Scaling Crowd+AI Sidewalk Accessibility Assessments: Initial Experiments Examining Label Quality and Cross-city Training on Performance

MICHAEL DUAN, Allen School of Computer Science, University of Washington, USA
SHOSUKE KIAMI, Allen School of Computer Science, University of Washington, USA
LOGAN MILANDIN, Allen School of Computer Science, University of Washington, USA
JOHNSON KUANG, Allen School of Computer Science, University of Washington, USA
MICHAEL SAUGSTAD, Allen School of Computer Science, University of Washington, USA
MARYAM HOSSEINI, Visualization Imaging and Data Analysis (VIDA) Center, New York University, USA
JON E. FROEHLICH, Allen School of Computer Science, University of Washington, USA

Increasingly, crowds plus machine learning techniques are being used to semi-automatically analyze the accessibility of built environments; however, open questions remain about how to effectively combine the two. We present two experiments examining the effect of crowdsourced data in automatically classifying sidewalk accessibility features in streetscape images. In Experiment 1, we investigate the effect of validated data—which has been voted correct by the crowd but is more expensive to collect—compared with a larger but noisier aggregate dataset. In Experiment 2, we examine whether crowdsourced labeled data gathered in one city can be used as effective training data for another. Together, these experiments contribute to the growing literature in Crowd+AI approaches for semi-automatic sidewalk assessment and help identify pertinent challenges.

ACM Reference Format:
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 . ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Despite broad policy measures and an increasing emphasis on the design of inclusive cities [13], urban streets and sidewalks remain largely inaccessible. The problem is not just a lack of accessibility but a lack of data on sidewalk location and their condition, which fundamentally impacts policy making, urban planning, and the design of accessibility features in tools like Apple or Google Maps [5, 15]. Traditionally, sidewalk data is gathered via in-person inspections administered by local governments; however, this approach is laborious, expensive, and infrequent [2, 5, 12]. Thus, emerging work has explored alternative methods using remote crowdsourcing [7, 16], computer vision [1, 9, 10, 19], or both [8, 11].

While vision-based solutions are more scalable, they require copious amounts of training data. Large-scale, standardized streetscape datasets like *CityScapes* [3] or *Mapillary Vistas* [14] exist and have dramatically accelerated research in autonomous navigation. However, these labeled datasets do not include sidewalk accessibility features. Furthermore, while automatic walkability [1] and surface condition assessment [10] have benefited from including streetscape imagery, these experiments used expensive, researcher-labeled datasets for training, which limits experimental scope and size (*e.g.,* to only single neighborhoods or cities). To achieve larger training sets, others have explored crowdsourced

Manuscript submitted to ACM

labeling techniques [8, 19], which can engage community members and gather diverse perspectives on streetscape accessibility but potentially create noisier datasets.

Recently, Weld *et al.* [19] showed how Project Sidewalk's crowdsourced sidewalk accessibility feature dataset [16] could be used to train and test deep learning methods for automatic sidewalk assessment. While promising, questions remain about whether validated data—labels with correctness determined by crowd-voting—results in better model performance compared to unvalidated but larger datasets. Additionally, Weld concludes by calling for cross-city experiments that demonstrate the potential of model generalization—that is, can we train an AI model on labeled data from one city and have it assess sidewalks in another? This is an important but difficult task: a generalizable AI model would allow fast and broad assessment of sidewalk accessibility across the world; however, urban designs, pedestrian pathways, and sidewalk degradations can be geographically and culturally specific [6].
In this paper, we examine:

R1: What is the effect of using unfiltered crowdsourced training data *vs.* a validated subset on model performance?
R2: How does model performance vary as a function of per-city training set composition?

To address these questions, we present two experiments using Project Sidewalk's open sidewalk dataset, which has grown to over 700,000 image-based labels and 400,000 validations across 10 cities [16]. Similar to Weld *et al.* [19], we seek to classify the following sidewalk accessibility features: *curb ramps*, *missing curb ramps*, *obstacles*, and *surface problems*. In Experiment 1, we compare a model trained on a large, unvalidated training dataset *vs.* a subset composed of only positively validated labels (R1). Surprisingly, our results do not show a uniform improvement across each feature type when using only validated training data–perhaps due to noise in the crowdsourced validations themselves and/or the smaller dataset. In Experiment 2, extending Weld *et al.* [19], we conduct a series of experiments examining the effect of cross-city model generalization. Here, our results show the promise of cross-city generalization—for example, we were able to achieve an average of 76.8% recall and 77.1% precision over all label types in the city of Columbus *without* any training data from Columbus itself. However, significant work is needed to improve results around generalizing internationally.

In summary, our work helps advance the growing literature in Crowd+AI urban accessibility assessments by highlighting tradeoffs in using larger but noisier training datasets *vs.* smaller but more expensive crowd-validated datasets and examining the potential of cross-city model generalization.

2 DATASET

For our training and test datasets, we used the publicly available crowd-sourced sidewalk accessibility data provided by Project Sidewalk [16]. The data consists of point labels on Google Streetview Panoramas of various feature types. Note that, while we sought to classify the aforementioned four feature types, our data also contains labels of other feature types supported in the Project Sidewalk labeling interface, such as *crosswalks* and *missing sidewalks*. A subset of labels also include crowdsourced validation (agree/disagree) counts. For each label, we extract a 1000×1000px image crop from the panorama the label belongs to centered around the label—this crop size was determined experimentally. Since any given crop may contain multiple features simultaneously (*e.g.,* a cracked curb ramp), we created a label set for each crop containing all labels on the same panorama within 300px (determined experimentally) of the center.

Table 1. The total counts for *CR*: curb ramp, *MR*: missing curb ramp, *O*: obstacle, and *SP*: surface problem labels in our training, positively validated training, and test sets. Note that "Crops" does not represent a total label count since crops may contain more than one label or no labels at all.

	Unfiltered Train Data					Positively Validated Train Data					Test Data				
City	CR	MR	O	SP	**Crops**	CR	MR	O	SP	**Crops**	CR	MR	O	SP	**Crops**
Seattle	72,186	36,269	11,320	28,400	**171,349**	37,199	25,926	5,940	16,033	**82,049**	309	255	138	344	**967**
SPGG	4,528	23,202	53,361	24,967	**100,033**	2,269	8,814	5,764	9,231	**22,022**	210	336	356	470	**969**
Columbus	13,546	942	3,360	5,726	**28,934**	4,902	374	1,027	3,385	**9,065**	389	183	210	394	**984**
DC	147,736	22,765	26,593	9,348	**233,029**	N/A	N/A	N/A	N/A	**N/A**	507	121	152	403	**966**
All	**237,996**	**83,178**	**94,634**	**68,441**	533,345	**44,370**	**35,114**	**12,731**	**28649**	113136	**1,415**	**895**	**856**	**1,611**	**3,886**

Project Sidewalk is deployed in 10 cities across the United States, the Netherlands, and Mexico. For this paper, we used data from three US cities: Seattle, Columbus, and Washington DC as well as one from Mexico: San Pedro Garza Garcia, MX (SPGG). See Table 1. For each city, we created a test set by randomly selecting 250 crops from positively validated labels of each feature type, manually correcting label sets when necessary and removing any indeterminable crops. Note that this process led to a few test crops with empty label sets (*e.g.*, a crop of a passable sidewalk we determined didn't have any of the four feature types present). No validation data exists for Washington DC as the validation pipeline did not exist during DC's deployment.

3 EXPERIMENTS

Framework. Our classification framework consists of four independently trained binary classifiers that each check an image for the presence of one of four sidewalk accessibility features: *curb ramps*, *missing curb ramps*, *obstacles*, and *surface problems*. For each classification model, we employ *HRNet-W48* [18]–one of the top performing models in streetscape vision benchmarks [3, 14]. The model is fine-tuned using pre-trained weights from the HRNet trunk used by Hosseini *et al.* in their *CitySurfaces* framework [10]. Fine-tuning with CitySurfaces' weights led to faster convergence and better performance than using out-of-the-box models like *EfficientNet* [17] pre-trained on *ImageNet* [4].

Training Strategy. For each feature type, we binarized the train and test set such that each crop is considered positive if its label set contains the feature and negative otherwise. We then reserved 20% of the train set for the validation set as is standard in the CV literature. We observed that our post-binarization train sets were often imbalanced, so given p positive and n negative train crops, we sampled $\min(p, n)$ positive and negative crops without replacement per epoch to avoid overpredicting the majority class. We trained our classifiers on the crops using cross-entropy loss as our criterion, optimizing the loss with stochastic gradient descent. The hyperparameters used were a base learning rate of 0.01, a momentum of 0.9, and a weight decay of 1e-6 — similar to [10]. We trained for 10 epochs, using a stepwise learning rate scheduler that halved the learning rate every 3 epochs to ensure convergence.

3.1 Baseline

To establish a baseline for our two experiments, we trained/tested binary classifiers per city on their respective unfiltered datasets (Table 2). We use three primary measures of performance across experiments: *precision*, which is the fraction of positive predictions (model predicts the feature is present) that are correct; *recall*, which is the fraction of actual positives (feature is present) correctly identified; and *accuracy*, which is simply the fraction of correct predictions out of all predictions.

Table 2. Baseline precision, recall, and accuracy for curb ramps, missing curb ramps, obstacles, and surface problems for each city.

	Precision (%)				Recall (%)				Accuracy (%)			
City	CR	MR	O	SP	CR	MR	O	SP	CR	MR	O	SP
Seattle	**89.1**	**85.1**	53.0	**92.1**	85.1	**89.8**	**77.4**	**64.4**	**91.9**	**93.2**	**87.1**	**85.4**
SPGG	79.6	82.5	**68.1**	81.9	74.6	88.4	60.8	52.2	90.4	89.5	75.2	71.3
Columbus	81.5	62.4	57.0	88.9	**86.3**	62.1	73.7	61.3	86.9	86.1	82.6	81.5
DC	88.5	45.1	42.6	85.1	85.0	76.7	74.8	41.3	86.3	85.5	80.3	72.6

Overall, we found that our binary classification approach resulted in an of average 73.9% precision, 72.1% recall, and 84.1% accuracy across the four cities. The cities with the most training data performed the best. For example, Seattle yielded 79.8% precision, 79.1% recall, and 89.4% accuracy. In terms of label types, we could detect curb ramps most accurately with 84.7% precision, 82.7% recall, and 88.7% accuracy across the four cities—similar to prior work [19].

To more closely examine model performance, we visually assessed a subset of false positives/negatives for each binary classifier (Figure 1). A common source of error was the presence of other features in the crops; about 25% of all false negatives for curb ramps, missing curb ramps and surface problem involve an obstacle either blocking the view or nearby. We speculate that this is due to label sparseness; some train crops include other features that weren't caught by crowdsourcers, so they become negative examples for features they actually contain. Other mistakes include labeling driveways as curb ramps, labeling features not on a sidewalk as obstacles or surface problems, and failing to label features due to shadows and discoloration—as also identified in [19].

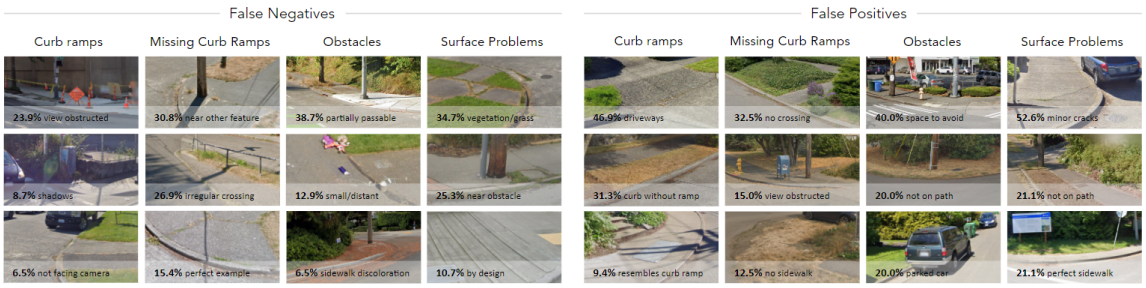

Fig. 1. To qualitatively examine model performance, we assessed up to 50 false positives and 50 false negatives of each label type (294 total). We group the mistakes and present the three most common types for each category above.

3.2 Experiment 1: Effect of Validated Data on Model Performance

To address R1, we constructed a classification framework per city, trained the models only on positively validated data, and evaluated on an individual city's test set (Table 3). As noted above, no validated data exists for DC, so this city was excluded from Experiment 1.

Table 3. Precision, recall, and accuracy from models trained on positively validated data, given in percentages per city for curb ramps, missing curb ramps, obstacles, and surface problems.

	Precision (%)				Recall (%)				Accuracy (%)			
City	CR	MR	O	SP	CR	MR	O	SP	CR	MR	O	SP
Seattle	**93.5**	**87.5**	52.7	**92.8**	84.4	**90.9**	**78.1**	63.6	**93.2**	**94.2**	**87.0**	**85.3**
SPGG	86.6	81.3	**68.4**	77.2	74.2	88.1	47.6	55.0	92.0	88.9	72.8	70.4
Columbus	82.7	61.8	64.5	82.0	**85.1**	70.3	67.0	**69.7**	87.1	86.5	85.2	81.8
DC	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Overall, we found that compared to baseline, precision and accuracy for curb ramps increased but there were minimal performance differences for other label types. We speculate that the lack of improvement is due to two challenges inherent to our methodology: (1) subjectivity/nuance in validations and (2) decreased training set size.

Regarding the first challenge, obstacles require spatial contextualization (*i.e.,* is the object actually blocking the path?), which is often difficult to assess with only panoramic imagery. Additionally, some label types, like surface problems, can vary widely in appearance and severity due to weathering, root uplifts, or other degradations, which makes it challenging for both the classifier and human validation. An example of the second challenge was observed in filtering the SPGG train data, which reduced the 53,361 obstacle labels to only 5,764 positively validated obstacles (~10.8%). Our model may not have had enough positive data to properly learn the variation in obstacles, resulting in lower performance.

3.3 Experiment 2: Cross-city Model Generalization

To address R2, we ran two sub-experiments: (1) Train models on the aggregated train set over all cities and evaluate on each city's test set; (2) For each city C, train models on the train sets for all cities except C then evaluate on C's test set. The results are in Table 4.

Table 4. Precision, recall, and accuracy from models trained on all cities as well as with the city of interest excluded, given in percentages per city for curb ramps, missing curb ramps, obstacles, and surface problems.

	All Cities Combined Results												One City Excluded Results											
	Precision (%)				Recall (%)				Accuracy (%)				Precision (%)				Recall (%)				Accuracy (%)			
City	CR	MR	O	SP	CR	MR	O	SP	CR	MR	O	SP	CR	MR	O	SP	CR	MR	O	SP	CR	MR	O	SP
Seattle	**90.0**	**78.9**	51.1	85.3	87.7	**94.1**	81.8	**71.1**	**93.0**	**91.8**	86.3	**85.4**	87.1	78.4	49.3	77.9	80.8	82.7	**75.9**	**68.8**	**90.1**	89.5	85.5	82.0
SPGG	83.7	73.5	61.9	76.6	83.3	92.5	78.6	64.4	92.9	85.9	74.4	73.3	64.0	**82.5**	**71.6**	76.2	76.6	46.6	42.0	62.0	85.7	78.1	72.7	72.2
Columbus	84.9	73.3	**64.9**	85.3	**88.1**	81.3	80.4	71.0	89.1	91.1	**86.6**	83.5	85.9	71.6	64.6	**86.2**	83.2	**83.0**	74.2	66.7	88.0	**90.8**	**85.9**	**82.4**
DC	89.5	42.7	42.5	**88.2**	87.7	83.3	**84.8**	40.8	88.2	84.1	79.7	73.1	**90.9**	46.6	45.8	80.3	**84.4**	63.3	60.9	42.5	87.4	86.4	82.6	71.7

Regarding sub-experiment 1, we noticed an overall increase in recall compared to the baseline. This suggests that when models are exposed to more diverse examples of a given feature, they become more sensitive to the presence of that feature and predict it more often. In general, this increase in recall was accompanied by a drop in precision, which is a common trade-off when the model becomes too sensitive to the positive class, thus overpredicting it and resulting in more potential false positives.

As expected in sub-experiment 2, we observed an overall decrease in performance across all label types. This is likely due to the varying design elements and urban fabrics across cities. Consider Figure 2, an example false negative from

the exclude-SPGG obstacle classifier. Terraced sidewalks are a frequent obstacle in SPGG. The exclude-SPGG model, however, classifies the below as not containing an obstacle because most sidewalk obstacles in US cities are not stairs.

Fig. 2. A terraced sidewalk (a frequent obstacle in SPGG).

Promisingly, Columbus doesn't experience the reduced performance mentioned above in either experiment. The exclude-Columbus framework performs better than the baseline only-Columbus framework, perhaps due to Columbus's relatively small dataset and the similar infrastructure in cities with more data. This suggests that for cities with little to no feature data, we can supplement training data from other cities, particularly those with relatively similar built environments to the test city. The all-cities framework performs better on Columbus than both the baseline and the exclude-Columbus framework, demonstrating how even a little data can be fed into the training process to create a more generalizable framework that performs better on the city of interest.

4 LIMITATIONS AND FUTURE WORK

While these experiments yielded promising results, significant work remains.

Re-imagining Validation. Many feature types require more context for accurate classification (*e.g.,* predicting missing curb ramps requires inferences about where pedestrians are intended to cross streets). Fixed-size crops often lead to features being cut off. The need for more context and analysis of our unprocessed data (panoramas with labeled feature coordinates) suggests instead training on entire panoramas with coordinate metadata. Outputs would then be predicted labels for unlabeled feature coordinates. Semantic segmentation can also be implemented as an improvement to the reformulated object detection problem above to achieve more reliable and accurate results.

Ethical Bias. Can varying urban infrastructure in neighborhoods of differing socioeconomic standings lead to within-city biases in automated assessment? Would less well-maintained sidewalk infrastructure in lower-income neighborhoods lead to less accurate model predictions, resulting in a worse representation of the accessibility issues in comparison to wealthier neighborhoods? Such questions call for a study of model performance within varying socioeconomic regions to discover such biases and investigate solutions to address them if needed.

Training Crowdsourcers. A lack of validated data/noisy validations call for proper training and community outreach. For example, initiatives like Project Sidewalk mapathons are training grounds that have the potential to generate a large number of clean validations and promote awareness/knowledge of accessibility issues within communities.

Active Learning with Deployed Classifiers. User quality scoring on crowdsourcing platforms can be used to weight the impact of any particular training data point in training, with higher scores giving a user's labels/validations greater weight. Deployed classifiers can then be actively trained by using real-time audits/validations, weighting them based on user's score, and feeding them into the training pipeline to improve deployed classifiers.

ACKNOWLEDGMENTS

This work was funded in part by the National Science Foundation under grant SCC-IRG 2125087, the Pacific Northwest Transportation Consortium (PacTrans), UW CREATE, and the Google Cloud Research Credits Program.

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), 4548.

[2] Kelly Clifton, A Livi, and DA Rodriguez. 2005. Pedestrian Environment Data Scan (PEDS) Tool. *Planning* 80 (2005), 95–110.

[3] 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 (CVPR)*.

[4] 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*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[5] 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 2019), 78–81. Issue 2. https://doi.org/10.1145/3301657

[6] Jon E. Froehlich, Mikey Saugstad, Edgar Martínez, and Rebeca de Buen Kalman. 2020. Sidewalk Accessibility in the US and Mexico: Policies, Tools, and A Preliminary Case Study.

[7] Kotaro Hara, Vicki Le, and Jon Froehlich. 2013. Combining crowdsourcing and google street view to identify street-level accessibility problems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 631–640.

[8] 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*. 189–204.

[9] Maryam Hosseini, Iago B Araujo, Hamed Yazdanpanah, Eric K Tokuda, Fabio Miranda, Claudio T Silva, and Roberto M Cesar Jr. 2021. Sidewalk measurements from satellite images: Preliminary findings. In *Spatial Data Science Symposium*.

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

[11] Maryam Hosseini, Mikey Saugstad, Fabio Miranda, Andres Sevtsuk, Claudio T. Silva, and Jon E. Froehlich. 2022. Towards Global-Scale Crowd+AI Techniques to Map and Assess Sidewalks for People with Disabilities. In *CVPR2022 Workshop: Accessibility, Vision, and Autonomy (AVA)* (New Orleans, LA). 6 pages.

[12] Wesley E Marshall and Norman W Garrick. 2010. Effect of street network design on walking and biking. *Transportation Research Record* 2198, 1 (2010), 103–115.

[13] United Nations. 2020. The New Urban Agenda. , 194 pages.

[14] 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.

[15] Manaswi Saha, Devanshi Chauhan, Siddhant Patil, Rachel Kangas, Jeffrey Heer, and Jon E. Froehlich. 2021. Urban Accessibility as a Socio-Political Problem: A Multi-Stakeholder Analysis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3, Article 209 (jan 2021), 26 pages. https://doi.org/10.1145/3432908

[16] 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.

[17] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[18] 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* (2020).

[19] 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 *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 196–209.