS WITH ACCESSIBILITY IMPACT: A FOCUS GROUP STUDY

We conducted a focus group study with Six participants to revise our initial list of 80 objects and determine whether any objects were missing or redundant. This section presents the study's methodology, results, and implications for design.

4.1 Procedure

Participants. Among the six study participants: two were blind, two were low-vision, and two were sighted individuals. The two sighted participants were experts in non-visual navigation because of their professions and proximity to blind individuals. Participants were chosen based on their familiarity with AI-based applications, such as "Seeing AI" [1], "Be My Eyes" [2], and "Oko" [25]. Among the six, two were males. The average age was 49.17 (SD 7.45).

Two researchers conducted the study; while one researcher asked predefined questions on possibly important objects (from the initial list) for accessible navigation, the other researcher took notes and identified new objects of interest. The session lasted around an hour.

Prompts. We established common ground by describing an imaginary AI app similar to "Seeing AI" [1] and "Aira" [4] that can detect and read out surrounding objects. We generated the following questions/prompts: ***Prompt 1***: *Let us imagine we have an AI app that can read out a scene in front of you. What objects would you want the AI to describe proactively?* ***Prompt 2***: *We have compiled a list of around 80 objects that we felt would be important in navigation for blind individuals. Would you please mention whether an object is relevant or not when we read it out?* ***Prompt 3***: *Please add more objects that are not on our list but you feel would be important for the scenario (i.e., non-visual navigation). Please also provide a rationale for your choice.*

4.2 Feedback on the Initial List

Below, we summarize study participants' object-specific opinions and the relative priority of the objects in our preliminary list.

Sidewalk Objects. Participants unanimously agreed on the importance of detecting objects that share the sidewalk, such as *bicycles*, *wheelchairs*, *pets*, and *water hoses*. *Bicycles* and *pets* were highlighted as top priorities. While one participant noted that a loose *water hose* is a tripping hazard, two others felt their cane could detect it, preferring the AI to focus on other objects.

Sidewalk Obstructions. Participants marked all sidewalk obstructions as important, with *growing tree branches* and *barrier posts* receiving the highest priority.

Traffic Signals and Signs. Participants unanimously agreed on the importance of detecting *traffic signals* and *signs*, except for speed limit signs, which they deemed unnecessary for blind individuals. They emphasized the need for accurate detection of *pedestrian crossing signals*, especially with aural warnings, citing apps like "Oko" [25]. However, they expressed concerns about over-reliance on AI tools, with one participant highlighting the importance of including legal disclaimers, referencing the lengthy terms of the Oko app.

Indoor Objects. Participants emphasized the need for accurate indoor object detection, particularly for *chairs, tables, elevators*, and *moving walks*, which are crucial in venues like restaurants, theaters, and public transport. While *escalators* and *stairs* were noted as important, they were considered less critical due to their prominence and ease of detection.

4.3 Revised Taxonomy and Design Implications

This section presents the revised object taxonomy and discusses design implications for an AI application supporting navigation for blind and low-vision individuals.

Newly Identified Objects. Throughout our study, participants frequently named items they considered important for blind and low-vision navigation. Many items matched our list, often with slight name variations. Discussions also revealed new objects, which we added to our revised taxonomy, such as *doorways*, *black ice*, *cobblestone pavements*, and *moving walks*. Participants found some items redundant, which we removed. We finalized a list of 90 objects, categorized in Table 1.

Group	Parent Concept	Accessibility-related objects
1	Attributes of a sidewalk and driveway	Accent Paving, Driveway (flat), Puddle, Raised Entryway, Sidewalk, Sidewalk Pits, Sloped Driveway, Tactile Paving, Brick Paving, Cobblestone Paving, Unpaved Sidewalk, Wet Surface
2	Obstructions likely to be detected by a white cane	Fire hydrant, Gutter, Vegetation, Tree, Brick Wall, Fence, Trash Bins, Lamp Post, Pole, Mailbox
3	Obstructions less likely to be detected by a white cane	Closed Sidewalk, Barrier Post, Barrier Stump, Foldout Sign, Bench
4	Objects that are too late to be detected by a white cane	Train Tracks, Train Platform
5	Objects that pick you before you pick them	Overhanging Tree Branches
6	Objects that provide navigational guidance	Retaining Wall, Railing, Wall, Curved Railing
7	Objects not supposed to be on the sidewalk	Hose, Maintenance Vehicle, Trash on Roads, Snow, Water Leakage, Yard Waste, Water Pipes
8	Moving objects sharing the sidewalk	Person, Bicycle, Wheelchair, Person with a Disability, White Cane, Dog, Guide Dog, Street Vendor
9	Intersection	Pedestrian Crossing, Slopped Curb, Intersection, Crosswalk, Curb, Bridge, Uncontrolled Crossing
10	Objects on the road shoulder	Road Shoulder, Roadside Parking, Parallel Parking Spot, Paratransit Vehicle
11	Objects on the road	Road, Unpaved Road, Bus, Car, Motorcycle, Road Divider
12	Traffic signals and street signs	Traffic Signals, Stop Sign, Sign, Sign Post, Push Button, "Use the Other Door" Sign, Toilet Sign
13	Objects related to building exits and entrances	Gate, Flush Door, Doorway
14	Indoor objects	Counter, Elevator, Escalator, Fountain, Stairs, Uneven Stairs, Table, Building, Moving Walk, Pillar, Chair
15	Objects related to public transit	Bus Stop, Turnstile

Table 1. Key accessibility-related objects, classified into different groups.

AI Tools Are Not a Replacement of Physical Assistance Devices. Our study participants emphasized that AI tools should not replace common physical assistance devices such as white canes. One said: *"I have my concerns about using a phone [all the time] for lack of reliability. Having a physical object such as a white cane to guide is always preferred."* Our participants also reported that while they would appreciate the detection of some objects via AI beforehand, they want their cane, which they trust more, to take care of most of the detection tasks.

Information Priority. Our participants agreed that navigation-related objects should be prioritized. For instance, they noted that elevators are more important than escalators in complex indoor setups due to difficulty locating elevators. They suggested ignoring trash smaller than four inches and small pits that do not disrupt walking. Wet surface warnings are redundant if it is already raining. The presence of a driveway is crucial, but whether it is flat or sloped is less important. Vehicles on the road should not be detected. And if an object is within reach of a white cane, AI assistance should remain silent.

Configurable Information Presentation. Participants desire the ability to choose which objects they are alerted to and for the AI tool to learn from their usage habits, such as turning off specific object detection frequently opted out by the user. They also want the tool to adjust its detection and recommendations based on their mode of travel (e.g., with a guide dog, with a white cane, or in specific weather conditions).

One participant wanted the tool to provide weather-appropriate dressing advice, including recommendations for heavy jackets or raincoats. Others believed the tool should issue cautions if the weather is ideal for icing and emphasized the importance of detecting ice beforehand, though they acknowledged the difficulty for AI, given the challenge for humans. One participant suggested the tool could use recent weather data and temperature warnings near the icing point to predict which roads might have ice.

Objects that Pick You Before the Cane Picks Them. Participants reported specific objects creating significant navigational challenges for blind and low-vision individuals. Tree branches extending over sidewalks pose a major threat, often existing at head level or higher and undetectable by canes, leading to frequent injuries. Another challenging object is a train track. One participant noted that detecting a train track with a cane when a train is approaching is already too late.

5 ANALYZING THE OBJECT LIST: COVERAGE IN PROMINENT DATASETS

An AI model's effectiveness in assisting BLV individuals' navigation hinges on accurately detecting objects from our proposed list. However, object recognizers are limited to items in their training datasets, such as ImageNet [9], MS COCO [20], Mapillary Vistas [24], Kitti [11], Cityscapes [7], Pascal VOC [10], PFB [28], and ADE20K [40, 41]. Among these, Mapillary Vistas is the most advanced, offering detailed annotations for 66 outdoor object categories [24], but it and other datasets lack coverage for many objects on our list. Figure 1 highlights this gap in these datasets.

Our study findings indicate that objects in groups such as "*Obstructions less likely to be detected by a white cane (group 3)*", "*Objects that pick you before you pick them (group 5)*", and "*Objects not supposed to be on the sidewalk (group 7)*" are the most significant. Failure to detect these objects can lead to accidents or injuries for blind individuals.

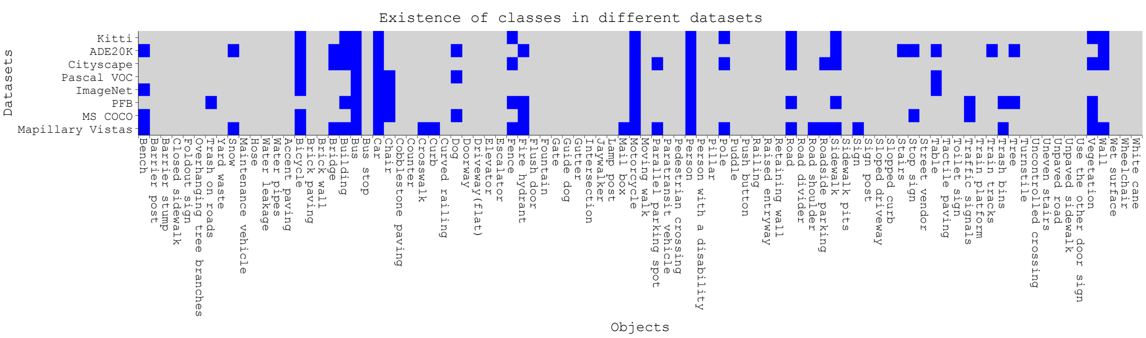

Fig. 1. A heatmap representing the existence of different objects of our list in prominent datasets. in a cell means the corresponding object exists in the corresponding dataset. In contrast, means the object does not exist in the corresponding dataset.

The first nine objects in Figure 1 (from *Bench* to *Water Pipes*) are the most significant objects that belong to the aforementioned groups 3, 5, and 7. Only three of these objects exist in some datasets, such as Mapillary Vistas, ADE20K, ImageNet, and PFB—highlighting their limitations in this aspect.

6 OBJECT LABELING

We picked 31 video segments from 21 videos based on the diversity of content, object density, and key transitions, and created their object labeling manually. The authors visually inspected and annotated keyframes by labeling objects' presence (1) or absence (0) in each frame, using a differential approach that took less than 60 seconds per frame after the first keyframe. We make these object-labeling, the 21 videos, and the 90 objects publicly available [16]. Appendix A.2 contains more details regarding the labeling process. Preliminary evaluations of seven well-known models using our object labeling are presented in Appendix B.

REFERENCES

[1] [n. d.]. Seeing AI. https://www.microsoft.com/en-us/seeing-ai/

[2] 2015. Be My Eyes: Bringing sight to blind and low vision people. https://www.bemyeyes.com/

[3] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. 2018. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision* 126, 9 (2018), 961–972.

[4] Aira. 2018. Aira. https://aira.io/. Retrieved May 23, 2020 from https://aira.io

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on computer vision*.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.

[8] Lennard J. Davis. 2016. *The Disability Studies Reader (5th ed.)*. Routledge. https://doi.org/10.4324/9781315680668

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88 (2010), 303–338.

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3354–3361.

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Tanmay Gupta, A. Kamath, Aniruddha Kembhavi, and Derek Hoiem. 2021. Towards General Purpose Vision Systems. *ArXiv* abs/2104.00743 (2021).

[14] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3608–3617.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision (ICCV)*.

[16] Md Touhidul Islam, Imran Kabir, Elena Ariel Pearce, Md Alimoor Reza, and Syed Masum Billah. 2024. A Dataset for Crucial Object Recognition in Blind and Low-Vision Individuals' Navigation. arXiv:2407.16777 [cs.CV] https://arxiv.org/abs/2407.16777

[17] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven CH Hoi. 2022. Lavis: A library for language-vision intelligence. *arXiv preprint arXiv:2209.09019* (2022).

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.

[19] Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. 2021. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2042–2051.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[21] Simi Linton. 1998. *Claiming disability: Knowledge and identity*. NYU Press.

[22] Meredith Ringel Morris. 2020. AI and Accessibility. *Commun. ACM* 63, 6 (2020), 35–37.

[23] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. 2017. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 5000–5009.

[24] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. 2017. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*. 4990–4999.

[25] OKO. 2023. OKO makes every intersection accessible. https://www.ayes.ai/oko

[26] Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. 2021. Designing an Online Infrastructure for Collecting AI Data From People With Disabilities. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 52–63.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[28] Stephan R. Richter, Zeeshan Hayder, and Vladlen Koltun. 2017. Playing for Benchmarks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*.

[29] Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, et al. 2019. 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*. 1–14.

[30] Woosuk Seo and Hyunggu Jung. 2017. Exploring the community of blind or visually impaired people on YouTube. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 371–372.

[31] Woosuk Seo and Hyunggu Jung. 2018. Understanding blind or visually impaired people on youtube through qualitative analysis of videos. In *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video*. 191–196.

[32] Woosuk Seo and Hyunggu Jung. 2021. Understanding the community of blind or visually impaired vloggers on YouTube. *Universal Access in the Information Society* 20 (2021), 31–44.

[33] Daniel J Simons and Daniel T Levin. 1997. Change blindness. *Trends in cognitive sciences* 1, 7 (1997), 261–267.

[34] Lida Theodorou, Daniela Massiceti, Luisa Zintgraf, Simone Stumpf, Cecily Morrison, Edward Cutrell, Matthew Tobias Harris, and Katja Hofmann. 2021. Disability-first Dataset Creation: Lessons from Constructing a Dataset for Teachable Object Recognition with Blind and Low Vision Data Collectors. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–12.

[35] Garreth W Tigwell. 2021. Nuanced perspectives toward disability simulations from digital designers, blind, low vision, and color blind people. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[36] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7464–7475.

[37] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* 43, 10 (2020), 3349–3364.

[38] Jingyi Xie, Na Li, Sooyeon Lee, and John M Carroll. 2022. YouTube Videos as Data: Seeing Daily Challenges for People with Visual Impairments During COVID-19. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*. 218–224.

[39] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. 2023. Recognize Anything: A Strong Image Tagging Model. *arXiv preprint arXiv:2306.03514* (2023).

[40] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 633–641.

[41] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision* 127 (2019), 302–321.

A APPENDIX: VIDEO ANALYSIS AND OBJECT LABELING DETAILS

A.1 Video Analysis

To thoroughly analyze the collected videos, we split each video into small clips of variable lengths between five and ninety-five seconds. Each clip revolves around the appearance of objects that are significant in people's navigation on roads and sidewalks. We refer to each clip as a video segment. Table 2 shows the number of segments created from each video. Using a keyframe extraction tool called *Katna*[2], we further segmented the clips into keyframes. The keyframes are characterized as the representative frames of a video stream that serve a precise and concise summary of the video content, considering transitions in the scene and changes in lighting conditions and activities. The number of keyframes (i.e., images) extracted from the video segments was between three and ninety-three. Afterward, we manually annotated the presence and absence of objects (of our final list) for a subset of keyframes extracted from the video segments. Let us denote the final list of objects as L_u for future reference. Afterward, we manually labeled each video frame for each object in L_u.

A.2 Object Labeling

All authors of this paper visually inspected the keyframes to generate object labeling. Each author annotated a subset of video segments by observing the changes between two consecutive keyframes. For each frame, the existence (denoted by 1) or absence (denoted by 0) of all the objects that belong to L_u was annotated. If In a keyframe F_k the existence of

[2]https://katna.readthedocs.io/en/latest/

ID	Title/Context	Duration	# Segments	# Annotated Seg.	Year	Location	URL
V1	Blind Man Walking	2:24	5	2	2011	London	https://youtu.be/RmsoHyMRtbg
V2	following a blind person for a day | JAYKEEOUT	7:02	1	1	2021	Seoul	https://youtu.be/dPisedvLKQQ
V3	Orientation & Mobility for the Blind-1*	0:00-10:00	8	2	2012	—	https://youtu.be/Gkf5tEbP-oo
V4	Orientation & Mobility for the Blind-2*	10:01-19:10	4	3	2012	—	https://youtu.be/Gkf5tEbP-oo?t=602
V5	My First Blind Cane Adventure to Get Coffee | Did I Succeed or Give Up*	10:00	3	1	2019	Caribbean Cruise Ship	https://youtu.be/SZM-Le6MEE0
V6	Using A White Cane | Legally Blind*	10:00	2	1	2018	—	https://youtu.be/TxUxbXyh7Y4
V7	How a Blind Person Uses a Cane	4:18	4	1	2013	—	https://youtu.be/xi0JMS1rulo
V8	Orientation mobility	9:36	2	1	2022	—	https://youtu.be/6u53Q7IvVIY
V9	TAKING THE METRO AND WALKING THROUGH MADRID ALONE AND BLIND-1*	9:19	4	1	2020	Madrid	https://youtu.be/Vx3-ltp9p-Y
V10	TAKING THE METRO AND WALKING THROUGH MADRID ALONE AND BLIND-2*	10:00 - 19:00	1	1	2020	Madrid	https://youtu.be/Vx3-ltp9p-Y?t=600
V11	Mobility and Orientation Training for Young People with Vision Impairment	5:48	3	1	2019	Edinburgh	https://youtu.be/u-3GlbJ5RMc
V12	Mobility and Orientation	8:49	4	1	2018	New York City	https://vimeo.com/296488214
V13	Traveling with the white cane	2:14	3	1	2009	Maryland	https://vimeo.com/2851243
V14	Blindness Awareness Month - Orientation and Mobility with ELC and 1st Grade Students	5:52	5	2	2022	—	https://vimeo.com/758153786
V15	The White Cane documentary	5:40	3	1	2021	—	https://vimeo.com/497359578
V16	Craig Eckhardt takes the subway on Vimeo	4:43	4	1	2010	New York	https://vimeo.com/17293270
V17	Guide Techniques for people who are blind or visually impaired*	10:00	3	2	2015	—	https://youtu.be/iJfxkBOekvs
V18	Russia: Blind Commuter Faces Obstacles Every Day	3:20	6	2	2013	Moscow	https://youtu.be/20W2ckx-BcE
V19	The "Challenges" you may not know about "Blind" People | A Day in Bright Darkness	8:00	6	2	2016	Malaysia	https://youtu.be/xdyj1Is5IFs
V20	Blind Challenges in a Sighted World	3:54	5	2	2017	—	https://youtu.be/3pRWq8ritc8
V21	What to expect from Orientation & Mobility Training (O&M) at VisionCorps	2:21	7	2	2012	Pennsylvania	https://youtu.be/wU7b8rwr2dM

Table 2. List of our collected videos. We cropped the YouTube videos using https://streamable.com, which has a crop limit of 10 mins.

the object $O \in L_u$ is denoted by a function E, the E can be defined as follows:

$$E(O) = \begin{cases} 1, & \text{if } O \text{ exists in } F_k \\ 0, & \text{otherwise} \end{cases}$$

Now the annotation of keyframe F_k can be written as, $(E(O) \mid O \in L_u)$. Since visually inspecting changes can be subject to "*change blindness*"—a phenomenon when the visual feed is momentarily interrupted by a blank screen [33]—we avoided this by viewing two consecutive keyframe pairs on the screen side-by-side and glancing between them.

For each video segment, annotating the first keyframe took the longest, typically 5-7 minutes. For subsequent frames, we only looked at what objects newly appeared/disappeared from the previous frame and adjusted the annotations accordingly. Such differential annotation took less than 60 seconds in most cases. The annotation task required more time if a new frame appeared with a completely different background or camera viewport. With an average of around 15 keyframes per video segment, we needed around ≈ 20 minutes to label the objects a video. Each video segment was annotated by at least two researchers independently. Later, we resolved the conflicts in annotation through collaboration.

A.3 Analyzing Annotated Data

The bar chart in Figure 2 depicts the distribution of objects in our annotated dataset.

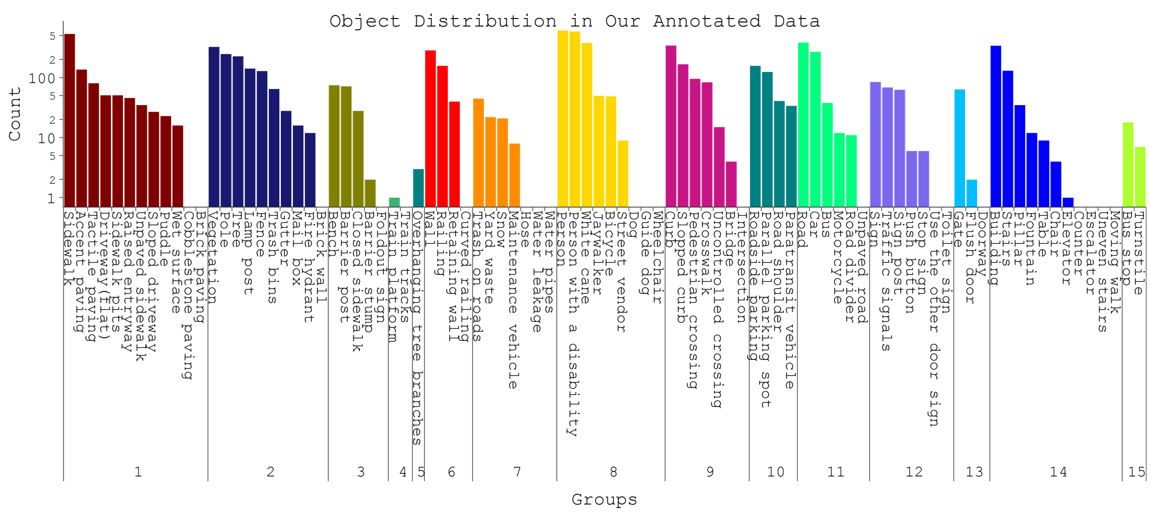

Fig. 2. Bar chart representing the object distribution in our annotated data. Each bar represents the number of keyframes in which an object (as labeled on the x-axis) was present. The X-axis also shows the id of the parent concept or group (as described in Table 1) to which each object belongs. The Y-axis is in logarithmic scale.

B APPENDIX: PRELIMINARY EVALUATIONS USING OUR OBJECT LABELING

B.1 Model Selection

We evaluated the preparedness of several state-of-the-art computer vision models for potential application in aiding blind and low-vision individuals utilizing our object labeling. More specifically, we selected models specialized in addressing four distinct computer vision tasks: i) object recognition, ii) object detection, iii) semantic segmentation, and iv) visual question answering (VQA). Table 3 shows the chosen models alongside their respective computer vision genre. We chose a total of seven models from these four types. For our experiments, all models were adapted to solve the recognition task. The adaptation process is explained in Sec. B.2.

Type	Models
Recognition Model	RAM (Recognize Anything Model) [39]
Detection Model	Faster R-CNN [27], YOLO V7 [36]
Segmentation Model	HRNet V2 [37], Mask R-CNN [15]
VQA (Visual Question Answering) [5] Model	GPV-1 [13], BLIP [17, 18]

Table 3. Selected models for evaluation and their type.

While a recognition model only predicts the object categories present in a given image, detection models go beyond category prediction and additionally provide rectangular bounding boxes encompassing the objects. On the other hand, semantic segmentation models detect fine-grained object masks (exact shapes of the objects) along with their categories. Traditionally, these models are trained on a set of pre-defined object categories and, hence, can not predict objects that were not present in the vocabulary set of training images. More recently, the Recognize Anything Model (RAM) can predict open vocabulary categories due to its language decoder [39]. Finally, VQA models consist of language encoders and decoders, enabling them to predict answers to open-ended questions. VQA models can also perform tasks such as image captioning, object localization, and image descriptions. Among all the models, RAM and VQA models are suitable for blind and low-vision individuals because of their open-ended capabilities.

B.2 Experiments and Observations

Experiment Procedure. Recognition models provide a list of predicted objects for a given image, which can be directly used for evaluation. However, other model types provide output in different formats requiring some pre-processing.

Recognition from Detection Model. Traditionally, detection and segmentation models predict bounding boxes and masks, respectively, for each of the categories. If an object is present in a given image, then the detection model will provide us with the pixel coordinates of bounding boxes and the categories of detected objects. We ignored the bounding boxes in this experiment since we only needed the object categories.

Recognition from Segmentation Model. In contrast, the segmentation model predicts a mask for each object category. If an object exists in a frame, its corresponding mask will contain some nonzero values; otherwise, the whole mask will be a matrix of zeros. Thus, We easily created a list of objects that were present in a given image by analyzing the nonzero values of the masks.

Recognition from VQA Model. For VQA models, we asked them a question for each of the objects regarding whether the object is present or not. Generally, visual questions are formed from broad categories of question types such as (i) recognition, (ii) common sense, (iii) reasoning, (iv) OCR, and (v) counting. State-of-the-art VQA models perform well when asked visual recognition questions; however, these models are prone to failure when questions from other categories [19] are asked. To test our selected VQA models, we asked simple visual recognition questions. We generated a set of questions, each focusing on one accessibility-related object as shown in Table 1. All of our questions followed a predominant structure like "Is there object X in the scene?", where X is an object that came from Table 1. Consequently, the answer from the VQA model was either "yes" or "no". By analyzing the answers, we constructed the list of objects that were in the given image. We passed all the extracted keyframes (mentioned in Section **??**) to all the models we chose for this experiment and generated predictions following the steps mentioned earlier in this paragraph. Later, we matched the prediction of each model against our annotation.

Model	N	Precision (Micro Avg.)	Recall (Micro Avg.)	F1 (Micro Avg.)
BLIP	90	0.3366	0.8263	0.4783
GPV-1	90	0.2273	0.8070	0.3547
RAM	54	0.8175	0.2715	0.4077
YOLO V7	12	0.9437	0.1570	0.2692
HRNet V2	15	0.6110	0.3998	0.4834
Mask R-CNN	12	0.6077	0.1801	0.2779
Faster R-CNN	12	0.6029	0.1837	0.2815

Table 4. Precision, Recall, and F1 score (For all three metrics, higher is better) of all the selected models (shown in Table 3) over our annotated keyframes. *N* column shows the number of object categories the corresponding model can predict.

Results and Observations. We calculated the micro average precision, recall, and F1 score of each of the chosen models. All the scores are reported in Table 4. The table also shows the number of objects each model can predict under the *N* column, out of L_u objects in our dataset. From the table, we can observe that the detection and segmentation models can only predict labels for 12 to 15 objects from the list L_u—which explains their relatively poor performance. The lack of representative annotations in their training dataset explains their inability to predict a larger number of object categories.

In contrast, BLIP, GPV-1, and RAM have a built-in language encoder/decoder. Therefore, they were able to predict more object categories. As we asked questions about the existence of each of the 90 object categories of list L_u to BLIP and GPV-1, they generated answers for each question. On the other hand, RAM, being an open-ended object recognition model, could predict 54 object categories from our list. BLIP achieved the highest F1 score over 90 object categories among these three models.

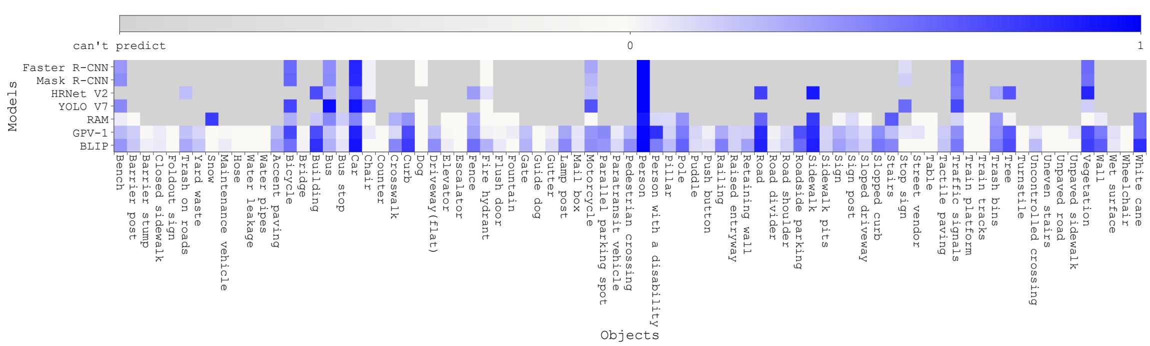

Fig. 3. A heatmap representing the classwise F1 score of all the selected models (shown in Table 3).

We also calculated classwise F1 scores of all the selected models and plotted a heatmap using these scores to visualize the classwise performance of all the models. Figure 3 shows the heatmap representing the classwise F1 score of all seven models. In this figure, in a cell means the corresponding model can not predict the corresponding object. The

F1 score is within the [0, 1] range and is represented by the intensity of the blue color ( ) in Figure 3. The heatmap shows that RAM, BLIP, and GPV-1 can recognize more objects than the detection and segmentation models. These models can accurately predict a subset of object categories listed in L_u, such as persons, cars, buildings, roads, sidewalks, etc. (object columns where the intensity of blue is higher).

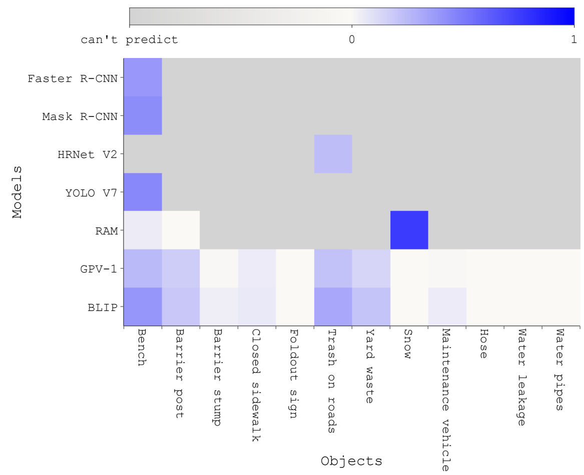

Fig. 4. A heatmap representing the classwise F1 score of all the selected models for the objects of groups 3, 5, and 7 (shown in Table 1).

Moreover, none of the models can recognize the objects of the most significant groups (groups 3, 5, and 7) well. Figure 4 represents the F1 scores of all the models for the objects of those three groups. The experiments described in this section suggest that any AI tool based on the selected models is not ready to be directly used in blind or low-vision individuals' navigation.