Towards Fine-Grained Sidewalk Accessibility Assessment with Deep Learning: Initial Benchmarks and an Open Dataset

Xinlei Liu*
Kevin Wu*
xllegion@cs.uw.edu
kwu2004@cs.uw.edu
Allen School of Computer Science
University of Washington, USA

Minchu Kulkarni
Allen School of Computer Science
University of Washington, USA
minchu@cs.uw.edu

Michael Saugstad
Allen School of Computer Science
University of Washington, USA
saugstad@cs.uw.edu

Peyton Rapo
Allen School of Computer Science
University of Washington, USA
peytor01@cs.uw.edu

Jeremy Freiburger
Allen School of Computer Science
University of Washington, USA
jeremyfr@uw.edu

Maryam Hosseini
City and Regional Planning
University of California, Berkeley
maryamh@berkeley.edu

Chu Li
Allen School of Computer Science
University of Washington, USA
chuchuli@cs.uw.edu

Jon E. Froehlich
Allen School of Computer Science
University of Washington, USA
jonf@cs.uw.edu

Figure 1: We examine whether deep learning models can classify sidewalk accessibility conditions from pre-cropped 640x640 streetscape images—*e.g.,* whether a curb ramp is *too steep*, *too narrow*, or *missing a tactile indicator* or if a sidewalk panel is *uneven*, *bumpy*, or composed of *brick/cobblestone*. The grid above showcases all 33 conditions we attempt to infer.

Abstract

We examine the feasibility of using deep learning to infer 33 classes of sidewalk accessibility conditions in pre-cropped streetscape images, including *bumpy*, *brick/cobblestone*, *cracks*, *height difference (uplifts)*, *narrow*, *uneven/slanted*, *pole*, and *sign*. We present two experiments: first, a comparison between two state-of-the-art computer vision models, Meta's *DINOv2* and OpenAI's *CLIP-ViT*, on a cleaned dataset of ~24k images; second, an examination of a larger but noisier crowdsourced dataset (~87k images) on the best performing model from Experiment 1. Though preliminary, Experiment 1 shows that certain sidewalk conditions can be identified with high precision and recall, such as missing tactile warnings on curb ramps and grass grown on sidewalks, while Experiment 2 demonstrates that larger but noisier training data can have a detrimental effect on performance. We contribute an open dataset and classification benchmarks to advance this important area.

*Both authors contributed equally to this research.

ASSETS '24, October 27–30, 2024, St. John's, NL, Canada

ACM ISBN 979-8-4007-0677-6/24/10
https://doi.org/10.1145/3663548.3688531

CCS Concepts

• **Human-centered computing → Accessibility technologies**;
• **Computing methodologies → Computer vision**.

Keywords

Sidewalk accessibility, computer vision, human mobility, obstacle detection, DINOv2, ViT-CLIP

ACM Reference Format:
Xinlei Liu, Kevin Wu, Minchu Kulkarni, Michael Saugstad, Peyton Rapo, Jeremy Freiburger, Maryam Hosseini, Chu Li, and Jon E. Froehlich. 2024. Towards Fine-Grained Sidewalk Accessibility Assessment with Deep Learning: Initial Benchmarks and an Open Dataset. 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/10.1145/3663548.3688531

1 Introduction

Ensuring that sidewalks are safe and accessible to all is a key US priority [19] and a goal of the UN's *New Urban Agenda* [25]. A looming challenge, however, is the lack of scalable data collection techniques to assess and map the condition of pedestrian environments throughout the world [9]. Emerging work in urban studies and accessibility have trained state-of-the-art computer vision models to find and identify pedestrian-related features using streetscape imagery, such as crosswalks, curb ramps, and obstacles [1, 8, 10, 12, 15, 26]. While promising and scalable, these models only detect features, they do not assess condition—for example, they detect curb ramps but not whether the ramp has a tactile warning strip or whether there is sufficient landing space for a wheelchair. Some recent work has examined sidewalk condition assessment; however, it has taken a narrower scope, such as classifying the sidewalk material (*e.g.,* asphalt, cobblestone) [13].

In this paper, we explore the feasibility of classifying pre-cropped streetscape images into 33 sidewalk conditions (or tags) using state-of-the-art deep learning models. For training and testing, we use data derived from Project Sidewalk, an open-source sidewalk accessibility labeling tool currently deployed in 21 cities across eight countries [22]. We present two experiments: first, a comparison between two state-of-the-art computer vision models, Meta's *DINOv2* and OpenAI's *CLIP-ViT*, on a cleaned dataset of ∼24k images; second, an examination of a larger but noisier crowd-sourced dataset (∼87k images) on the best performing model from Experiment 1. Though preliminary, Experiment 1 shows that certain sidewalk conditions can be identified with high precision and recall, such as missing tactile warning on curb ramps and grass grown on sidewalks, while Experiment 2 demonstrates that larger but noisier training data can have a detrimental effect on performance. Both our datasets and analysis code are released as open source on GitHub[1].

Our overarching goal is twofold: first, to advance the field of automated streetscape analysis and establish performance benchmarks for sidewalk condition assessment; second, inspired by the *VizWiz Challenge* [6, 11, 18], to provide two open datasets to spur future research and enable performance comparisons.

[1]https://github.com/ProjectSidewalk/sidewalk-tagger-ai

2 Dataset

Our datasets derive from the open source crowdsourcing tool, Project Sidewalk (https://projectsidewalk.org) [22]. In Project Sidewalk (PS), online users are given interactive missions to locate, label, and tag sidewalk and crosswalk accessibility features and problems in interactive Google Street View (GSV) images. Currently, Project Sidewalk is deployed in 21 cities across eight countries with over 1 million image-based sidewalk accessibility labels and 693k validations across 11k street miles. For validations, users are shown labels by other users and vote on their correctness by selecting *agree*, *disagree*, or *unsure*.

Project Sidewalk uses a hierarchical labeling approach. Users first apply one of seven high-level label types: *curb ramp*, *pedestrian signal*, *crosswalk*, *missing curb ramp*, *obstacle*, *surface problem*, and *missing sidewalk*. Each label has an associated set of 5-11 tags, which can optionally be applied. For example, surface problem tags include *grass*, *cracks*, *uneven/slanted*, *sand/gravel*, *etc.*—see Tables 2-5 in the Appendix. In this paper, we attempt to automatically infer these tags given a label type and a pre-cropped 640×640 image around the center position of that label. We aim to create new Human-AI interfaces in Project Sidewalk that recommend tags to the user, help automatically validate previously applied tags, or back-fill missing tags for labels already in the Project Sidewalk database.

For our experiments, we attempt to classify 33 tags across four label categories: *curb ramp*, *crosswalk*, *obstacle*, and *surface problem*. We created two datasets drawn from 10 and 12 cities, respectively: (1) a cleaned dataset (Dataset 1) of 24,009 labels and 29,311 tags and (2) a raw dataset (Dataset 2) of 87,495 labels and 66,875 tags—see Table 1. For Dataset 1, four research assistants iteratively cleaned and verified each label and tag. In total, 16,424 tags were changed (7,988 tags added), suggesting an originally noisy dataset (Table 1). For Dataset 2, we subsampled raw labels directly from Project Sidewalk with a positive crowdsourced validation score (*i.e.,* # agree votes > # disagree votes) across the 12 cities.

In summary, each data point in our training and test set contains: (1) a 640×640 streetscape image center-cropped around the user's label belonging to one of the four PS categories (*curb ramp*, *crosswalk*, *surface problem*, or *obstacle*); and (2) PS category-specific tags (Figure 1 and Tables 2-5). Download the dataset on our GitHub.

3 Experiment 1

In Experiment 1, we examine the feasibility of using custom-trained, state-of-the-art deep learning models to classify sidewalk accessibility conditions given a 640×640 image crop of one of four categories (*curb ramp*, *crosswalk*, *surface problem*, or *obstacle*). We selected three open source models for our early experiments: (1) Ultralytics' YOLOv8 [2] [14] designed for fast, real-time applications, (2) Meta's DINOv2 [3] [20], a recent advancement in *Vision Transformer*-based models (ViT) specifically designed for self-supervised learning; and (3) OpenAI's CLIP ViT (pretrained on LAION-2B, ImageNet-12k, fine-tuned on ImageNet-1k) [4] [5, 7, 23], which combines a *Contrastive Language-Image Pre-training* with ViT for image encoding. In our initial experiments we noticed that even the largest YOLOv8

[2]https://github.com/ultralytics/ultralytics
[3]https://github.com/facebookresearch/dinov2
[4]https://huggingface.co/timm/vit_base_patch16_clip_224.laion2b_ft_in12k_in1k

	Dataset 1						Dataset 2		Training/Test Sets		
	Raw Labels	Lbl Cat Changed	Cleaned Labels	Raw Tags	Tags Changed	Cleaned Tags	Labels	Tags	Train Set Exper 1	Train Set Exper 2	Test Set
Curb Ramp	11,061	204	10,857	5,784	6,953	9,459	43,352	16,685	8,674	43,352	2,183
Surface Problem	9,654	562	9,092	12,592	6,704	14,540	26,370	36,840	7,282	26,370	1,810
Obstacle	2,497	64	2,433	2,336	1,878	3,972	10,150	10,363	1,954	10,150	479
Crosswalk	1,638	11	1,627	611	889	1,340	7,623	2,987	1,306	7,623	321
Total	**24,850**	**841**	**24,009**	**21,323**	**16,424**	**29,311**	**87,495**	**66,875**	**19,216**	**87,495**	**4,793**

Table 1: An overview of the two datasets. The cleaned dataset (Dataset 1) consists of 24,009 labels and 29,311 tags; The raw crowdsourced dataset (Dataset 2) consists of 87,495 labels and 66,875 tags. Both Experiment 1 and 2 used the same test dataset to enable comparison. *Lbl cat changed* stands for "Label categories changed" and indicates the number of instances where the RAs did not agree with the label category and removed it from the dataset.

model did not perform as well as the other two models, hence, we chose to use DINOv2 and CLIP ViT for our subsequent analysis.

Implementation. We adopted a multi-label classification approach, as each image crop could possess zero, one, or multiple tags. Because each PS label type has its own unique set of tags, we trained separate models for each label type and split the data into 80% training and 20% test sets. To train the DINOv2 model, we used the B/14 Distilled backbone and pre-trained weights, Adam optimizer with a learning rate of 1e-6, binary cross entropy as the loss function, and a batch size of four. The 640×640 crops were resized to 256×256 for optimizing computation and each model was trained for 100 epochs. For the CLIP-ViT model, we used the *ViT-B/16* pre-trained weights and followed the same training protocol as DINOv2. Since CLIP was pre-trained on 224×224 pixel images, we also resized the 640×640 crops accordingly to ensure compatibility. For both DINOv2 and CLIP-ViT, we saved the best model at each epoch with the highest accuracy, prioritizing lower loss in cases of ties. All training was done using Pytorch framework on an Alienware m18 R2 with NVIDIA® GeForce RTX™ 4080, 12 GB GPU.

Results. We present Experiment 1 results using standard metrics including precision, recall, mean average precision (mAP), and F1 scores. To compute the optimal confidence level with balanced precision and recall, we identified the confidence threshold that maximized the F1 score, with a minimum threshold of 0.3. Tags with fewer than 10 instances in the test set were excluded. To account for the imbalance in our tags, we computed macro, micro and weighted averaged F1 scores [17, 24]. See Appendix A.1 for derivation details.

As shown in Figure 2, DINOv2 slightly outperformed CLIP-ViT across all key metrics. For example, Obstacle tags achieved a mAP of 0.71 with DINO *vs.* 0.68 with CLIP as well as a weighted-F1 of 0.73 *vs.* 0.70. The most significant performance was observed in the crosswalk category, with the sharpest difference in the macro-F1 score (0.60 *vs.* 0.48). Within category, the macro is generally lower than the micro and weighted F1 scores since it treats all tags equally, regardless of frequency. This difference highlights the impact of tag imbalance, where minority classes underperform. However, the *obstacles* model shows more consistent performance, as indicated by the close macro and micro F1 scores in both DINOv2 (0.68 *vs.* 0.64) and CLIP-ViT (0.64 *vs.* 0.62).

Diving into DINOv2, the best performing model overall, 13 of the 33 tags (40%) had weighted F1 scores above 0.7. The highest performing tag for each label type included: *missing tactile warning* (*F1*=0.94) for curb ramps, *brick/cobblestone* (0.91) for surface problems, *parked car* (0.93) for obstacles, and *paint fading* (0.8) for crosswalks. The tags with the lowest scores were *steep* (*F1*=0) for curb ramps; *utility panel* (0) for surface problems; *narrow* (0.3) for obstacles; and *bumpy* (0.45) for crosswalks. See detailed Experiment 1 result tables in the Appendix (Tables 7-10).

To more deeply understand DINOv2's performance, we qualitatively analyzed classification errors. We selected the top two most frequently occurring tags for each label type in our test set—*e.g., pole* (*N*=114) and *trash/recycling* cans (*N*=92) for obstacles—and analyzed the top 30 false positive (FP) and false negative (FN) classifications (as sorted by classification confidence). Similar to related work [8, 12, 26], we found image-related issues such as shadows, overexposure, low contrast, and faint/distant features as well as interclass similarity (*e.g.,* tree appears like a pole), viewpoint occlusion, and atypical forms/textures. More work is needed to address these limitations.

4 Experiment 2

While Experiment 1 helps establish a performance baseline using a manually-cleaned dataset, Experiment 2 explores the impact of a larger but noisier crowdsourced dataset. Because DINOv2 outperformed CLIP-ViT above, we focus solely on the former here. Data quality is, of course, essential for training robust models [3, 4, 21] but collecting high-quality data is expensive and laborious—*e.g.,* to create Dataset 1, four research assistants spent over 100 hours.

Implementation. In Experiment 2, we trained an additional DINOv2 model following the same protocol as Experiment 1 but using the larger, uncleaned Dataset 2 for training (Table 1). To enable comparison across the two experiments, the test dataset was the same as Experiment 1.

Results. Overall, with the larger but noisier dataset, performance dropped across all four key metrics—for example, the weighted F1 score dropped from 0.62 to 0.3 for curb ramp tags and 0.68 to 0.36 for crosswalk tags. Interestingly, surface problem and obstacle tags experienced a smaller decline: 0.71 to 0.66 and 0.73 to 0.66, respectively. With the cleaned training dataset (Dataset 1), 13 tags achieved weighted F1 scores $\geq$ 0.7. In Experiment 2, this drops to 8. While some tags were largely unaffected (*e.g., grass* dropped from 0.9 to 0.88, *trash* from 0.88 to 0.84) or even improved (*e.g., tree*

A

		Curb Ramp	Cross walk	Surface Problem	Obstacle
	tags cnt	8	4	10	11
DINOv2 (Dataset 1)	mAP	0.34	0.68	0.54	**0.71**
	macro-F1	0.31	0.60	0.50	**0.68**
	micro-F1	0.67	0.67	**0.70**	0.64
	weighted-F1	0.62	0.68	0.71	**0.73**
CLIP-ViT (Dataset 1)	mAP	0.30	0.60	0.51	**0.68**
	macro-F1	0.29	0.48	0.48	**0.64**
	micro-F1	0.63	0.61	**0.67**	0.62
	weighted-F1	0.58	0.61	0.68	**0.70**
DINOv2 (Dataset 2)	mAP	0.26	0.48	0.45	**0.68**
	macro-F1	0.14	0.22	0.43	**0.59**
	micro-F1	0.32	0.4	0.65	**0.71**
	weighted-F1	0.3	0.36	**0.66**	**0.66**

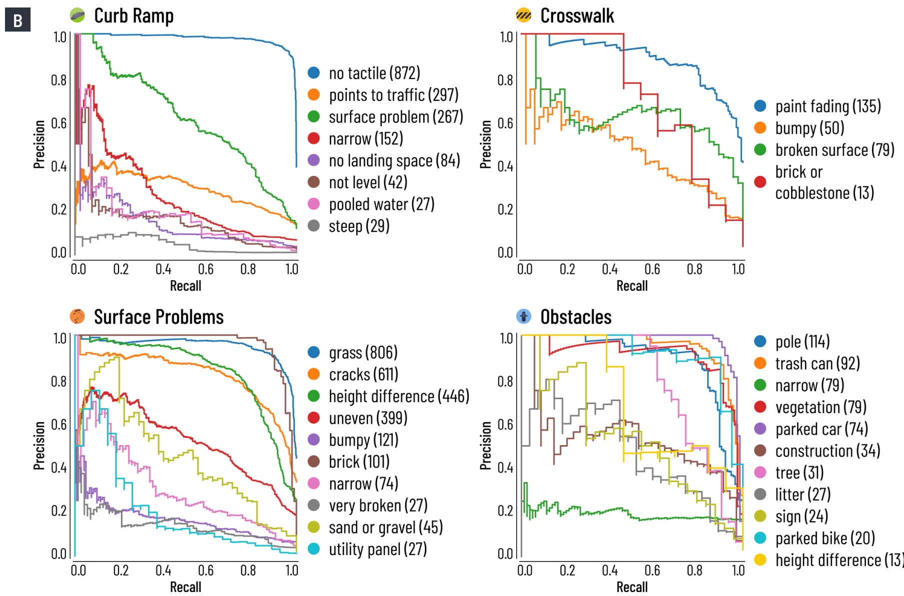

Figure 2: (A) Overall classification results for Experiment 1 (Dataset 1) and Experiment 2 (Dataset 2). F1 scores computed at a 0.3 confidence threshold. (B) The Experiment 1 precision-recall curves across the four label type categories and 33 tags. The legend shows tags sorted by frequency in the test set (the parenthetical shows the occurrence count of the tag in the test set).

Figure 3: To better understand DINOv2 performance, we visually analyzed Experiment 1 errors. We selected the top two most frequently occurring tags for each label type in our test set and analyzed the top 30 FPs and FNs (as sorted by classification confidence). One exception: for curb ramps, we selected *missing tactile strip* (*N*=872) and the third most common tag *surface problem* (*N*=297) because the second most common *points into traffic* (*N*=297) had a low F1 score (0.25).

from 0.75 to 0.79, *vegetation* from 0.84 to 0.89), others decreased significantly (*e.g., brick/cobblestone* went from 0.91 to 0.51, *paint fading* dropped from 0.8 to 0.58). These results suggest that more training data alone is not better.

5 Discussion and Conclusion

In this paper, we investigate the feasibility of assessing sidewalk and crosswalk conditions using state-of-the-art CV models. Our primary contribution is in establishing an open image dataset and initial performance benchmarks to enable future research in sidewalk condition classification. Below, we contextualize our findings, enumerate limitations, and outline directions for future work.

Similar to prior work [2, 21], our findings suggest that investing in obtaining high-quality training data is important. Our results show that a smaller (~24k) but cleaner dataset outperforms a much larger but noisier (~87k) training dataset. Still, even with the best performing model (DINOv2) and the clean training dataset (Dataset 1), only 13 of 33 tags achieved weighted F1 scores of 0.7 or better. So, while we have seen remarkable CV improvements in applications related to autonomous driving, face/pose classification, and other high interest areas, the same is not yet so for pedestrian-related infrastructure and disability. Our hope is that our paper provides a positive step in drawing attention to this area and establishing benchmarks to spur future research. Future work should also conduct more in-depth analyses of trade-offs between the dataset size and quality to optimize curation strategies.

In both Experiment 1 and 2, we trained individual multi-label binary classification models for each label category (curb ramp, crosswalk, obstacle, and surface problem). Future research should develop a unified multi-class and multi-label model capable of simultaneously classifying multiple accessibility issues given a precropped image. In addition, PS includes other metadata such as severity; the ideal classification model would infer not just condition but also severity—which would help cities better triage and prioritize problems to fix and enable more personalized routing algorithms in mapping tools.

Our dataset exhibits a long-tail tag distribution. Future work should focus on techniques to handle such imbalanced data effectively to improve robustness and generalizability. While we believe our open dataset and initial custom-trained CV models are an important contribution to the urban studies and accessibility fields, a longer-term aim is to incorporate these models back into Project Sidewalk itself. Like the recent *LabelAId* system [16], our CV models could provide crowdworkers with real-time labeling and validation suggestions—*e.g.,* by recommending a tag as they are labeling.

Acknowledgments

This project was funded by the Pacific Northwest Transportation Consortium (PacTrans), a U.S. DOT University Transportation Center, and the NSF #2125087. We thank Favyen Bastani and Ranjay Krishna for their discussions and feedback about this work.

References

[1] Marc A. Adams, Christine B. Phillips, Akshar Patel, and Ariane Middel. 2022. Training Computers to See the Built Environment Related to Physical Activity: Detection of Microscale Walkability Features Using Computer Vision. *International Journal of Environmental Research and Public Health* 19, 8 (2022). https://doi.org/10.3390/ijerph19084548

[2] Breck, Polyzotis, Roy, Whang, and Zinkevich. 2019. Data Validation for Machine Learning. In *Proceedings of Machine Learning and Systems*. https://arxiv.org/pdf/2203.02155

[3] Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. 2022. The Effects of Data Quality on Machine Learning Performance. arXiv:2207.14529

[4] Haihua Chen, Jiangping Chen, and Junhua Ding. 2021. Data Evaluation and Enhancement for Quality Improvement of Machine Learning. *IEEE Transactions on Reliability* 70, 2 (2021), 831–847. https://doi.org/10.1109/TR.2021.3070863

[5] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2022. Reproducible scaling laws for contrastive language-image learning. arXiv:2212.07143

[6] Pierre Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A. Young, and Brian Belgodere. 2022. Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge. *Journal of Artificial Intelligence Research* 73 (Jan. 2022), 437–459. https://doi.org/10.1613/jair.1.13113

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929

[8] Michael Duan, Shosuke Kiami, Logan Milandin, Johnson Kuang, Michael Saugstad, Maryam Hosseini, and Jon E. Froehlich. 2022. Scaling Crowd+AI Sidewalk Accessibility Assessments: Initial Experiments Examining Label Quality and Cross-city Training on Performance. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 82, 5 pages. https://doi.org/10.1145/3517428.3550381

[9] Jon E. Froehlich, Anke M. Brock, Anat Caspi, João Guerreiro, Kotaro Hara, Reuben Kirkham, Johannes Schöning, and Benjamin Tannert. 2019. Grand challenges in accessible maps. *Interactions* 26, 2 (feb 2019), 78–81. https://doi.org/10.1145/3301657

[10] Jon E. Froehlich, Yochai Eisenberg, Maryam Hosseini, Fabio Miranda, Marc Adams, Anat Caspi, Holger Dieterich, Heather Feldner, Aldo Gonzalez, Claudina De Gyves, Joy Hammel, Reuben Kirkham, Melanie Kneisel, Delphine LabbÉ, Steve J. Mooney, Victor Pineda, ClÁUdia PinhÃO, Ana RodrÍGuez, Manaswi Saha, Michael Saugstad, Judy Shanley, Ather Sharif, Qing Shen, Claudio Silva, Maarten Sukel, Eric K. Tokuda, Sebastian Felix Zappe, and Anna Zivarts. 2022. The Future of Urban Accessibility for People with Disabilities: Data Collection, Analytics, Policy, and Tools. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 102, 8 pages. https://doi.org/10.1145/3517428.3550402

[11] 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 (CVPR)*.

[12] Kotaro Hara, Jin Sun, Robert Moore, David Jacobs, and Jon Froehlich. 2014. 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* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 189–204. https://doi.org/10.1145/2642918.2647403

[13] Maryam Hosseini, Fabio Miranda, Jianzhe Lin, and Claudio T. Silva. 2022. CitySurfaces: City-scale semantic segmentation of sidewalk materials. *Sustainable Cities and Society* 79 (2022), 103630. https://doi.org/10.1016/j.scs.2021.103630

[14] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. *Ultralytics YOLO*. https://github.com/ultralytics/ultralytics

[15] Bon Woo Koo, Subhrajit Guhathakurta, Nisha Botchwey, and Aaron Hipp. 2023. Can good microscale pedestrian streetscapes enhance the benefits of macroscale accessible urban form? An automated audit approach using Google street view images. *Landscape and Urban Planning* 237 (2023), 104816. https://doi.org/10.1016/j.landurbplan.2023.104816

[16] Chu Li, Zhihan Zhang, Michael Saugstad, Esteban Safranchik, Chaitanyashareef Kulkarni, Xiaoyu Huang, Shwetak Patel, Vikram Iyer, Tim Althoff, and Jon E. Froehlich. 2024. LabelAId: Just-in-time AI Interventions for Improving Human Labeling Quality and Domain Knowledge in Crowdsourcing Systems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 643, 21 pages. https://doi.org/10.1145/3613904.3642089

[17] Christopher D Manning. 2008. *Introduction to information retrieval.* Syngress Publishing,.

[18] Daniela Massiceti, Samreen Anjum, and Danna Gurari. 2022. VizWiz grand challenge workshop at CVPR 2022. *SIGACCESS Access. Comput.* 133, Article 1 (aug 2022), 1 pages. https://doi.org/10.1145/3560232.3560233

[19] U.S. Department of Transportation. 2024. Safe Streets and Roads for All Grant Program. https://www.transportation.gov/grants/SS4A Accessed June 20, 2024.

[20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193

[21] F. Prior, J. Almeida, P. Kathiravelu, T. Kurc, K. Smith, T.J. Fitzgerald, and J. Saltz. 2020. Open access image repositories: high-quality data to enable machine learning research. *Clinical Radiology* 75, 1 (2020), 7–12. https://doi.org/10.1016/j.crad.2019.04.002

[22] Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, and Jon Froehlich. 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300292

[23] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson,

Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402
[24] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management* 45, 4 (2009), 427–437.
[25] United Nations Human Settlements Programme (UN-Habitat). 2016. The New Urban Agenda. https://habitat3.org/the-new-urban-agenda/ Accessed June 20, 2024.
[26] Galen Weld, Esther Jang, Anthony Li, Aileen Zeng, Kurtis Heimerl, and Jon E. Froehlich. 2019. Deep Learning for Automatically Detecting Sidewalk Accessibility Problems Using Streetscape Imagery. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 196–209. https://doi.org/10.1145/3308561.3353798

A Appendix

A.1 Metric Definitions

Our datasets exhibit long-tail label distribution, as shown in Tables 2-5. Hence, the cross-label performance metrics can differ significantly. To account for such cases, we report macro, micro and weighted averaged F1 scores. While macro-averaged F1 score is an unweighted average that treats all labels the same, micro-averaged F1 score is a label agnostic measure that is more impacted by the performance of the majority label and weighted-average F1 uses true instance frequency as weights.

In multi-label binary classification, each instance can be assigned multiple labels. The F1 score can be calculated in different ways depending on how the individual label results are aggregated. Below, we define the Micro F1, Weighted F1, and Macro F1 scores. Let L be the number of labels and i denote a specific label. Note that, for our case, a label here is a tag.

Macro F1 Score

The Macro F1 score calculates the F1 score for each label and then takes the average (unweighted) of these scores.

$$\text{Macro F1} = \frac{1}{L}\sum_{i=1}^{L} \text{F1}_i$$

Micro F1 Score

The Micro F1 score aggregates the contributions of all labels to compute the average F1 score. It is calculated using the total True Positives (TP), False Positives (FP), and False Negatives (FN) across all labels, following Sokolova and Lapalme [24] alternative definition.

$$\text{Micro Precision} = \frac{\sum_{i=1}^{L} \text{TP}_i}{\sum_{i=1}^{L} (\text{TP}_i + \text{FP}_i)}$$

$$\text{Micro Recall} = \frac{\sum_{i=1}^{L} \text{TP}_i}{\sum_{i=1}^{L} (\text{TP}_i + \text{FN}_i)}$$

$$\text{Micro F1} = \frac{2 \cdot \text{Micro Precision} \cdot \text{Micro Recall}}{\text{Micro Precision} + \text{Micro Recall}}$$

Weighted F1 Score

The Weighted F1 score calculates the F1 score for each label and takes a weighted average based on the number of true instances (support) for each label.

$$\text{Weighted F1} = \sum_{i=1}^{L} w_i \times \text{F1 Score}_i$$

Where,

$$w_i = \frac{\text{No. true instances for label } i}{\text{Total number of samples}}$$

A.2 Validation UI

For Dataset 1, we designed and implemented a custom validation user interface to clean Project Sidewalk label and tag data. We show two example screenshots of this interface in Figure 4. Four research assistants used this UI to iteratively clean and verify 24,009 labels and 29,311 tags. In total, 16,424 tags were changed (7,988 tags added)—see Table 1.

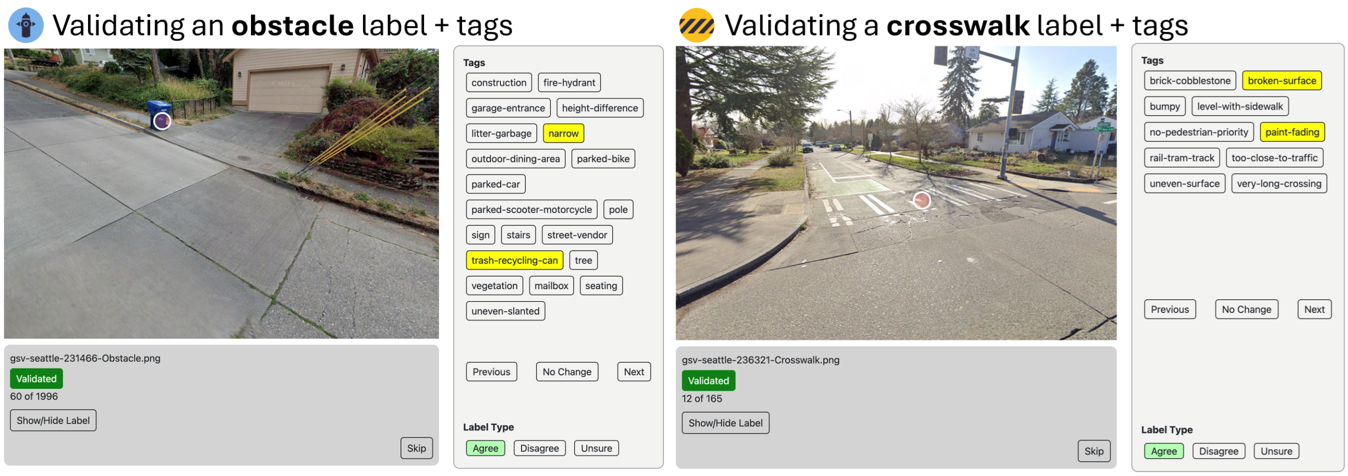

Figure 4: Our custom built validation UI to clean Project Sidewalk label and tag data. (left) The user validating an *obstacle* label and tags: there is a recycling can blocking the sidewalk, which is tagged with *trash-recycling-can* and *narrow*. (right) The user validating a *crosswalk* label and tags: there is a painted crosswalk but it has a *broken surface* and *paint fading*.

A.3 Tags Frequency by Category

Below, we present the frequency of all tags in the training and test sets for each PS category. Tags listed below the gray horizontal rule were present in the training set but were excluded from the test results because their frequency count was < 10. Download the dataset here: https://github.com/ProjectSidewalk/sidewalk-tagger-ai.

Table 2: The *curb ramp* dataset for both Experiments 1 and 2. The table is sorted in descending order by the tag count in the test set. Labels with *No Tags* are last. The gray line indicates tags with counts < 10, which were excluded from the experiments

	Dataset 1			Experiment 1	Experiment 2	Test Set
Curb Ramp	**Raw #**	**Cleaned #**	**# Tags Changed**	**# Tags Training Set**	**# Tags Training Set**	**# Tags**
Missing-tactile-warning	2,286	4,225	2,053	3,353	5,791	872
Points-into-traffic	954	1,384	1,118	1,087	3,539	297
Surface-problem	431	1,076	835	809	1,011	267
Narrow	1,026	927	1,055	775	2,602	152
Not-enough-landing-space	590	631	685	547	1,447	84
Not-level-with-street	341	446	381	381	1,409	65
Pooled-water-debris	15	149	134	107	200	42
Steep	140	150	220	121	522	29
No tag	6,687	4,719	2,904	3,748	31,149	971
Tactile-warning	1	471	472	471	164	0
Total	**5784**	**9459**	**6953**	**7651**	**16685**	**1808**

Table 3: The *surface problem* dataset for both Experiments 1 and 2. The table is sorted in descending order by the tag count in the test set. Labels with *No Tags* are last. The gray line indicates tags with counts < 10, which were excluded from the experiments

	Dataset 1			Experiment 1	Experiment 2	Test Set
Surface Problem	**Raw #**	**Cleaned #**	**# Tags Changed**	**# Tags Training Set**	**# Tags Training Set**	**# Tags**
Grass	3,233	4,025	894	3,219	7,906	806
Cracks	3,572	3,524	1,468	2,913	11,021	611
Height-difference	1,202	1,694	642	1,248	2,626	446
Uneven-slanted	1,844	1,672	816	1,333	5,284	339
Bumpy	1,004	1,763	1,579	1,642	4,083	121
Brick-cobblestone	371	461	100	360	272	101
Narrow-sidewalk	696	611	417	537	3,011	74
Very-broken	496	400	488	329	1,783	71
Sand-gravel	112	252	196	207	490	45
Utility-panel	7	71	64	44	98	27
No tag	574	37	559	37	1,902	0
Construction	35	32	23	26	177	6
Rail-tram-track	20	26	8	21	73	5
Uncovered-manhole	0	9	9	9	16	0
Total	**12,592**	**14,540**	**6,704**	**11,888**	**36,840**	**2,652**

Table 4: The *obstacle* dataset for both Experiments 1 and 2. The table is sorted in descending order by the tag count in the test set. Labels with *No Tags* are last. The gray line indicates tags with counts < 10, which were excluded from the experiments. *Parked-motor* is the "parked-scooter-motorcycle" tag where people park their scooters/motorcycles on the sidewalk, which become accessibility barriers.

	Dataset 1			Experiment 1	Experiment 2	Test Set
Obstacle	**Raw #**	**Cleaned #**	**# Tags Changed**	**# Tags Training Set**	**# Tags Training Set**	**# Tags**
Pole	470	575	129	461	2,713	114
Trash-recycling-can	350	383	91	291	1,117	92
Narrow	143	1,260	1,135	1,181	736	79
Vegetation	393	434	65	355	1,343	79
Parked-car	386	400	24	326	829	74
Construction	122	208	104	174	545	34
Tree	145	153	30	122	1,230	31
Litter-garbage	36	100	76	73	103	27
Sign	90	169	91	145	624	24
Parked-bike	62	70	18	50	147	20
Height-difference	25	72	53	59	168	13
No tag	322	2	322	0	1,339	2
Garage-entrance	34	43	15	37	254	6
Parked-moto	14	22	8	16	51	6
Stairs	20	23	21	20	132	3
Fire-hydrant	43	49	8	46	257	3
Street-vendor	3	6	5	4	114	2
Total	**2,336**	**3,972**	**1,878**	**3,365**	**10,363**	**607**

Table 5: The *crosswalk* dataset for both Experiments 1 and 2. The table is sorted in descending order by the tag count in the test set. Labels with *No Tags* are last. The gray line indicates tags with counts < 10, which were excluded from the experiments.

	Dataset 1			Experiment 1	Experiment 2	Test Set
Crosswalk	**Raw #**	**Cleaned #**	**# Tags Changed**	**# Tags Training Set**	**# Tags Training Set**	**# Tags**
Paint-fading	384	561	255	426	1,852	135
Bumpy	46	232	200	182	145	50
Broken-surface	78	361	291	315	207	46
Brick-cobblestone	16	54	38	41	138	13
No tag	1,116	781	413	623	5,159	158
Uneven-surface	47	67	56	60	148	7
Rail-tram-track	16	34	34	30	126	4
Very-long-crossing	24	30	14	28	356	2
Level-with-sidewalk	0	1	1	0	12	1
No-pedestrian-priority	0	0	0	0	3	0
Total	**611**	**1,340**	**889**	**1,082**	**2,987**	**258**

A.4 Datasets by City

Experiment 1 (clean data only) used Project Sidewalk data from 10 cities across three countries (US, Mexico, Netherlands) while Experiment 2 (crowdsourced data only) added two additional cities (St. Louis, MO; Teaneck, NJ).

Table 6: The distribution of our two datasets by city sorted by the num of tags in our test set. Dataset 1 is composed of 10 cities across three countries while Dataset 2 adds two additional cities (St. Louis and Teaneck). *SPGG* stands for San Pedro Garza García in Mexico; *CDMX* is Mexico City, Mexico.

	Experiment 1 Training Set		Experiment 2 Training Set		Both Experiments: Same Test Set	
Cities	**Num Labels**	**Num Tags**	**Num Labels**	**Num Tags**	**Num Labels**	**Num Tags Test Set**
Seattle, WA	4,417	5,125	30,842	18,810	1,113	1,111
Chicago, IL	3,626	4,602	18,489	10,552	921	982
Oradell, NJ	3,185	3,866	1,653	1,736	806	883
SPGG, MX	1,304	2,256	8,576	10,187	316	562
Columbus, OH	2,496	2,303	4,856	3,078	612	506
Pittsburgh, PA	1,191	1,667	3,569	3,485	295	405
Newberg, OR	1,526	2,018	1,523	1,085	372	391
CDMX, MX	840	1,467	7,825	9,753	209	350
Amsterdam, NL	478	599	3,929	2,562	110	116
Walla Walla, WA	153	83	522	560	39	19
St. Louis, MO	N/A	N/A	2,846	4,044	N/A	N/A
Teaneck, NJ	N/A	N/A	2,865	1,023	N/A	N/A
Total	**19,216**	**23,986**	**87,495**	**66,875**	**4,793**	**5,325**

A.5 Experiment 1: DINOv2 Results

Details of frequency of tag in the test set, the selected confidence (maximizing the F1 score with a minimum threshold of 0.3), and precision, recall, and F1 score of that threshold for each tag of the label category. Tags with less than 10 instances in the test set are excluded.

Table 7: DINOv2 Experiment 1 *curb ramp* tag classification results. Results are sorted by F1 score.

Curb Ramp Tags	N	Confidence	Precision	Recall	F1
Missing-tactile-warning	872	0.3	0.92	0.96	0.94
Surface-problem	267	0.3	0.67	0.44	0.53
Narrow	152	0.3	0.4	0.28	0.33
Points-into-traffic	297	0.3	0.42	0.18	0.25
Not-enough-landing-space	84	0.3	0.27	0.15	0.20
Not-level-with-street	65	0.3	0.21	0.14	0.17
Pooled-water-debris	42	0.3	0.67	0.05	0.09
Steep	29	0.3	0	0	0

Table 8: DINOv2 Experiment 1 *surface problem* tag classification results. Results are sorted by F1 score.

Surface Problem Tags	N	Confidence	Precision	Recall	F1
Brick-cobblestone	101	0.3	0.97	0.86	0.91
Grass	806	0.53	0.91	0.9	0.90
Cracks	611	0.82	0.74	0.8	0.77
Height-difference	446	0.3	0.87	0.63	0.73
Uneven-slanted	339	0.3	0.52	0.53	0.52
Sand-gravel	45	0.3	0.46	0.38	0.41
Narrow-sidewalk	74	0.32	0.4	0.34	0.37
Bumpy	121	0.3	0.17	0.47	0.25
Very-broken	71	0.3	0.21	0.14	0.17
Utility-panel	27	0.3	0	0	0

Table 9: DINOv2 Experiment 1 *obstacle* tag classification results. Results are sorted by F1 score.

Obstacle Tags	N	Confidence	Precision	Recall	F1
Parked-car	74	0.3	0.97	0.89	0.93
Parked-bike	20	0.32	0.9	0.9	0.90
Trash-recycling-can	92	0.3	0.9	0.86	0.88
Pole	114	0.3	0.92	0.78	0.84
Vegetation	79	0.3	0.88	0.81	0.84
Tree	31	0.55	0.95	0.61	0.75
Sign	24	0.36	0.52	0.62	0.57
Height-difference	13	0.3	1	0.38	0.56
Litter-garbage	27	0.75	0.71	0.44	0.55
Construction	34	0.3	0.59	0.5	0.54
Narrow	79	1	0.22	0.46	0.3

Table 10: DINOv2 Experiment 1 *crosswalk* tag classification results. Results are sorted by F1 score.

Crosswalk Tags	N	Confidence	Precision	Recall	F1
Paint-fading	135	0.3	0.86	0.76	0.80
Broken-surface	46	1	0.66	0.72	0.69
Brick-cobblestone	13	0.3	1	0.38	0.56
Bumpy	50	0.3	0.53	0.4	0.45

A.6 Experiment 1: CLIP-ViT Results

Details of frequency of tag in the test set, the selected confidence (maximizing the F1 score with a minimum threshold of 0.3), and precision, recall, and F1 score of that threshold for each tag of the label category. Tags with less than 10 instances in the test set are excluded.

Table 11: CLIP-ViT Experiment 1 *curb ramp* tag classification results. Results are sorted by F1 score.

Curb Ramp	N	Confidence	Precision	Recall	F1
Missing-tactile-warning	872	0.72	0.89	0.95	0.92
Surface-problem	267	0.3	0.6	0.34	0.44
Narrow	152	0.3	0.26	0.24	0.25
Points-into-traffic	297	0.3	0.4	0.13	0.20
Not-level-with-street	65	0.3	0.28	0.15	0.20
Not-enough-landing-space	84	0.3	0.22	0.15	0.18
Pooled-water-debris	42	0.3	0.5	0.05	0.09
Steep	29	0.3	0.07	0.03	0.05

Table 12: CLIP-ViT Experiment 1 *surface problem* tag classification results. Results are sorted by F1 score.

Surface Problem	N	Confidence	Precision	Recall	F1
Grass	806	0.71	0.91	0.87	0.89
Brick-cobblestone	101	0.3	0.94	0.74	0.83
Cracks	611	0.84	0.7	0.75	0.72
Height-difference	446	0.3	0.88	0.58	0.70
Uneven-slanted	339	0.3	0.51	0.5	0.50
Sand-gravel	45	0.3	0.52	0.31	0.39
Narrow-sidewalk	74	0.3	0.34	0.31	0.33
Bumpy	121	0.7	0.18	0.45	0.25
Very-broken	71	0.3	0.24	0.11	0.15
Utility-panel	27	0.3	1	0.04	0.07

Table 13: CLIP-ViT Experiment 1 *obstacle* tag classification results. Results are sorted by F1 score.

Obstacle	N	Confidence	Precision	Recall	F1
Parked-car	74	0.84	0.97	0.95	0.96
Trash-recycling-can	92	0.87	0.95	0.85	0.90
Pole	114	0.91	0.88	0.78	0.83
Vegetation	79	0.3	0.84	0.8	0.82
Parked-bike	20	0.3	0.87	0.65	0.74
Tree	31	0.36	0.89	0.55	0.68
Height-difference	13	0.48	0.67	0.62	0.64
Construction	34	0.94	0.82	0.41	0.55
Sign	24	0.3	0.5	0.58	0.54
Litter-garbage	27	0.3	0.41	0.26	0.32
Narrow	79	0.88	0.21	0.61	0.31

Table 14: CLIP-ViT Experiment 1 *crosswalk* tag classification results. Results are sorted by F1 score.

Crosswalk	N	Confidence	Precision	Recall	F1
Paint-fading	135	0.3	0.79	0.69	0.74
Broken-surface	46	0.96	0.5	0.76	0.60
Bumpy	50	0.3	0.5	0.34	0.40
Brick-cobblestone	13	0.3	0.67	0.15	0.25

A.7 Experiment 2: DINOv2 Results

Details of frequency of tag in the test set, the selected confidence (maximizing the F1 score with a minimum threshold of 0.3), and precision, recall, and F1 score of that threshold for each tag of the label category. Tags with less than 10 instances in the test set are excluded.

Table 15: DINOv2 Experiment 2 *curb ramp* tag classification results. Results are sorted by F1 score.

Curb Ramp	N	Confidence	Precision	Recall	F1
Missing-tactile-warning	872	0.3	0.99	0.35	0.51
Narrow	152	0.3	0.33	0.15	0.21
Not-level-with-street	65	0.3	0.21	0.08	0.11
Points-into-traffic	297	0.3	0.28	0.06	0.10
Surface-problem	267	0.3	0.93	0.05	0.10
Not-enough-landing-space	84	0.3	0.14	0.04	0.06
Steep	29	0.3	0.12	0.03	0.05
Pooled-water-debris	42	0.3	0	0	0

Table 16: DINOv2 Experiment 2 *surface problem* tag classification results. Results are sorted by F1 score.

Surface Problem	N	Confidence	Precision	Recall	F1
Grass	806	0.3	0.95	0.82	0.88
Cracks	611	0.77	0.65	0.78	0.71
Height-difference	446	0.3	0.86	0.54	0.66
Uneven-slanted	339	0.3	0.45	0.58	0.51
Brick-cobblestone	101	0.3	1	0.35	0.51
Narrow-sidewalk	74	0.3	0.26	0.31	0.28
Bumpy	121	0.95	0.25	0.23	0.24
Sand-gravel	45	0.3	0.54	0.16	0.24
Very-broken	71	0.3	0.23	0.21	0.22
Utility-panel	27	0.3	1	0.04	0.07

Table 17: DINOv2 Experiment 2 *obstacle* tag classification results. Results are sorted by F1 score.

Obstacle	N	Confidence	Precision	Recall	F1
Parked-car	74	0.3	1	0.86	0.93
Vegetation	79	0.98	0.94	0.85	0.89
Trash-recycling-can	92	0.3	0.93	0.76	0.84
Parked-bike	20	0.3	0.84	0.8	0.82
Tree	31	0.55	0.78	0.81	0.79
Pole	114	0.3	0.89	0.67	0.76
Sign	24	0.44	0.46	0.75	0.57
Construction	34	0.3	0.71	0.44	0.55
Height-difference	13	0.3	1	0.15	0.27
Narrow	79	0.3	0.5	0.05	0.09
Litter-garbage	27	0.3	1	0.04	0.07

Table 18: Experiment 2 Test Results of DINOv2 on Crosswalk Data

Crosswalk	N	Confidence	Precision	Recall	F1
Paint-fading	135	0.3	0.9	0.42	0.58
Brick-cobblestone	13	0.3	1	0.08	0.14
Broken-surface	46	0.3	1	0.07	0.12
Bumpy	50	0.3	1	0.02	0.04