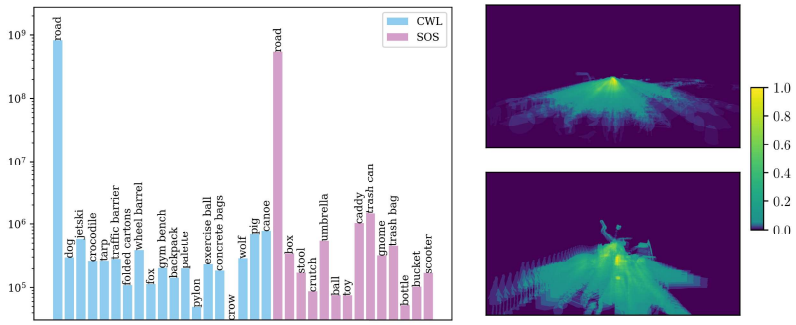
Appendix

A Details on the Data Sets



number of annotated pixels per class OOD pixel heatmaps

Fig. 5: Visualization of the pixel distributions of SOS and CWL on a class-level (left) and as a heatmap of OOD pixels (right) for CWL (top) and SOS (bottom).

The real-world images in SOS were labeled using the LabelMe tool[5]. For the synthetic CWL data set the labels are provided automatically by the CARLA software. CWL was generated with the driving simulator CARLA [23] 0.9.13. The OOD objects used are not part of the original CARLA repository and were hand placed by the Unreal Editor using freely available assets from the Unreal Engine webpage. The ego-vehicle (Audi TT) to which the sensors are attached was spawned into 8 different maps. It is spawned near OOD objects and drives towards them at a maximum speed of 50 km/h, recorded with 10 fps. Each vehicle can be placed on predefined road points and move in the global coordinate system of the selected map, possessing its own vehicle coordinate system with the zero point at its center. In addition to the spatial coordinates $(y, z, x) = (0, 1.7, 1.6)$, the rotation angles (pitch, yaw, roll)$= (0, 0, 0)$ of an object/sensor can be specified. During each simulation step, the program waits until the scene has been completely rendered and then records each sensor in a queued manner before proceeding to the next simulation step. Except for the *motion blur intensity*$= 0$, the default value was selected for all other intrinsic camera parameters which are listed on the CARLA documentation webpage[6].

Our data sets are **not** intended to be used as training data in order to develop new deep learning methods. Methods could overfit the data, which is undesirable in the field of OOD detection. The purpose of our proposed datasets is rather to

[5] https://github.com/wkentaro/labelme
[6] https://carla.readthedocs.io/en/latest/ref_sensors/#rgb-camera

Fig. 6: Some examples images of the SOS (top), CWL (middle) and WOS (bottom) data sets.

validate generalization capabilities of new approaches for the new task of OOD tracking.

For a better understanding of the SOS and CWL data sets, we provide some statistics in fig. 5 and more example images in fig. 6. SOS contains 0.21% OOD, 23.29% road and 76.50% void pixels, where the top five OOD classes, i.e., the classes that constitute the most pixels, are 1) *trash can*, 2) *caddy*, 3) *umbrella*, 4) *trash bag* and 5) *box*. CWL contains 0.20% OOD, 32.82% road and 66.98% void pixels with top five OOD classes 1) *canoe*, 2) *pig*, 3) *jetski*, 4) *wheel barrel* and 5) *dog*.

B Training of the Meta Classifier

In addition to the experiments presented in the main paper, we train the meta classifier per (sequence-wise) leave-one-out cross-validation on the respective dataset, i.e., one image sequence is used for testing and the remaining ones for training, denoted by M_1. Note that this procedure however requires in domain OOD ground truth data.

Note that despite single instances of OOD objects occur in more than one video sequence in both data sets, their uncertainty features used for meta classification are distinct. In this sense, a proper split between the training and test data set is maintained during leave-one-out cross validation.

In the main article, the meta classifier was trained on one dataset and evaluated on the other one, e.g. for experiments on SOS the meta classification model is trained on CWL, denoted by M_2. This procedure did not require any in domain OOD ground truth data and, e.g., real world OOD meta classification can be be trained on synthetic OOD ground truth, which is easily obtained.

In the following, we benchmark both approaches.

Table 2: OOD object segmentation results on segment-level for the SOS and the CWL dataset obtained by two differently trained meta classifiers.

dataset	$\bar{F}_1(M_1) \uparrow$	$\bar{F}_1(M_2) \uparrow$
SOS	50.27	35.84
CWL	47.60	45.46

Table 3: Object tracking results for the SOS and the CWL dataset obtained by two differently trained meta classifiers.

dataset	model	$MOTA \uparrow$	$\overline{mme} \downarrow$	$MOTP \downarrow$	GT	MT	PT	ML	$l_t \uparrow$
SOS	M_1	0.3116	0.0639	12.5042	26	10	13	3	0.5635
	M_2	−0.0826	0.0632	12.3041	26	9	14	3	0.5510
CWL	M_1	0.4869	0.0266	16.2387	62	34	22	6	0.6689
	M_2	0.4043	0.0282	16.4965	62	24	30	8	0.5389

Table 4: Object clustering results for the SOS and the CWL dataset with two differently trained meta classifiers. We report results for clustering with and without incorporating the object tracking information.

		without tracking ($\ell = 0$)			with tracking ($\ell = 10$)		
dataset	model	$CS_{\text{inst}} \uparrow$	$CS_{\text{imp}} \downarrow$	$CS_{\text{frag}} \downarrow$	$CS_{\text{inst}} \uparrow$	$CS_{\text{imp}} \downarrow$	$CS_{\text{frag}} \downarrow$
SOS	M_1	0.8779	2.1818	2.7273	0.8992	1.7143	2.0909
	M_2	0.8652	2.5217	2.8182	0.8955	1.7917	1.9091
CWL	M_1	0.8426	2.5455	2.9500	0.8627	2.5161	2.6500
	M_2	0.8637	2.8181	2.2500	0.8977	2.1739	1.8000

The OOD segmentation results are given in table 2. We observe that the M_1 model achieves higher values as the meta classifier performs better trained on the respective dataset via leave-one-out cross-validation than under domain shift using the other dataset. There is only a small gap between the $\bar{F}_1$ scores for the CWL dataset while this gap is comparatively large for SOS. It follows that training the meta model on SOS and testing it on CWL is more effectively than vice versa.

The object tracking results are shown in table 3 for both dataset and the two meta classifiers. We observe similar results for each dataset for the two different meta classifiers. The only exception is the $MOTA$ metric for the SOS dataset, with a comparatively poor performance for model M_2.

The results for object clustering are provided in table 4, also for both dataset and meta classifiers. Additionally, we analyze the impact of OOD tracking on the clustering results. We observe, that incorporating the tracking information has a positive effect on all clustering metrics. In general, both models M_1 and M_2 produce similar results, however, for CWL with $\ell = 10$, model M_2 performs significantly better. A visual comparison of these results is provided in fig. 7 for SOS and in fig. 8 for CWL.

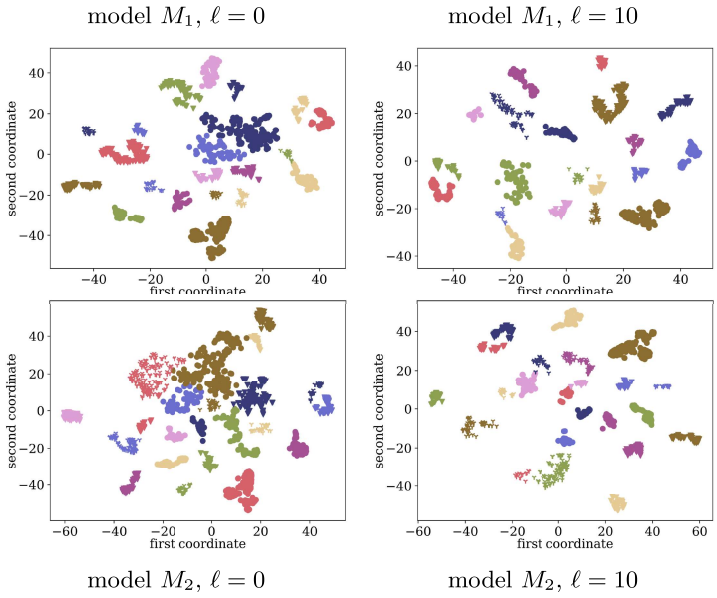


Fig. 7: Clustering of SOS OOD segments via DBSCAN in the embedding space for different experimental setups. Note that tSNE produces non-deterministic, hence different embeddings for each setup.

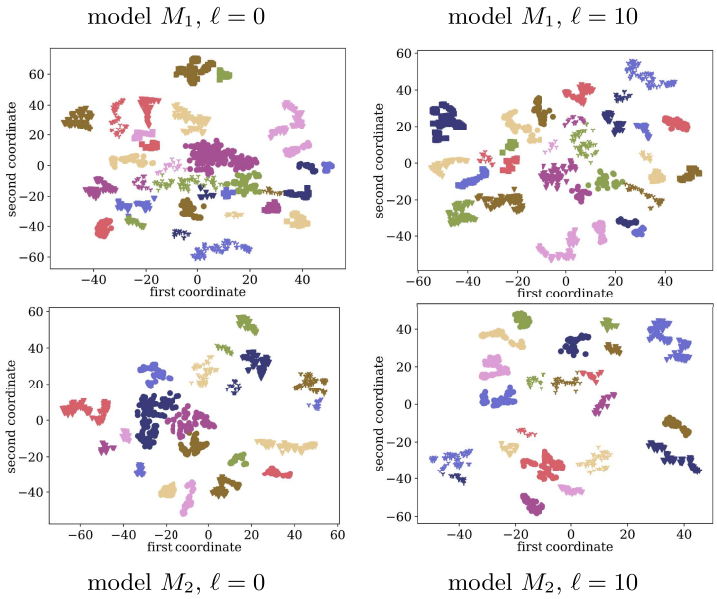


Fig. 8: Clustering of CWL OOD segments via DBSCAN in the embedding space for different experimental setups. Note that tSNE produces non-deterministic, hence different embeddings for each setup.

C Numerical Results for Depth Binnings

From a safety point of view, it is crucial to detect objects that are in short distance to the ego-car as they are a more immediate hazard than long-distance objects. For this reason, our datasets (SOS and CWL) provide meta information like depth, i.e., distance between ground truth OOD objects and camera. In this section, we apply the segmentation metrics (see section 4.1) on different depth intervals and report the results in table 5.

We separate the depth values in 5 equally sized binnings having a typical size of a compact car (4 meters) and two greater intervals for the CWL dataset for far distances. With respect to the pixel-wise metrics (AuPRC and FPR_{95}) as well as the segment-wise metric ($\bar{F}_1$) the best performance is mostly achieved for distances between 4 and 12 meters. The values degrade, on the one hand, when the OOD objects are very close to the vehicle due to partial occlusion. On the other hand, the OOD objects are poorly detected at greater distances, due to the smallness of the area covered in the image.

This same behavior can also be observed in fig. 9 for the SOS dataset and in fig. 10 for the CWL dataset.

Table 5: OOD object segmentation results for the SOS and the CWL dataset obtained by two differently trained meta classifiers (M_1 and M_2) separated into depth intervals, i.e., the difference between ground truth segments and the ego-vehicle (in m).

	SOS				CWL			
depth [m]	AuPRC ↑	FPR_{95} ↓	$\bar{F}_1(M_1)$ ↑	$\bar{F}_1(M_2)$ ↑	AuPRC ↑	FPR_{95} ↓	$\bar{F}_1(M_1)$ ↑	$\bar{F}_1(M_2)$ ↑
$(0-4]$	**82.49**	1.50	49.20	23.56	54.59	1.38	7.72	14.36
$(4-8]$	57.42	0.77	**49.98**	**47.71**	**71.08**	1.30	**50.56**	**46.57**
$(8-12]$	45.79	**0.69**	40.80	39.94	55.63	1.38	47.98	45.06
$(12-16]$	31.61	1.01	30.61	31.94	38.66	**1.23**	40.57	37.12
$(16-20]$	24.26	2.86	29.54	31.20	23.16	1.30	33.77	30.72
$(20-40]$	-	-	-	-	22.18	02.13	36.78	30.66
$(40-65]$	-	-	-	-	01.74	02.35	11.67	10.26

These plots show the correlation between the IoU (of ground truth and predicted objects using meta classifier M_1) and the distance of the ground truth objects to the camera. For most objects, the segment-wise IoU increases the closer the objects are, i.e. we observe a negative correlation between the distance and OOD segmentation performance.

D Numerical Results per Class

Up to now, the presented results are aggregated over all OOD classes, here we present results for these classes separately. The OOD segmentation results are given in table 6 for the SOS dataset and in table 7 for the CWL dataset.

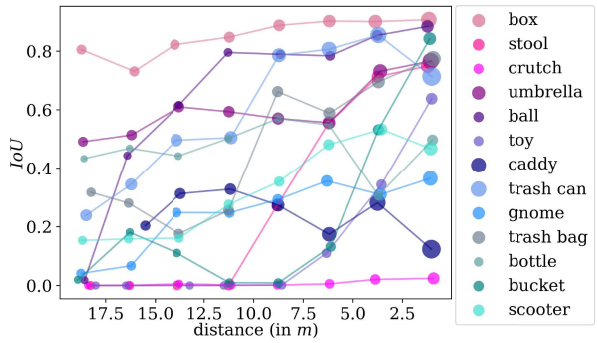


Fig. 9: Discretized distance between ground truth objects and camera vs. mean IoU for the different object types of the SOS dataset and meta classifier M_1. The dot size is proportional to mean segment size.

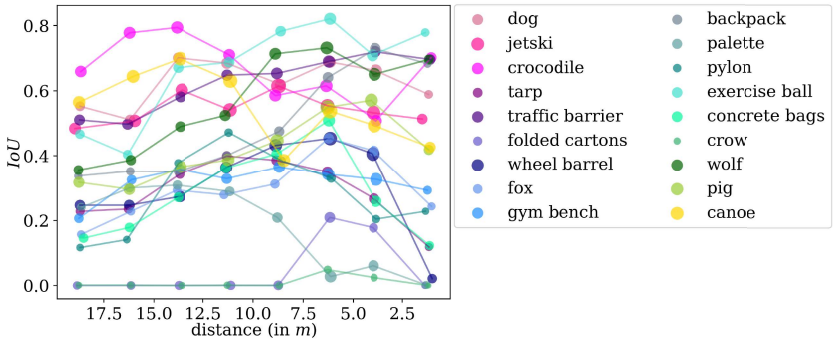


Fig. 10: Discretized distance between ground truth objects and camera vs. mean IoU for the different object types of the CWL dataset and meta classifier M_1. The dot size is proportional to mean segment size.

We observe strong results for classes like box and umbrella in SOS. In CWL objects like jetski and dog are segmented best. The values decrease for flat and narrow obstacles like the folded cartons, palette (CWL), or crutch (SOS). This observation can also be seen in fig. 9 and fig. 10 as these objects are rarely detected (IoU values equal to or slightly greater than zero). Furthermore, unlike observed in the previous section, there is no correlation between segment size and IoU, i.e., both large and small OOD objects are well detected and tracked. Moreover, there are also performance gaps for different animals in the CWL dataset. With respect to dog, pig, crocodile and wolf, we observe better results than for fox and crow.

The tracking results separated by classes are shown in table 8 for the SOS dataset and in table 9 for the CWL dataset. We obtain good tracking performance for classes that also performed well in the OOD segmentation task. This

can be observed for objects such as umbrella and box (SOS) or jetski, crocodile and wolf (CWL). Besides that, other classes can be tracked reliably as well. For the SOS dataset, the best results are achieved for OOD objects of class ball, yielding the highest *MOTA* and comparatively small *MOTP* values. For the CWL dataset, our method performs best for the backpack objects in terms of the *MOTP* metric, i.e., high tracking precision. Moreover, all traffic barriers objects are tracked consistently, yielding high tracking length l_t scores.

Table 6: OOD object segmentation results per class for the SOS dataset obtained by two differently trained meta classifiers (M_1 and M_2).

class	AuPRC ↑	FPR_{95} ↓	$\bar{F}_1(M_1)$ ↑	$\bar{F}_1(M_2)$ ↑
box	55.92	2.28	**39.17**	**14.01**
stool	39.41	0.90	16.82	5.95
crutch	00.46	10.46	0.00	0.00
umbrella	**87.21**	**0.06**	30.79	11.97
ball	31.78	0.59	31.37	11.19
toy	16.48	4.49	9.88	2.64
caddy	23.92	4.49	9.80	3.12
trash can	86.49	0.20	32.75	9.16
gnome	40.93	0.41	13.90	4.45
trash bag	67.73	0.23	28.91	10.24
bottle	3.08	1.59	25.90	8.56
bucket	18.07	2.04	14.10	3.80
scooter	18.58	0.65	15.09	6.11

Table 7: OOD object segmentation results per class for the CWL dataset obtained by two differently trained meta classifiers (M_1 and M_2).

class	AuPRC ↑	FPR_{95} ↓	$\bar{F}_1(M_1)$ ↑	$\bar{F}_1(M_2)$ ↑
dog	49.36	0.57	**29.42**	**38.94**
jetski	**69.68**	**0.19**	18.00	30.51
crocodile	22.03	0.62	12.94	20.64
tarp	22.67	3.52	12.74	16.39
traffic barrier	17.82	0.52	20.63	30.02
folded cartons	1.86	18.45	1.13	1.55
wheel barrel	53.58	0.28	8.58	13.88
fox	8.94	0.57	7.49	11.65
gym bench	7.98	2.23	8.99	14.72
backpack	20.53	2.23	21.53	22.12
palette	1.44	4.71	5.84	6.69
pylon	1.02	1.46	12.34	8.97
exercise ball	48.69	0.36	23.31	32.95
concrete bags	15.08	2.35	10.53	11.95
crow	0.46	10.30	0.34	1.14
wolf	23.26	1.02	17.36	24.09
pig	67.24	**0.19**	20.53	29.47
canoe	37.38	1.30	18.57	24.93

Table 8: Object tracking per class for the SOS dataset obtained by two differently trained meta classifiers (M_1 and M_2).

class	model	$MOTA \uparrow$	$\overline{mme} \downarrow$	$MOTP \downarrow$	GT	MT	PT	ML	$l_t \uparrow$
box	M_1	0.6117	0.0194	1.8829	2	2	0	0	**0.9596**
	M_2	0.3689	0.0291	**1.8780**	2	2	0	0	0.9339
stool	M_1	0.2233	**0.0097**	4.6991	2	0	2	0	0.3689
	M_2	−0.3981	0.0097	5.0206	2	0	2	0	0.4375
crutch	M_1	0.0159	0.0794	85.0229	2	0	0	2	0.0986
	M_2	0.1190	0.0476	49.4550	2	0	1	1	0.1624
umbrella	M_1	0.5041	0.0661	7.6697	2	2	0	0	0.8945
	M_2	0.3388	**0.0000**	9.8624	2	2	0	0	**0.9958**
ball	M_1	**0.6893**	0.0485	**1.8242**	2	1	1	0	0.8136
	M_2	**0.7184**	0.0680	1.8902	2	2	0	0	0.9148
toy	M_1	0.2255	0.0980	6.5618	2	0	2	0	0.2696
	M_2	0.0882	0.0588	6.7381	2	0	2	0	0.2757
caddy	M_1	−0.3402	0.1443	54.3125	2	1	1	0	0.7373
	M_2	−0.3299	0.1959	57.0536	2	0	2	0	0.6392
trash can	M_1	0.5000	0.0726	12.3672	2	1	1	0	0.7223
	M_2	0.1210	0.1532	11.8235	2	0	2	0	0.4505
gnome	M_1	0.2761	0.0672	9.3688	2	0	1	1	0.3437
	M_2	0.2761	0.0149	7.8730	2	0	1	1	0.3108
trash bag	M_1	−0.0569	0.0732	5.2728	2	1	1	0	0.5799
	M_2	−1.5691	0.0813	4.8609	2	1	1	0	0.6070
bottle	M_1	0.0325	0.0488	4.6618	2	1	1	0	0.7799
	M_2	−1.8862	0.0569	4.4068	2	1	1	0	0.6866
bucket	M_1	0.0547	0.0625	3.5966	2	0	2	0	0.2631
	M_2	−0.0781	0.0078	3.6122	2	0	1	1	0.2049
scooter	M_1	0.2522	0.0435	15.9011	2	1	1	0	0.6106
	M_2	−3.5739	0.1217	18.6407	2	1	1	0	0.6894

Table 9: Object tracking results per class for the CWL dataset obtained by two differently trained meta classifiers (M_1 and M_2).

class	model	$MOTA \uparrow$	$\overline{mme} \downarrow$	$MOTP \downarrow$	GT	MT	PT	ML	$l_t \uparrow$
dog	M_1	0.8730	0.0106	4.9561	5	5	0	0	0.9153
	M_2	0.7143	0.0159	3.9630	5	3	2	0	0.7407
jetski	M_1	0.9223	0.0097	53.0367	5	5	0	0	0.9806
	M_2	**0.9417**	0.0097	63.7777	5	5	0	0	**0.9515**
crocodile	M_1	0.8493	0.0137	3.0580	1	1	0	0	0.8767
	M_2	0.7397	**0.0000**	2.8612	1	0	1	0	0.7534
tarp	M_1	0.4298	0.0165	7.0243	3	0	3	0	0.6860
	M_2	0.3140	0.0165	7.0676	3	0	3	0	0.5372
traffic barrier	M_1	−0.8394	**0.0000**	5.2763	2	2	0	0	0.9854
	M_2	−0.2263	0.0073	5.1370	2	2	0	0	0.8978
folded cartons	M_1	−1.2722	**0.0000**	10.2944	3	0	0	3	0.0278
	M_2	−0.7667	**0.0000**	11.5456	3	0	0	3	0.0222
wheel barrel	M_1	0.0463	0.0093	8.8480	3	1	1	1	0.4167
	M_2	0.1944	0.0093	8.9901	3	1	1	1	0.4537
fox	M_1	0.1269	0.0224	5.3693	3	1	1	1	0.4925
	M_2	0.2164	0.0149	5.6093	3	1	1	1	0.4403
gym bench	M_1	0.4876	0.0248	9.2609	3	1	2	0	0.5537
	M_2	0.4132	0.0331	9.7580	3	1	2	0	0.5455
backpack	M_1	0.3966	0.0168	**2.2138**	4	2	2	0	0.5307
	M_2	0.2235	0.0223	**2.4282**	4	0	4	0	0.3743
palette	M_1	0.2958	0.0493	9.6678	3	0	3	0	0.4014
	M_2	0.1338	0.0211	13.1479	3	0	2	1	0.2465
pylon	M_1	−0.0676	0.0743	2.7237	3	0	3	0	0.4932
	M_2	−0.2365	0.0946	3.3018	3	0	2	1	0.2297
exercise ball	M_1	0.4333	0.0533	5.9323	4	3	1	0	0.8600
	M_2	0.3733	0.0533	5.4039	4	2	2	0	0.7667
concrete bags	M_1	0.3704	0.0222	7.3404	3	0	3	0	0.4889
	M_2	0.1852	0.0519	8.3914	3	0	2	1	0.3259
crow	M_1	−1.4920	0.0053	5.9448	4	0	0	4	0.0214
	M_2	−0.9198	0.0107	7.2169	4	0	0	4	0.0374
wolf	M_1	**0.9633**	**0.0000**	5.5236	3	3	0	0	**0.9908**
	M_2	0.8624	0.0092	7.5521	3	3	0	0	0.9083
pig	M_1	0.6173	0.0051	31.1796	6	5	1	0	0.8061
	M_2	0.6633	0.0204	32.2780	6	4	2	0	0.7704
canoe	M_1	0.4545	0.0210	16.7630	4	3	1	0	0.8671
	M_2	0.3776	0.0210	18.4700	4	1	3	0	0.6853

E Retrieval of OOD Objects for WOS

In addition to the labeled data sets SOS and CWL, we applied our toolchain to another data set which we abbreviate as WOS. As this data set does not include any annotated data, it serves as a test scenario, only. This is, we do not provide any evaluation results, but some visualizations of the retrieved clusters. We trained two meta classifiers on SOS and CWL, respectively. Since the results for both meta classification models are similar and the domain shift between SOS and WOS is less, we limit our visualizations onto this respective meta model, while increasing the minimal tracking length to $\ell = 25$.

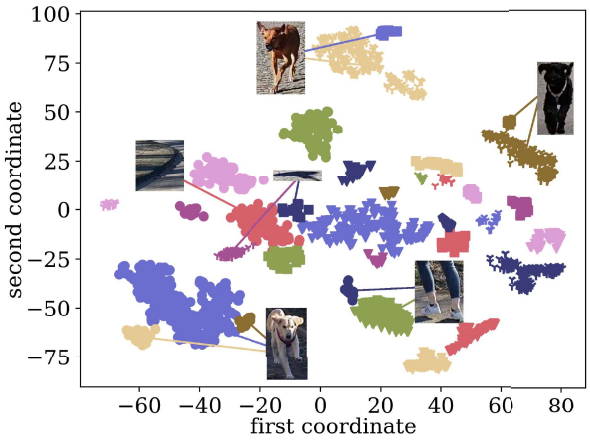


Fig. 11: Clustering of WOS OOD segments (plus some example images) via DBSCAN in the embedding space for a meta classifier trained on SOS and minimum tracking length $\ell = 25$.

As illustrated in fig. 11, we are able to retrieve clusters constituted of OOD objects, e.g. dogs (see fig. 12). Our data set includes three different dogs, that are visible in multiple scenes. We observe that these three dogs do not constitute one overall dog cluster, however, each of them forms a cluster containing multiple sequences, as well as different postures, sizes/distances, backgrounds and perspectives. Moreover, some of the retrieved clusters represent OOD objects like balls, bags or skateboards.

Further, we discover many false positive OOD predictions, that are partly represented in fig. 13, e.g. humans, sidewalks, manhole covers or shadows.

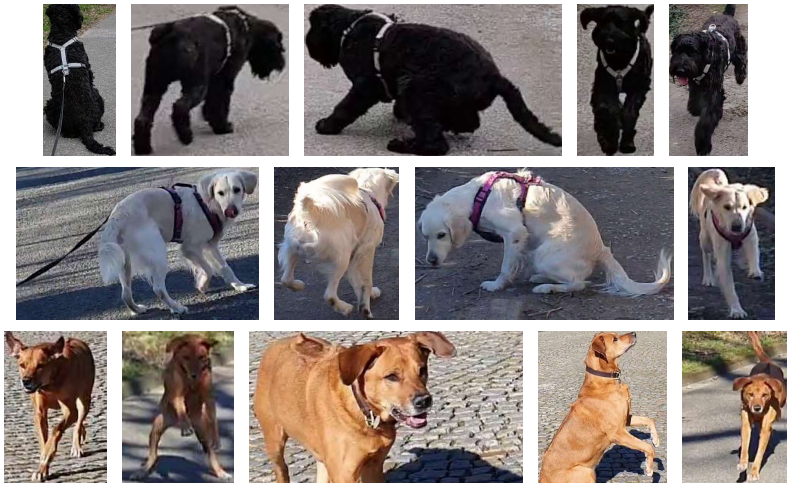


Fig. 12: Example images taken from three different clusters (one cluster per row), all representing the overall category *dog*.

Fig. 13: False positive OOD predictions forming three different clusters, namely legs, sidewalks and shadows.