Towards Unsupervised Open World Semantic Segmentation (Supplementary material)

Svenja Uhlemeyer[1] **Matthias Rottmann**[1] **Hanno Gottschalk**[1]

[1]Faculty of Mathematics and Natural Sciences, University of Wuppertal, Germany,

A EVALUATED MODELS

We performed six experiments that differ in terms of underlying datasets, network architectures and novelties. In this section we provide a class-wise evaluation of each initial and extended DNN, as well as example images for all evaluated models, *i.e.,* also for the baseline and the oracle DNNs. For the extended models, we report the mean and standard deviation of the evaluation metrics for five runs, respectively, using the random seeds 14, 123, 666, 375 and 693.

A.1 EXPERIMENT 1

For the first experiment, we trained a DeepLabV3+ on the Cityscapes dataset, excluding the classes *pedestrian* and *rider*, both together constituting the class *human*. This novelty is well separable from all the known classes as these belong to different, non-organic categories. As there are no similar classes, humans are either totally "overlooked" by the segmentation DNN, *i.e.,* assigned to the class predicted in their background, or predicted as related classes, *e.g.,* as *bicycle*, *motorcycle* or *car* (cf. Fig. 1). Since our anomaly detection method fails to spot overlooked persons, these remain mislabeled even in the pseudo ground truth, thus negatively affecting the incremental training procedure. For an example we refer to Fig. 2, where a cyclist is assigned to the background classes *road* and *car*. To prevent this issue, we ignore all known classes $c \in \mathcal{C}$ present in the pseudo labels. Our newly collected data $\mathcal{D}^{C+1}$ contains 76 pseudo-labeled images. The replayed training data is selected such that at least 25% - 35% of the images contain cars, motorcycles and bicycles, respectively.

We evaluated the initial and the extended DNN on the Cityscapes validation data. Class-wise results are provided in Tab. 1. Besides the novel class, which achieves an IoU value of nearly 40% with approximately 50-60% precision and recall, the incremental training has only little impact on previously-known classes. For many classes, however, we observe an improvement in precision at the expense of the corresponding recall values, *e.g.,* for the classes *fence*, *truck* and *train*. This is also reflected in the mean precision and recall values over $\mathcal{C}$, *i.e.,* while precision increases by 3.53%, recall decreases by 3.77%. Especially the classes *motorcycle* and *bicycle* gain performance regarding the IoU and precision, which is mainly due to human pixels initially assigned to those classes, while the proportion of bikes (motor- or bicycles) that are predicted correctly drops significantly.

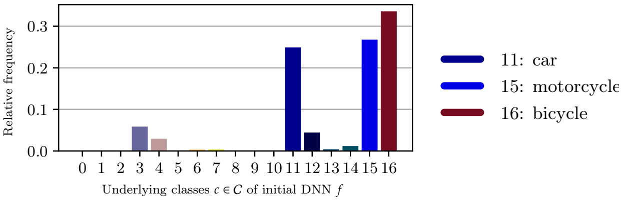


Figure 1: Bar plot showing the relative frequencies of predicted classes for instances of the novel class *human*.

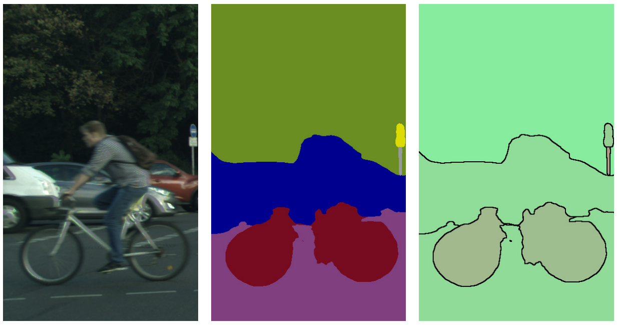


image patch

predicted segmentation

quality estimation

Figure 2: Image patch, semantic segmentation and prediction quality estimation for a scene, where a cyclist is overlooked by the initial DNN.

A comparison of all evaluated models in the first experiment is illustrated for an example image in Fig. 3. We observe

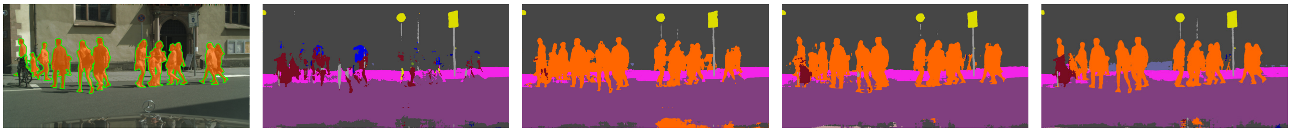


image & annotation initial DNN extended DNN baseline oracle

Figure 3: Comparison of the semantic segmentation predictions of all DNNs evaluated in the first experiment for an exemplary scene from the Cityscapes validation data.

1. experiment Cityscapes, human	DeepLabV3+					
	initial			extended		
Class	IoU	precision	recall	IoU	precision	recall
road	97.34	98.35	98.96	97.43 ± 0.05	98.54 ± 0.12	98.86 ± 0.08
sidewalk	80.63	89.39	89.16	80.51 ± 0.23	89.50 ± 0.50	88.91 ± 0.67
building	88.91	92.80	95.50	89.40 ± 0.05	93.42 ± 0.20	95.42 ± 0.24
wall	47.24	74.57	56.32	47.74 ± 0.57	78.92 ± 0.49	54.71 ± 0.77
fence	51.03	66.76	68.41	49.20 ± 0.44	70.06 ± 1.55	62.33 ± 1.26
pole	52.90	72.68	66.02	53.30 ± 0.39	74.42 ± 1.41	65.31 ± 1.64
traffic light	55.44	75.04	67.98	55.33 ± 0.19	75.49 ± 1.24	67.47 ± 1.21
traffic sign	66.66	86.22	74.61	66.32 ± 0.62	87.54 ± 1.41	73.27 ± 1.67
vegetation	89.95	93.60	95.85	90.15 ± 0.03	94.01 ± 0.22	95.65 ± 0.22
terrain	56.29	77.66	67.17	55.29 ± 0.47	75.88 ± 1.67	67.14 ± 1.77
sky	93.76	96.38	97.18	93.60 ± 0.11	96.01 ± 0.26	97.39 ± 0.19
human	00.00	00.00	00.00	39.80 ± 0.73	60.60 ± 1.20	53.72 ± 1.42
car	90.61	92.97	97.27	91.16 ± 0.21	95.25 ± 0.50	95.50 ± 0.47
truck	69.66	80.23	84.09	68.98 ± 0.56	84.92 ± 2.35	78.70 ± 1.97
bus	76.90	88.59	85.35	71.57 ± 0.60	87.25 ± 1.33	79.95 ± 1.15
train	70.35	83.33	81.87	63.11 ± 3.17	89.63 ± 1.61	68.13 ± 3.93
motorcycle	24.45	28.57	62.92	32.92 ± 1.13	53.91 ± 2.07	45.89 ± 2.21
bicycle	54.57	59.30	87.24	59.01 ± 0.61	71.62 ± 2.43	77.20 ± 3.38
mean over $\mathcal{C}$	68.63	79.79	80.94	68.53 ± 0.27	83.32 ± 0.28	77.17 ± 0.60
mean over $\mathcal{C}^+$	64.82	75.36	76.44	66.94 ± 0.27	82.05 ± 0.25	75.86 ± 0.55

Table 1: In-depth evaluation on the Cityscapes validation data for the first experiment, where we incrementally extend a DeepLabV3+ by the novel class *human* on the Cityscapes dataset. We provide IoU, precision and recall values obtained for both, the initial and the extended DNN, on a class-level as well as averaged over the classes in $\mathcal{C}$ and $\mathcal{C}^+$, respectively.

2. experiment Cityscapes, bus	DeepLabV3+					
	initial			extended		
Class	IoU	precision	recall	IoU	precision	recall
road	97.63	98.81	98.80	97.57 ± 0.03	98.76 ± 0.09	98.79 ± 0.08
sidewalk	81.60	89.65	90.09	81.57 ± 0.10	90.07 ± 0.46	89.63 ± 0.45
building	90.19	94.50	95.19	89.90 ± 0.10	94.22 ± 0.26	95.15 ± 0.25
wall	48.77	78.07	56.51	44.89 ± 3.11	79.23 ± 1.36	50.94 ± 4.20
fence	53.86	70.97	69.08	51.74 ± 0.81	71.82 ± 0.62	64.92 ± 1.27
pole	55.03	75.71	66.83	54.05 ± 0.61	77.62 ± 1.11	64.06 ± 1.54
traffic light	55.87	77.29	66.84	54.70 ± 0.92	80.15 ± 2.02	63.35 ± 2.46
traffic sign	68.21	87.02	75.94	67.88 ± 0.32	87.87 ± 0.98	74.91 ± 1.08
vegetation	90.35	93.98	95.91	90.21 ± 0.09	93.70 ± 0.33	96.04 ± 0.26
terrain	54.03	79.90	62.53	52.77 ± 0.46	75.06 ± 1.14	64.00 ± 1.01
sky	93.64	96.14	97.30	93.26 ± 0.29	95.55 ± 0.63	97.49 ± 0.36
person	71.65	83.27	83.70	71.02 ± 0.21	82.22 ± 0.87	83.92 ± 0.65
rider	48.77	68.86	62.58	47.15 ± 0.73	70.85 ± 1.32	58.55 ± 1.99
car	91.90	94.65	96.94	91.76 ± 0.11	95.35 ± 0.61	96.07 ± 0.62
truck	47.51	51.19	86.87	54.14 ± 1.85	69.81 ± 4.17	71.09 ± 5.25
bus	00.00	00.00	00.00	44.73 ± 1.46	58.33 ± 3.13	66.15 ± 5.16
train	43.57	48.58	80.88	55.46 ± 1.64	74.35 ± 5.75	69.19 ± 5.46
motorcycle	44.35	61.76	61.13	41.66 ± 1.17	71.22 ± 1.70	50.16 ± 2.38
bicycle	68.00	77.42	84.82	67.52 ± 0.28	76.38 ± 0.64	85.35 ± 0.44
mean over $\mathcal{C}$	66.94	79.32	79.55	67.07 ± 0.12	82.46 ± 0.56	76.31 ± 0.46
mean over $\mathcal{C}^+$	63.42	75.15	75.36	65.89 ± 0.10	81.19 ± 0.54	75.78 ± 0.34

Table 2: In-depth evaluation on the Cityscapes validation data for the second experiment, where we incrementally extend a DeepLabV3+ by the novel class *bus* on the Cityscapes dataset. We provide IoU, precision and recall values obtained for both, the initial and the extended DNN, on a class-level as well as averaged over the classes in $\mathcal{C}$ and $\mathcal{C}^+$, respectively.

a reduction of noise in the model's predictions, starting from the initial DNN, to the extended DNN, the baseline and the oracle. Nonetheless, the predicted segmentation of our extended DNN comes close to those predicted by the comparative models that both require ground truth for the novel class.

A.2 EXPERIMENT 2

The setup of the second experiment is the same as in the first one (DeepLabV3+, Cityscapes dataset), but excluding busses from the set of known classes instead of humans. This novelty belongs to the vehicle category, thus being akin to other vehicle classes as *train* or *truck*. These are also the classes the objects declared as novel were predicted for the most part, as we illustrated in Fig. 4. On that account, at least 50% of the 55 images in $\mathcal{D}^{C+1}$ contain trucks, 30% trains. As a consequence of the visual relatedness, trucks and trains that exhibit a low prediction quality, *i.e.,* that are treated as anomalies, contaminate the cluster of busses in the two-dimensional embedding space. We observed, that the segmentation network predicts most of these "detected" trucks and trains correctly, while it assigns multiple classes, *i.e.,* multiple segments in the semantic segmentation prediction, to a bus. Thus, we delete anomalies from the embedding space, whose predicted segmentation consists of only one segment (ignoring segments with less than 500 pixels).

Again, we provide a class-wise evaluation on the Cityscapes validation split in Tab. 2 and present a comparison of different models for one exemplary street scene in Fig. 4. Here, large parts of the bus in the foreground are predicted correctly by our extended DNN. The bus in the background is even better recognized by our network than by the baseline and oracle. Analogous to the first experiment, the most similar classes *truck* and *train* show increasing IoU and precision, but decreasing recall values. Averaged over the known classes $c \in \mathcal{C}$, we again observe improvement in IoU and precision with a concurrent drop in recall. Averaged over the extended class set $\mathcal{C}^+$, all three performance measures increase after class-incremental learning.

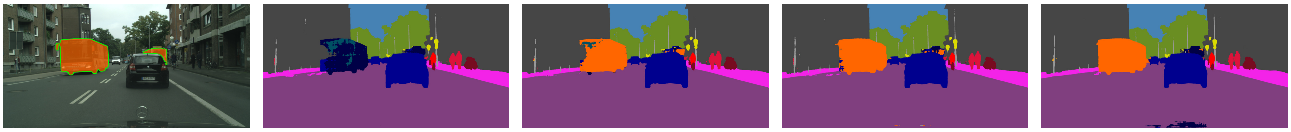


image & annotation

initial DNN

extended DNN

baseline

oracle

Figure 4: Comparison of the semantic segmentation predictions of all DNNs evaluated in the second experiment for an example image from the Cityscapes validation data.

3. experiment Cityscapes, multi	DeepLabV3+					
	initial			extended		
Class	IoU	precision	recall	IoU	precision	recall
road	95.43	96.41	98.95	96.62 ± 0.07	98.29 ± 0.20	98.27 ± 0.22
sidewalk	77.23	83.84	90.74	76.42 ± 0.26	84.27 ± 0.98	89.16 ± 0.91
building	87.21	91.05	95.39	87.42 ± 0.12	92.66 ± 0.30	93.92 ± 0.40
wall	45.86	68.38	58.20	40.36 ± 0.59	76.67 ± 1.57	46.03 ± 1.07
fence	47.86	59.63	70.79	41.15 ± 1.47	69.23 ± 2.40	50.44 ± 2.54
pole	51.63	69.15	67.09	48.68 ± 0.48	73.74 ± 1.13	58.93 ± 1.42
traffic light	55.61	77.70	66.17	45.62 ± 0.47	72.64 ± 0.85	55.09 ± 1.07
traffic sign	64.84	80.37	77.04	58.34 ± 0.74	86.84 ± 0.70	64.01 ± 1.23
vegetation	88.26	91.27	96.40	88.61 ± 0.22	91.80 ± 0.43	96.22 ± 0.21
terrain	53.22	72.42	66.74	45.43 ± 0.77	79.11 ± 1.55	51.66 ± 1.67
sky	93.58	96.11	97.27	92.41 ± 0.16	95.56 ± 0.19	96.56 ± 0.10
human	00.00	00.00	00.00	40.22 ± 1.77	68.74 ± 4.84	49.65 ± 4.80
car	00.00	00.00	00.00	81.27 ± 1.16	86.56 ± 2.20	93.05 ± 1.12
truck	9.31	9.41	89.35	25.59 ± 7.41	61.27 ± 5.50	30.77 ± 9.90
train	41.70	45.05	84.87	49.87 ± 5.21	60.85 ± 8.56	73.99 ± 2.61
motorcycle	4.03	4.12	66.09	14.30 ± 2.72	63.79 ± 3.44	15.64 ± 3.31
bicycle	39.13	41.30	88.15	51.97 ± 1.58	71.26 ± 1.98	65.95 ± 4.30
mean over $\mathcal{C}$	56.99	65.75	80.88	57.52 ± 0.80	78.53 ± 1.20	65.78 ± 1.00
mean over $\mathcal{C}^+$	50.29	58.01	71.37	57.90 ± 0.68	78.43 ± 1.10	66.43 ± 0.94

Table 3: In-depth evaluation on the Cityscapes validation data for the third experiment, where we incrementally extend a DeepLabV3+ by the novel classes *human* and *car* on the Cityscapes dataset. We provide IoU, precision and recall values obtained for both, the initial and the extended DNN, on a class-level as well as averaged over the classes in $\mathcal{C}$ and $\mathcal{C}^+$, respectively.

A.3 EXPERIMENT 3

In the next experiment we extend the previous ones by enlarging the set of novel classes, withholding the classes *pedestrian&rider*, *bus* and *car*. Again, we trained a DeepLabV3+ network on the Cityscapes dataset to learn the remaining, non-novel classes. We reconsidered our approach to reject possibly known objects from the embedding space to improve the purity of novel object clusters. Instead of rejecting anomalous segments that consist of only one predicted segment in the semantic segmentation mask, we include a random choice of objects / segments from each known class into the embedding space. If an anomalous object can be assigned to an existing class, it is no longer taken into account in the further procedure. To decide whether an object is novel or known, we consider its 2.75-neighborhood. If this contains at least 10 known objects from which at least 80% belong to the most frequent class, we assume the anomaly belongs to even this class, *i.e.,* we reject it. Consequently, we discard the detected bus segments since these are closely related to the classes *truck* and *train*. However, we obtain two clusters, one for the class *car* (1375 segments) and one for the class *human* (135 segments). We incrementally expand the model by these classes, achieving a similar IoU value (around 40%) for the *human* class as in experiment 1, where we only learned a single class. For the *bus* class, we even get an IoU value of more than 80%. Detailed results are provided in Tab. 3.

4. experiment (a) A2D2, guardrail	DeepLabV3+					
	initial			extended		
Class	IoU	precision	recall	IoU	precision	recall
road	95.59	97.21	98.29	95.93 ± 0.06	97.94 ± 0.18	97.91 ± 0.15
sidewalk	72.01	86.73	80.92	72.08 ± 0.41	85.29 ± 0.84	82.33 ± 1.28
building	87.82	93.58	93.44	85.75 ± 0.67	93.13 ± 0.53	91.54 ± 1.01
fence	59.35	81.59	68.53	56.76 ± 0.37	79.89 ± 2.40	66.29 ± 1.63
pole	56.13	76.39	67.91	54.31 ± 0.24	77.86 ± 0.52	64.23 ± 0.66
traffic light	68.41	85.10	77.72	65.48 ± 0.19	84.21 ± 0.77	74.65 ± 0.83
traffic sign	76.34	86.78	86.38	74.53 ± 0.38	89.98 ± 1.11	81.30 ± 1.19
vegetation	91.61	94.01	97.29	92.00 ± 0.23	94.81 ± 0.38	96.89 ± 0.17
sky	97.96	98.72	99.22	97.81 ± 0.03	98.57 ± 0.07	99.22 ± 0.04
person	67.60	79.28	82.11	64.27 ± 0.58	87.70 ± 0.87	70.65 ± 1.21
car	93.19	96.73	96.22	92.42 ± 0.11	96.04 ± 0.35	96.08 ± 0.35
truck	84.99	88.51	95.53	80.98 ± 2.66	84.75 ± 3.29	94.82 ± 0.69
motorcycle	48.68	84.71	53.37	26.05 ± 2.72	90.18 ± 2.09	26.85 ± 3.04
bicycle	61.08	80.65	71.57	50.65 ± 3.27	85.78 ± 2.10	55.43 ± 4.78
guardrail	00.00	00.00	00.00	46.10 ± 4.79	80.41 ± 2.12	52.09 ± 6.42
mean over $\mathcal{C}$	75.77	87.86	83.47	72.07 ± 0.39	89.01 ± 0.48	78.44 ± 0.52
mean over $\mathcal{C}^+$	70.72	82.00	77.90	70.34 ± 0.50	88.44 ± 0.40	76.69 ± 0.47

Table 4: In-depth evaluation on the A2D2 validation data for the fourth experiment, where we first fine-tune and then incrementally extend a DeepLabV3+ by the novel class *guardrail* on the A2D2 dataset. We provide IoU, precision and recall values obtained for both, the initial and the extended DNN, on a class-level as well as averaged over the classes in $\mathcal{C}$ and $\mathcal{C}^+$, respectively.

A.4 EXPERIMENT 4(A)

The fourth experiment involves two different network architectures. Results for the first one are shown in experiment 4(a), results for the other one in 4(b). We start with a DeepLabV3+ network trained on the Cityscapes dataset and aim to detect and learn the *guardrail* class using images taken from the A2D2 dataset. To mitigate a performance drop caused by the domain shift from Cityscapes to A2D2, we first fine-tune the decoder for 70 epochs on our A2D2 training split, applying the same hyperparameters we used for the incremental training (see Sec. 5). By that, we improve the mean IoU of the initial network from 59.38% to 75.77%. The classes which suffer the most are *person*, *motorcycle* and *bicycle*, which is presumably due to their rare occurrence on country roads and highways, and therefore,

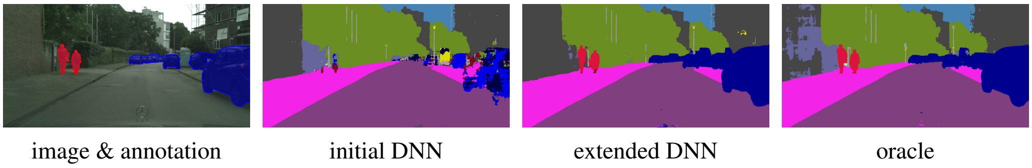


image & annotation initial DNN extended DNN oracle

Figure 5: Comparison of the semantic segmentation predictions of all DNNs evaluated in the third experiment for an example image from the Cityscapes validation data.

4. experiment (b) A2D2, guardrail	PSPNet					
	initial			extended		
Class	IoU	precision	recall	IoU	precision	recall
road	95.18	97.10	97.96	94.93 ± 0.21	96.94 ± 0.55	97.86 ± 0.34
sidewalk	66.15	83.68	75.94	62.19 ± 2.28	82.28 ± 2.09	71.99 ± 4.75
building	84.32	92.46	90.54	82.38 ± 0.46	90.78 ± 0.86	89.91 ± 1.04
fence	54.48	76.84	65.18	50.67 ± 1.24	80.91 ± 1.85	57.62 ± 2.33
pole	44.60	63.94	59.59	42.15 ± 0.91	65.52 ± 2.19	54.31 ± 2.89
traffic light	58.94	81.14	68.30	56.07 ± 0.17	80.65 ± 1.85	64.83 ± 1.37
traffic sign	71.30	87.71	79.22	67.63 ± 0.47	87.61 ± 0.71	74.79 ± 0.56
vegetation	90.68	93.12	97.18	90.65 ± 0.11	93.71 ± 0.41	96.53 ± 0.32
sky	97.57	98.44	99.10	97.21 ± 0.12	98.06 ± 0.19	99.12 ± 0.10
person	59.17	82.53	67.64	46.20 ± 1.13	82.99 ± 0.99	51.04 ± 1.60
car	89.39	94.36	94.44	86.82 ± 0.34	93.90 ± 0.57	92.01 ± 0.60
truck	77.83	84.05	91.31	73.53 ± 1.91	82.11 ± 2.40	87.58 ± 1.25
motorcycle	19.73	76.72	20.99	7.00 ± 2.02	94.92 ± 3.73	7.04 ± 2.07
bicycle	53.49	71.82	67.70	46.05 ± 1.37	79.31 ± 2.49	52.44 ± 2.71
guardrail	00.00	00.00	00.00	32.79 ± 3.47	70.75 ± 2.04	38.04 ± 4.90
mean over $\mathcal{C}$	68.77	84.57	76.79	64.54 ± 0.28	86.41 ± 0.77	71.22 ± 0.69
mean over $\mathcal{C}^+$	64.19	78.93	71.67	62.42 ± 0.42	85.36 ± 0.78	69.01 ± 0.94

Table 5: In-depth evaluation on the A2D2 validation data for the fourth experiment, where we first fine-tune and then incrementally extend a PSPNet by the novel class *guardrail* on the A2D2 dataset. We provide IoU, precision and recall values obtained for both, the initial and the extended DNN, on a class-level as well as averaged over the classes in $\mathcal{C}$ and $\mathcal{C}^+$, respectively.

5. experiment A2D2, guardrail	DeepLabV3+					
	initial			extended		
Class	IoU	precision	recall	IoU	precision	recall
road	89.88	92.18	97.30	93.15 ± 0.19	94.89 ± 0.23	98.07 ± 0.12
sidewalk	47.91	76.22	56.33	35.28 ± 2.43	86.95 ± 0.98	37.26 ± 2.67
building	70.94	86.88	79.45	71.25 ± 1.46	90.51 ± 0.89	77.03 ± 2.21
fence	26.08	35.30	49.94	26.20 ± 0.49	37.25 ± 1.46	46.99 ± 1.26
pole	42.59	59.24	60.25	42.77 ± 0.37	62.91 ± 0.73	57.21 ± 0.85
traffic light	47.59	85.85	51.64	52.52 ± 0.70	89.21 ± 1.15	56.10 ± 1.19
traffic sign	54.89	82.49	62.13	57.23 ± 0.25	87.34 ± 1.03	62.42 ± 0.43
vegetation	69.15	96.68	70.83	73.42 ± 0.41	95.05 ± 0.62	76.35 ± 0.34
sky	94.96	98.25	96.59	96.92 ± 0.09	97.81 ± 0.13	99.08 ± 0.05
person	59.77	71.00	79.08	59.58 ± 1.23	84.68 ± 2.45	66.88 ± 2.89
car	90.47	95.72	94.28	90.72 ± 0.16	96.14 ± 0.39	94.16 ± 0.53
truck	62.64	83.61	71.40	71.10 ± 0.24	89.44 ± 0.51	77.62 ± 0.36
motorcycle	28.39	70.82	32.15	32.77 ± 3.05	79.50 ± 3.43	35.96 ± 4.24
bicycle	46.04	78.74	52.57	43.84 ± 1.01	85.43 ± 1.50	47.41 ± 1.56
guardrail	00.00	00.00	00.00	20.90 ± 1.73	77.12 ± 3.95	22.32 ± 2.07
mean over $\mathcal{C}$	59.38	79.50	68.14	60.48 ± 0.47	84.08 ± 0.49	66.61 ± 0.64
mean over $\mathcal{C}^+$	55.42	74.20	63.60	57.84 ± 0.48	83.61 ± 0.68	63.66 ± 0.63

Table 6: In-depth evaluation on the A2D2 validation data for the fifth experiment, where we incrementally extend a DeepLabV3+ (trained on Cityscapes) by the novel class *guardrail* on the A2D2 dataset. We provide IoU, precision and recall values obtained for both, the initial and the extended DNN, on a class-level as well as averaged over the classes in $\mathcal{C}$ and $\mathcal{C}^+$, respectively.

low frequency in the re-training data, which involves only 30 pseudo-labeled and 30 replayed images. Further details are provided in Tab. 4.

A.5 EXPERIMENT 4(B)

In experiment 4(b), we employ a PSPNet instead of a DeepLabV3+, for the rest we proceed as in the previous subsection. Again, the training data consists of 30 images with pseudo ground truth and 30 labeled, replayed images (containing only old classes) from the A2D2 training split. Note that these 30 images are not the same as in experiment 4(a) due to the different network providing predictions of estimated low quality on different images. In total, the initial and the extended PSPNet are outperformed by DeepLabV3+, however, both architectures show similar patterns:

- extended DNN exhibits a high $\text{precision}_{\text{guardrail}}$ and a low $\text{recall}_{\text{guardrail}}$
- classes that are mostly affected by re-training: *person*, *motorcycle*, *bicycle*
- averaged over $\mathcal{C}$ and $\mathcal{C}^+$, respectively, IoU and recall values decrease, precision values increase

For more detailed information we refer to Tab. 5.

A.6 EXPERIMENT 5

Finally, we perform the same experiment as in 4(a) without prior fine-tuning the initial DNN on A2D2. Consequently, the domain shift causes many noisy predictions, exhibiting low prediction quality estimates. We exclude such images from the further process based on two criteria:

1. mean quality score (averaged over pixels) less than 0.7
2. more than 1/3 of all pixels with quality estimate less than 0.9.

If at least one criterion holds, we reject the image, as illustrated in the bottom row of Fig. 7.

Applying our method, we obtain 70 pseudo-labeled images. The incorporation of data seen during training of the initial DNN, *i.e.*, the Cityscapes training data, restrains the network from adapting onto the new domain. We therefore decided to extend the model only on $\mathcal{D}^{C+1}$.

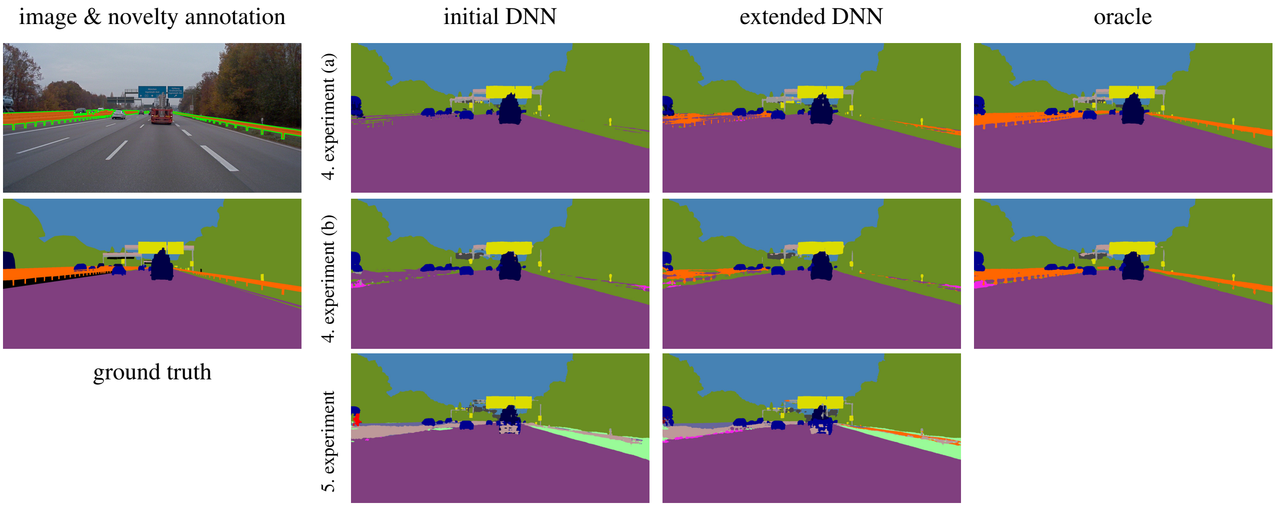


Figure 6: Comparison of the semantic segmentation predictions of all models incrementally extended by the *guardrail* class for an example image from the A2D2 validation split.

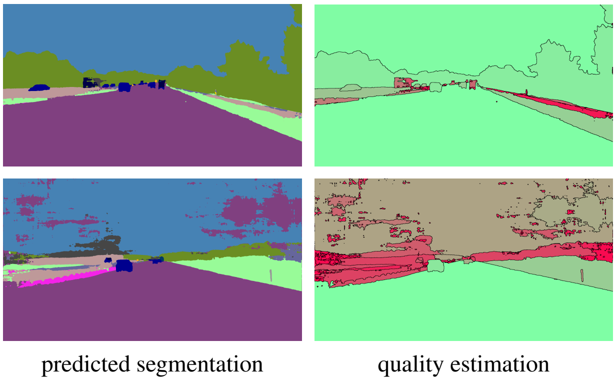


Figure 7: Illustration of prediction quality differences (green color indicates high, red color low prediction quality), caused by the domain shift from Cityscapes to A2D2, mainly due to weather conditions.

Class-wise evaluation results are reported in Tab. 6. Even with a domain shift, we achieve an IoU of 20.90 ± 1.73% for the novel class. This is less than the value obtained with prior fine-tuning. However, this DNN still outperforms the PSPNet from the previous experiment considering only the precision. The low recall values are tolerable since many guardrails are still assigned to the "supercategory" *fence*. For most other classes, the IoU values increase or remain roughly the same. In contrast to the other experiments, the *motorcycle* class improves in IoU, precision and recall values. Only classes that are rare in rural street scenes, *e.g., sidewalk* or *bicycle*, suffer from the incremental training.

A visual comparison of the experiments 4(a), 4(b) and 5 is provided in Fig. 6. All three extended DNNs have learned to predict the novel class to some extent. The prior fine-tuned networks show similar predictions, though DeepLabV3+ is much more precise than the PSPNet and better recognizes the guardrail on the right. The model from the fifth experiment predicts the left guardrail as *fence* (which is not totally mistaken), though it performs better on the right-hand guardrail than the others. Both oracles illustrate, that the *guardrail* class is learnable with high accuracy, still leaving room for improvement of unsupervised methods.

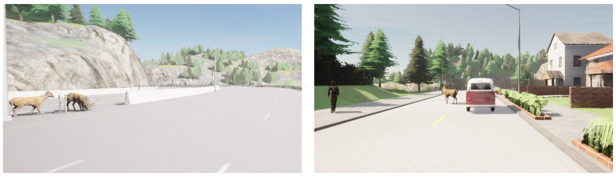


Figure 8: Two examples from our CARLA test dataset including the novel class *deer*.

B SYNTHETIC DATASET

We generated a synthetic dataset with the CARLA simulator, that contains novel classes such as *deer* in the test data. Two examples are provided in Fig. 8. All classes considered as novel are never seen before, *i.e.,* they are not contained in the training data. Besides that, the street scenes for training and testing are recorded under identical conditions, *i.e.,* on the same maps, with the same weather conditions, camera angles etc., so that the segmentation network is not distracted

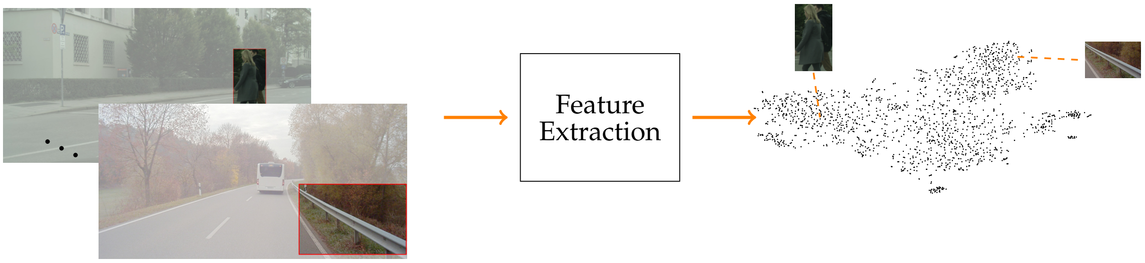


Figure 9: Coarse illustration of the feature extraction process. Detected unknown objects (here: human and guardrail) are cropped out (indicated by the red box). The image patches are fed into an encoder, the resulting feature vectors are then projected into a two dimensional space.

experiment	#metrics	#segments in training set
1	71	608,906
2	73	571,853
3	67	946,318
4a	75	492,210
4b	75	313,720
5	75	535,457

Table 7: Overview about the training data of the meta regressor for each experiment. We report the number of metrics per segment k (that depends on the number of classes $|\mathcal{C}|$) as well as the number of segments produced by the initial network during inference of the training data.

by anything different than the novel objects.

C MODULES

We present a modular procedure, this is, the individual modules can be modified or exchanged. In this section, we provide a deeper insight into the modules **meta regressor** and **feature extractor**.

C.1 UNCERTAINTY METRICS & META REGRESSION

For every segment $k \in \mathcal{K}(\mathcal{D}^{\text{train}})$ we compute the following metrics:

- the size of the segment k, its interior k^{o} and its boundary ∂k:

$$S(k) = |k|,\ S^{\text{o}}(k) = |k^{\text{o}}|,\ \partial S(k) = |\partial k|$$

- the relative sizes:

$$\tilde{S}(k) = S(k)/\partial S(k),\ \tilde{S}^{\text{o}}(k) = S^{\text{o}}(k)/\partial S(k)$$

- several dispersion measures aggregated over k, k^{o} and ∂k, respectively:

$$\bar{D}(k) = \frac{1}{S}\sum_{z\in k} D_z(x),\ \bar{D}^{\text{o}}(k) = \frac{1}{S^{\text{o}}}\sum_{z\in k^{\text{o}}} D_z(x),$$
$$\partial\bar{D}(k) = \frac{1}{\partial S}\sum_{z\in\partial k} D_z(x)$$

where $D \in \{E, M, V\}$, *i.e.,* softmax entropy E, probability margin M and variation ration V.

- the relative dispersion measures:

$$\tilde{\bar{D}}(k) = \bar{D}(k)S(k),\ \tilde{\bar{D}}^{\text{o}}(k) = \bar{D}^{\text{o}}(k)\tilde{S}^{\text{o}}(k)$$

$D \in \{E, M, V\}$.

- the variance of the dispersion measures
- the predicted class $c \in \mathcal{C}$
- the mean softmax probabilities for each class $c \in \mathcal{C}$
- the pixel position of the segment's geometric center
- the ratio of the amount of pixels in the neighborhood of segment k predicted to belong to class $c \in \mathcal{C}$ to the neighborhood size for each class $c \in \mathcal{C}$

Further, we compute the IoU (averaged over each segment), which is the only metric that requires ground truth and serves as target value for the meta regressor. The number of training metrics, *i.e.,* explanatory variables, is reported in Tab. 7 for each experiment. This is, the training data for the meta regressor has a dimension of $|\mathcal{K}(\mathcal{D}^{\text{train}})| \times \#$metrics.

C.2 FEATURE EXTRACTOR

We apply an image classification CNN, pre-trained on ImageNet, without the final classification layer to extract features of image patches as illustrated in Fig. 9. This feature extraction CNN can be exchanged arbitrarily, as long as the resulting feature vectors equally sized for different input dimensions. In Tab. 8 we compare the results for experiment 1, using three different feature extractors, namely DenseNet201, ResNet18 and ResNet152.

model	DenseNet201			ResNet18			ResNet152		
metric	IoU	precision	recall	IoU	precision	recall	IoU	precision	recall
human	39.80 ± 0.73	**60.60** ± 1.20	53.72 ± 1.42	**40.56** ± 0.95	54.80 ± 4.50	61.50 ± 4.12	40.30 ± 0.94	52.17 ± 1.59	**63.97** ± 1.71
mean over C	**68.53** ± 0.27	83.32 ± 0.28	**77.17** ± 0.60	68.19 ± 0.56	84.44 ± 0.28	75.84 ± 0.90	67.44 ± 0.36	**84.73** ± 0.36	74.58 ± 0.48
mean over C^+	**66.94** ± 0.27	82.05 ± 0.25	**75.86** ± 0.55	66.65 ± 0.58	82.80 ± 0.22	75.05 ± 0.68	65.94 ± 0.31	**82.92** ± 0.25	73.99 ± 0.38

Table 8: Ablation study for the feature extractor: we provide the IoU, precision and recall values for the first experiment, where we incrementally extend a DeepLabV3+ by the novel class *human* on the Cityscapes dataset, using three different architectures for the feature extraction. For each feature extractor, we report the mean and standard deviation over five runs, respectively.

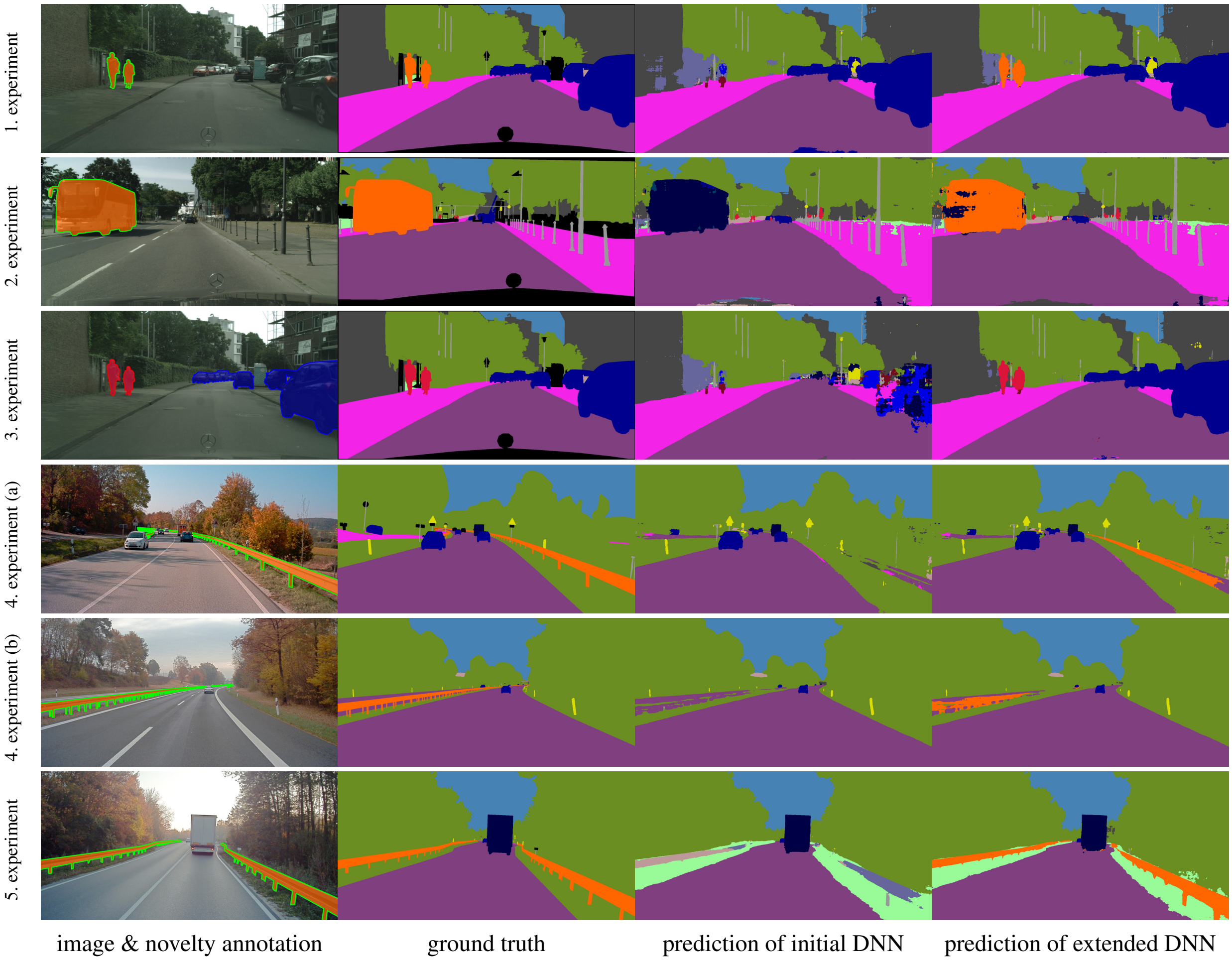
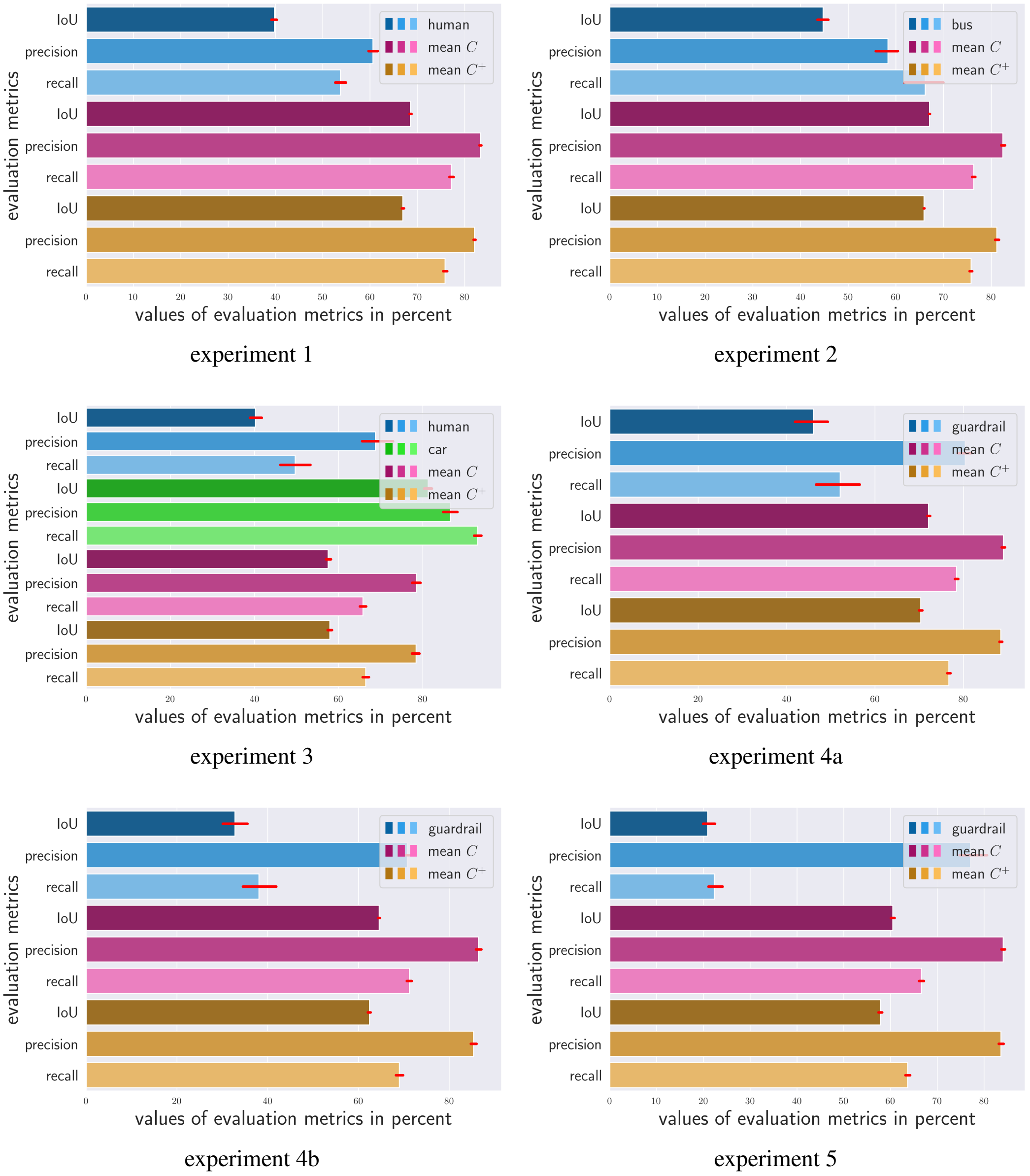


Figure 10: Example images from the validation data for all conducted experiments, respectively.

D RESULTS - VISUALIZATION

In Fig. 10 we provide an overall visualization of all conducted experiments. Our approach predicts the novel objects with adequate accuracy while the predictions of the initial and the extended DNNs remain similar on previously-known objects. Note that in the fifth experiment, the A2D2 ground truth consists of coarser classes than the segmentation DNN, which is trained on Cityscapes. Further, Fig. 11 illustrates the mean and standard deviation of the main evaluation metrics for each experiment, respectively. We observe, that the standard deviation values regarding the mean over $\mathcal{C}$ are at the maximum 1.20%, and besides that $\leq 1\%$. This is, our method is robust considering the initially known classes. In experiment 4 (a) and (b), we observe the highest standard deviation for the IoU values of the novel class with 4.80% and 3.48%, respectively, which is $< 2\%$ for all other experiments.



experiment 1

experiment 2

experiment 3

experiment 4a

experiment 4b

experiment 5

Figure 11: Bar plots showing the evaluation metrics averaged over five runs per experiment. The standard deviation is indicated by the red lines.