Towards Unsupervised Open World Semantic Segmentation

Abstract

For the semantic segmentation of images, state-of-the-art deep neural networks (DNNs) achieve high segmentation accuracy if that task is restricted to a closed set of classes. However, as of now DNNs have limited ability to operate in an open world, where they are tasked to identify pixels belonging to unknown objects and eventually to learn novel classes, incrementally. Humans have the capability to say: "I don't know what that is, but I've already seen something like that". Therefore, it is desirable to perform such an incremental learning task in an unsupervised fashion. We introduce a method where unknown objects are clustered based on visual similarity. Those clusters are utilized to define new classes and serve as training data for unsupervised incremental learning. More precisely, the connected components of a predicted semantic segmentation are assessed by a segmentation quality estimate. Connected components with a low estimated prediction quality are candidates for a subsequent clustering. Additionally, the component-wise quality assessment allows for obtaining predicted segmentation masks for the image regions potentially containing unknown objects. The respective pixels of such masks are pseudo-labeled and afterwards used for re-training the DNN, *i.e.,* without the use of ground truth generated by humans. In our experiments we demonstrate that, without access to ground truth and even with few data, a DNN's class space can be extended by a novel class, achieving considerable segmentation accuracy.

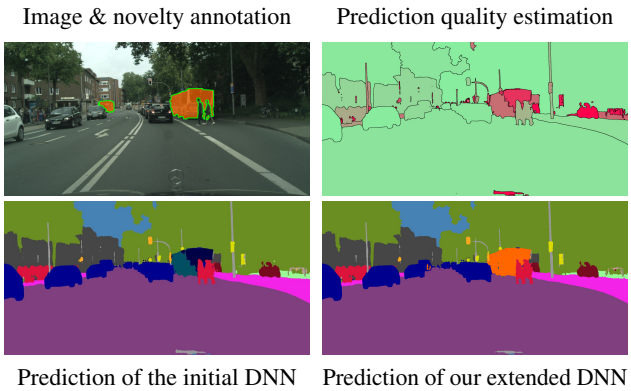


Figure 1: Comparison of the semantic segmentation predictions of an initial DNN (bottom left) whose semantic space does not include the category *bus* and a DNN which is incrementally extended by this novel class (bottom right, novel class in orange) for an image from the Cityscapes dataset. The novel class is highlighted in orange (top left). Further, the initial prediction exhibits a low prediction quality (top right) on pixels belonging to the novel objects, which is indicated by red color.

1 INTRODUCTION

Semantic segmentation is a computer vision task that terms the classification of image data on pixel level. State-of-the-art approaches are based on deep convolutional neural networks (DNNs) [Chen et al., 2018b, Wang et al., 2021, Zhao et al., 2017], benefiting from finely annotated datasets, *e.g.,* for automated driving [Cordts et al., 2016, Geyer et al., 2020, Neuhold et al., 2017, Yu et al., 2020]. However, DNNs for semantic segmentation are usually trained on a predefined, closed set of classes. This closed world setting assumes, that all classes present during testing were already included in the training set. In an open world setting, this assumption does not hold. In particular for safety-critical open-world applications like perception systems for automated driving, it is indispensable that neural networks recognize previously unseen objects instead of wrongly assigning them to *one-of-the-known* classes. In addition, they must constantly adapt to evolving environments.

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

Some terms often used interchangeably for anomaly are *outlier*, *out-of-distribution* (OoD) object and *novelty*. As there is no clear convention on how to distinguish these terms, we define them as subcategories of anomalies: outliers and OoD objects denote noise or samples drawn from another distribution than the model was trained on, respectively. In this work, we are seeking novelties, which we define as previously-unseen objects that constitute a new concept, *i.e.,* objects of the same category appear frequently. In automated driving, detecting and learning those novel classes becomes necessary, *e.g.,* due to new appearances like e-scooters or due to local specialities like boat trailers near the sea. The concept of detecting and learning novelties was first introduced in Bendale and Boult [2015] as *open world recognition*. Open world recognition for different computer vision tasks is an emerging research area [Bendale and Boult, 2015, Joseph et al., 2021, Cen et al., 2021, Shu et al., 2018], still only little explored for unsupervised methods [He and Zhu, 2021, Nakajima et al., 2019], yet.

We propose a new and modular procedure for learning new classes of novel objects without any handcrafted annotation:

1. Anomaly segmentation to detect suspicious objects,
2. clustering of potentially novel objects,
3. creation of so-called *pseudo labels*, and
4. incremental learning of novel classes.

In the following, we will outline each of these four steps in more detail.

For the first step, we post-process the predictions of an underlying semantic segmentation DNN via a *meta regressor*, that estimates the quality of the predicted segments (connected components of pixels in the segmentation mask), similar as proposed in Rottmann and Schubert [2019], Rottmann et al. [2020], Maag et al. [2020]. The segment-wise quality score is obtained on the basis of aggregated dispersion measures and geometrical information, *i.e.,* without requiring ground-truth. The predicted segmentation mask on anomalous objects is often split into several segments. To this end, we first aggregate neighboring segments, *i.e.,* segments that have at least one adjacent pixel each, with quality estimates below some threshold, into (potentially) anomalous objects.

For the second step, we adapt the idea introduced in Oberdiek et al. [2020] to gather segments with poor prediction quality and to cluster them into visually related neighborhoods. Therefore, all anomalies (of sufficient size) are cropped out in the RGB images and the resulting image patches are fed into a convolutional neural network (CNN), *e.g.,* for image classification. To obtain comparable information about the anomalies, we then extract the features provided by the penultimate layer of the CNN, *i.e.,* right before the final classification layer. By reducing the dimensionality of these features up to two, we enable the use of low-dimensional, unsupervised clustering techniques, such as Ester et al. [1996], MacQueen [1967].

As third, we obtain pseudo labels for novel classes in an automated manner: each (large / dense enough) cluster constitutes a novel category, and each pixel belonging to a clustered object is assigned to the appropriate (not necessarily named) class. More precisely, the prediction of the segmentation model is updated at those pixel positions to the next "free" label ID.

Finally, the segmentation network is incrementally extended by these novel classes (see Fig. 1 for an example). To this end, we apply established incremental learning methods [Hinton et al., 2015, Robins, 1995]. However, these are mainly examined for supervised learning tasks, while we do not include any hand-labeled new data. This last two steps were never done in literature so far.

To outline our contributions, we demonstrate in our experiments that our method is able to incrementally extend a neural network by novel classes without collecting or annotating novelties manually. To the best of our knowledge, we are the first to introduce an unsupervised approach for open world semantic segmentation with DNNs. Fine-tuning neural networks on automatically created pseudo-labels instead of human-made annotations is economically valuable. We observe in all experiments, that even a poor labeling quality is sufficient to learn novel classes, achieving IoU values around 40%. Further, the amount of new data was less than 100 images, respectively. Unsupervised open world semantic segmentation therefore is a powerful tool for open world applications, that provides an enormous potential for future improvement.

2 RELATED WORK

In this section, we first review anomaly detection methods and briefly go into class discovery approaches. Then we describe different strategies for class-incremental learning. Finally, we give an overview of existing work on open world computer vision tasks.

Novelty Detection. The detection of anomalous objects in general is a key task in many machine learning applications. Early works estimate the prediction uncertainty, *e.g.,* by uncertainty measures derived from the softmax probability [Hendrycks and Gimpel, 2017, Liang et al., 2018]. Uncertainty-based approaches can be further improved by integrating anomalous data into the training procedure [Devries and Taylor, 2018, Chan et al., 2021b]. Another line of works employs generative models such as autoencoders (AEs) or generative adversarial models (GANs) to reconstruct or synthesise images and measure the reconstruction quality. Various of those novelty detection methods are described in Vasilev et al. [2018], not only reconstruction-, but also density- or distance-based. A benchmark for anomaly

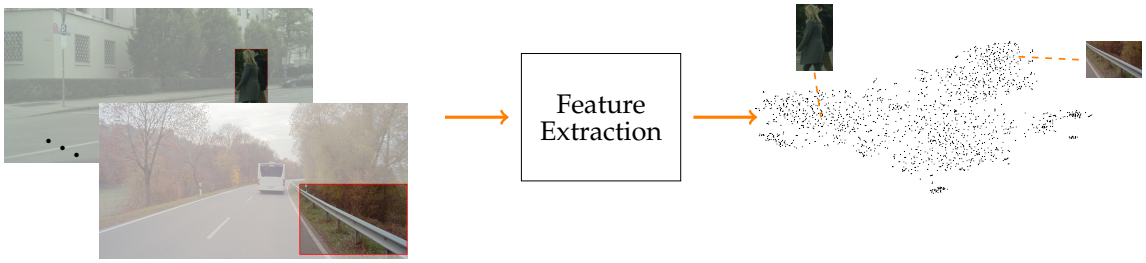


Figure 2: In novelty detection, to discover novel classes, anomalies belonging to the same class must be grouped together. Therefore, we use an image classification model to extract features of the detected image patches (outlined in red) and reduce their dimension up to two. By that, we obtain a two-dimensional feature space with visually related neighborhoods, were we employ clustering methods to discover novel classes.

segmentation, *i.e.,* anomaly detection methods for semantic segmentation, was recently published in Chan et al. [2021a], providing a cleaner comparison of proposed methods. Given a set of anomalies, the prevailing approach for class discovery is to form clusters based on some similarity measure or intrinsic features with traditional clustering methods. A detailed survey of image clustering has been published in Liu et al. [2021].

Class-Incremental Learning. Class-incremental learning refers to the extension of a neural network's semantic space by further, previously unknown, classes. This extension is achieved by fine-tuning a model on additional, usually human-annotated data [Jung et al., 2018, Li and Hoiem, 2018, Klingner et al., 2020, Michieli and Zanuttigh, 2019], whereas in this work we only provide pseudo labels for these new images. The primary issue to tackle when re-training a neural network is to mitigate the performance loss on previously learned classes, commonly known as catastrophic forgetting [McCloskey and Cohen, 1989]. To this end, we employ two different strategies: first, we penalize large variations of the softmax output (compared to the one of the original network) [Hinton et al., 2015], second we utilize a subset of the previously-seen training data [Robins, 1995].

The first strategy belongs to the category of regularization based approaches, or more specifically to knowledge distillation methods. These were originally developed to distill knowledge from sophisticated into simpler models [Hinton et al., 2015], *i.e.,* for model compression. Thereupon, distillation methods have been evolved for incremental learning in image classification [Li and Hoiem, 2018, Yao et al., 2019, Kim et al., 2019, Jung et al., 2018, Lee et al., 2019], some of which were later adapted to semantic segmentation [Klingner et al., 2020, Michieli and Zanuttigh, 2019, Tasar et al., 2019].

The second approach belongs to so-called rehearsal methods [Robins, 1995], were old training data is included in the retraining process [Rebuffi et al., 2017, Castro et al., 2018].

Open World. The open world setting was first introduced in Bendale and Boult [2015] for image classification. The authors formally define the solution of open world recognition problems as a tuple, consisting of a recognition function, a novelty detector, a labeling process and an incremental learning function. Ideally, these steps should be automated, however, most approaches presume a supervised setting, *i.e.,* , they require ground-truth for detected novelties. In summary, open world recognition covers the entire process from discovering up to learning novel classes.

A supervised solution for open world object detection is presented in Joseph et al. [2021], based on contrastive clustering, an unknown-aware proposal network and energy based unknown identification. A similar approach was proposed in Cen et al. [2021] for open world semantic segmentation, where novel classes are learned via few-shot learning. In He and Zhu [2021], an unsupervised method to obtain pseudo labels for image classification based on cluster assignments is introduced. There exists also some prior work for unsupervised open world semantic segmentation [Nakajima et al., 2019], however, the segmentation mask is obtained via agglomerative clustering of superpixels and there is no update of the neural network at all. While it is capable of creating ad hoc novel classes unsupervisedly on given images, it does not create a consistent semantic category over multiple images.

Our work introduces an open world semantic segmentation framework, where a neural network is incrementally extended by novel classes. These classes are discovered **and** labeled without any human effort. Therefore, our work goes beyond all existing approaches in this research area.

3 DISCOVERY OF UNKNOWN SEMANTIC CLASSES

Whether a class is novel or not depends on the neural network's underlying set of known classes $\mathcal{C} = \{1, \dots, C\}$. Let $f : \mathcal{X} \to (0,1)^{|\mathcal{H}| \times |\mathcal{W}| \times |\mathcal{C}|}$ be a semantic segmentation

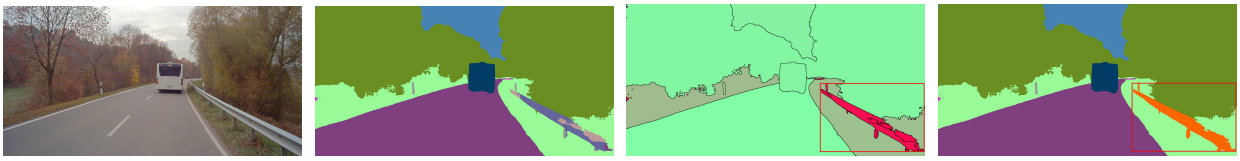


Image from A2D2 | Semantic segmentation prediction | Prediction quality estimation | Pseudo ground-truth

Figure 3: Novelty segmentation: Example for obtaining pseudo ground-truth with regard to some image patch (outlined in red). If segments inside the red box exhibit quality estimates below some predefined threshold, they are "re-labeled" in the segmentation mask.

DNN which is trained on the classes in $\mathcal{C}$, mapping an image $x \in \mathcal{X} \subseteq [0,1]^{|\mathcal{H}|\times|\mathcal{W}|\times 3}$ onto its softmax probabilities for each pixel $z \in \mathcal{H} \times \mathcal{W}$. Then, $f_{z,c}(x) \in (0,1)$ denotes the probability with which the model f assigns some pixel z to a class $c \in \mathcal{C}$. As decision rule, we apply the $\arg\max$ function, *i.e.,* we obtain the semantic segmentation mask $m(x) \in \mathcal{C}^{|\mathcal{H}|\times|\mathcal{W}|}$ with $m_z(x) = \arg\max_{c\in\mathcal{C}} f_{z,c}(x)$. In the following, we will estimate the prediction quality on a segment-level instead of pixel-wise, employing a meta regression approach that was first introduced in Rottmann et al. [2020]. On that account, we denote a segment, *i.e.,* a connected component of pixels that share the same class in $m(x)$, as $k \in \mathcal{K}(x)$.

Uncertainty Metrics and Prediction Quality Estimation. We consider novelties as *none-of-the-known* objects, *i.e.,* they differ semantically from the model's training data. Assuming that the segmentation DNN produces unstable predictions on these unexplored entities, various measurable phenomena occur. For instance, the model exhibits a high prediction uncertainty. This is quantified by dispersion measures as the softmax entropy, probability margin or variation ratio, which we compute pixel-wise via

$$E_z(f(x)) = -\frac{1}{\log(|\mathcal{C}|)} \sum_{c\in\mathcal{C}} f_{z,c}(x) \log(f_{z,c}(x)) \ , \quad (1)$$

$$D_z(f(x)) = 1 - \max_{c\in\mathcal{C}} f_{z,c}(x) + \max_{c\in\mathcal{C}\setminus\{m_z(x)\}} f_{z,c}(x) \ , \quad (2)$$

$$V_z(f(x)) = 1 - \max_{c\in\mathcal{C}} f_{z,c}(x) \ , \quad (3)$$

respectively. These are then averaged over the segments $k \in \mathcal{K}(x)$. Moreover, we examine some geometrical properties of the segments, such as their size, i.e., the number of pixels $|k|$ contained in k, their shape or their position in the image. For in-depth details on the constructed metrics, we refer to Rottmann et al. [2020]. By feeding these metrics into a meta regression model, we obtain prediction quality estimates for each segment $k \in \mathcal{K}(x)$, which we denote by $s(k) \in [0,1]$. These quality estimates approach the true segment-wise *Intersection over Union* (IoU) with reasonably high accuracy [Rottmann et al., 2020]. To fit the meta regressor, we compute the metrics plus the true IoU values of all segments included in the training data of the segmentation network. This meta model is then applied to unseen data, *i.e.,* data that was not included in the training of f, for the purpose of anomaly segmentation. Here, we consider a segment k to be anomalous, if its quality score is below some predefined threshold $\tau \in [0,1]$, *i.e.,* if $s(k) < \tau$. By that, we identify individual segments as unknown, however, the segmentation mask of unknown objects usually consists of several segments, *i.e.,* of different predicted classes. As we can uniquely assign each pixel z to a segment $k(z)$, we obtain a binary pixel-wise classification mask $a \in \{0,1\}^{|\mathcal{H}|\times|\mathcal{W}|}$ via

$$a_z = \mathbb{1}_{\{s(k(z))<\tau\}} \ \ \forall z \in \mathcal{H} \times \mathcal{W} \ , \quad (4)$$

where the class label $\mathbb{1}_{\{s(k(z))<\tau\}} = 1$ indicates anomalous pixels. Finally, the connected components in the anomaly mask a merge adjacent anomalous segments into anomalous objects. Under ideal conditions,

1. the semantic segmentation network performs perfectly on in-distribution data,
2. the meta model detects all (but only) unknowns, and
3. novel objects of different classes are separable.

Embedding and Clustering of Image Patches. Image clustering usually takes place in a lower dimensional latent space due to the curse of dimensionality. To this end, we feed image patches tailored to the anomalies into an image classification CNN, which is trained on the ImageNet dataset [Deng et al., 2009] with 1000 classes. Their feature representations are further compressed, resulting in a two-dimensional embedding space as illustrated in figure 2 (left). We apply two commonly used dimensionality reduction techniques. For complexity reasons, we compute the first 50 principal components [F.R.S.] before deploying the better performing *t-SNE* method [van der Maaten and Hinton, 2008] with Euclidean distance as similarity measure.

This procedure for image embedding is adopted from Oberdiek et al. [2020], where the authors evaluated sev-

eral feature extractors, distance metrics and feature dimensions. We employ the best performing setup in this quantitative analysis to obtain clusters of visually related image patches. Beyond that, we identify these clusters using the *DBSCAN* [Ester et al., 1996] algorithm. This clustering method requires two hyperparameters, namely the radius $\varepsilon \in \mathbb{R}$ that defines a neighborhood $B_\varepsilon(\cdot)$ and a threshold $N_{\min} \in \mathbb{N}$ regarding the number of data points within this ε-neighborhood. Let $\mathcal{E} = \{e_1, e_2, \dots\} \subset \mathbb{R}^2$ denote the set of the embedded features. Then, an embedding is considered a core point, if and only if it has at least $N_{\min}$ neighbors, *i.e.,*

$$\begin{aligned} e_i \in \mathcal{E} \text{ is core point } &\Leftrightarrow \\ |\{e_j \in \mathcal{E} : \ e_j \in B_\varepsilon(e_i)\}| &\geq N_{\min} \ . \end{aligned} \tag{5}$$

The algorithm further distinguishes between border points, *i.e.,* embeddings that are not core points themselves, but belong to a core point's neighborhood, and noise else. To mitigate the risk of failures, *i.e.,* objects from a different category in the novel clusters, we only consider the core points. We further reject embeddings representing image patches that are smaller than some predefined size. The cluster with the most remaining core points (or all clusters that involve "enough" core points) will be used to extend the segmentation network by new classes (figure 2, right).

Novelty Segmentation. Using pseudo labels instead of manually annotated targets is a cost-efficient (in the sense of human effort) method of training neural networks on unlabeled data. For the sake of simplicity we assume that exactly one cluster is returned by the aforementioned procedure. For some image $x \in \mathcal{X}$, we denoted the predicted segmentation mask by $m(x)$ and the respective segments by $\mathcal{K}(x)$. Let $\mathcal{K}^{\text{novel}}(x) \subseteq \mathcal{K}(x)$ describe the set of segments $k \in \mathcal{K}(x)$ that are also included in the considered cluster. If $\mathcal{K}^{\text{novel}}(x) \neq \emptyset$, *i.e.,* image x (probably) contains the novel class, we include the tuple $(x, \tilde{y}(x)) \in \mathcal{X} \times \{1, \dots, C+1\}^{|\mathcal{H}| \times |\mathcal{W}|}$ into the re-training data $\mathcal{D}^{C+1}$ for learning the novel class $C+1$. Here, $\tilde{y}(x)$ denotes the pseudo label, where

$$\tilde{y}_z(x) = \begin{cases} C+1 & \text{, if } k(z) \in \mathcal{K}^{\text{novel}}(x) \\ m_z(x) & \text{, otherwise} \end{cases} \ . \tag{6}$$

An example for acquiring pseudo ground-truth for one image is given in Fig. 3. In the following section we extend the segmentation DNN f by fine-tuning it on $\mathcal{D}^{C+1}$.

4 EXTENSION OF THE MODEL'S SEMANTIC SPACE

In this section we describe our approach to semantic incremental learning with the pseudo ground-truth acquired by novelty segmentation. Starting from our initial segmentation model f, we are seeking an extended model

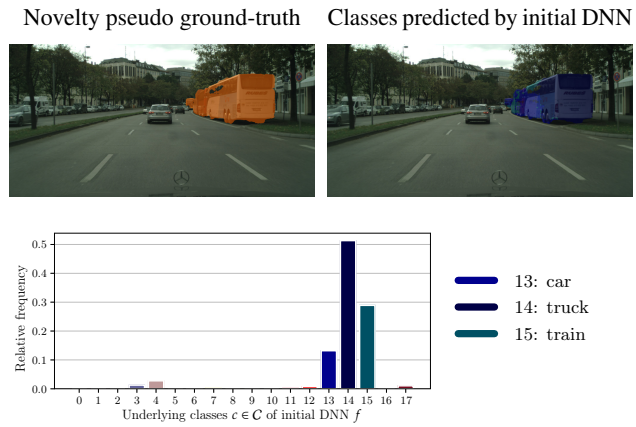


Figure 4: Histogram plot showing the relative frequencies of predicted classes for instances of the novel class, together with an exemplary image.

$g : \mathcal{X} \to (0,1)^{|\mathcal{H}| \times |\mathcal{W}| \times (C+1)}$ that retains the knowledge of f while additionally learning the novel class $C+1$. Denote the extended semantic space by $\mathcal{C}^+ = \mathcal{C} \cup \{C+1\}$. In more detail, we replace the ultimate layer of f and reinitialize only the affected weights to obtain the initial model g for re-training, *i.e.,* the model we train on the newly collected data $\mathcal{D}^{C+1}$. As loss function we apply a weighted cross entropy loss [Yi-de et al., 2004], denoted by $l_{\text{ce},\omega}$. The class-wise weights $\omega_c \in (0,1]$, $c \in \mathcal{C}^+$, are recalculated for each batch based on the inverse class frequency to alleviate class imbalances.

To mitigate the problem of catastrophic forgetting [McCloskey and Cohen, 1989], we pursue two strategies, namely knowledge distillation [Hinton et al., 2015] and rehearsal [Robins, 1995].

Knowledge distillation in class-incremental learning aims at minimizing variations of the softmax output restricted to only the old classes $c \in \mathcal{C}$. This is realized by an additional distillation loss function [Michieli and Zanuttigh, 2021] l_{d}, where

$$\begin{aligned} & l_{\text{d}}(g(x), f(x)) \\ & := -\frac{1}{|\mathcal{H}||\mathcal{W}|} \sum_{z \in \mathcal{H} \times \mathcal{W}} \sum_{c \in \mathcal{C}} f_{z,c}(x) \log(g_{z,c}(x)) \ . \end{aligned} \tag{7}$$

Overall, we aim at minimizing the objective

$$\begin{aligned} L := \ & \lambda \, \mathbb{E}[l_{\text{ce},\omega}(g(x), \tilde{y}(x))] \\ & + (1-\lambda) \, \mathbb{E}[l_{\text{d}}(g(x), f(x))], \quad \lambda \in [0,1] \end{aligned} \tag{8}$$

with λ regulating the impact of the distillation loss.

Rehearsal methods propose to replay (some of) the data $\mathcal{D}^{\text{train}} \subset \mathcal{X} \times \mathcal{C}^{|\mathcal{H}| \times |\mathcal{W}|}$ seen during the training of the initial model f. We select a subset $\mathcal{D}^{\text{known}} \subseteq \mathcal{D}^{\text{train}}$ that contains as much data as $\mathcal{D}^{C+1}$. This subset is chosen largely at random, but in such a way that it involves classes, that are

1. not or rarely present in $\mathcal{D}^{C+1}$ (class frequency), or
2. similar or related to the novel class.

As there is no measure for the second case, we identify those classes by considering the frequency, with which a class is predicted by f on pixels assigned to the novel class. This is, for all data $(x, \tilde{y}(x)) \in \mathcal{D}^{C+1}$ and classes $c \in \mathcal{C}$, we sum up the number of pixels $z \in \mathcal{H} \times \mathcal{W}$ where $\tilde{y}_z(x) = C+1 \wedge m_z(x) = c$. An example is given in Fig. 4, where the classes *truck*, *train* and *car* are the most frequently predicted classes for instances of the novel class *bus*.

5 EXPERIMENTAL SETUP & EVALUATION

We evaluate our approach on the task of detecting and incrementally learning novel classes in traffic scenes, for which there exist large datasets such as Cityscapes [Cordts et al., 2016] and A2D2 [Geyer et al., 2020]. To this end, all evaluated segmentation DNN's were trained on a training split and only on a subset of all available classes. We then perform our experiments on a test split of the same dataset on which the DNN was trained in order to extent it by exactly one or even multiple novel classes. We measure the performance of the extended models computing the evaluation metrics *intersection over union* (IoU), *precision* and *recall* for a validation set.

Experimental Setup. As segmentation DNNs we employ the DeepLabV3+ [Chen et al., 2018b] and the PSPNet [Zhao et al., 2017] The first is trained for different subsets of known classes on the Cityscapes dataset. Moreover, both models are pretrained on Cityscapes with all 19 classes and then fine-tuned on the A2D2 dataset. Here we use a label mapping between both datasets through which 14 classes remain.

We perform five experiments: For the first three experiments, a DeepLabv3+ with a WideResNet38 backbone is trained on the Cityscapes dataset, where 1) the classes *person* & *rider*, 2) the class *bus* and 3) the classes *person* & *rider*, *bus* and *car* are excluded. In a fourth experiment, a DeepLabv3+ as well as a PSPNet based on a ResNet50 backbone are fine-tuned on the A2D2 dataset, for which we specified subsets for training, testing and validation, including 2975, 1355 and 451 annotated images, respectively. Then, we also apply our method to the A2D2 dataset without prior fine-tuning, *i.e.,* under a domain shift, employing a DeepLabV3+ trained on Cityscapes. Our experiments follow a hierarchical structure with increasing complexity:

1. Construction of a "well" separated category (*human*),
2. Construction of a category in the midst of known similar categories (*bus*),
3. Construction of multiple novel categories (*human*, *bus* and *car*),
4. Construction of a new category under domain shift with ground truth for known classes (*guardrail*, with fine-tuning),
5. Construction of a new category under domain shift without ground truth (*guardrail*, without fine-tuning).

Each of those initial DNNs is employed to predict the semantic segmentation masks for the images contained in the respective test set. For the segment-wise prediction quality estimation introduced in Sec. 3, we apply a gradient boosting model to obtain the quality scores $s(k) \in [0, 1]$ for each segment $k \in \mathcal{K}(x)$ and image x in the test set. The threshold in Eq. (4) is set to $\tau = 0.5$, *i.e.,* a segment $k \in \mathcal{K}$ is considered as anomalous, if $s(k) < 0.5$. To extract features of the detected anomalies, we employ a DenseNet201 [Huang et al., 2017], trained on the ImageNet dataset [Deng et al., 2009] with 1000 classes. Note that the DBSCAN hyperparameters have to be selected dependent on the density of the desired clusters.

For the class-incremental extension of an initial DNN f, we replace its final layer to obtain a larger DNN g (see Sec. 4). Only the decoder of this model is trained for 70 epochs on the newly collected data $\mathcal{D}^{C+1}$ together with the replayed data $\mathcal{D}^{\text{known}}$. We use random crops of size 1000×1000 pixels, the Adam optimizer with a learning rate of $5 \cdot 10^{-5}$ and a weight decay of 10^{-4}. Further, the learning rate is adjusted after every iteration via a polynomial learning rate policy [Chen et al., 2018a]. The distillation loss and the cross-entropy loss are weighted equally in the overall loss function defined in Eq. (8), *i.e.,* $\lambda = 0.5$ (analogously to Michieli and Zanuttigh [2019]).

As the five experiments struggle with different issues, the experimental setup slightly differs. For the first case, we construct the novel category *human*, which is "well" separable from all known classes, to enhance the purity of the "human cluster" and to simplify the learning of novel objects. However, we observe that the DNN tends to "overlook" many humans, *i.e.,* they are assigned to the class predicted in the background, *e.g.,* to the *road* class. As a consequence, the segment-wise anomaly detection fails to detect such persons, which is why these will be assigned to other classes in our acquired pseudo ground-truth. To not distract the extended segmentation network, we modify the pseudo labels by ignoring all known classes $c \in \mathcal{C}$ during the incremental training procedure. The *bus* class added in the second experiment is closely related to other classes in the vehicle category, such as *truck*, *train* and *car*, which complicates the construction of pure clusters. We mitigate the impact of objects from similar classes by discarding all objects from the cluster that consist of only one segment in the predicted segmentation. Experiment three extends the previous ones by facing multiple unknown classes, namely *human*, *bus* and *car*. The last two experiments deal with an addi-

Model	$\text{mIoU}_{\mathcal{C}}$	$\text{IoU}_{\text{novelty}}$	$\text{mIoU}_{\mathcal{C}^+}$
1. experiment: Cityscapes, human	DeepLabV3+		
initial DNN	68.63	00.00	64.82
extended DNN (ours)	68.24	41.42	66.52
extended DNN (baseline)	69.43	59.33	68.87
oracle	71.05	72.85	71.15
2. experiment: Cityscapes, bus	DeepLabV3+		
initial DNN	66.94	00.00	63.42
extended DNN (ours)	67.05	41.85	65.72
extended DNN (baseline)	66.74	41.40	65.41
oracle	69.48	76.66	69.86
3. experiment: Cityscapes, multi	DeepLabV3+		
initial DNN	56.99	00.00 & 00.00	50.29
extended DNN (ours)	58.12	41.14 & 82.05	58.53
oracle	77.28	81.90 & 94.94	78.59
4. experiment (a): A2D2, guardrail	DeepLabV3+ (fine-tuned)		
initial DNN	75.77	00.00	70.72
extended DNN (ours)	71.73	46.31	70.03
oracle	75.23	74.58	75.19
4. experiment (b): A2D2, guardrail	PSPNet (fine-tuned)		
initial DNN	68.77	00.00	64.19
extended DNN (ours)	65.64	18.71	62.51
oracle	67.71	69.08	67.80
5. experiment: A2D2, guardrail	DeepLabV3+ (not fine-tuned)		
initial DNN	59.38	00.00	55.42
extended DNN (ours)	60.69	38.20	59.19

Table 1: Comparing overview of all evaluated models, where the results for our extended DNNs are highlighted in gray. As performance metrics, we provide the mean IoU over the old and new classes, denoted by $\text{mIoU}_{\mathcal{C}}$ and $\text{mIoU}_{\mathcal{C}^+}$, respectively, and the IoU value of the novel class(es), $\text{IoU}_{\text{novelty}}$.

	IoU	precision	recall	IoU	precision	recall
1. experiment:	DeepLabV3+					
Cityscapes, human	initial			extended		
human	00.00	00.00	00.00	41.42	59.73	57.48
mean over $\mathcal{C}$	68.63	79.79	80.94	68.24	84.28	76.08
mean over $\mathcal{C}^+$	64.82	75.36	76.44	66.75	82.91	75.05
2. experiment:	DeepLabV3+					
Cityscapes, bus	initial			extended		
bus	00.00	00.00	00.00	41.85	53.99	65.06
mean over $\mathcal{C}$	66.94	79.32	79.55	67.05	82.03	76.50
mean over $\mathcal{C}^+$	63.42	75.15	75.36	65.72	80.56	75.90
3. experiment:	DeepLabV3+					
Cityscapes, multi	initial			extended		
human	00.00	00.00	00.00	41.14	58.66	57.93
car	00.00	00.00	00.00	82.05	88.94	91.37
mean over $\mathcal{C}$	56.99	65.75	80.88	58.12	80.69	65.51
mean over $\mathcal{C}^+$	50.29	58.01	71.37	58.53	79.88	66.59
4. experiment (a):	DeepLabV3+					
A2D2, guardrail	initial			extended		
guardrail	00.00	00.00	00.00	46.31	81.61	51.70
mean over $\mathcal{C}$	75.77	87.86	83.47	71.73	89.10	78.01
mean over $\mathcal{C}^+$	70.72	82.00	77.90	70.03	88.60	76.26
4. experiment (b):	PSPNet					
A2D2, guardrail	initial			extended		
guardrail	00.00	00.00	00.00	18.71	66.37	20.67
mean over $\mathcal{C}$	68.77	84.57	76.79	65.64	85.97	72.50
mean over $\mathcal{C}^+$	64.19	78.93	71.67	62.51	84.66	69.05
5. experiment:	DeepLabV3+					
A2D2, guardrail	initial			extended		
guardrail	00.00	00.00	00.00	38.20	58.30	52.57
mean over $\mathcal{C}$	59.38	79.50	68.14	60.69	83.91	66.71
mean over $\mathcal{C}^+$	55.42	74.20	63.60	59.19	82.20	65.77

Table 2: Direct comparison of the initial and the extended DNNs for all conducted experiments. We report the IoU, precision and recall values for the novel class (highlighted with gray rows), respectively, as well as averaged over the previously-known and the extended class spaces $\mathcal{C}$ and $\mathcal{C}^+$.

tional domain shift from urban street scenes in Cityscapes to countryside and highway scenes in A2D2. To bridge this gap, we fine-tune the initial DNN on our A2D2 training set, which, however, requires A2D2 ground-truth for the known classes. Without fine-tuning, the prediction quality and thereby the quality of our pseudo ground-truth suffers. On that account, we discard images that are generally rated as badly predicted, *i.e.,* where the relative amount of pixels with a low quality estimate exceeds $1/3$ of the image in total. Moreover, we renounce the replay of previously-seen data, since this prevents the DNN from adapting to the new domain.

Evaluation of Results. We provide a qualitative comparison of different models for all four experiments in Tab. 1, reporting the mean IoU over the known classes and over the extended class set, denoted as $\text{mIoU}_{\mathcal{C}}$ and $\text{mIoU}_{\mathcal{C}^+}$, respectively, as well as the IoU value of the novel class ($\text{IoU}_{\text{novelty}}$). The models considered in this comparison are the initial and the extended DNN, where the class space is extended via our method. For the first and second experiment we further compare our approach with a baseline, where a DNN is extended using a self-training approach. That is, we employ a so-called teacher network, which is already trained on the extended semantic space $\mathcal{C}^+$, to produce pseudo labels for some student network. Thereby, we obtain a high quality pseudo ground-truth. Apart from this, the baseline DNN is extended analogously to ours. In addition, for the first four experiments we provide results of an *oracle*, *i.e.,* a DNN, that is initially trained on the extended class set $\mathcal{C}^+$ and only with human-annotated ground-truth In the fifth experiment, we extend the initial DNN by a novel class derived from a different dataset. To some extent, the oracle from experiment four (a) can serve as a coarse reference for experiment five. In Tab. 2 we give a more detailed overview about all experiments, reporting not only the IoU, but also the precision and recall values of the novel class as well as averaged over $\mathcal{C}$ and $\mathcal{C}^+$. Note that the fourth experiment is evaluated twice, once for (a) the DeepLabV3+ and once for (b) the PSPNet. For class-wise evaluation results and visualizations, we refer to Appendix A.

In general, we observe that our approach succeeds in incrementally extending a DNN by a novel class, while the performance on previously-known classes remains stable. On Cityscapes, we achieve IoU values for the novel classes human and bus of $\text{IoU}_{\text{human}} = 41.42\%$ and $\text{IoU}_{\text{bus}} = 41.85\%$, respectively. While these IoU values are a considerable achievement for a method working without ground truth, the distinct gaps to the oracle's IoU val-

ues still leave room for further improvement. Compared to the baseline DNN, we do not achieve competitive performance in the first experiment, while in the second experiment, our approach actually performs slightly better. This is explained by the fact, that the pseudo ground-truth for the *human* class incorporates much more noise than that for the *bus* class. In the fourth experiment we mitigate the domain shift from Cityscapes to A2D2 by prior fine-tuning of the networks, using A2D2 ground-truth. By that, we obtain IoU values of $\mathrm{IoU}_{\mathrm{guardrail}} = 46.31\%$ for the DeepLabV3+ and $\mathrm{IoU}_{\mathrm{guardrail}} = 18.71\%$ for the PSPNet. We conclude, that our approach achieves better results for models which are initially better-performing. Without fine-tuning the DeepLabV3+ on A2D2, we obtain an $\mathrm{IoU}_{\mathrm{guardrail}} = 38.20\%$, while the mean IoU over the previously-known classes $\mathcal{C}$ slightly increases from 59.38% to 60.69%.

6 CONCLUSION & OUTLOOK

In this work, we have introduced a new and modular procedure for the class-incremental extension of a semantic segmentation network, were novel classes are detected, annotated and learned in an unsupervised fashion. While there already exists an unsupervised open world approach for semantic segmentation [Nakajima et al., 2019], we are the first in this field to extend a neural network's semantic space by robust novel classes. We performed five hierarchically structured experiments with an increasing level of difficulty. We demonstrated that our approach can deal with novelties that are either "well" separated or related to known categories, and that it is even applicable when the test data is sampled from a slightly different distribution than the DNN was trained on. Moreover, we applied two different models in the fourth experiment, where the initial DeepLabV3+ already outperformed the initial PSPNet. This performance gap is also reflected in the model's ability to learn the novel class, thus we conclude that our method benefits significantly from high performance networks.

For future work, we plan to improve the extension of a neural network by multiple classes at once. On that account, suitable datasets are in demand. Two datasets for the task of anomaly segmentation were recently published in Chan et al. [2021a], however, these show a wide variety of anomalous objects. To advance the research in class-incremental learning, it requires datasets where novel objects, *i.e.,* objects that do not appear in the training data, appear frequently in the test data.

We are currently working on a synthetic dataset tailored to our approach. This data is generated using the CARLA 0.9.12 simulator Dosovitskiy et al. [2017], similar as extensively described in Kowol et al. [2022]. The data include annotated street scene images, generated on the same maps for training and testing. Since we aim at detecting novel classes in the test data, these images are enriched by several **never-seen** object classes, *e.g., deer*, *construction vehicle* or *portable toilet* (examples provided in Appendix B).

Besides, we plan to adapt our approach to video instead of image data, where anomaly detection includes anomaly tracking over multiple frames.

Our source code is publicly available under `https://tba`.

7 LIMITATIONS & NEGATIVE IMPACT

With the procedure presented in this work, we are taking a first step towards a new machine learning problem. This first step is highly experimental and our method has not the technology readiness level to be applied to real-world problems in a fully automated fashion. Especially from the safety point of view, a neural network should not be modified without any supervision, since we can not guarantee to avoid significant performance drops.

References

Abhijit Bendale and Terrance E. Boult. Towards open world recognition. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1893–1902, 2015.

F. M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil Mata, C. Schmid, and Alahari Karteek. End-to-end incremental learning. *ArXiv*, abs/1807.09536, 2018.

Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15333–15342, October 2021.

Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021a.

Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5128–5137, October 2021b.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. Deeplab: Semantic image segmentation with deep convolutional nets,

atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40: 834–848, 2018a.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *ArXiv*, abs/1802.02611, 2018b.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

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

Terrance Devries and Graham W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *ArXiv*, abs/1802.04865, 2018.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 1*, 2:559–572.

Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Ganapati Mirashi, Chiragkumar Savani, M. Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2d2: Audi autonomous driving dataset. *ArXiv*, abs/2004.06320, 2020.

Jiangpeng He and Feng Zhu. Unsupervised continual learning via pseudo labels. *ArXiv*, abs/2104.07164, 2021.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5830–5840, June 2021.

Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetful learning for domain expansion in deep neural networks. In *AAAI*, 2018.

Dahyun Kim, Jihwan Bae, Yeonsik Jo, and Jonghyun Choi. Incremental learning with maximum entropy regularization: Rethinking forgetting and intransigence. *ArXiv*, abs/1902.00829, 2019.

Marvin Klingner, Andreas Bär, Philipp Donn, and Tim Fingscheidt. Class-incremental learning for semantic segmentation re-using neither old data nor old labels. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8, 2020.

Kamil Kowol, Stefan Bracke, and Hanno Gottschalk. A-eye: Driving with the eyes of ai for corner case generation, 2022.

Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 312–321, 2019.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2935–2947, 2018.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Jiaxin Liu, Dongwei Wang, Siquan Yu, Xueliang Li, Zhi Han, and Yandong Tang. A survey of image clustering: Taxonomy and recent methods. In *2021 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pages 375–380, 2021.

Kira Maag, Matthias Rottmann, and Hanno Gottschalk. Time-dynamic estimates of the reliability of deep semantic segmentation networks. *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 502–509, 2020.

J. MacQueen. Some methods for classification and analysis of multivariate observations. 1967.

M. McCloskey and N. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

Umberto Michieli and Pietro Zanuttigh. Incremental learning techniques for semantic segmentation. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3205–3212, 2019.

Umberto Michieli and Pietro Zanuttigh. Knowledge distillation for incremental learning in semantic segmentation. *Comput. Vis. Image Underst.*, 205:103167, 2021.

Yoshikatsu Nakajima, Byeongkeun Kang, H. Saito, and Kris Kitani. Incremental class discovery for semantic segmentation with rgbd sensing. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 972–981, 2019.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. pages 5000–5009, 2017.

Philipp Oberdiek, Matthias Rottmann, and Gernot A. Fink. Detection and retrieval of out-of-distribution objects in semantic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1331–1340, 2020.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, G. Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542, 2017.

Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.

Matthias Rottmann and Marius Schubert. Uncertainty measures and prediction quality rating for the semantic segmentation of nested multi resolution street scene images. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1361–1369, 2019.

Matthias Rottmann, Pascal Colling, Thomas-Paul Hack, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9, 2020.

Lei Shu, Hu Xu, and Bing Liu. Unseen class discovery in open-world classification. *ArXiv*, abs/1801.05609, 2018.

O. Tasar, Y. Tarabalka, and P. Alliez. Incremental learning for semantic segmentation of large-scale remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12:3524–3537, 2019.

Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Aleksei Vasilev, Vladimir Golkov, Ilona Lipp, Eleonora Sgarlata, Valentina Tomassini, Derek K. Jones, and Daniel Cremers. q-space novelty detection with variational autoencoders. *ArXiv*, abs/1806.02997, 2018.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, D. Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3349–3364, 2021.

X. Yao, Tianchi Huang, Chenglei Wu, Ruixiao Zhang, and L. Sun. Adversarial feature alignment: Avoid catastrophic forgetting in incremental task lifelong learning. *Neural Computation*, 31:2266–2291, 2019.

Ma Yi-de, Liu Qing, and Qian Zhi-bai. Automated image segmentation using improved pcnn model based on cross-entropy. In *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, pages 743–746, 2004. doi: 10.1109/ISIMP.2004.1434171.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2633–2642, 2020.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2017.

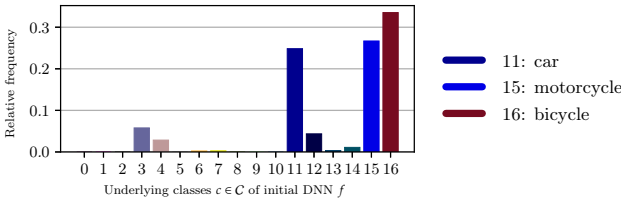


Figure 5: Histogram plot showing the relative frequencies of predicted classes for instances of the novel class *human*.

A EVALUATED MODELS

We performed five 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.

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. 5). 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. 6, 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. 3. Besides the novel class, which achieves an IoU value of about 40% with nearly 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 4.49%, recall decreases by 4.86%. 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.

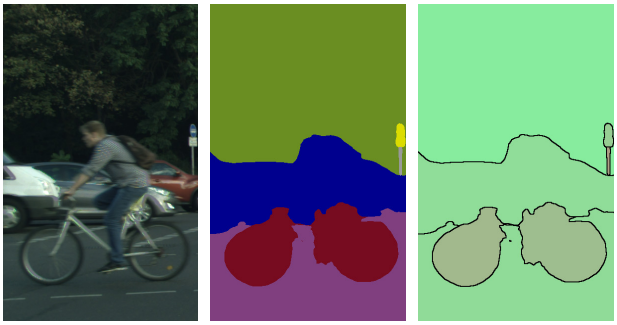


Image patch | Predicted segmentation | Quality estimation

Figure 6: 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. 7. We observe 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. 4 and present a comparison of different models for one exemplary street scene in Fig. 8. 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 base-

Image & annotation | Initial DNN | Extended DNN | Baseline | Oracle

Figure 7: 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	97.46	98.68	98.75
sidewalk	80.63	89.39	89.16	80.78	89.31	89.43
building	88.91	92.80	95.50	89.30	93.11	95.62
wall	47.24	74.57	56.32	47.48	78.33	54.67
fence	51.03	66.76	68.41	49.31	69.48	62.95
pole	52.90	72.68	66.02	53.25	73.74	65.70
traffic light	55.44	75.04	67.98	55.28	76.02	66.96
traffic sign	66.66	86.22	74.61	65.72	88.99	71.54
vegetation	89.95	93.60	95.85	90.17	94.21	95.46
terrain	56.29	77.66	67.17	54.53	75.66	66.13
sky	93.76	96.38	97.18	93.47	95.69	97.57
human	00.00	00.00	00.00	41.42	59.73	57.48
car	90.61	92.97	97.27	91.21	95.26	95.54
truck	69.66	80.23	84.09	69.30	84.88	79.06
bus	76.90	88.59	85.35	72.52	87.26	81.11
train	70.35	83.33	81.87	62.06	91.68	65.76
motorcycle	24.45	28.57	62.92	30.45	64.38	36.61
bicycle	54.57	59.30	87.24	57.72	76.01	70.57
mean over $\mathcal{C}$	68.63	79.79	80.94	68.24	84.28	76.08
mean over $\mathcal{C}^+$	64.82	75.36	76.44	66.75	82.91	75.05

Table 3: 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	97.51	98.85	98.63
sidewalk	81.60	89.65	90.09	81.40	89.70	89.79
building	90.19	94.50	95.19	90.12	94.35	95.26
wall	48.77	78.07	56.51	44.67	78.75	50.80
fence	53.86	70.97	69.08	52.78	69.68	68.50
pole	55.03	75.71	66.83	54.34	77.36	64.61
traffic light	55.87	77.29	66.84	55.46	78.44	65.44
traffic sign	68.21	87.02	75.94	67.64	88.81	73.94
vegetation	90.35	93.98	95.91	90.21	93.83	95.90
terrain	54.03	79.90	62.53	51.60	71.71	64.79
sky	93.64	96.14	97.30	93.56	96.43	96.91
person	71.65	83.27	83.70	71.11	82.19	84.05
rider	48.77	68.86	62.58	46.60	71.87	57.00
car	91.90	94.65	96.94	91.67	94.91	96.40
truck	47.51	51.19	86.87	53.08	71.51	67.32
bus	00.00	00.00	00.00	41.85	53.99	65.06
train	43.57	48.58	80.88	55.14	71.35	70.83
motorcycle	44.35	61.76	61.13	42.37	70.25	51.63
bicycle	68.00	77.42	84.82	67.61	76.62	85.19
mean over $\mathcal{C}$	66.94	79.32	79.55	67.05	82.03	76.50
mean over $\mathcal{C}^+$	63.42	75.15	75.36	65.72	80.56	75.90

Table 4: 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.

line 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.

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%.

Image & annotation | Initial DNN | Extended DNN | Baseline | Oracle

Figure 8: 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	96.67	98.25	98.37
sidewalk	77.23	83.84	90.74	76.68	84.89	88.80
building	87.21	91.05	95.39	87.55	92.72	94.01
wall	45.86	68.38	58.20	38.88	78.84	43.41
fence	47.86	59.63	70.79	39.80	72.20	47.00
pole	51.63	69.15	67.09	48.53	74.74	58.05
traffic light	55.61	77.70	66.17	46.00	69.61	57.56
traffic sign	64.84	80.37	77.04	57.46	88.24	62.22
vegetation	88.26	91.27	96.40	88.34	91.37	96.38
terrain	53.22	72.42	66.74	45.01	78.12	51.51
sky	93.58	96.11	97.27	92.52	96.20	96.03
human	00.00	00.00	00.00	41.14	58.66	57.93
car	00.00	00.00	00.00	82.05	88.94	91.37
truck	9.31	9.41	89.35	40.66	64.32	52.50
train	41.70	45.05	84.87	52.40	75.20	63.34
motorcycle	4.03	4.12	66.09	11.51	73.19	12.01
bicycle	39.13	41.30	88.15	49.83	72.51	61.44
mean over $\mathcal{C}$	56.99	65.75	80.88	58.12	80.69	65.51
mean over $\mathcal{C}^+$	50.29	58.01	71.37	58.53	79.88	66.59

Table 5: 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.

Detailed results are provided in Tab. 5.

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,

4. experiment (a) A2D2, guardrail	DeepLabV3+					
	initial			extended		
Class	IoU	precision	recall	IoU	precision	recall
road	95.59	97.21	98.29	95.83	97.85	97.89
sidewalk	72.01	86.73	80.92	71.84	85.62	81.70
building	87.82	93.58	93.44	85.22	93.76	90.34
fence	59.35	81.59	68.53	56.61	76.74	68.34
pole	56.13	76.39	67.91	54.12	78.61	63.47
traffic light	68.41	85.10	77.72	64.84	85.33	72.97
traffic sign	76.34	86.78	86.38	74.37	90.71	80.51
vegetation	91.61	94.01	97.29	91.90	94.45	97.15
sky	97.96	98.72	99.22	97.87	98.63	99.22
person	67.60	79.28	82.11	63.73	86.91	70.49
car	93.19	96.73	96.22	92.34	96.20	95.84
truck	84.99	88.51	95.53	81.50	85.28	94.84
motorcycle	48.68	84.71	53.37	23.51	92.29	23.98
bicycle	61.08	80.65	71.57	50.48	85.00	55.42
guardrail	00.00	00.00	00.00	46.31	81.61	51.70
mean over $\mathcal{C}$	75.77	87.86	83.47	71.73	89.10	78.01
mean over $\mathcal{C}^+$	70.72	82.00	77.90	70.03	88.60	76.26

Table 6: 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.

low frequency in the re-training data, which involves only 30 pseudo-labeled and 30 replayed images. Further details are provided in Tab. 6.

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*,

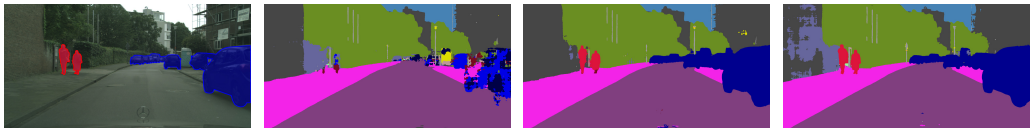


Image & annotation | Initial DNN | Extended DNN | Oracle

Figure 9: 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)	PSPNet					
A2D2, guardrail	initial			extended		
Class	IoU	precision	recall	IoU	precision	recall
road	95.18	97.10	97.96	95.14	97.08	97.94
sidewalk	66.15	83.68	75.94	62.13	84.04	70.45
building	84.32	92.46	90.54	82.56	94.00	87.15
fence	54.48	76.84	65.18	52.87	75.93	63.52
pole	44.60	63.94	59.59	43.33	63.02	58.10
traffic light	58.94	81.14	68.30	56.07	82.39	63.70
traffic sign	71.30	87.71	79.22	70.19	87.85	77.74
vegetation	90.68	93.12	97.18	89.87	91.99	97.50
sky	97.57	98.44	99.10	97.41	98.38	99.00
person	59.17	82.53	67.64	50.87	82.47	57.03
car	89.39	94.36	94.44	87.84	94.69	92.39
truck	77.83	84.05	91.31	74.64	83.73	87.31
motorcycle	19.73	76.72	20.99	07.76	88.79	07.84
bicycle	53.49	71.82	67.70	48.33	79.23	55.34
guardrail	00.00	00.00	00.00	18.71	66.37	20.67
mean over $\mathcal{C}$	68.77	84.57	76.79	65.64	85.97	72.50
mean over $\mathcal{C}^+$	64.19	78.93	71.67	62.51	80.24	69.05

Table 7: 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	DeepLabV3+					
A2D2, guardrail	initial			extended		
Class	IoU	precision	recall	IoU	precision	recall
road	89.88	92.18	97.30	93.66	95.24	98.26
sidewalk	47.91	76.22	56.33	50.84	82.17	57.14
building	70.94	86.88	79.45	67.38	89.21	73.36
fence	26.08	35.30	49.94	27.52	44.52	41.87
pole	42.59	59.24	60.25	39.69	58.41	55.32
traffic light	47.59	85.85	51.64	44.87	94.30	46.12
traffic sign	54.89	82.49	62.13	54.82	88.13	59.19
vegetation	69.15	96.68	70.83	74.71	94.09	78.39
sky	94.96	98.25	96.59	96.16	96.79	99.33
person	59.77	71.00	79.08	60.88	85.72	67.75
car	90.47	95.72	94.28	90.26	95.07	94.69
truck	62.64	83.61	71.40	67.90	92.92	71.61
motorcycle	28.39	70.82	32.15	35.49	78.73	39.26
bicycle	46.04	78.74	52.57	45.54	79.37	51.65
guardrail	00.00	00.00	00.00	38.20	58.30	52.57
mean over $\mathcal{C}$	59.38	79.50	68.14	60.69	83.91	66.71
mean over $\mathcal{C}^+$	55.42	74.20	63.60	59.19	82.20	65.77

Table 8: 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.

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. 7.

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:

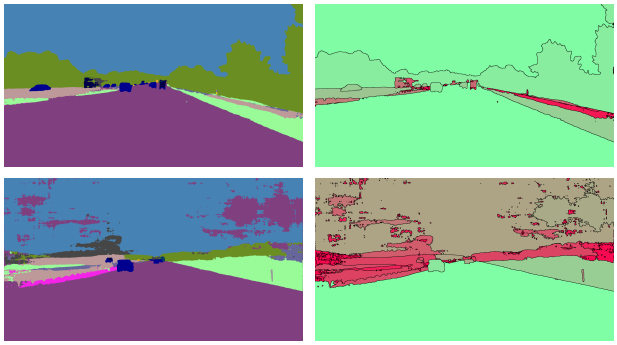
1. mean quality score (averaged over pixels) less than 0.7
2. more than 1/3 of all pixels with quality estimate less than 0.9.

If at least one criterion holds, we reject the image, as illustrated in the bottom row of Fig. 10.

Applying our method, we obtain 29 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 $\mathcal{D}^{C+1}$.

Class-wise evaluation results are reported in Tab. 8. Despite the domain shift, we still achieve an IoU of 38.20% for the novel class, which is "only" 8.11% less than the value obtained with prior fine-tuning. However, this DNN still outperforms the PSPNet from the previous experiment, although no A2D2 ground truth is involved at all. 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.

A visual comparison of the experiments 4(a), 4(b) and 5 is provided in Fig. 13. 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



Predicted segmentation Quality estimation

Figure 10: 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.

Figure 11: Two examples from our Carla test dataset including the novel class *deer*.

the fourth 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.

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. 11. 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 by anything different than the novel objects.

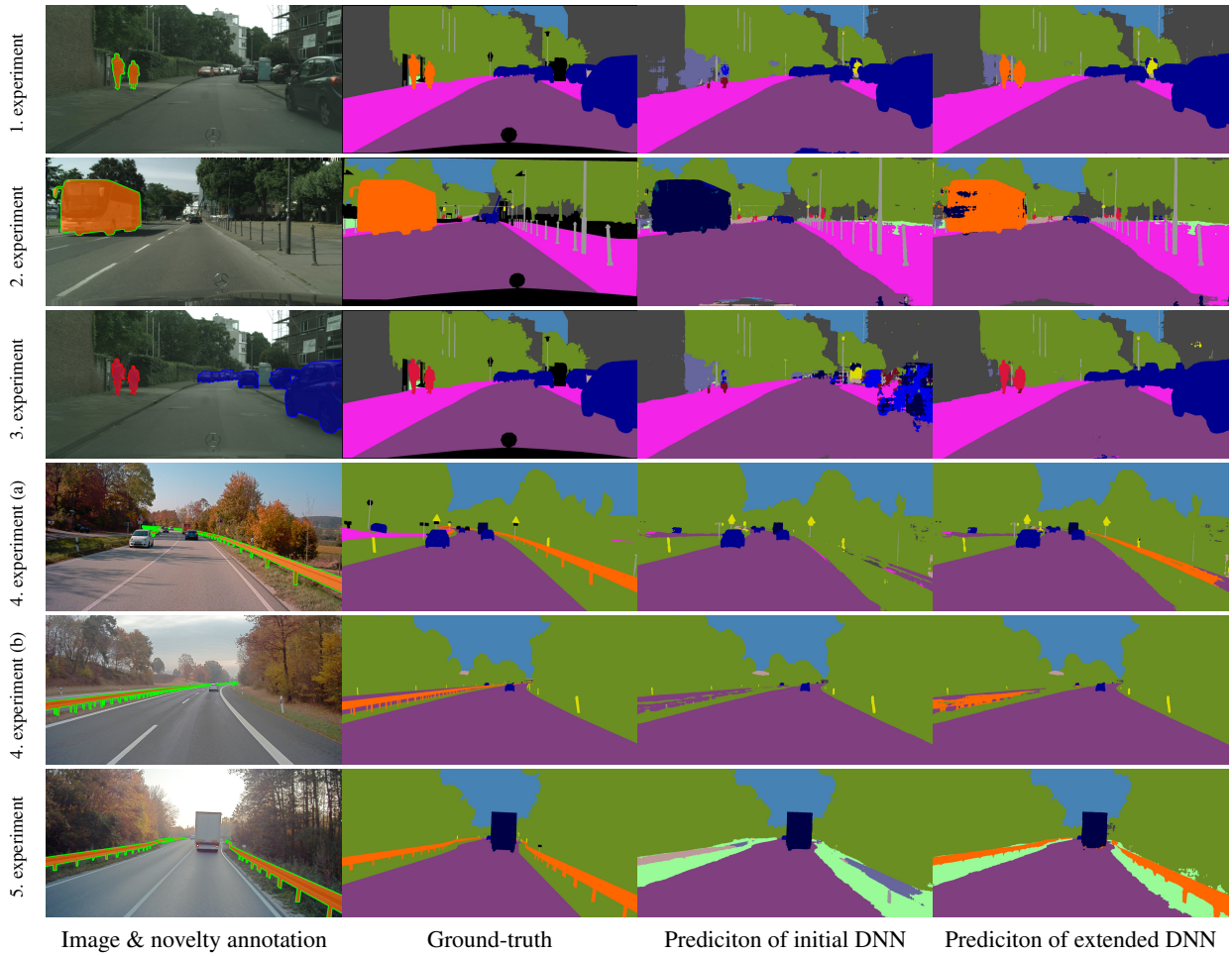


Figure 12: Visualization of the results obtained for the performed experiments, where a DNN is extended by a novel class, respectively. The top row depicts an example image & ground-truth and the predictions of the initial & extended DNN for the novelty *human*, the second row for the novelty *bus*, the third for *human* & *car* and the bottom rows for the novelty *guardrail* with prior fine-tuning. 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.

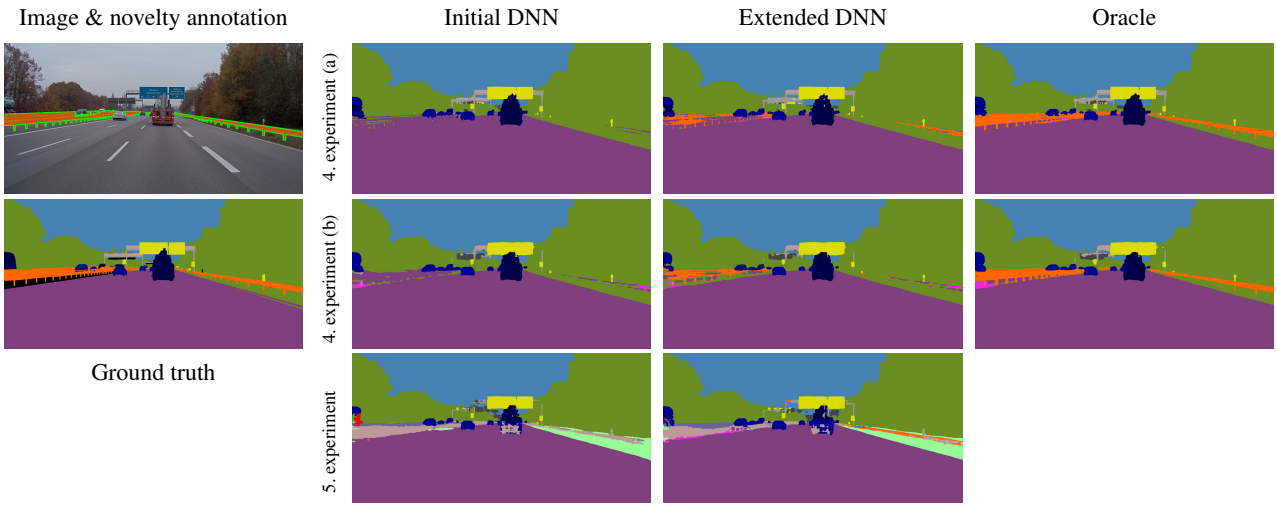


Figure 13: Comparison of the semantic segmentation predictions of all models incrementally extended by the *guardrail* class for an example image from the A2D2 validation split.

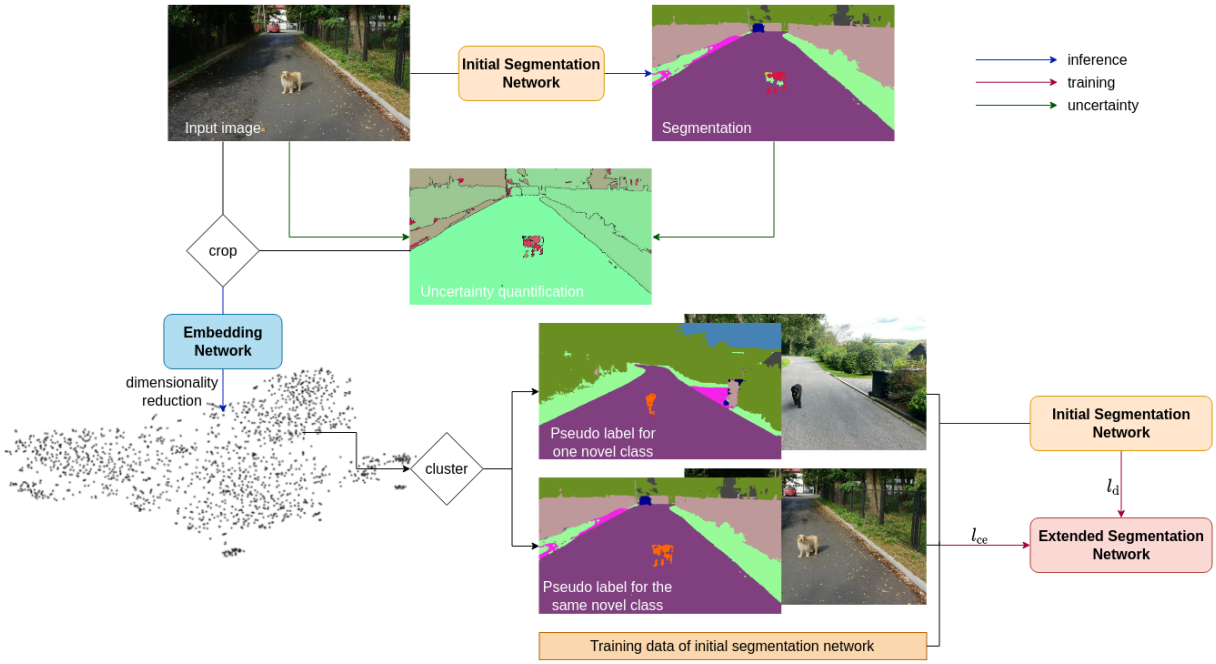


Figure 14: Illustration of the entire framework.