X-World: Accessibility, Vision, and Autonomy Meet

Jimuyang Zhang* Minglan Zheng* Matthew Boyd Eshed Ohn-Bar
Boston University
{zhangjim, mzheng27, mcboyd, eohnbar}@bu.edu

Figure 1: **The X-World Platform.** Our goal is to facilitate accessibility-driven autonomous systems, i.e., systems that can interact with diverse pedestrians with disabilities who may be rarely observed and represented in data collected by real-world fleets. Towards this goal, we introduce a novel simulation environment for the training and development of vision-based systems. A subset of the mobility aids used in the experiments is visualized.

Abstract

An important issue facing vision-based intelligent systems today is the lack of accessibility-aware development. A main reason for this issue is the absence of any large-scale, standardized vision benchmarks that incorporate relevant tasks and scenarios related to people with disabilities. This lack of representation hinders even preliminary analysis with respect to underlying pose, appearance, and occlusion characteristics of diverse pedestrians. What is the impact of significant occlusion from a wheelchair on instance segmentation quality? How can interaction with mobility aids, e.g., a long and narrow walking cane, be recognized robustly? To begin addressing such questions, we introduce X-World, an accessibility-centered development environment for vision-based autonomous systems. We tackle inherent data scarcity by leveraging a simulation environment to spawn dynamic agents with various mobility aids. The simulation supports generation of ample amounts of finely annotated, multi-modal data in a safe, cheap, and privacy-preserving manner. Our analysis highlights novel challenges introduced by our benchmark and tasks, as well as numerous opportunities for future developments. We further broaden our analysis using a complementary real-world evaluation benchmark of in-situ navigation by pedestrians with disabilities. Our contributions provide an initial step towards widespread deployment of vision-based agents that can perceive and model the interaction needs of diverse people with disabilities.

1. Introduction

As prototypical vision-based machines move from their controlled development labs into the real-world, their impact on people with disabilities becomes discernible. People with various abilities (e.g., sighted, visually-impaired, mobility-impaired) may each react fairly differently when interacting with an autonomous platform (e.g., ground robot, autonomous vehicle, wearable system), the dynamic context of the surrounding scene (e.g., traffic, crowds), and infrastructure components (e.g., intersection type, potholes, ramps, stairs) [37, 2, 28, 26]. Hence, such factors are related

*Contributed equally.

to increased risk of traffic fatalities among pedestrians with disabilities [36, 1, 38, 42], notably during critical navigation junctions, e.g., intersections [37].

Take, for instance, the case of Starship Technologies which paused operations of its food and package delivery robot within only a matter of days after its deployment at the University of Pittsburgh in October of 2019 [43]. The pause came as a result of an adverse encounter between one of the delivery robots and a mobility-impaired doctoral student who uses a wheelchair. In the reported scenario, the delivery robot, which routinely occupies the curb's ramp when waiting to cross at intersections, blocked the student from being able to safely board the sidewalk. As a result, the student reportedly had to wait, in dangerously moving traffic, at the intersection. Clearly, failing to account for the diversity of interaction needs among people with disabilities can have dire consequences. Unfortunately, there is a current lack of shared and principled development tools for accessibility-driven, vision-based autonomous systems. ***How can we advance the state-of-the-art of systems that understand and seamlessly interact with all people in their environment?***

A fundamental barrier to realizing accessibility-aware autonomous systems is the access to data. In particular, people with disabilities can be both *highly diverse*, for instance in appearance and mobility characteristics [44], as well as *quite rare*, even in extensive large-scale data collection efforts by instrumented real-world fleets. Thus, common issues related to the "long tail" distribution of events can compound in our context, i.e., due to the combinatorial rarity of joint safety-critical events (e.g., person crossing an intersection in low visibility at nighttime) with accessibility-related events (e.g., a wheelchair user). To emphasize, due to their rarity, people with disabilities are currently entirely absent from large-scale datasets in computer vision and robotics, even within heavily studied tasks such as pedestrian detection [17, 60, 13] and path prediction [49]. Yet, the presence of a mobility aid can potentially impact the underlying perception algorithm, e.g., due to occlusion, as well as be used to infer the intent and future state of a pedestrian, e.g., a visually-impaired pedestrian that may take longer to explore tactile cues when crossing an intersection. In this paper, we take a crucial step forward towards developing perception models that are both *robust*, i.e., operate at high accuracy under appearance variations of pedestrians with disabilities, and *functional*, i.e., provide a sufficiently fine-grained representation of pedestrians' states for any subsequent decision-making modules.

We support the development of vision-based systems with comprehensive understanding of the needs of diverse people with disabilities through ***four main contributions***: (1) We present the X-World platform, which includes an interactive photo and behavior-realistic simulation module integrated into CARLA [18] with support for spawning agents that use diverse mobility aids, various sensor and environmental configurations, and extensive ground truth generation for numerous visual-semantic reasoning tasks. (2) To rigorously uncover issues in perception of diverse people, we explore the task of segmenting people and their mobility aids. We leverage the simulation environment to generate the first large-scale accessibility-oriented instance segmentation dataset. By incorporating images across varying perspectives, towns, environmental conditions, and mobility aids, we use the dataset to highlight new challenges and opportunities in developing robust and broad-impact machine vision models. (3) Although collecting a large-scale dataset for our task in the physical world is difficult, we accompany the simulation-based benchmark with a diverse and challenging real-world dataset obtained from public internet videos of in-situ navigation. The real-world dataset provides complementary analysis and produces generalization insights related to our fine-grained instance segmentation task. (4) By publicly releasing tools and data for closely integrating computer vision and accessibility research, we contribute towards improving the quality-of-life of individuals with disabilities.

2. Related Work

Our work focuses on obtaining fine-grained visual recognition models of humans. We build on several recent advances in computer vision, specifically work in person detection and segmentation and interactive simulated worlds.

Benchmarks for Visual Understanding of People: Our main objective is to provide machines with a rich visual understanding of the people in their environment, including people with disabilities. Robust perception of people has steadily progressed with the introduction of increasingly large and diverse benchmarks, including INRIA [14], Daimler [19], ImageNet [16], Caltech [17], COCO [39], KITTI [25], ApolloScape [32], BDD100K [58], Cityscapes [13], CityPersons [60], HICO-DET [9], EuroCity [4], Argoverse [8], 100DOH [52], Audi A2D2 [27], NuScenes [5], A*3D [46], Waymo [53] and more. Nonetheless, recent large-scale benchmarks are still substantially narrow in their *scope*, i.e., in contrast to the full range of scenarios and conditions that the real-world may present, and *tasks*, i.e., compared to detailed and contextual visual reasoning that is performed by humans. Despite exponential growth in development of benchmarks and methods for machine vision, based on our survey, people with disabilities are currently poorly represented in visual learning benchmarks. Datasets also tend to be ambiguous in their annotations of carried or worn objects, e.g., these may be annotated as part of a generic "stuff" class, as a separate object (e.g., COCO), or as part of the person mask (e.g.,

Cityscapes). By developing a novel platform and benchmark emphasizing rare pedestrians with mobility aids, our work is complementary to ongoing efforts towards more diverse data and detailed learning tasks.

Simulation for Machine Vision: Our approach is inspired by recent advancements in realistic three-dimensional virtual worlds. Simulated worlds have been previously used to cheaply collect vast amounts of diverse, finely-annotated data and benchmark various tasks, e.g., pedestrian detection and segmentation [48, 23, 47] and sensorimotor control [15, 35, 18, 45, 51]. In particular, the CARLA simulation [18] has gained popularity as a realistic and easy-to-use environment. CARLA includes support for state-of-the-art sensing and is often used for benchmarking vision-to-action pipelines by researchers. Nonetheless, as in the former case of large-scale real-world benchmarks, current simulation environments lack support for dynamic simulation of people with disabilities. The applicability of various generalization techniques [11, 56] has also not been evaluated in the context of accessibility. To broaden the impact of our contributions on future research and development in the computer vision and robotics communities, we chose to integrate our novel accessibility modules into CARLA. By building upon CARLA's simulation capabilities, we are also able to go beyond many current real-world benchmarks to test instance segmentation models across drastically varying weathers, perspectives, and geographical locations.

Vision for Accessibility: Systems for assisting people with disabilities are generally user-centered in their operation [21, 20]. For example, systems may use computer vision to identify curb ramps [29] or intersection crossings [40] from the perspective of a user. Yet, very few assistive or autonomous systems have been trained to *perceive other surrounding pedestrian with disabilities*. Since data collection requires involving people with specific disabilities, the cost of coordinating user studies is prohibitive. Moreover, benchmarks are rarely shared due to privacy concerns. Therefore, in contrast to large-scale vision benchmarks such as COCO and Cityscapes, accessibility-oriented studies tend to focus on small-scale settings in simplified environments (e.g., a fixed indoor room [34, 57, 41, 3]). Kollmitz et al. [34] and Vasquez et al. [54] use Fast R-CNN to detect mobility aids in a single indoor hospital environment. FSOD [22] formulates training a wheelchair detector in a few-shot learning framework [55]. However, modeling diverse accessibility events goes beyond few-shot learning and involves addressing multiple fundamental issues in model generalization, as we demonstrate in our analysis. For instance, we find a general failure to detect canes and reasoning over person-aid context even with ample training data. Related to our topic are methods for sign language recognition (e.g., Koller et al. [33] and Camgoz et al. [6, 7]), but these generally assume a controlled setup with a centralized person. To scale such efforts beyond controlled labs and simplified settings, our study emphasizes learning broad and general-purpose accessibility-related vision models. Here, we seek to analyze for the full range of realistic environments, i.e., dense urban settings with people at various image scales, a range of weather and lighting conditions, and with diverse mobility aids. Our implementation can also easily extend to incorporate additional visual and learning tasks by researchers and developers in the future.

3. The X-World Platform and Benchmark

In this section, we describe X-World, our comprehensive data generation and model evaluation platform. X-World comprises (1) an expansive set of mobility aid designs integrated with realistic person-aid interaction into the open-source, three-dimensional virtual CARLA environment (Section 3.1), (2) a large, multi-perspective, fine-grained instance segmentation dataset collected by spawning navigating agents in the simulation (Section 3.2), and (3) accompanying real-world benchmark for comparative analysis to the simulation benchmark (Section 3.3).

3.1. Diverse Person Simulation Module

In building our realistic accessibility world, we sought to leverage recent advances in simulation rendering and physics, in particular the widely used CARLA environment [18] implemented with Unreal Engine 4 (UE4) [24]. While CARLA includes a rich array of environmental conditions and digital assets, it does not provide any support for accessibility-related events. Given the highly active research and development communities built around CARLA, integration provides a crucial step towards dissemination and future extensibility of our contributions. To enable the field to tackle important problems in accessibility, we introduce novel pedestrian types based on realistic visual and kinematic aspects of people with disabilities.

Mobility Aid Design: As a first step towards teaching machines to perceive pedestrians with disabilities, e.g., recognize a walking cane as an oncoming obstacle and a salient cue for a pedestrian's future trajectory, we introduce a diverse set of 28 new models (some are visualized in Fig. 1). The digital assets include canes with different visual patterns and various wheelchair designs, inspired by real-world instances (Section 3.3). While we focus on two representative categories of mobility aids, i.e., wheelchairs and walking canes, additional asset models can be integrated in the future. The extensive set enables novel model generalization analysis in Section 4. In our implementation, the various wheelchair and cane assets can be attached to any of the existing pedestrians in CARLA thereby changing their pedestrian type and behavior, as described next.

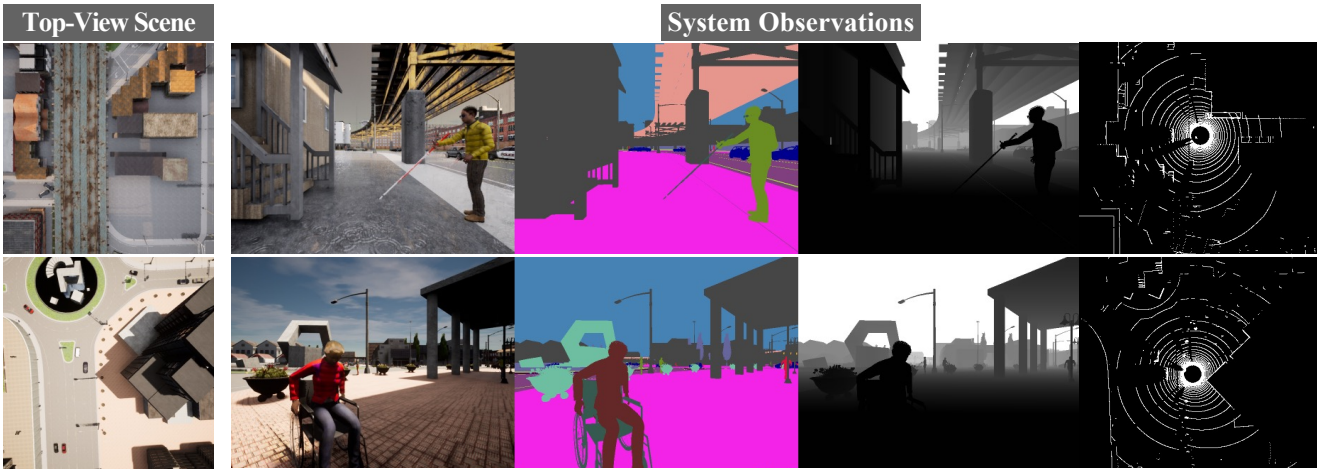

Figure 2: **Accessibility-Centered Urban Simulation World**. Our novel open-source simulation module enables to dynamically and realistically simulate people with disabilities. The simulation can be used for various use-cases, including an on-road autonomous vehicle or a sidewalk delivery robot, for training broad and general-purpose perception and interaction models. The interactive environment supports continuously collecting multi-modal data, e.g., RGB camera, semantic segmentation, depth, LIDAR.

Pedestrian Dynamics: We leverage CARLA 0.9.10 which includes a total of 28 pedestrian models (22 adult and six child assets). While CARLA pedestrians exhibit very basic walking skills, we utilize the in-built skeleton control to tune the poses, motion, and path selection of pedestrians based on their mobility aid. Examples of pedestrian navigation behavior are visualized in Fig. 2 together with the various perspectives used in the experiments and recorded sensor data. To ensure realistic pedestrian appearance and dynamics, we manually set kinematic parameters, (e.g., gaits, damping factor), and incorporated multiple cane and wheelchair techniques (e.g., side-to-side tactile scanning, tapping) to match common real-world mobility and orientation behavior. Our integration includes realistic obstacle avoidance, crowd navigation, and collision handling for both the pedestrians and their mobility aids. Moreover, we can incorporate rare safety-critical scenarios, e.g., near-crashes and accidents, into our visual reasoning benchmark. While realistic simulation of agent dynamics is an open research question, our carefully tuned behaviors can be used to generate highly realistic scenarios, as shown in Fig. 2. We emphasize an instance segmentation task as it provides a crucial step towards disability-aware intelligent systems.

3.2. Person-Aid Instance Segmentation Benchmark

By employing the simulation module from Section 3.1, we address inherent concerns in our visual reasoning task, i.e., issues relating to the significant rarity of events in the real-world, the cost of obtaining fine-grained annotation, and privacy. We now use the simulation to advance over previous work in instance segmentation. Specifically, we analyze an instance segmentation task which incorporates a detailed and functional understanding of person-object interactions in the context of our application. The task and benchmark will then be used to analyze current limitations of instance segmentation models in Section 4.

Generating Instance Segmentation Annotations: Instance segmentation involves an essential visual task by a robot when perceiving a scene with person-aid interactions. CARLA provides support for a rich set of sensors, including RGB and depth cameras, LIDAR, radar, dynamic vision, and inertial measurements. However, it does not provide built-in support for generating 2D instance segmentation ground truth. We extract such annotations by post-processing the ground truth depth map and the 3D bounding box annotations. Specifically, we back-project pixels in the semantic segmentation map into the 3D point cloud, and use the 3D bounding box annotations to determine an instance ID per-pixel. As visualized in Fig. 3, we found this solution to provide accurate annotations.

Simulation Data Collection Methodology: We collect an extensive dataset by spawning navigating agents and recording their interactions with the simulation throughout pre-specified routes. We utilize data collected across six perspectives, five towns, six ambient settings with various weather conditions and times of day (similarly to [12]), novel pedestrian types, and the extensive set of mobility aids. Thus, beyond addressing the current lack of pedestrians with disabilities and aids in large-scale datasets, our dataset contains challenging environmental and situational variability. To provide perspective variability, we also address two realistic use cases of an autonomous vehicle and a ground sidewalk delivery robot. This is done by spawning four cameras around an on-road vehicle and two cameras on pedestrians with different heights for a sidewalk per-

Figure 3: **Instance Segmentation Ground Truth in Simulation.** We visualize two examples with RGB camera view and corresponding instance segmentation ground truth.

Figure 4: **Real-World Examples.** A diverse real-world benchmark is used to ensure generalization.

spective (see Fig. 2). The dataset, collected from continuous video sequences of forty hours of navigation in various towns and ambient settings, is then sub-sampled to a total of 72,000 images containing 224,951 pedestrians (out of which 18,803 and 25,575 are wheelchair and cane users, respectively), 28,808 riders, and 254,560 vehicles.

3.3. Diverse Real-World Benchmark

We sought to provide a comprehensive validation of our findings by analyzing models both in simulation (with diverse weathers, towns, aids, perspectives) and in the real-world (prohibitive to collect in our application, but with complementary naturalistic long-tail cases). We therefore accompany the simulation benchmark from Section 3.2 with a smaller, but diverse annotated real-world benchmark. As collecting a representative dataset in the physical world is difficult due to the rarity of instances, we leverage in-situ navigation videos and images from publicly available internet sources. Altogether, we identified 80 videos (primarily from YouTube). Next, to ensure meaningful evaluation and model generalization, we deliberately selected a subset of the images involving dense urban settings and a high variability of scenarios, geographical locations, weathers, pedestrian poses, and camera perspectives. Finally, the resulting 383 images have been annotated with instance segmentation labels. In total, the challenging real-world benchmark contains 1,494 pedestrians, including 264 in wheelchairs and 209 cane users, and 787 vehicles. Examples from the dataset can be seen in Fig. 4.

4. Analysis

4.1. Dataset Statistics

To provide context for the experimental analysis, we first discuss the statistics of the two introduced benchmarks. Statistics from Cityscapes and COCO are also shown for reference. We only consider vehicle and person categories to perform a meaningful comparison to the human-centered categories in our study.

Class Statistics: Fig. 5 compares the overall pixel statistics with respect to different categories. Note that existing benchmarks, such as COCO and Cityscapes, do not include mobility aid or person type samples and annotations. Moreover, different datasets may contain inconsistently labeled person instances, for example Cityscapes includes any handheld or worn bags as part of a pedestrian mask, in contrast to COCO which does not. Furthermore, suitcases and strollers are labeled as the 'stuff' category in Cityscapes. As Cityscapes is captured in on-road settings, we observe similar person and vehicle statistics in Fig. 5 to our simulated dataset. In our real-world dataset, internet-sourced images and videos tend to incorporate centralized pedestrians resulting in larger instances.

Density and Size Statistics: Fig. 6 quantifies the density of categories and instances in the datasets, per image. While both introduced datasets contain dense and challenging urban scenes, our generated simulation dataset contains more complex images with crowded pedestrian and vehicle traffic. We observe similar trends in terms of density statistics to Cityscapes. Fig. 7 analyzes the pedestrian size distribution among datasets. Despite our inclusion of many safety-critical, near-crash events in the large simulated dataset, a significant portion of the instances in this dataset is found at greater distances from the camera. This is to be expected, as traffic scenes with busy intersections can contain many instances along the background. In contrast, internet-sourced images and videos may contain a bias towards larger instances with fewer categories (e.g., as also shown in Fig. 6). Nonetheless, the two complementary benchmarks cover a significant range of scales and object sizes.

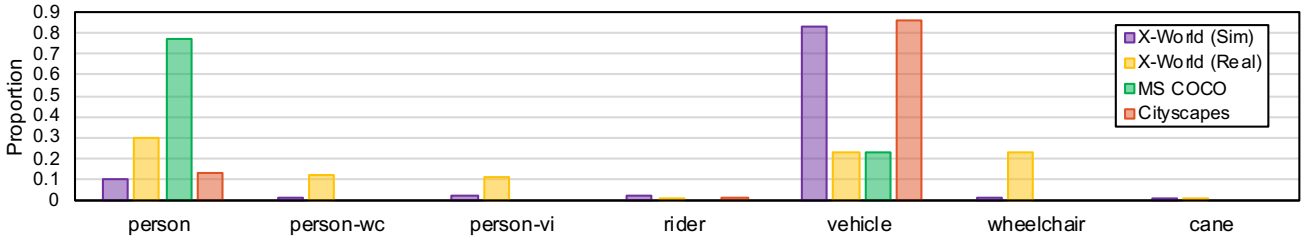

Figure 5: **Semantic Annotation Statistics.** Comparison of the proportion of per-category annotated pixels in our simulation and real-world benchmarks. 'person-wc' and 'person-vi' refer to person in wheelchair and visually-impaired person, respectively. COCO and Cityscapes do not include mobility aid annotations.

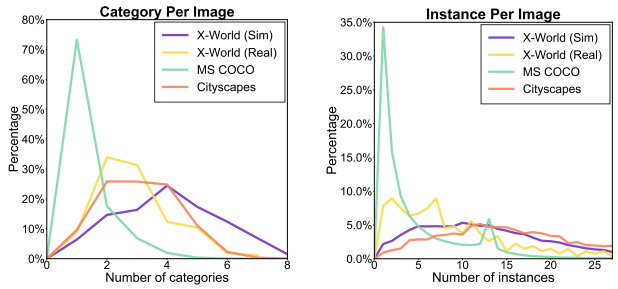

Figure 6: **Density Statistics.** Our simulated and real-world benchmarks include dense urban settings, as shown by the number of object categories per image (left) and number of overall object instances per image (right).

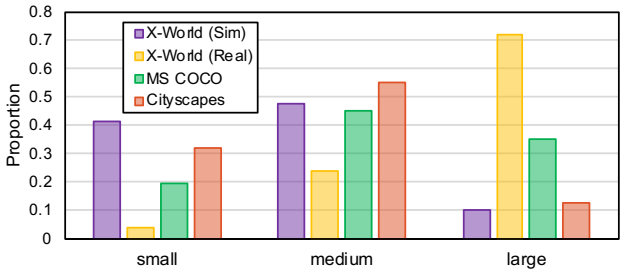

Figure 7: **Person Size Statistics.** Complementary size distribution between simulated and real-world data.

4.2. Experiments

To gain insights into our novel perception task, we conduct three main experiments. First, we motivate our study by analyzing the limitations of current models on our introduced pedestrians and settings in simulation. Second, we explore the ability of models to robustly recognize fine-grained accessibility-related categories in diverse settings. Third, we expand the analysis by investigating model performance in the real-world. Throughout the experiments, we follow He et al. [30] and train a standard Mask R-CNN instance segmentation model with a ResNet-50 [31] backbone. We compare model pre-training using either COCO or Cityscapes.

Baseline Performance and Challenges: We begin by evaluating the performance of traditional person detection models in Table 1. Common approaches [13, 39, 60] train a singular 'person' category model, i.e., agnostic to the different underlying 'person-wc' and 'person-vi' categories. While such models are not explicitly supervised to recognize finer-grained 'person' categories, we can leverage our annotations to uncover their performance in testing. Here, we produce a breakdown over each of the introduced 'person' categories, by ignoring all remaining categories. Overall, we find that our diverse simulated benchmark provides a highly challenging test setting. Notably, off-the-shelf models are unable to handle persons in wheelchairs. After fine-tuning the two-category ('person' and 'vehicle') model with the simulated dataset, we still find significantly degraded performance on 'person-wc' and 'person-vi' instances. Interestingly, despite the similarity in domain and statistics to Cityscapes, the pre-trained COCO model outperforms a Cityscapes model. This improved generalization to our simulated data could be due to the broader scope of COCO images, e.g., in terms of perspective and ambient settings. After fine-tuning, we find both models to perform comparatively. Overall, the 'person-wc' class is shown to be the most challenging class, with unsatisfactory performance. Motivated by the impact of different 'person' categories on model performance, we continue to analyze additional aspects of our rich benchmarks in subsequent experiments.

Fine-Grained Perception Experiment: In this experiment, we introduce additional accessibility-related classes, such as mobility aids and person types, during training of the model. Consequently, we analyze the model's ability to segment the eight key object categories that can be found in our simulation benchmark (Table 2). In particular, we analyze challenges in segmenting the fine-grained categories under harsh generalization settings of new towns, new weathers and ambient settings, and previously unseen mobility aids. Our benchmark also includes a two-wheeled vehicle category aimed to analyze whether its addition can help the model better discern wheelchair in-

Training	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	person	person-wc	person-vi	vehicle
COCO	35.1	62.0	33.5	14.4	54.7	76.6	30.2	8.9	36.0	40.2
COCO & X-World (Sim)	59.6	86.5	66.8	28.1	60.7	82.7	62.8	31.3	47.4	59.2
Cityscapes	25.1	48.9	21.7	9.6	43.9	63.1	19.4	1.2	24.4	29.8
Cityscapes & X-World (Sim)	60.0	86.6	67.5	28.2	61.0	83.1	63.0	31.5	48.0	59.8

Table 1: **Uncovering Challenges Inherent to Our Task.** We follow traditional instance segmentation model training with two disability-agnostic object classes, person and vehicle. Results are shown **in simulation** for pre-trained and fine-tuned versions of the Mask R-CNN model. We leverage pedestrian type annotations to analyze the models under our novel task, i.e., breakdown of the model performance over the three person categories obtained by setting other person types as ignore regions. Results are shown for segmentation on the simulation-based test set ('wc' - wheelchair, 'vi' - visually-impaired).

Training	T	W	A	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	person	person-wc	person-vi	rider	v-4w	v-2w	wc	cane
	✓			43.9	73.4	45.0	22.0	50.8	65.3	63.8	38.0	54.8	49.5	64.0	47.2	33.2	0.61
COCO &	✓	✓		43.7	72.7	45.7	22.6	50.5	62.5	65.5	37.3	55.1	51.0	62.5	44.3	33.9	0.12
X-World (Sim)	✓		✓	39.9	68.1	40.9	19.8	45.2	61.4	64.7	26.0	54.1	45.5	61.6	45.7	20.9	0.86
	✓	✓	✓	40.6	68.9	41.3	20.7	46.5	62.7	66.1	30.0	51.8	46.9	60.9	45.4	23.5	0.53
	✓			44.2	73.2	45.8	22.0	51.2	64.9	64.4	38.1	55.3	50.4	63.8	47.2	33.9	0.53
Cityscapes &	✓	✓		43.9	72.8	46.0	22.5	50.9	62.6	66.1	37.1	55.4	51.6	62.0	45.0	34.1	0.20
X-World (Sim)	✓		✓	40.5	68.7	41.6	20.1	46.2	61.7	65.4	26.5	54.7	45.9	61.8	46.7	22.1	0.98
	✓	✓	✓	41.3	69.7	42.4	21.1	47.4	63.9	66.5	30.6	52.4	46.7	60.9	46.9	25.6	0.85

Table 2: **Fine-Grained Instance Segmentation.** Fine-tuned models trained with eight object classes, including different person types, mobility aids ('wc' - wheelchair, 'vi' - visually-impaired), and vehicles ('v-4w' - vehicle with four wheels, 'v-2w' - vehicle with two wheels). Results are shown **in simulation**. Evaluation is performed using different test settings, with 'T' - new towns test setting, 'W' - new weathers test settings, and 'A' - new mobility aids test settings.

Training	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	person	person-wc	person-vi	rider	v-4w	v-2w	wc	cane
COCO	36.8	59.7	40.6	13.9	31.4	45.9	48.8	22.5	69.0	13.3	46.6	12.4	-	-
COCO & X-World (Sim)	6.6	13.8	6.1	1.6	4.2	8.7	9.2	5.1	19.9	0.0	12.9	0.0	5.7	0.10
COCO & X-World (Real)	29.8	57.3	28.9	11.1	24.2	34.7	39.0	28.2	57.6	17.6	47.6	14.4	33.7	0.66
COCO & X-World (Sim+Real)	28.6	49.8	28.4	15.4	26.2	32.6	38.5	30.3	55.7	9.5	42.9	14.2	36.1	1.83
Cityscapes	21.7	40.2	21.3	6.0	19.9	27.6	35.7	4.5	53.0	3.6	38.2	7.4	-	-
Cityscapes & X-World (Sim)	5.8	11.9	4.8	1.7	3.3	7.9	8.5	5.1	15.0	0.0	10.9	0.0	6.8	0.09
Cityscapes & X-World (Real)	26.6	47.9	26.7	6.3	23.2	31.8	37.6	29.4	55.9	0.0	45.7	13.1	30.3	0.70
Cityscapes & X-World (Sim+Real)	27.5	50.4	26.2	6.4	22.5	32.8	37.0	30.0	55.3	4.0	43.4	15.3	33.9	1.41

Table 3: **Real-World Benchmark Analysis.** Baseline and fine-tuned instance segmentation model performance for the **real-world benchmark**. The Mask R-CNN model is trained using either simulated only, real-world only, or a combined simulated and real-world dataset. The APs for the baseline model (highlighted in red) are shown for reference as they are computed without the 'wc' and 'cane' classes.

stances. Surprisingly, even with ample training data for wheelchair and cane instances, we demonstrate reduced performances. Specifically, we find Mask R-CNN to exhibit general failure on the 'cane' category (relevant to obstacle avoidance) with poor reasoning over the geometry, appearance, and person-aid context. This failure leads to an abysmal performance of less than 1% accuracy. Moreover, classes such as 'wheelchair' may be confused with the two-wheeled vehicle 'v-2w' class, which also occurs in our real-world benchmark experiment (see Fig. 8). When considering the four generalization test cases, we find segmentation under new weather conditions to be an overall easier task compared to testing on the new mobility aids (e.g., wheelchair types unseen in training). For consistency with existing work on CARLA, in particular the recent no-crash benchmark [12, 10], we used a similar weather training and testing split. It seems that for our novel task, the new testing weathers only present a mild challenge. This is also consistent with previous studies for other tasks on CARLA [12, 10]. Nonetheless, as state-of-the-art models improve their generalization performance in the future, so can our platform be used to explicitly generate increasingly difficult benchmarks, i.e., under more drastic weathers, occlusion, appearance, and pose variations. We therefore reiterate our findings of novel challenges introduced by our platform and benchmarks.

Real-World Analysis: We perform a final experiment in Table 3 using the real-world benchmark. We find the results on the real-world benchmark to be generally consistent with

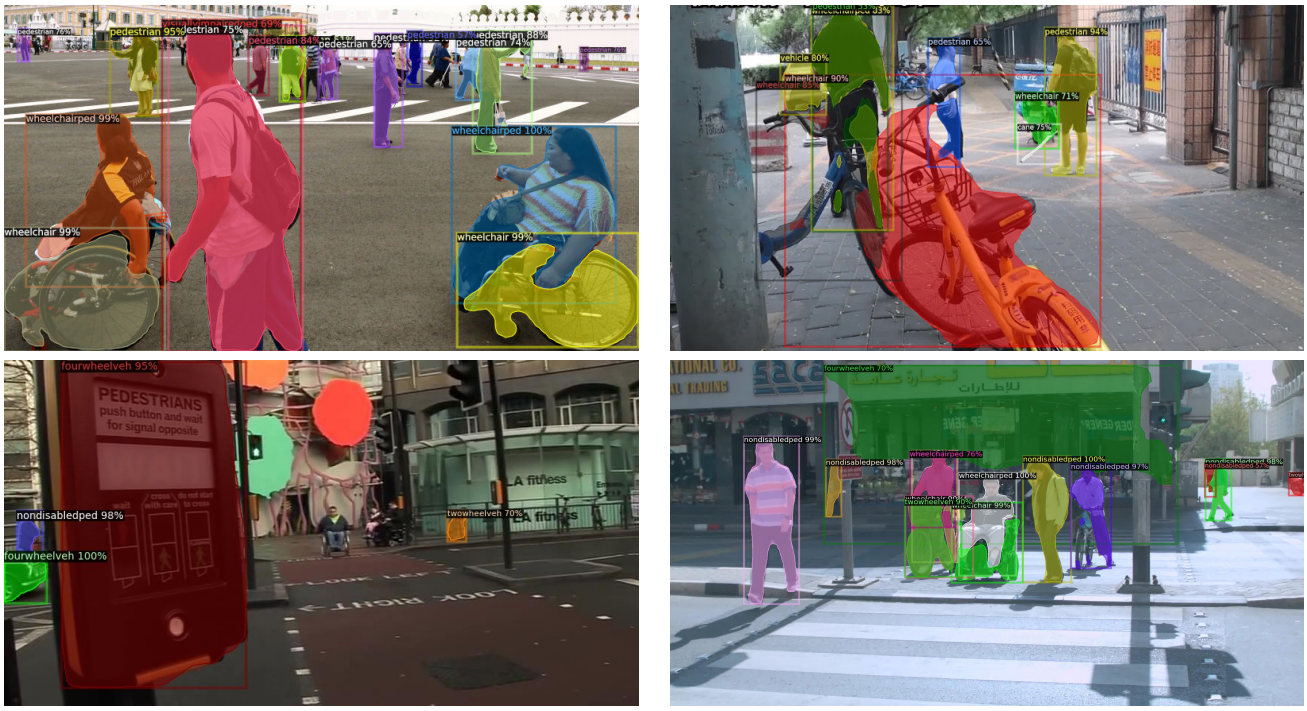

Figure 8: **Qualitative Results on the Real-World Benchmark.** Top-left: successful detection and segmentation of people and mobility aids. Top-right: a failure case with wheelchair false positive instances (often confused with two-wheeled vehicles and bicycles). Bottom-left: a failure case with wheelchair pedestrians and wheelchair false negative instances. Bottom-right: a failure case with two-wheeled vehicle false negative instances.

the results in simulation (Tables 1 and 2). For instance, the performance on the novel pedestrian types of 'person-wc' and 'person-vi' show similar trends in terms of absolute performance. One difference of note is the 'person-vi' class outperforms the 'person' class in the real-world data (Table 3). As real-world imagery contains mostly centralized and isolated instances of visually-impaired pedestrians, such cases provide easier settings for detection and segmentation, thereby leading to higher overall performances. Similar trends between the tables are shown for the mobility aid classes as well. This finding serves as affirmation of the simulation's utility. To further validate the simulation environment, we also leverage the real-world data and analyze simulation-to-real generalization. In general, Table 3 shows poor generalization performance when training in simulation and testing in the real-world. While this is difficult due to a domain shift, we also explore boosting the performance of models trained on real data using simulated data [48]. By combining the datasets, we find improved performances for the challenging mobility aid classes in Table 3, achieving state-of-the-art for the 'cane' and 'wc' categories. The simulation dataset also benefits performance over the most difficult person class of 'person-wc'. Fig. 8 demonstrates several failure cases, where a 'rider' may be misclassified as a 'person-wc' as well as cases of false negative instances.

5. Conclusion

This paper takes an initial step toward learning robust and detailed models for perceiving diverse people with disabilities. By centering our analysis on often neglected accessibility-related tasks and events, we contribute towards safe and effective operation of autonomous systems. We find that the novel X-World benchmark is complimentary to existing benchmarks. We hope that the uncovered research challenges and opportunities can spark the interest of computer vision and robotics researchers, thereby broadly influencing future development and evaluation of visual-reasoning models. In our paper, we focused on a fundamental perception task of fine-grained instance segmentation. Nonetheless, our benchmark and tools support analysis of additional learning tasks and use-cases, e.g., policy learning [59], simulation-to-real adaptation [50, 11], and transfer learning [55, 56]. By releasing X-World, we hope such tasks can be studied by researchers tackling the multifaceted problem of accessibility-aware vision-based machines.

Acknowledgments: This work was supported in part by the Department of Transportation (the Inclusive Design Challenge), the BU Center for Information and Systems Engineering, and a BU Undergraduate Research Opportunities Program Grant.

References

[1] National Highway Traffic Safety Administration. Wheelchair users injuries and deaths associated with motor vehicle related incidents. In *Research Note. Department of Transportation*, 1997. 2

[2] Daniel H Ashmead, David Guth, Robert S Wall, Richard G Long, and Paul E Ponchillia. Street crossing by sighted and blind pedestrians at a modern roundabout. *Journal of Transportation Engineering*, 131(11):812–821, 2005. 1

[3] Lucas Beyer, Alexander Hermans, and Bastian Leibe. Drow: Real-time deep learning-based wheelchair detection in 2-d range data. *IEEE Robotics and Automation Letters*, 2(2):585–592, 2017. 3

[4] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M Gavrila. The EuroCity persons dataset: A novel benchmark for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1844–1861, 2019. 2

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. Nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2

[6] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *ICCV*, 2017. 3

[7] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*, 2020. 3

[8] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3D tracking and forecasting with rich maps. In *CVPR*, 2019. 2

[9] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 2

[10] Dian Chen, Brady Zhou, and Vladlen Koltun. Learning by cheating. In *CoRL*, 2019. 7

[11] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 2018. 3, 8

[12] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *ICCV*, 2019. 4, 7

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 6

[14] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[15] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. RoboTHOR: An Open Simulation-to-Real Embodied AI Platform. In *CVPR*, 2020. 3

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[17] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012. 2

[18] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: an open urban driving simulator. In *CoRL*, 2017. 2, 3

[19] Markus Enzweiler and Dariu M Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE transactions on pattern analysis and machine intelligence*, 31(12):2179–2195, 2008. 2

[20] Navid Fallah, Ilias Apostolopoulos, Kostas Bekris, and Eelke Folmer. The user as a sensor: navigating users with visual impairments in indoor spaces using tactile landmarks. In *CHI*, 2012. 3

[21] Navid Fallah, Ilias Apostolopoulos, Kostas Bekris, and Eelke Folmer. Indoor human navigation systems: A survey. *Interacting with Computers*, 25(1):21–33, 2013. 3

[22] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *CVPR*, 2020. 3

[23] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 3

[24] Epic Games. Unreal engine 4. https://www.unrealengine.com. 3

[25] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012. 2

[26] Duane R Geruschat and Shirin E Hassan. Driver behavior in yielding to sighted and blind pedestrians at roundabouts. 99(5):286–302, 2005. 1

[27] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2D2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020. 2

[28] David Guth, Daniel Ashmead, Richard Long, Robert Wall, and Paul Ponchillia. Blind and sighted pedestrians' judgments of gaps in traffic at roundabouts. 47(2):314–331, 2005. 1

[29] Kotaro Hara, Jin Sun, Robert Moore, David Jacobs, and Jon Froehlich. Tohme: detecting curb ramps in google street view using crowdsourcing, computer vision, and machine learning. In *UIST*, 2014. 3

[30] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 6

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[32] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. 42(10):2702–2719. 2

[33] Oscar Koller, Hermann Ney, and Richard Bowden. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *CVPR*, 2016. 3

[34] Marina Kollmitz, Andreas Eitel, Andres Vasquez, and Wolfram Burgard. Deep 3d perception of people and their mobility aids. 114:29–40, 2019. 3

[35] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv preprint arXiv:*, 1712.05474, 2017. 3

[36] John D. Kraemer. Epidemiology of non-fatal us emergency room visits for road crashes involving pedestrians in wheelchairs. *Injury prevention*, 21(5):331–334, 2015. 2

[37] John D. Kraemer and Connor S. Benton. Disparities in road crash mortality among pedestrians using wheelchairs in the usa: results of a capture–recapture analysis. *BMJ open*, 5(11):e008396, 2015. 1, 2

[38] Myron M LaBan and Thomas S Nabity Jr. Traffic collisions between electric mobility devices (wheelchairs) and motor vehicles: Accidents, hubris, or self-destructive behavior? *American journal of physical medicine & rehabilitation*, 89(7):557–560, 2010. 2

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 6

[40] Sergio Mascetti, Dragan Ahmetovic, Andrea Gerino, and Cristian Bernareggi. Zebrarecognizer: Pedestrian crossing recognition for people with visual impairment or blindness. *Pattern Recognition*, 60:405–419, 2016. 3

[41] Amir Mukhtar, Michael J Cree, Jonathan B Scott, and Lee Streeter. Mobility aids detection using convolution neural network (cnn). In *IVCNZ*, 2018. 3

[42] Kenneth Nemire. Case study: Wheelchair conspicuity at night. *HFES*, 2010. 2

[43] The Pitt News. Pitt pauses testing of starship robots due to safety concerns. https://pittnews.com/article/151679/news/pitt-pauses-testing-of-starship-robots-due-to-safety-concerns/, 2019. 2

[44] Eshed Ohn-Bar, Kris Kitani, and Chieko Asakawa. Personalized dynamics models for adaptive assistive navigation systems. In *CoRL*, 2018. 2

[45] Eshed Ohn-Bar, Aditya Prakash, Aseem Behl, Kashyap Chitta, and Andreas Geiger. Learning situational driving. In *CVPR*, 2020. 3

[46] Quang-Hieu Pham, Pierre Sevestre, Ramanpreet Singh Pahwa, Huijing Zhan, Chun Ho Pang, Yuda Chen, Armin Mustafa, Vijay Chandrasekhar, and Jie Lin. A*3D dataset: Towards autonomous driving in challenging environments. In *ICRA*, 2020. 2

[47] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *ICCV*, 2017. 3

[48] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 3, 8

[49] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020. 2

[50] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *CVPR*, 2018. 8

[51] Axel Sauer, Nikolay Savinov, and Andreas Geiger. Conditional affordance learning for driving in urban environments. In *CoRL*, 2018. 3

[52] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 2

[53] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 2

[54] Andres Vasquez, Marina Kollmitz, Andreas Eitel, and Wolfram Burgard. Deep detection of people and their mobility aids for a hospital robot. In *ECMR*, 2017. 3

[55] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. 3, 8

[56] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *ICCV*, 2019. 3, 8

[57] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *ECCV*, 2020. 3

[58] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 2

[59] Jimuyang Zhang and Eshed Ohn-Bar. Learning by watching. In *CVPR*, 2021. 8

[60] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. CityPersons: A diverse dataset for pedestrian detection. In *CVPR*, 2017. 2, 6