Supplemental Material

In this supplemental material, we provide additional analyses for the classification experiments outlined in the main paper. Furthermore, we provide confusion matrices for the detection experiment on the full image resolution presented in the main paper, as well as a further detection experiment done on a reduced image resolution of 640×480 pixels.

Classification

Figure 6 depicts sample crops used in classification, showing the padding of 20 pixels added to capture relevant context for classification.

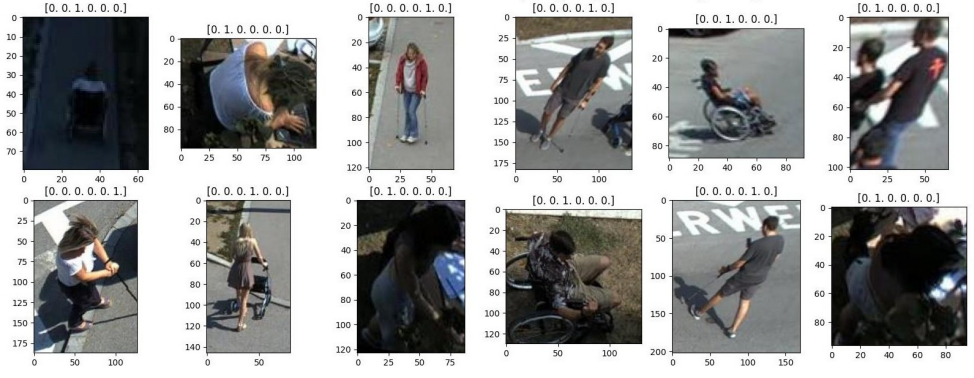


Figure 6: Example crops used for classification, showing the added padding to capture relevant context. The respective title shows the one-hot encoding of the class label (*[pedestrian, wheelchair, rollator, crutch, cane]*).

We provide a comparison of the feature backbones used in the classification experiment in Table 5, providing information on their performance on ImageNet, their complexity as well as final validation accuracy reached in fine-tuning on our dataset. Although the difference in model complexity spans over a decade in terms of number of parameters and almost two decades in floating point operations per second, their final difference in validation accuracy is less than 2 percentage points, allowing to chose smaller, faster implementations.

Figure 7 shows the confusion matrices of the test split for all classifiers. Confusion wise, all classifiers perform the same, with the biggest mix-up existing between classes *crutch* and *cane*. This is not surprising, considering that the visual structures for walking canes and crutches are both very thin and highly similar, differing mostly in the handle.

Detection

We present the confusion matrices for the detection experiments on image size 1280×960 in Figure 8. In each cell, the first line shows the percentage of ground truth labels classified into this cell, the second line denotes the respective number of instances. In general, the biggest confusion exists between the classes *crutch* and *cane*, which both represent small structures mostly distinguished by the grip. A further source for misclassification is these two classes being classified as *pedestrian*, *i.e.* not using any mobility aid. This is not a surprising result, since both classes represent thin structures, challenging to be captured by a

	Backbone				Fine-tuned
	Acc@1↑	Acc@5↑	Params↓	GFLOPs↓	ValAcc↑
MobileNetV3 L	0.740	0.913	**5.5M**	**0.22**	0.954
MobileNetV3 L (V2)	0.740	0.913	**5.5M**	**0.22**	0.958
ResNet18	0.698	0.891	*11.7M*	*1.81*	0.951
ResNet34	0.733	0.914	21.8M	3.66	0.953
ResNet50	0.761	0.929	25.6M	4.09	0.963
ResNet50 (V2)	*0.809*	*0.954*	25.6M	4.09	0.962
DenseNet201	0.770	0.934	20.0M	4.29	*0.967*
ResNet152	0.783	0.940	60.2M	11.51	0.965
ResNet152 (V2)	**0.823**	**0.960**	60.2M	11.51	**0.968**
ViT-B/16	0.811	0.953	86.6M	17.56	0.955
VGG16	0.716	0.904	138.4M	15.47	0.953

Table 5: Comparison of the classification backbones used with their original performance on ImageNet in terms of top-1 accuracy (Acc@1) and top-5 accuracy (Acc@5), model complexity in terms of millions of model parameters (Params) and billions of floating point operations per second (GFLOPs) as well as final accuracy on the validation split of our dataset (ValAcc). Best values are marked **bold**, second best *italic*.

deep neural network. Furthermore, we can see that the hierarchical training approach vastly reduces the number of false positive detections, especially so for mobility aid use and the thin mobility aids *crutch* and *cane*, irrespective of model size.

The results for detection on smaller images of size 640×480 are listed in Table 6 in terms of average precision (AP), and in Table 7 in terms of accuracy (ACC), misclassification rate (MIS) and missed detections (MIS). We observe a general performance drop of about $4-5$ percentage points over all models. Classes *crutch* and *cane* are affected more severe, with *cane* faring even worse than *crutch*. This observation can be explained largely due to the respective mobility aids having only a long and thin footprint within the images, thereby making detection and classification even more challenging with a smaller image resolution. This interpretation is consistent with the fact that classification of class *wheelchair*, which has a comparatively large footprint, does not show any significant performance degradation. Model performance in terms of missed detections (false negatives, FN) and wrong detections (false positives FP in Table 7) stays more or less the same for training with a class hierarchy, models trained with independent classes fare worse for wrong detections, yet a bit better for missed detections. The small model trained with class hierarchy seems to be an outlier in this regard. The confusion matrices presented in Figure 9 show the same qualitative results observed for detectors trained on the large image size, again underlining the benefit of providing the detectors with a safe fall-back by using class hierarchies.

	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	0.811	0.565	0.979	0.661	0.912	0.639	0.651	0.525	0.383	0.311	0.747	0.540
medium	0.842	0.612	0.980	0.693	0.912	0.648	*0.725*	*0.604*	0.471	0.396	0.786	0.591
large	**0.857**	**0.631**	0.977	0.693	*0.931*	0.672	**0.742**	**0.617**	**0.499**	0.413	**0.801**	*0.605*
xlarge	*0.849*	*0.626*	**0.982**	0.697	**0.938**	0.679	0.716	0.599	*0.490*	0.413	*0.795*	0.603
h small	0.758	0.508	0.965	0.667	0.868	0.612	0.629	0.529	0.364	0.311	0.717	0.525
h medium	0.812	0.584	0.979	*0.716*	0.883	0.640	0.661	0.569	0.486	*0.420*	0.764	0.586
h large	0.812	0.600	*0.981*	**0.734**	0.923	**0.695**	0.671	0.583	0.489	**0.428**	0.775	**0.608**
h xlarge	0.799	0.597	0.973	0.689	0.920	*0.690*	0.632	0.547	0.462	0.401	0.757	0.585

Table 6: Detector performance for models trained on images resolution 640×480 pixels. 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	**0.858**	*0.089*	*0.054*	0.958	0.016	0.026	0.912	0.075	**0.013**	0.491	0.500	*0.009*	0.332	0.652	0.017	0.796	0.171	629	235
medium	0.842	0.097	0.061	**0.969**	**0.008**	0.023	0.877	0.103	0.020	*0.587*	*0.400*	0.012	0.461	0.522	0.017	0.811	0.152	561	262
large	*0.849*	0.094	0.057	0.962	*0.010*	0.028	0.919	0.064	0.018	0.577	0.409	0.014	**0.519**	**0.463**	0.019	**0.820**	*0.143*	511	259
xlarge	0.842	**0.087**	0.071	0.964	0.015	*0.022*	**0.935**	**0.047**	0.018	**0.590**	**0.397**	0.012	0.454	0.520	0.026	*0.816*	**0.142**	521	292
h small	0.802	0.134	0.064	0.947	0.024	0.030	0.856	0.127	0.017	0.486	0.504	0.010	0.332	0.656	0.013	0.762	0.199	353	274
h medium	0.829	0.118	**0.053**	0.962	0.022	**0.016**	0.869	0.118	**0.013**	0.538	0.455	**0.007**	0.448	0.541	*0.011*	0.795	0.175	*256*	**212**
h large	0.837	0.107	0.056	*0.965*	0.019	**0.016**	*0.920*	0.065	*0.015*	0.579	0.407	0.014	*0.487*	*0.504*	**0.009**	0.814	0.154	266	*229*
h xlarge	0.804	0.115	0.080	0.961	0.015	0.024	0.917	*0.057*	0.025	0.515	0.469	0.016	0.457	0.528	0.015	0.787	0.166	**227**	330

Table 7: 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 for detectors trained on image size 640×480.

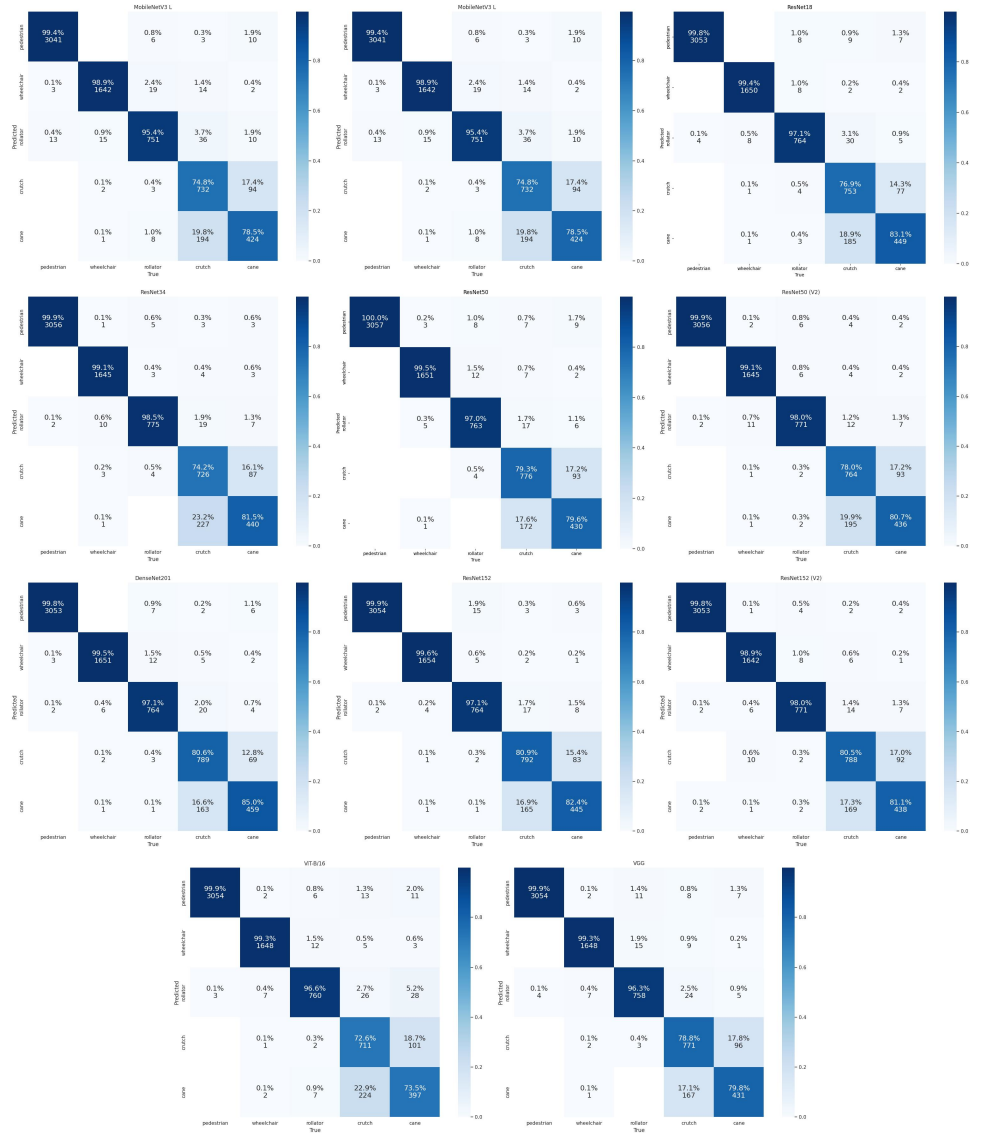
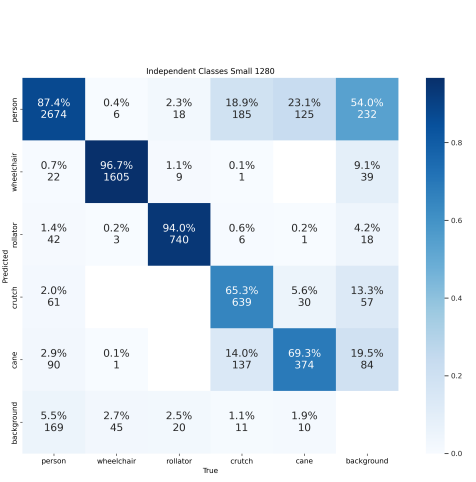
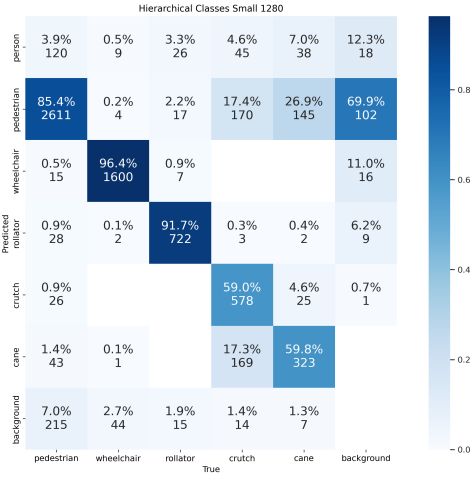
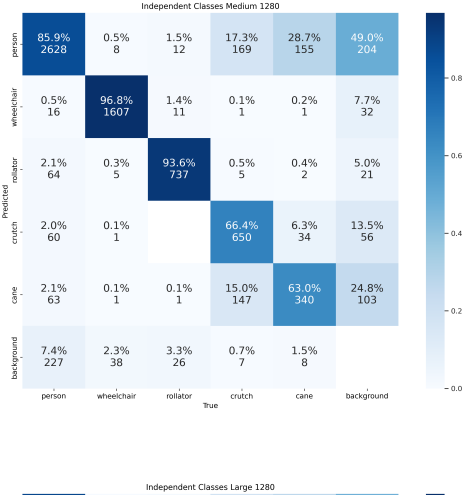
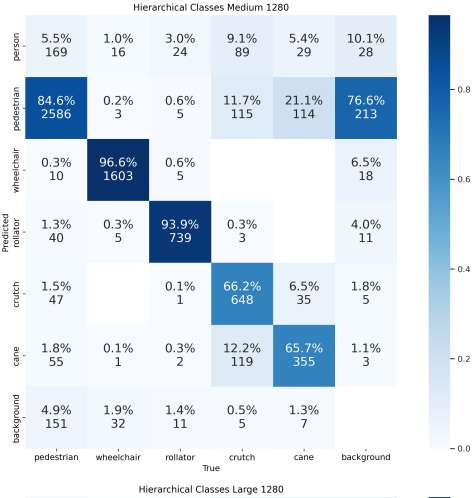


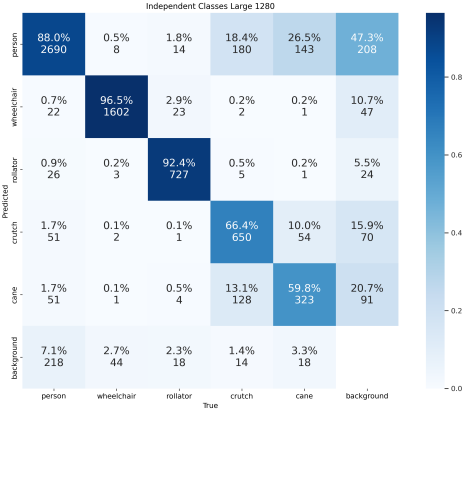
Figure 7: Confusion matrices of the classifiers compared. The biggest mix-up exists for the classes *Crutch* and *Cane* across all classifiers.

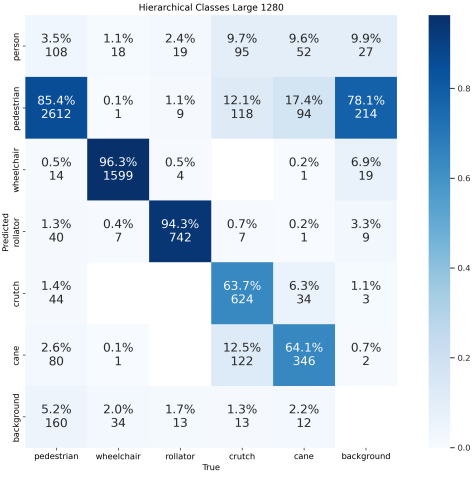

Independent Classes Small 1280
person 87.4% 2674 | 0.4% 6 | 2.3% 18 | 18.9% 185 | 23.1% 125 | 54.0% 232
wheelchair 0.7% 22 | 96.7% 1605 | 1.1% 9 | 0.1% 1 | | 9.1% 39
rollator 1.4% 42 | 0.2% 3 | 94.0% 740 | 0.6% 6 | 0.2% 1 | 4.2% 18
crutch 2.0% 61 | | | 65.3% 639 | 5.6% 30 | 13.3% 57
cane 2.9% 90 | 0.1% 1 | | 14.0% 137 | 69.3% 374 | 19.5% 84
background 5.5% 169 | 2.7% 45 | 2.5% 20 | 1.1% 11 | 1.9% 10 |
person wheelchair rollator crutch cane background
Predicted
True
0.0 0.2 0.4 0.6 0.8


Hierarchical Classes Small 1280
person 3.9% 120 | 0.5% 9 | 3.3% 26 | 4.6% 45 | 7.0% 38 | 12.3% 18
pedestrian 85.4% 2611 | 0.2% 4 | 2.2% 17 | 17.4% 170 | 26.9% 145 | 69.9% 102
wheelchair 0.5% 15 | 96.4% 1600 | 0.9% 7 | | | 11.0% 16
rollator 0.9% 28 | 0.1% 2 | 91.7% 722 | 0.3% 3 | 0.4% 2 | 6.2% 9
crutch 0.9% 26 | | | 59.0% 578 | 4.6% 25 | 0.7% 1
cane 1.4% 43 | 0.1% 1 | | 17.3% 169 | 59.8% 323 |
background 7.0% 215 | 2.7% 44 | 1.9% 15 | 1.4% 14 | 1.3% 7 |
pedestrian wheelchair rollator crutch cane background
Predicted
True
0.0 0.2 0.4 0.6 0.8


Independent Classes Medium 1280
person 85.9% 2628 | 0.5% 8 | 1.5% 12 | 17.3% 169 | 28.7% 155 | 49.0% 204
wheelchair 0.5% 16 | 96.8% 1607 | 1.4% 11 | 0.1% 1 | 0.2% 1 | 7.7% 32
rollator 2.1% 64 | 0.3% 5 | 93.6% 737 | 0.5% 5 | 0.4% 2 | 5.0% 21
crutch 2.0% 60 | 0.1% 1 | | 66.4% 650 | 6.3% 34 | 13.5% 56
cane 2.1% 63 | 0.1% 1 | 0.1% 1 | 15.0% 147 | 63.0% 340 | 24.8% 103
background 7.4% 227 | 2.3% 38 | 3.3% 26 | 0.7% 7 | 1.5% 8 |
person wheelchair rollator crutch cane background
Predicted
True
0.0 0.2 0.4 0.6 0.8


Hierarchical Classes Medium 1280
person 5.5% 169 | 1.0% 16 | 3.0% 24 | 9.1% 89 | 5.4% 29 | 10.1% 28
pedestrian 84.6% 2586 | 0.2% 3 | 0.6% 5 | 11.7% 115 | 21.1% 114 | 76.6% 213
wheelchair 0.3% 10 | 96.6% 1603 | 0.6% 5 | | | 6.5% 18
rollator 1.3% 40 | 0.3% 5 | 93.9% 739 | 0.3% 3 | | 4.0% 11
crutch 1.5% 47 | | 0.1% 1 | 66.2% 648 | 6.5% 35 | 1.8% 5
cane 1.8% 55 | 0.1% 1 | 0.3% 2 | 12.2% 119 | 65.7% 355 | 1.1% 3
background 4.9% 151 | 1.9% 32 | 1.4% 11 | 0.5% 5 | 1.3% 7 |
pedestrian wheelchair rollator crutch cane background
Predicted
True
0.0 0.2 0.4 0.6 0.8


Independent Classes Large 1280
person 88.0% 2690 | 0.5% 8 | 1.8% 14 | 18.4% 180 | 26.5% 143 | 47.3% 208
wheelchair 0.7% 22 | 96.5% 1602 | 2.9% 23 | 0.2% 2 | 0.2% 1 | 10.7% 47
rollator 0.9% 26 | 0.2% 3 | 92.4% 727 | 0.5% 5 | 0.2% 1 | 5.5% 24
crutch 1.7% 51 | 0.1% 2 | 0.1% 1 | 66.4% 650 | 10.0% 54 | 15.9% 70
cane 1.7% 51 | 0.1% 1 | 0.5% 4 | 13.1% 128 | 59.8% 323 | 20.7% 91
background 7.1% 218 | 2.7% 44 | 2.3% 18 | 1.4% 14 | 3.3% 18 |
person wheelchair rollator crutch cane background
Predicted
True
0.0 0.2 0.4 0.6 0.8


Hierarchical Classes Large 1280
person 3.5% 108 | 1.1% 18 | 2.4% 19 | 9.7% 95 | 9.6% 52 | 9.9% 27
pedestrian 85.4% 2612 | 0.1% 1 | 1.1% 9 | 12.1% 118 | 17.4% 94 | 78.1% 214
wheelchair 0.5% 14 | 96.3% 1599 | 0.5% 4 | | 0.2% 1 | 6.9% 19
rollator 1.3% 40 | 0.4% 7 | 94.3% 742 | 0.7% 7 | 0.2% 1 | 3.3% 9
crutch 1.4% 44 | | | 63.7% 624 | 6.3% 34 | 1.1% 3
cane 2.6% 80 | 0.1% 1 | | 12.5% 122 | 64.1% 346 | 0.7% 2
background 5.2% 160 | 2.0% 34 | 1.7% 13 | 1.3% 13 | 2.2% 12 |
pedestrian wheelchair rollator crutch cane background
Predicted
True
0.0 0.2 0.4 0.6 0.8

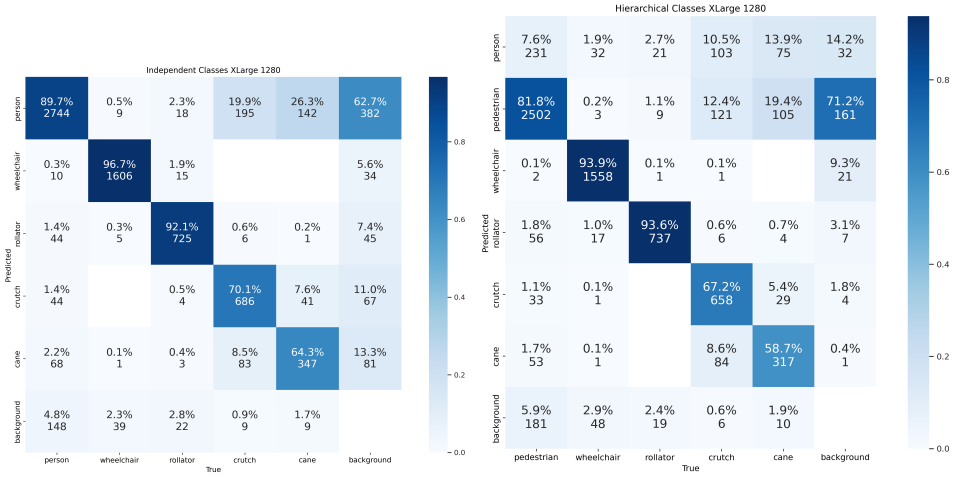


Figure 8: Confusion matrices for YOLOv5 detectors trained on image size 1280×960. The left column shows the results for models trained with independent classes, the right column shows the results for those trained with hierarchical classes.

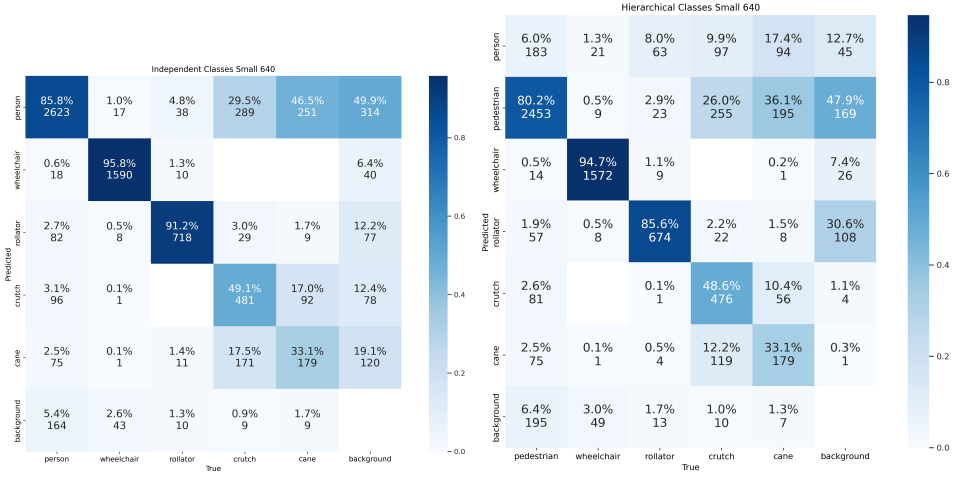


Figure 9: Confusion matrices for YOLOv5 detectors trained on image size 640×480. The left column shows the results for models trained with independent classes, the right column shows the results for those trained with hierarchical classes.