A Comprehensive Crossroad Camera Dataset of Mobility Aid Users

Ludwig Mohr
ludwig.mohr@icg.tugraz.at
Nadezda Kirillova
nadezda.kirillova@icg.tugraz.at
Horst Possegger
possegger@icg.tugraz.at
Horst Bischof
bischof@icg.tugraz.at

Institute of Computer Graphics and Vision
Graz University of Technology
Graz, Austria

Abstract

Improving the safety of traffic participants and reducing the severity of injuries as well as the number of fatalities in the event of accidents is becoming ever more important in the development of vehicles and transportation infrastructure. The most vulnerable group of road users is unquestionably pedestrians, of which people with mobility impairments are especially at risk due to reduced reaction speed or reduced visibility due to smaller silhouettes or unusual postures. Successful strategies for increasing safety by reducing the likelihood of accidents include architectural improvements in planning of pedestrian crossings, as well as advancements in their operation. These strategies can benefit from camera based pedestrian detection systems, yet pedestrians using mobility aids are highly underrepresented in common datasets for object detection and classification, if present at all. To fill this gap and enable researchers to develop methods considering pedestrians in their mobility, we present a novel dataset of pedestrians using mobility aids, together with evaluations of state-of-the-art methods for classification and detection.

1 Introduction

In the effort to reduce the number of traffic accidents as well as their impact, traffic safety research puts particular emphasis on safety improvements for pedestrians, since they usually suffer the most serious injuries in traffic accidents [25, 39]. Among this already vulnerable group of road users, however, pedestrians requiring mobility aids are even more at danger. For example, pedestrians using wheelchairs report a 36% higher mortality rate than standing pedestrians in car-to-pedestrian collisions [9, 16]. Potential reasons for such higher risks can be partially attributed to longer reaction times due to impaired mobility, or their unusual posture, leading to potentially reduced visibility.

Strategies to improve pedestrian safety range from wider integration of Pedestrian Protection Systems (PPS) into Advanced Driver Assistance Systems (ADAS), which are already

becoming mandatory in some areas [5], to increasing safety of traffic interactions by specific measures in planning and operation of pedestrian crossings [36]. These road safety strategies can benefit from camera-based pedestrian detection systems, which enable visual perception of the environment as a crucial component in ADAS, or allow us to perform detailed analysis of traffic scenarios in order to optimize crossing layouts and traffic flow in general. Both approaches to improved road safety require a detailed understanding of which road users are present. Knowledge about the presence of pedestrians using mobility aids may help to further improve these approaches in several ways, for instance by avoiding bias in detection systems towards upright walking and standing pedestrians, by helping city planners to develop roads according to specific needs and to incorporate this knowledge into active traffic management. For instance, through the use of traffic lights enabled for on-demand and online scheduling of light cycles, the presence of pedestrians requiring mobility aids can be taken into account by increasing the duration of their green phase, allowing for a possibly longer time required to safely clear a crossing.

While there has been significant progress regarding the robustness and reliability of pedestrian detection models, the consideration of pedestrians reliant on mobility aids is lacking behind. Incorporating them is a difficult task currently, since most of the recent progress is driven by deep learning approaches and therefore depends on the availablibility of adequate training data, yet the representation of pedestrians using mobility aids is lacking in public datasets, being underrepresented at best or not represented at all at worst. However, such data would be needed to foster research on detecting the presence of mobility aids and to further improve the safety of these especially vulnerable road users.

To help fill this gap and to enable researchers to consider the presence of mobility aids in the development of pedestrian detectors, in this work we present a dataset for the detection and classification of pedestrians using four different types of mobility aids (wheelchairs, rollators, crutches and walking canes) in images taken from standard traffic monitoring viewpoints, *i.e*. cameras mounted on traffic lights or poles. Furthermore, we provide detailed evaluations demonstrating the performance of state-of-the-art models for both, classification and detection and consider robust classification and detection, formulating a hierarchical class-dependency among all pedestrian classes, thereby allowing an explicit safe fall-back for the classification in case of uncertainty.

2 Related Work

In object detection, approaches can be classified roughly into two different categories. The first, traditional category treats object *detection* and object *classification* as two separate and independent steps, *i.e*. performing classification of the object proposals as a second step on the result of the object detector. Well-known members of this category are the R-CNN variants [7, 8], depending on object proposals as input, and methods like Faster R-CNN [28] and MSC-MultiBox [32], which integrate the generation of object proposals into their pipeline. The more recent class of so called *one-shot detectors* like SSD [23] and YOLO (You Only Look Once) [35] treat object detection and classification as a compound task, solved by a single forward pass through a deep neural network. Since these methods yield object proposals jointly with classification scores in a single forward pass of a deep neural network, their speed is vastly increased while simultaneously reducing computational needs compared to traditional two-stage detectors. Recent YOLO variants [15] comprise the state-of-the-art among the one-shot detectors, due to their favorable performance, ease of use

and readily available pre-trained model weights.

Alternatively, pedestrian attribute recognition approaches can also be used to identify mobility aid use in a top-down approach, *i.e*. performing classification after detection of a generic person/pedestrian class. Such approaches typically leverage the strong representational capabilities of convolutional neural networks to extract image features and apply a classification head to estimate the attribute categories. The underlying feature extractor significantly contributes towards the accuracy and inference speed and ranges from AlexNet backbones, in [17, 31], to more accurate VGG or Inception backbones, *e.g*. in [19, 21]. This approach is highly effective for fine-grained category recognition and has lead to a variety of applications, including Human Attribute Recognition (HAR), *e.g*. [14], [17], [21], or [41], and Person Re-identification (Re-ID), *e.g*. [10], [18], [30], [38], or [42]. To improve accuracy even further, recent approaches such as [29, 40] adopt deeper backbones (ResNet and DenseNet) or try to learn feature representations tailored to pedestrian attributes, *e.g*. [2, 34]. We demonstrate the capabilities of classifiying mobility aids with a typical pedestrian attribute recognition pipeline using a wide variety of state-of-the-art feature extraction backbones.

As previously discussed, representation of mobility aids in existing datasets is sparse to non-existent. Of datasets geared towards autonomous driving, nuScenes [1] is the only one providing annotations for wheelchairs, yet still offers only very little annotations (35 instances in nuImages). In the common object detection and semantic segmentation datasets publicly available, mobility aids usually are not present at all. To our knowledge, the only dataset specifically targeted at people using mobility aids is the MobilityAids dataset [37], which is focused at the application of autonomous robots in hospitals, featuring a low view point and detection and classification of 3D data from depth cameras. In contrast to these our focus is on providing a dataset captured from a standard traffic monitoring viewpoint. This can facilitate research on detection and classification models that will help city and traffic planners to properly consider the needs of mobility aid users when designing crossroads or optimizing traffic light schedules.

3 Dataset

The aim of this work is to facilitate research into detection and classification of especially vulnerable traffic participants, specifically pedestrians with mobility impairments. To this end and to complement the little already available data [37], we chose a standard traffic monitoring viewpoint for recording, as can be found in CCTV cameras or when monitoring pedestrian crossings. Figure 1 shows examples of the recording setup and annotations. Besides persons not using any aids (class *pedestrian*), the dataset contains people using *wheelchairs*, *rollators*, one or two *crutches* and walking *canes*, covering a wide spectrum of the most common mobility aids. The dataset is available online at [24].

A major concern in the creation of datasets containing people is the protection of privacy and the non-infringement of personal rights, especially so when dealing with impairments or medical conditions. Image alterations such as blurring or cropping regions which could lead to identification would thwart the purpose of the dataset creation, since any alterations may give hints to object detection algorithms, disclosing information on object position or class. These constraints increase the difficulty of creating publicly available datasets focused on detection and classification of pedestrians using mobility aids, placing additional obstacles to data acquistion at rehabilitation centers where a high number of people using mobility

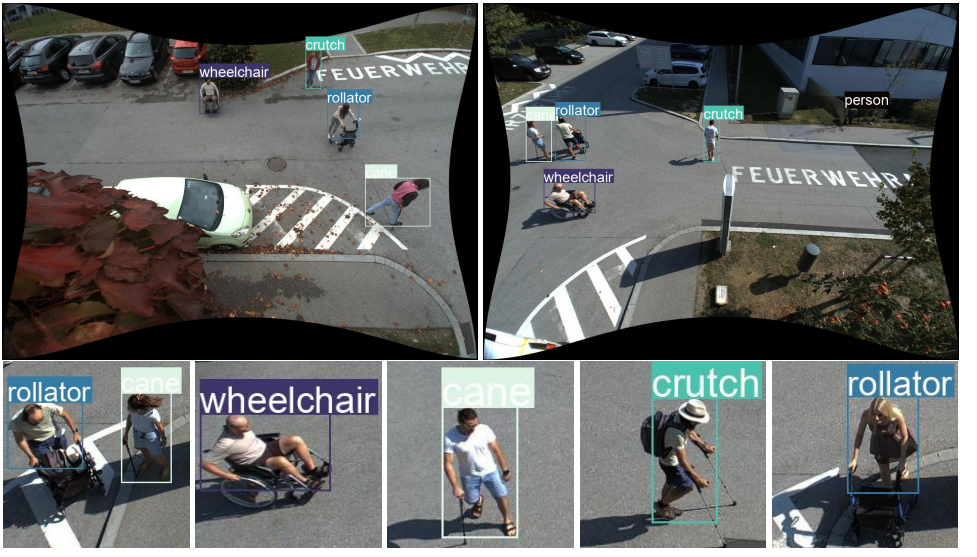

Figure 1: Exemplary dataset images with annotations. In the top row, the two camera viewpoints (image resolution is 1280×960) can be seen. The bottom row shows sample crops of the available mobility aids.

aids can be encountered. To reconcile privacy concerns on one hand with the need for a large dataset with many individuals and sufficient variability on the other hand, we chose to conduct data acquisition with able-bodied volunteers using a calibrated camera setup at an intersection on the university campus. This constraint of course means that although video streams were captured, analysis and classification based on more high-level cues like movement patterns or trajectories is not possible in a meaningful way, since these cues will not be similar to those encountered *in the wild*. Furthermore, at this stage it was not possible to close off the road section where data capture was performed, meaning that during capture also people not participating in the dataset recording were present at some times. All such samples were removed during annotation. Since for these reasons the source data does not lend itself to processing as a time series, annotation was done for frames at equidistant intervals of 4 seconds on a basis of pure visibility, meaning that especially for the thin mobility aids *cane* and *crutch*, annotations were only labeled as such if they were discernible for the human eye, and as *pedestrian* otherwise. Data was aqcuired on three days in summer and early autumn of 2022, resulting in a total of 17 hours of raw video. In total, 34 different people participated in data recording, some of them at several days. The data was divided into splits for training, validation and testing along full intervals to minimize subjects appearing in more than one split while trying to keep class frequencies representative over all splits. Each split is guaranteed to contain recordings of all three sessions. The validation split is about a tenth of the size of the training split, the testing split is about a third of the size of the training split. Data annotation has been realized by six human annotators, each sample was verified by at least two individuals. Keeping with the recommendations of YOLOv5 [15], each split contains about 9% of background images without any annotations. Table 1 lists the number of frames and total annotations per class for each split in the dataset.

Figure 2 shows the relative distribution of annotations for the whole dataset (*all*), as well as each split. Similarly, Figure 3 shows the histograms for the relative sizes of the

	frames			annotations									
split	total	background		pedestrian		wheelchair		rollator		crutch		cane	
train	8456	768	(9.0%)	10518	(48.1%)	3907	(17.9%)	3178	(14.5%)	2466	(11.3%)	1781	(8.2%)
val	784	71	(9.0%)	486	(30.8%)	418	(26.5%)	347	(22.0%)	153	(9.7%)	176	(11.1%)
test	2946	267	(9.0%)	3058	(43.5%)	1660	(23.6%)	787	(11.2%)	979	(13.9%)	540	(7.7%)

Table 1: Number of frames and annotations for each split in the dataset. The relative amount is given in percent in brackets.

bounding boxes. The dataset has only little bias towards preferred locations, the splits are a good representation for the whole data. Figure 4 shows the distribution of annotations per class. There exists no bias towards preferred locations for any class compared to the others. Lastly, Figure 5 shows the respective size and location of annotations. The bounding box size is color coded as the length of a square of same area. We can see small annotations of pedestrians on the sidewalk on the opposite side of the road. These were included since state-of-the-art detectors are able to detect such small structures, and ignoring them could lead to poor learning cues for detector training. Since in practice they are most likely not relevant, users may filter them before evaluation if the need arises.

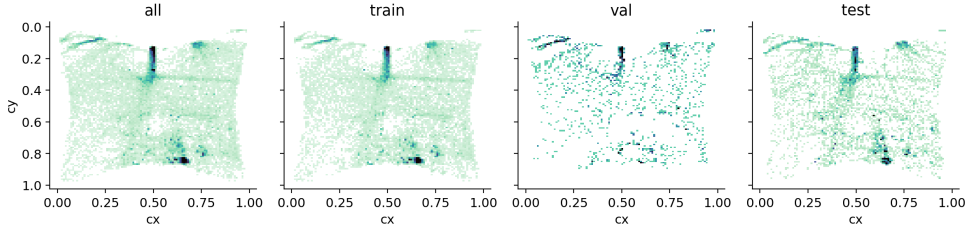

Figure 2: Distribution of bounding box centers for the complete dataset (left) and each split. Coordinates are relative to the respective image dimension.

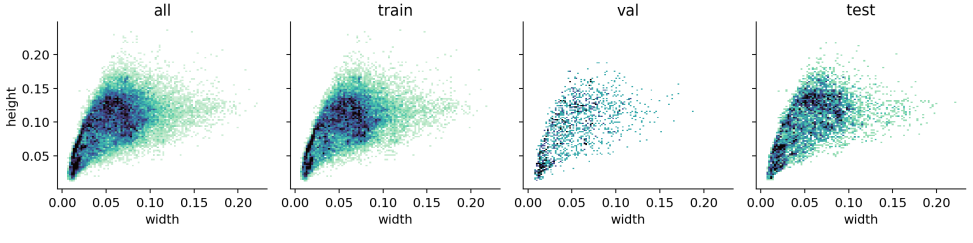

Figure 3: Distribution of bounding box sizes for the complete dataset (left) and the respective splits. Bounding box sizes are relative to the respective image dimension.

4 Classification

As the first baselines, we provide classification results of the methods detailed in Sec. 4.1. All approaches are trained, validated and evaluated using ground-truth annotations for detection, thereby eliminating the impact of object detectors in the dataset evaluation. This provides an estimate on the distinguishability of object classes, as well as hard cases and possibilities

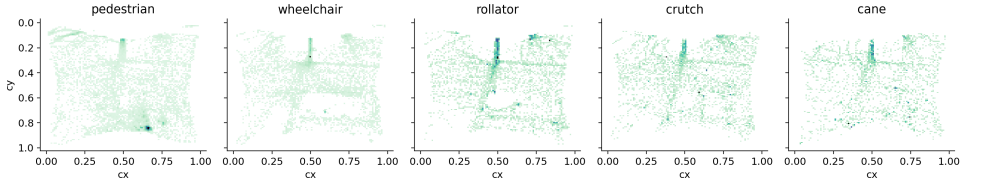

Figure 4: Distribution of labels (centers of respective annotation bounding boxes) per class.

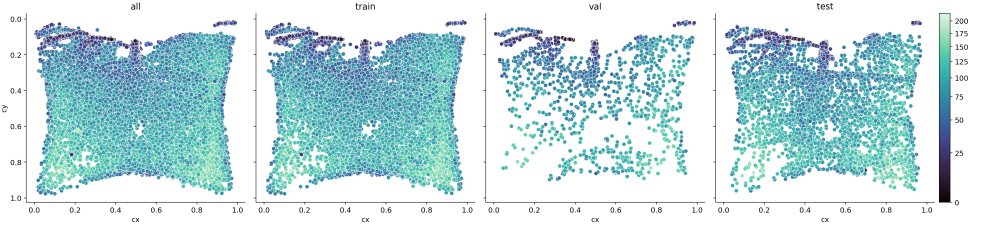

Figure 5: Distribution of annotations, color coded for bounding box size. The size is given as equivalent length of a square of same area. Again, coordinates are relative to the respective image dimension.

for confusion. Evaluation metrics and evaluations are decribed in Sec. 4.2 and Sec. 4.3, respectively.

4.1 Method

To classify pedestrians based on their use of mobility aids, we employ a standard whole-image classification method, based on a pre-trained convolutional neural network for feature extraction [33]. The classifier is learned with a transfer learning strategy, replacing and retraining the linear classification layer on top of the backbone for the proposed five classes, and fine-tuning the weights of the entire network.

We compare the classification performance of several commonly used feature backbones on our dataset, including MobileNetV3 Large (MobileNetV3 L) [12], different ResNet [11] variants, DenseNet201 [13], VGG16 [22], and ViT Base 16 (ViT-B/16) [4]. Our aim is to provide a diverse set of baselines with different characteristics, including shallow and deep models, light-weight as well as memory and compute-intensive models. All networks are pre-trained on the ImageNet-1K dataset [3] and fine-tuned on our dataset for 20 epochs using Sigmoid Cross-Entropy Loss with Adam optimizer and Step Decay every 5 epochs. For training we use a batch size of 32 and an initial learning rate of $1e^{-4}$. We employ the implementations provided in the PyTorch model zoo [26] with pre-trained weights, which reproduce the original results published by the respective authors (denoted ImageNet_1K_V1), as well as the improved TorchVision weights (denoted ImageNet_1K_V2), where available.

As a data pre-processing step, we extract image patches using bounding box annotations (examples can be found in the supplemental). Since the bounding boxes tightly enclose the person annotated, we provide the model with extended context by expanding the boundaries of the bounding boxes by 20 pixels on each side when extracting images patches. The extracted patches are resized to 224×224, their intensities are scaled to $[0,1]$ and normalized to be zero-mean, unit-variance.

4.2 Metrics

During fine-tuning on our dataset, we monitor the performance of the model after each epoch, focusing on the final validation accuracy score after 20 epochs. A more comprehensive comparison of the models in term of training performance and complexity can be found in the supplemental material. We report the classification performance in terms of the number of true positives (TP) and test accuracy (Acc).

4.3 Results

Table 2 reports the classification accuracy on the test set. All models predict classes *pedestrian* and *wheelchair* with very high accuracy of about 99%. Class *rollator* is also predicted with high accuracy above 94%, all ResNet models as well as DenseNet201 achieve very high accuracies between 97% to 98%. Unsurprisingly, the most challenging task for all models is to distinguish between classes *crutch* and *cane*, demonstrating the challenge of correct classification depending on thin and fine-graned visual cues.

An interesting observation is that the most complex networks do not necessarily perform the best on our dataset, allowing for selection of more lightweight and faster networks. Please refer to the supplemental material for a further analysis of the classification performance and confusion matrices.

	pedestrian GT 3058		wheelchair GT 1660		rollator GT 787		crutch GT 979		cane GT 540		all GT 7024	
	TP↑	Acc↑	TP↑	Acc↑	TP↑	Acc↑	TP↑	Acc↑	TP↑	Acc↑	TP↑	Acc↑
MobileNetV3 L	3041	0.9944	1642	0.9892	751	0.9543	732	0.7477	424	0.7852	6590	0.9382
MobileNetV3 L (V2)	3044	0.9954	1642	0.9892	744	0.9454	745	0.7610	421	0.7796	6596	0.9391
ResNet18	3053	0.9984	1650	0.9940	764	0.9708	753	0.7992	***449***	***0.8315***	6669	0.9495
ResNet34	***3056***	***0.9993***	1645	0.9910	**775**	**0.9848**	726	0.7416	440	0.8148	6642	0.9456
ResNet50	**3057**	**0.9997**	***1651***	***0.9946***	763	0.9695	776	0.7926	430	0.7963	6677	0.9506
ResNet50 (V2)	***3056***	***0.9993***	1645	0.9910	***771***	***0.9797***	764	0.7804	436	0.8074	6672	0.9499
DenseNet201	3053	0.9984	***1651***	***0.9946***	764	0.9708	***789***	***0.8059***	**459**	**0.8500**	**6716**	**0.9562**
ResNet152	3054	0.9987	**1654**	**0.9964**	764	0.9708	**792**	**0.8090**	445	0.8241	***6709***	***0.9552***
ResNet152 (V2)	3053	0.9984	1642	0.9892	***771***	***0.9797***	788	0.8049	438	0.8111	6692	0.9527
ViT-B/16	3054	0.9987	1648	0.9928	760	0.9657	711	0.7263	397	0.7352	6570	0.9354
VGG16	3054	0.9987	1648	0.9928	758	0.9632	771	0.7875	431	0.7981	6662	0.9485

Table 2: Classification performance on the test split of our dataset. Scores in bold represent the **best** results, scores in bold and italics indicate the ***second-best*** results, and underlined scores notate the third-best results. GT denotes the number of ground truth samples.

5 Detection

For use in applications, not only the classification of persons according to their use of mobility aids is essential, but processing of whole scenes with localization and simultaneous classification is necessary. To this end, we provide results for the state-of-the-art one-shot YOLOv5 [15] detector. An important aspect, especially in tasks geared towards providing increased safety for vulnerable road users, is to assure detection as reliably as possible. When treating all mobility classes as equivalent but independent, detections may be lost if, due to uncertainty, multiple classes are predicted with all having a low score below the detection threshold. To remedy this, we propose to model classification into mobility aids used in a hierarchical fashion inspired by YOLO9000 [27] as explained in Sec. 5.1. The used evaluation metrics, as well as the detection and classification results are presented in Sec. 5.2 and 5.3, respectively.

5.1 Hierarchical Class Prediction

The key idea behind hierarchical class training is to structure class labels in a semantic hierarchy with a tree-like structure [27], where all labels in lower branches also share common parent labels up to the respective semantic root. The classification head is then trained to not only predict the fine-grained class, but to also predict all classes further up in the hierarchy necessary to reach that specific label. When evaluating class predictions, we start at the root labels, at every point taking the branch with the highest score, up to the point where we either reach the end of the semantic tree or the score drops below a predefined threshold. In practice, we realize this by giving each of the 5 classes {*pedestrian*, *wheelchair*, *rollator*, *crutch*, *cane*} in our mobility dataset the common root class of *person* and training the detection model to also predict this class, jointly with the specific mobility class. The idea being, that through this formulation even in cases where all fine-grained labels score low due to uncertainty in class prediction, at least the common superclass of *person* will remain with a high enough score to yield a proper detection. For a further explanation we refer to the original publication of YOLO9000 [27]. The code used for training and evaluation of the detection models is made available at https://github.com/mordecaimalignatius/yolo-9k.

5.2 Metrics

We provide detection and classification results in terms of the well established *average precision* metric (AP), defined as the area under the precision-recall curve for each class separately, as well as the average over all classes. In addition to the common AP@50 measure introduced by Pascal VOC [6], *i.e.* average precision at 50% Intersection-over-Union (IoU), we also provide results averaged over 10 IoU thresholds from 50% to 95% as mAP@50-95, as proposed by COCO [20]. Furthermore, we compare detector performance in terms of recall (accuracy ACC), relative amount of misclassified (MCL) and missed detections (MIS) for each class, as well as overall number of false detections (FP) and missing detections (FN).

5.3 Results

All models have been fine-tuned until convergence for a maximum of 200 epochs, with early stopping if the performance on the validation set has not increased in the last 20 epochs, starting from the weights pre-trained on COCO [20]. We used AdamW as optimizer. For testing the IoU threshold for non-maximum supression was chosen as 0.45, the confidence threshold as 0.4.

Table 3 lists the performance of the trained models for each class separately, as well as the average performance over all classes. As expected, performance increases with model complexity, yet within a comparatively small margin relative to model complexity. Performance is best for easily distinguishable classes *wheelchair* and *rollator*, which have large footprints in the images. Correct classification of thin mobility aids like crutches and walking canes reduced, implying again a necessity of further research to reliably detect such instances. Notably, the hierarchically trained models suffer a performance penalty in this measure of around 2-3 percentage points. However, we see from Table 4 that particularly for mobility aids training with the hierarchical class structure reduces the number of false positives as well as false negatives by up to a factor of 2, independent of model complexity. This

emphasizes the benefit of a hierarchical approach for increased safety, providing the network with an explicit fallback in case of uncertainty. For an in-depth comparison we refer to the confusion matrices available in the supplemental material, together with a further evaluation of detection on a smaller image size of 640×480 pixels.

	pedestrian		wheelchair		rollator		crutch		cane		all	
YOLOv5	AP@50	mAP@50-95	AP@50	mAP@50-95	AP@50	mAP@50-95	AP@50	mAP@50-95	AP@50	mAP@50-95	mAP@50	mAP@50-95
small	0.881	0.650	0.983	0.721	0.952	0.693	0.821	0.695	*0.674*	0.569	0.862	0.666
medium	0.886	0.669	**0.986**	0.719	*0.960*	0.706	*0.843*	0.719	0.639	0.545	0.863	0.672
large	**0.891**	*0.671*	0.982	0.727	**0.972**	0.712	0.840	0.720	0.669	0.568	*0.871*	0.679
xlarge	*0.887*	**0.682**	0.982	0.735	0.954	0.708	**0.858**	**0.732**	**0.716**	**0.612**	**0.880**	**0.694**
h small	0.860	0.637	0.981	0.722	0.941	0.699	0.760	0.666	0.583	0.508	0.825	0.647
h medium	0.853	0.642	*0.985*	0.734	0.949	*0.724*	0.796	0.699	0.660	0.582	0.849	0.676
h large	0.859	0.652	0.981	*0.736*	0.953	**0.726**	0.788	0.689	0.632	0.548	0.843	0.670
h xlarge	0.852	0.647	0.976	**0.747**	0.942	0.717	0.820	*0.723*	0.670	*0.597*	0.852	*0.686*

Table 3: Detector performance on the full resolution of the test split for different model sizes of YOLOv5. Models in the top have been trained with independent classes, models prepended with 'h' in the bottom part have been trained with the hierarchical class structure. The **best** model has been marked in **bold face**, the *second best* in *italic*.

	pedestrian			wheelchair			rollator			crutch			cane			all			
YOLOv5	ACC$^{\uparrow}$	MCL$^{\downarrow}$	MIS$^{\downarrow}$	ACC$^{\uparrow}$	MCL$^{\downarrow}$	MIS$^{\downarrow}$	ACC$^{\uparrow}$	MCL$^{\downarrow}$	MIS$^{\downarrow}$	ACC$^{\uparrow}$	MCL$^{\downarrow}$	MIS$^{\downarrow}$	ACC$^{\uparrow}$	MCL$^{\downarrow}$	MIS$^{\downarrow}$	ACC$^{\uparrow}$	MCL$^{\downarrow}$	FP$^{\downarrow}$	FN$^{\downarrow}$
small	0.874	0.070	0.055	*0.967*	**0.006**	0.027	*0.940*	**0.034**	0.025	0.653	0.336	0.011	**0.693**	**0.289**	*0.019*	0.859	0.105	430	255
medium	*0.880*	**0.049**	0.071	0.965	*0.008*	0.027	0.924	0.053	0.023	0.664	*0.322*	0.014	0.598	0.369	0.033	*0.863*	*0.103*	440	312
large	*0.880*	**0.049**	0.071	0.965	*0.008*	0.027	0.924	0.053	0.023	0.664	*0.322*	0.014	0.598	0.369	0.033	0.853	*0.103*	440	312
xlarge	**0.897**	*0.054*	0.048	**0.968**	0.009	0.024	0.921	0.051	0.028	**0.701**	**0.290**	0.009	0.643	0.341	*0.017*	**0.870**	**0.098**	609	*227*
h small	0.854	0.076	0.070	0.964	0.010	0.027	0.917	0.064	0.019	0.590	0.395	0.014	0.598	0.389	**0.013**	0.831	0.127	**146**	295
h medium	0.846	0.105	**0.049**	0.966	0.015	**0.019**	0.939	0.047	**0.014**	0.662	0.333	**0.005**	*0.657*	*0.330*	**0.013**	0.844	0.126	278	**206**
h large	0.854	0.094	*0.052*	0.963	0.016	*0.021*	**0.943**	0.041	*0.017*	0.637	0.349	0.013	0.641	0.337	0.022	0.843	0.124	274	232
h xlarge	0.818	0.123	0.059	0.939	0.033	0.029	0.937	*0.039*	0.024	*0.672*	*0.322*	*0.006*	0.587	0.394	0.019	0.822	0.141	*226*	264

Table 4: Detector performance in terms of accuracy (ACC), misclassification (MCL) and missed (MIS) detections as well as overall false negatives (FN) and false positive (FP) detections.

6 Conclusion

We presented a new dataset (available online at [24]) focused on the detection and classification of the most vulnerable road users, i.e. pedestrians using mobility aids. Our carefully designed data collection ensures that this dataset can be used without restrictions for future research. Improved detection and classification models for mobility aid users can improve traffic safety, as well as vision-based applications in convalescent homes. Since at least in Europe and the United States monitoring of public spaces is a highly sensitive task demanding processing of the data directly at the monitoring site, this requires lightweight detection and classification architectures which can be deployed on embedded devices efficiently. These architectures will need to be able to reliably classify mobility aid users based on thin structures, a task at which even current, large and powerful architectures struggle. Our detailed evaluations with commonly used classification and detection approaches provide strong baselines for future comparisons.

References

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuScenes: A multimodal dataset for autonomous driving.

In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[2] W. Chen, X. Xu, J. Jia, H. Luo, Y. Wang, F. Wang, R. Jin, and X. Sun. Beyond Appearance: a Semantic Controllable Self-Supervised Learning Framework for Human-Centric Visual Tasks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[3] J. Deng, W. Dong, R. Socher, L. J. Li, L. Kai, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2021.

[5] European Parliament and Council of European Union. Regulation (EU) no 2018/858. Official Journal of the European Union, 2018.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2010.

[7] R. Girshick. Fast R-CNN. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2015.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[9] D. Grindle, A. Balubaid, and C. Untaroiu. Investigation of traffic accidents involving seated pedestrians using a finite element simulation-based approach. *Computer Methods in Biomechanics and Biomedical Engineering*, 26(4):484–497, 2023.

[10] Y. Guo and N.-M. Cheung. Efficient and deep person re-identification using multi-level similarity. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] A. Howard, M. Sandler, B. Chen, W. Wang, L. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le. Searching for MobileNetV3. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2019.

[13] G. Huang, Z. Liu, L. van Der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] Z. Ji, E. He, H. Wang, and A. Yang. Image-attribute Reciprocally Guided Attention Network for Pedestrian Attribute Recognition. *Pattern Recognition Letters*, 120:89–95, 2019.

[15] G. Jocher, A. Chaurasia, A. Stoken, et al. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, 2022.

[16] J. Kraemer and C. Benton. Disparities in road crash mortality among pedestrians using wheelchairs in the USA: results of a capture-recapture analysis. *BMJ Open*, 5 (e008396), 2015.

[17] D. Li, X. Chen, and K. Huang. Multi-attribute Learning for Pedestrian Attribute Recognition in Surveillance Scenarios. In *Proc. of the IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015.

[18] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning Deep Context-Aware Features over Body and Latent Parts for Person Re-identification. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] Y. Li, C. Huang, C. C. Loy, and X. Tang. Human Attribute Recognition by Deep Hierarchical Contexts. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2016.

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2014.

[21] P. Liu, X. Liu, J. Yan, and J. Shao. Localization Guided Learning for Pedestrian Attribute Recognition. *Proc. of the British Machine Vision Conference (BMVC)*, 2018.

[22] S. Liu and W. Deng. Very deep convolutional neural network based image classification using small training sample size. In *Proc. of the IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single Shot MultiBox Detector. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2016.

[24] L. Mohr, N. Kirillova, H. Possegger, and H. Bischof. Crossroad Camera Dataset - Mobility Aid Users. 09 2023. doi: 0.3217/2gat1-pev27. URL https://repository.tugraz.at/records/2gat1-pev27.

[25] United Nations Department of Safety and Security. Road Safety Strategy, 2019.

[26] A. Paszke, S. Gross, F. Massa, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[27] J. Redmon and A. Farhadi. YOLO9000: Better, Faster, Stronger. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[29] N. Sarafianos, X. Xu, and I.A. Kakadiaris. Deep Imbalanced Attribute Classification using Visual Attention Aggregation. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2018.

[30] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang. End-to-end deep kronecker-product matching for person re-identification. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[31] P. Sudowe, H. Spitzer, and B. Leibe. Person Attribute Recognition with a Jointly-trained Holistic CNN Model. In *Proc. ICCV Workshops*, 2015.

[32] C. Szegedy, S. E. Reed, D. Erhan, and D. Anguelov. Scalable, High-Quality Object Detection. *CoRR*, abs/1412.1441, 2014. URL http://arxiv.org/abs/1412.1441.

[33] R. Szeliski. *Computer Vision. Algorithms and Applications*, pages 278–295. Springer International Publishing, second edition, 2022.

[34] C. Tang, L. Sheng, Z.-X. Zhang, and X. Hu. Improving Pedestrian Attribute Recognition With Weakly-Supervised Multi-Scale Attribute-Specific Localization. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2019.

[35] J. Terven and D. Cordova-Esparza. A Comprehensive Review of YOLO: From YOLOv1 to YOLOv8 and Beyond, 2023.

[36] United Nations Economic Commision for Europe (UN-ECE). A foundational safety system concept to make roads safer in the decade 2021-2030, 2020.

[37] A. Vasquez, M. Kollmitz, A. Eitel, and W. Burgard. Deep Detection of People and their Mobility Aids for a Hospital Robot. In *Proc. of the IEEE European Conference on Mobile Robotics (ECMR)*, 2017.

[38] Y. Wang, Z. Chen, F. Wu, and G. Wang. Person re-identification with cascaded pairwise convolutions. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] World Health Organization. Global Status Report on Road Safety, 2018.

[40] M. Wu, D. Huang, Y. Guo, and Y. Wang. Distraction-aware Feature Learning for Human Attribute Recognition via Coarse-to-fine Attention Mechanism. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[41] H. Zeng, H. Ai, Z. Zhuang, and L. Chen. Multi-task Learning via Co-attentive Sharing for Pedestrian Attribute Recognition. In *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, 2020.

[42] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-Scale Feature Learning for Person Re-Identification. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2019.