Unsupervised Class Incremental Learning using Empty Classes

Svenja Uhlemeyer[1]
uhlemeyer@math.uni-wuppertal.de
Julian Lienen[2]
julian.lienen@upb.de
Youssef Shoeb[3,5]
youssef.shoeb@continental.com
Eyke Hüllermeier[4]
eyke@lmu.de
Hanno Gottschalk[5]
gottschalk@math.tu-berlin.de

[1] IZMD and Faculty of Mathematics and Natural Sciences
University of Wuppertal, Germany
[2] Department of Computer Science
Paderborn University, Germany
[3] Continental AG, Germany
[4] Institute for Informatics
LMU Munich, Germany
[5] Institute of Mathematics
Technical University Berlin, Germany

Abstract

For real-world applications, deep neural networks (DNNs) must recognize and adapt to previously unseen inputs and changing environments. To achieve this, we propose a novel method to augment DNNs with the capability to identify and incrementally learn novel classes that were not present in their initial training set. Our approach uses anomaly detection to retrieve out-of-distribution (OoD) samples as potential candidates for new classes and uses k empty classes to learn these novel classes incrementally in an unsupervised fashion. We introduce two loss functions, which 1) encourage the DNN to allocate OoD samples to the new empty classes and 2) minimize the inner-class feature distance between the newly formed classes. Unlike previous approaches that rely on labeled data for each class, our model uses a single label for all OoD data and a precomputed distance matrix to differentiate between them. Our experiments across image classification and semantic segmentation tasks show our method's ability to expand a DNN's understanding by several classes without requiring explicit ground truth labels.

1 Introduction

In a closed-world setting, where all possible classes are known during training, state-of-the-art DNNs achieve impressive accuracy when trained in a supervised manner. However, a significant challenge arises in practical scenarios when the DNN encounters concepts not seen during training. Current DNNs offer no performance guarantees on inputs outside of their training distribution [27]. As a result, open world recognition [2] has emerged as a practically more relevant problem formulation. Open-world recognition extends a DNN's capabilities by integrating out-of-distribution (OoD) detection [1] with class-incremental learning [5], where the model is continuously updated with new classes. Despite progress in this area, a

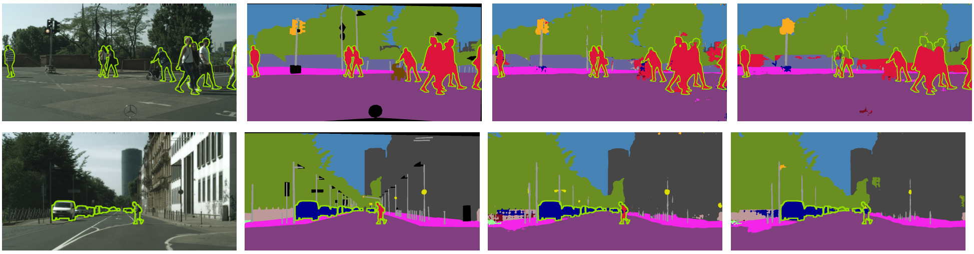
(e) image (f) ground truth (g) ours (h) baseline

Figure 1: Comparison of two segmentation DNNs which were extended by the classes human and car. While the segmentation masks are similar for the initial classes, the humans are much better segmented by the DNN, which was extended by our empty classes approach. The ground truth contours of the novel classes are highlighted in green.

predominant reliance on supervised learning methods persists. Typically, these approaches necessitate human annotation for updating models with new classes, which can be impractical in real-world settings.

The reliance on supervised learning is not the only challenge facing open-world recognition. To bridge the gap between current research and the practical implementation of DNNs in complex real-world scenarios, such as perception systems for automated driving, the field must evolve to tackle dense prediction tasks like semantic segmentation. Unlike image classification, which is predominantly object-centric and categorizes entire images as either in-distribution or OoD, semantic segmentation involves more complexity, dealing with scenes where only specific regions may be OoD. This necessitates careful consideration of OoD localization, retrieval of semantically meaningful features, and computational efficiency.

This work presents a novel unsupervised approach to extend DNNs with *empty classes* for identifying new concepts. Initially trained on known classes, the model incorporates an out-of-distribution (OoD) detection mechanism to separate known from unknown categories. We add auxiliary neurons to the DNN output layer to accommodate inputs with potentially unrecognized classes, forming *empty classes* for these new categories. We introduce two loss functions to enhance the model's ability to categorize similar OoD data, enabling dynamic adjustment of feature representations to differentiate between established and newly discovered classes effectively.

We evaluate our method on the task of semantic segmentation of street scenes using the Cityscapes dataset [10]. Our approach outperforms the current state-of-the-art in unsupervised class-incremental continual learning [37], particularly in detecting the novel *car* class and significantly improving the identification of humans (*c.f.* Fig. 1). Furthermore, we have included image classification results for CIFAR10 [18] and Animals10[1] datasets to validate the performance of our method on low- and medium-resolution images. These results illustrate that our loss term relies on general principles that work across complexity scales. Additionally, we show that combining clustering and continual learning in a single step leads to better performance than the common baseline approach of "cluster first, then learn".

[1] https://www.kaggle.com/datasets/alessiocorrado99/animals10

2 Related Work

Open world recognition [2] refers to the problem of adapting a learning system to a non-delimitable and potentially constantly evolving target domain. As such, it combines the disciplines of open set learning [33], where incomplete knowledge over the target domain is assumed at training time, with incremental learning [5], in which the model is updated by exploring additional target space regions at test time, thereby adapting to novel target information. Typically, open set recognition is formalized by specifying a novelty detector, a labeling process, and an incremental learning function, allowing for a generalized characterization of such systems [2].

Most previous approaches consider open-world recognition in the context of classification, where novel concepts are in the form of previously unseen classes. While a plethora of methods have been proposed to tackle the individual sub-problems for classification problems, for which we refer to [29] for a more comprehensive overview, literature on holistic approaches for open world classification is rather scarce. In [35], a metric learning approach is used to distinguish between pairs of instances belonging to the same classes, allowing the detection of instances that can not be mapped to known classes and thus used to learn novel class concepts. In [26], the likelihood ratio between known and proxy unknown objects is used to detect novel classes not included in the initial training set. Moreover, [28] suggests a semi-supervised learning approach that applies clustering on learned feature representations to reason about unknown classes. Related to this, [38] describes a kernel method using an alternative loss formulation to learn embeddings to be clustered for class discovery. Recently, similar concepts have been tailored to specific data modalities, such as tabular data [36].

In the domain of semantic segmentation, open world recognition is also covered under the term *zero-shot semantic segmentation* [3]. To predict unseen categories for classified pixels, a wide range of methods leverage additional language-based context information [3, 21, 39], or proxy. Besides enriching visual information by text, unsupervised methods, e.g. , employing clustering based on visual similarity [37] or contrastive losses [6, 12], have also been considered. More recently, [7] adopts semantic segmentation based on LiDAR point clouds by augmenting conventional classifiers with predictors recognizing unknown classes, thereby enabling incremental learning.

In a more general context, unsupervised representation learning [30] constitutes a major challenge to generalize learning methods to unseen concepts. Methods of this kind are typically tailored to data modalities, e.g. , by specifying auxiliary tasks to be solved [13, 40]. In the domain of images, self-supervised learning approaches have emerged recently [4, 19], which commonly apply knowledge distillation between different networks, allowing for learning in a self-supervised fashion. Other methods include ideas stemming from metric [15] or contrastive learning [9].

3 Method Description

In this section, we present our training framework for unsupervised class-incremental learning with empty classes. For the sake of brevity, all equations are introduced for image classification and adapted to semantic segmentation in Sec. 4. First, we give a motivating example in Fig. 2, where we enrich data stemming from the TwoMoons dataset[2] with OoD samples and extend the model by three novel classes.

[2] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.datasets

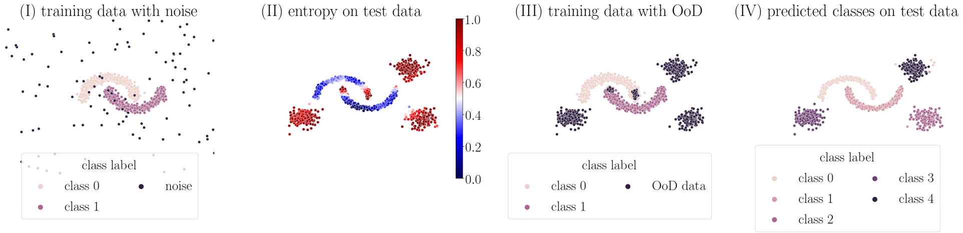

Figure 2: (I) A binary classification model is trained on two classes and additional noise data for entropy maximization. (II) OoD samples in the test data are obtained by entropy thresholding. (III) The training data is enriched with the OoD samples and a distance matrix, containing their pair-wise Euclidean distances. (IV) The model is class-incrementally extended by three novel classes.

I) Learning Model For an input image $x \in \mathcal{X}$, let $f(x) \in (0,1)^q$ denote the softmax probabilities of some image classification model $f : \mathcal{X} \to (0,1)^q$ with underlying classes $\mathcal{C} = \{1,\dots,q\}$. Consider a test dataset which includes images from classes $c \in \{1,\dots,q,q+1,\dots\}$. Note that our framework does not necessarily assume labels for the test data as these will be only used for evaluation and not during the training. Furthermore, let $u(f(x)) \in [0,1]$ denote some arbitrary uncertainty score which derives from the predicted class-probabilities $f(x)$. Thus, a test image x is considered to be OoD, if $u(f(x)) > \tau$ for some threshold $\tau \in [0,1]$.

Next, we extend the initial model f by $k \in \mathbb{N}$ empty classes in the final classification layer, which is then denoted as $f^k : \mathcal{X} \to (0,1)^{q+k}$, and fine-tune it on the OoD data $\mathcal{X}^{\mathrm{OoD}}$. Therefore, we compute pairwise distances $d_{ij} = d(x_i,x_j)$ for all $(x_i,x_j) \in \mathcal{X}^{\mathrm{OoD}} \times \mathcal{X}^{\mathrm{OoD}}$ as a pre-processing step, e.g. using the pixel-wise Euclidean distance or any distance metric in the feature space of some embedding network. The model f^k is then fine-tuned on (a subset of) the initial training data $\mathcal{X}^{\mathrm{train}}$, enriched with the OoD samples from the test data. For the in-distribution samples (x,y), we compute the cross-entropy loss

$$\ell_{\mathrm{ce}}(x,y) = -\sum_{c=1}^{q} \mathbb{1}_{\{c=y\}} \log(f_c^k(x)) \,. \tag{1}$$

Further, we entice the model to predict one of the empty classes $q+1,\dots,q+k$ for OoD data by minimizing the class-probabilities $f_1^k(x),\dots,f_q^k(x)$, $x \in \mathcal{X}^{\mathrm{OoD}}$, i.e. , by computing

$$\ell_{\mathrm{ext}}(x) = \frac{1}{q}\sum_{c=1}^{q} f_c^k(x) \,. \tag{2}$$

Finally, we aim to divide the data among the empty classes based on their similarity. Thus, our clustering loss is computed pair-wise as

$$\ell_{\mathrm{cluster}}(x_i,x_j) = \frac{\alpha}{q+k} \cdot d_{ij} \cdot \sum_{c=1}^{q+k} f_c^k(x_i) f_c^k(x_j) \,, \tag{3}$$

where $\alpha \in \mathbb{R}_{>0}$ can be adjusted to control the impact of the clustering loss function. Together, these three loss functions give the overall objective

$$\begin{aligned} L = {} & \lambda_1 \mathbb{E}_{(x,y)\sim\mathcal{X}^{\text{train}}}[\ell_{\text{ce}}(x,y)] \\ & + \lambda_2 \mathbb{E}_{x\sim\mathcal{X}^{\text{OoD}}}[\ell_{\text{ext}}(x)] \\ & + \lambda_3 \mathbb{E}_{x_i,x_j\sim\mathcal{X}^{\text{OoD}}}[\ell_{\text{cluster}}(x_i,x_j)] \ , \end{aligned} \tag{4}$$

where the hyperparameters λ_1, λ_2 and λ_3 can be adjusted to balance the impact of the objectives.

II) OoD Detection OoD detection is a pre-processing part of our framework, which can be exchanged in a plug and play manner. In our experiments, we implemented entropy maximization [16] for image classification and thus perform OoD detection by thresholding on the softmax entropy.

The idea of entropy maximization is the inclusion of *known unknowns* into the training data of the initial model in order to entice it to exhibit a high softmax entropy

$$u(x) = -\frac{1}{\log(q)} \sum_{c=1}^{q} f_c(x) \log(f_c(x)) \tag{5}$$

on OoD data $x \in \mathcal{X}^{\text{OoD}}$. Therefore, during training the initial model, we compute the entropy maximization loss

$$\ell_{\text{em}}(x) = -\sum_{c=1}^{q} \frac{1}{q} \log(f_c(x)) \tag{6}$$

for known unknowns $x \in \mathcal{X}^{\text{OoD}}$, giving the overall objective

$$\begin{aligned} L = {} & \lambda \ \mathbb{E}_{(x,y)\sim\mathcal{X}^{\text{train}}}[\ell_{\text{ce}}(x,y)] \\ & + (1-\lambda) \ \mathbb{E}_{x\sim\mathcal{X}^{\text{OoD}}}[\ell_{\text{em}}(x)] \ . \end{aligned} \tag{7}$$

In the Two Moons example, these OoD data was uniformly distributed noise. For image classification, we employ the domain-agnostic data augmentation technique mixup [41]. This is, an OoD image is obtained by computing the average of two in-distribution samples. Entropy maximization was also introduced for semantic segmentation of street scenes [8, 17], where the OoD samples originate from the COCO dataset [20]. Furthermore, the OoD loss and data was only included in the final training epochs, which means that existing networks can be fine-tuned for entropy maximization.

III) Distance Matrix Next, we compute pair-wise distances for the detected OoD samples, which constitute the OoD dataset for the incremental learning. For simple datasets such as TwoMoons or MNIST, the distance can be measured directly between the data samples. For MNIST, this is done by flattening the images and computing the Euclidean distance between the resulting vectors. For more complex datasets, we employ embedding networks to extract useful features of the images. These embedding networks are arbitrary image classification models, trained on large datasets such as ImageNet [11] or CIFAR100 [18], which need to be chosen carefully and individually for each experiment as the clustering loss strongly depends on their ability to extract separable features for the known and especially the novel classes.

The feature distances are either computed in the high-dimensional feature space directly, or, for the sake of transparency and better visual control, in a low-dimensional rearrangement. Applying the manifold learning technique UMAP [23] to the entire test data, we reduce the dimension of the feature space to two . The distance matrix is then computed as the Euclidean distances in the low-dimensional space for all pairs of OoD samples.

IV) Incremental Learning For class-incremental learning, we minimize three different loss functions defined in Eqs. (1) to (3). The cross-entropy loss (1) is computed for in-distribution to mitigate catastrophic forgetting [22]. The OoD samples are pushed towards the novel classes by the extension loss (2), which is minimized whenever the probability mass is concentrated in the empty classes, i.e. ,

$$\ell_{\text{ext}}(x) \to 0 \text{ for } \sum_{c=q+1}^{q+k} f_c^k(x) \to 1,\ x \in \mathcal{X}^{\text{OoD}} \,. \tag{8}$$

The cluster loss (3) is computed for all pairs of OoD candidates contained in a batch. Thus, it has a runtime complexity of $\mathcal{O}(n^2)$, as for n OoD candidates, we need to compute $\frac{n^2-n}{2}$ terms. Furthermore, the minimum of the cluster loss is probably greater than zero, as samples which belong to the same class rarely share exactly the same features. To reach this minimum for two OoD samples x_i, x_j with a large distance, they should be assigned to different classes, i.e. , whenever $f_c^k(x_i)$ is significantly different from zero, we desire that $f_c^k(x_j)$ becomes small.

4 Adjustments for Semantic Segmentation

Let $H \times W$ denote the resolution of the images $x \in \mathcal{X}$. Then, the softmax output of a semantic segmentation DNN $f : \mathcal{X} \to (0,1)^{H \times W \times q}$ provides class-probabilities for image pixels, denoted as $z = (h,w) \in \mathcal{Z}$. Thus, the OoD detector must not only identify OoD images, but also give information about their pixel positions. To store these information, we generate OoD instance masks by thresholding on the obtained OoD score and by distinguishing between connected components in the resulting OoD mask.

For semantic segmentation, the loss functions are computed for pixels of OoD objects instead of images. Let $\mathcal{Z}_s$ denote the set of pixel positions which belong to an OoD candidate $s \subseteq x$. The extension loss is computed equivalently to Eq. (2) as

$$\ell_{\text{ext}}(s) = -\frac{1}{|\mathcal{Z}_s|} \sum_{z \in \mathcal{Z}_s} \frac{1}{q} \sum_{c=1}^{q} f_{z,c}^k(x) \,. \tag{9}$$

For two OoD candidates $s_i \subseteq x_i, s_j \subseteq x_j$ with distance d_{ij}, the cluster loss is computed as

$$\ell_{\text{cluster}}(s_i, s_j) = \frac{\alpha}{q+k} d_{ij} \sum_{c=1}^{q+k} \overline{f_c^k(x_i)}\ \overline{f_c^k(x_j)} \,, \tag{10}$$

where

$$\overline{f_c^k(x)} = \frac{1}{|\mathcal{Z}_s|} \sum_{z \in \mathcal{Z}_s} f_{z,c}^k(x) \tag{11}$$

denotes the mean softmax probability over all pixels $z \in \mathcal{Z}_s$ for some class $c \in \{1, \ldots, q+k\}$.

For OoD detection in semantic segmentation, we adapt a meta-regression approach [31, 32], using uncertainty measures such as the softmax entropy and further information derived from the initial model's output, to estimate the prediction quality on a segment-level. Here, a segment denotes a connected component in the semantic segmentation mask, which the initial model predicts. That is, meta-regression is a post-processing approach to quantify uncertainty aggregated over segments, and considering that the model likely is highly uncertain if confronted with OoD objects, it can be applied for OoD detection. In contrast to image classification, where images are either OoD or not, semantic segmentation is performed on images that simultaneously contain in-distribution and OoD pixels. Aggregating uncertainty scores across segments simplifies the detection of OoD objects as contiguous OoD pixels since it removes the high uncertainty for class boundaries.

For an initial DNN, we use the training data to fit a gradient boosting model as meta regressor, which then estimates segment-wise uncertainty scores $u(s)$ for all segments $s \subseteq x \in \mathcal{X}$.

5 Numerical Experiments

We perform several experiments for image classification on CIFAR10 [18] and Animals10, as well as on Cityscapes [10] to evaluate our method for semantic segmentation. To this end, we extend the initial models by empty classes, i.e. , neurons in the final classification layer with randomly initialized weights, and fine-tune them on OoD data, retraining with a fixed encoder. For evaluation, we provide accuracy scores - separately for known and novel classes - for image classification, (mean) Intersection over Union (IoU), precision, and recall values for semantic segmentation.

The OoD classes in the following experiments were all chosen so that they are semantically far away from each other. For example, the Animals10 classes *horse* (1), *cow* (6) and *sheep* (7) are semantically related, as they are all big animals which are mostly on the pasture, whereas *elephant* (2) and *spider* (8) are well separable classes, which is also visible in the two-dimensional feature space. However, in the appendix, we will also provide evaluation metrics averaged over multiple runs with randomly picked OoD classes.

5.1 Experimental Setup

We consider the following dataset splits for each experiment: the *training data* denotes images with ground truth for the initially known classes. We train the initial model on these images and replay them during the training of the extended model to avoid catastrophic forgetting. The *test data* consist of unlabeled images which include both known and unknown classes. This dataset is fed into the OoD detector to identify *OoD data*, on which the model gets extended. The *evaluation dataset* includes images with ground truth for known and novel classes and is used to evaluate the models. If such labels are available for the test data, evaluation images may be the same as the test images.

Our approach requires prior OoD detection. Here, we only provide the experimental setup for fine-tuning the extended model. For all experiments, we tuned the weighting parameters $\lambda_1, \lambda_2, \lambda_3$ in Eq. (4) by observing all loss functions separately over several epochs using different parameter configurations to ensure that each loss term decreases. The following descriptions of the experiments include the network architecture, known and novel classes, information about the dataset splits, and the distance matrix generation. We refer to the appendix for further information about the experiments, including the TwoMoons experiment.

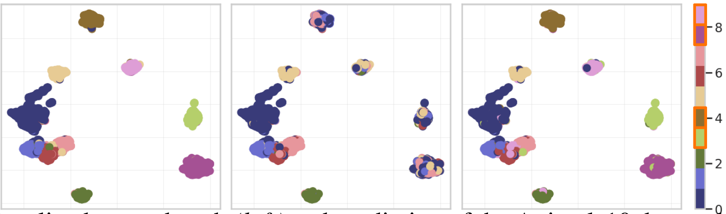

Figure 3: Visualized ground truth *(left)* and prediction of the Animals10 dataset by the initial *(middle)* and extended *(right)* model. The four novel classes $3, 4, 8$ and 9 are outlined in orange. The extended model's accuracy is $\sim 95\%$.

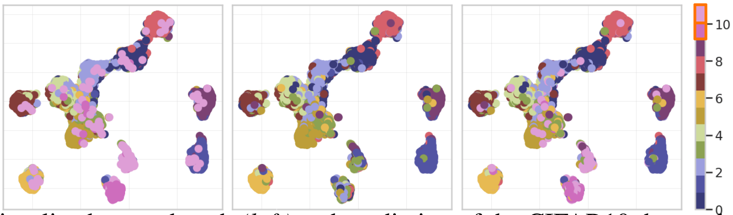

Figure 4: Visualized ground truth *(left)* and prediction of the CIFAR10 dataset by the initial *(middle)* and extended *(right)* model. The two novel classes 10 and 11 are outlined in orange. The extended model's accuracy is $\sim 89\%$.

Image Classification								
			supervised		unsupervised		ablation studies	
dataset	**OoD**	**accuracy**	**initial**	**oracle**	**ours**	**baseline**	**−detection**	**−−distance**
CIFAR10	10 11	known	91.45%	91.86%	**90.51%**	90.29%	88.90%	86.94%
		novel	-	89.53%	**70.00%**	33.40%	78.80%	87.00%
Animals10	3 4 8 9	known	96.29%	95.80%	**93.76%**	92.78%	94.46%	95.20%
		novel	-	97.65%	**96.68%**	72.59%	97.02%	97.90%

Table 1: Quantitative evaluation of the image classification experiments. For all evaluated models, the accuracy is stated separately for the previously known and the unlabeled novel classes. The highest scores for the unsupervised approaches are bolded.

CIFAR10 We employ a ResNet18, which is trained on the whole CIFAR10 training split, including all ten classes. For testing, we enrich the CIFAR10 test split with images from CIFAR100. Therefore, we split CIFAR100 into an unlabeled and a labeled subset: the classes $\{0, \ldots, 49\}$ are possible OoD candidates; thus, all samples belonging to these classes are considered to be unlabeled. We extend the CIFAR10 test data by the classes *apple* (0) and *clock* (22), mapping them onto the labels (10) and (11), respectively. We evaluate our models on the labeled test data. The labeled CIFAR100 subset includes the classes $\{50, \ldots, 99\}$ and is used together with the CIFAR10 training data to train a ResNet18 as an embedding network. To compute the distances, we feed the whole test data into this embedding network and extract the features of the penultimate layer. These are further projected into a 2D space with UMAP. Then, the distance matrix is computed as the pixel-wise Euclidean distance between the 2D representations of the OoD images.

Animals10 As an initial model, we employ a ResNet18, which is trained on six out of ten classes. As novel classes, we selected *butterfly* (3), *chicken* (4), *spider* (8), and *squirrel* (9). The distances are computed as for CIFAR10, but employing a DenseNet201, which is trained on ImageNet with 1,000 classes as an embedding network.

Cityscapes For comparison reasons with the baseline, we adapt the experimental setup from [37], where the class labels *human (person, rider), car* and *bus* are excluded from the 19 Cityscapes evaluation classes. Like the baseline, we extend the DNN by two empty classes and exclude the class *bus* from the evaluation. Thus, we train a semantic segmentation DeepLabV3+ with WideResNet38 backbone on 2,500 training samples with 15 trainable classes. We apply meta-regression to the Cityscapes test data and crop out image patches tailored to the predicted OoD segments, i.e., , connected component of OoD pixels. Afterward, we compute distances between these image patches analogously to Animals10 as the Euclidean distances between 2D representations of features, which we obtain by feeding the patches into a DenseNet201 trained on 1,000 ImageNet classes.

5.2 Evaluation & Ablation Studies

We compare our evaluation results to the following baselines. For image classification, we employ the k-means clustering algorithm to pseudo-label the OoD data samples and fine-tune the model on the pseudo-labeled data using the cross-entropy loss. For semantic segmentation, we compare with the method presented in [37], which also employs clustering algorithms in the embedding space to obtain pseudo-labels. Furthermore, to get an idea of the maximum achievable performance, we train oracle models that have learned all available classes in a fully supervised manner.

We evaluate our image classification approach for the ablation studies on "clean" OoD data (−detection). Therefore, we do not detect the OoD samples in the test data by thresholding on some anomaly score but by considering the ground truth. In this way, we simulate a perfect OoD detector. Since the results of our method are also affected by the quality of the distance matrix, we further analyze our method for a synthetic distance matrix (−−distance), where two OoD samples $x_i, x_j \in \mathcal{X}^{\mathrm{OoD}}$ have a distance $d(x_i, x_j) = 0$ if they stem from the same class, $d(x_i, x_j) = 1$ otherwise. Thus, the OOD samples are labeled by the distance matrix, and the fine-tuning is supervised, allowing a pure comparison of our loss functions with the cross-entropy loss. We do not provide ablation studies for semantic segmentation since the Cityscapes test data does not include publicly available annotations.

Image Classification As shown in Tab. 1 and visualized in Figs. 3 and 4, our approach exceeds the baseline's accuracy for novel classes by 36.60 and 24.09 percentage points (pp) for CIFAR10 and Animals10, respectively. This is mainly caused by in-distribution samples, which are false positive OoD predictions, or by OoD samples, which are embedded far away from their class centroids. Consequently, different OoD classes are assigned to the same cluster by the k-means algorithm. As our approach uses soft labels, the DNN is more likely to reconsider the choice of the OoD detector during fine-tuning.

In the ablation studies, we omit the OoD detector (−detection) and instead select the OoD samples based on their ground truth. Thereby, we observe an improvement in the accuracy of novel classes for the CIFAR10 and Animals10 datasets, while the performance remains constant for FashionMNIST and significantly decreases for MNIST. We further compute a ground truth distance matrix (−−distance) with distances 0 and 1 for samples belonging to the same or to different classes, respectively. Since this is supervised fine-tuning, these DNNs are comparable to oracles. We observe that the oracles tend to perform better on the initial and worse on the novel classes. However, this might be a consequence of class-incremental learning.

Semantic Segmentation

			supervised		unsupervised	
dataset	**class**	**metric**	**initial**	**oracle**	**ours**	**baseline**
Cityscapes	0,...,14	mean IoU	56.99%	77.28%	**59.72%**	57.52%
	15 (human)	IoU	-	81.90%	33.87%	**40.22%**
	16 (car)	IoU	-	94.94%	**84.14%**	81.27%
	0,...,14	mean precision	65.75%	88.03%	**84.63%**	78.53%
	15 (human)	precision	-	89.22%	37.80%	**68.74%**
	16 (car)	precision	-	96.83%	**87.11%**	86.56%
	0,...,14	mean recall	80.88%	85.38%	65.38%	**65.78%**
	15 (human)	recall	-	90.90%	**76.54%**	49.65%
	16 (car)	recall	-	97.99%	**96.11%**	93.05%

Table 2: Quantitative evaluation of the semantic segmentation experiment on the Cityscapes dataset. IoU, precision, and recall values are provided for both novel classes and averaged over the previously known classes. The highest scores for the unsupervised approaches are bolded.

Semantic Segmentation The quantitative results of our semantic segmentation method, reported in Tab. 2, demonstrate that the empty classes are "filled" with the novel concepts *human* and *car*. The performance on the previously-known classes is similar to the baseline even without including a distillation loss [24]. For the *car* class, our method outperforms the baseline with respect to IoU (+2.87 pp), precision (+0.55 pp) and recall (+3.06 pp). We lose performance in terms of IoU for the *human* class due to a higher tendency for false positives. However, the false negative rate is significantly reduced, which is indicated by an increase in the recall value of 26.89 pp.

When examining the OoD masks, we observed that the connected components are often very extensive, which is caused by neighboring OoD objects. Thus, the embedding space contains many large image patches that are not tailored to a single OoD object but rather to a number of parked cars, a crowd of people, or even a bicyclist riding next to a car, which appreciably impairs our results.

6 Conclusion & Outlook

In this work, we proposed a solution to open-world classification for image classification and semantic segmentation by learning novel classes in an unsupervised manner. We suggested postulating empty classes, which allows one to capture newly observed classes in an incremental learning approach. This way, the model can detect new classes in a flexible manner, potentially whitewashing mistakes of previous OoD detectors.

As our method employs several hyperparameters, e.g. , to specify the number of novel empty classes, we envision an automatic derivation of the optimal number of new classes as future work. In this regard, replacing the Elbow method in the eventual clustering by more suitable criteria appears desirable [34]. Moreover, we shall investigate approaches to improve the generalizability of our approach to embedding models of arbitrary kind to derive distance matrices that are not tailored to specific datasets. Furthermore, the semantic segmentation performance could be improved by incorporating mask-level information and obtaining OoD candidates based on mask- instead of segment-level [14, 25].

References

[1] Reza Averly and Wei-Lun Chao. Unified out-of-distribution detection: A model-specific perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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

[3] Maxime Bucher, Tuan-Hung VU, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021.

[5] Gert Cauwenberghs and Tomaso A. Poggio. Incremental and decremental support vector machine learning. In *NIPS*, 2000.

[6] Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15313–15322, 2021.

[7] Jun Cen, Peng Yun, Shiwei Zhang, Junhao Cai, Di Luan, Mingqian Tang, Ming Liu, and Michael Yu Wang. Open-world semantic segmentation for lidar point clouds. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, page 318–334, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19838-0.

[8] 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 2021.

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[10] 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.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[12] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10032–10042, 2021.

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[14] Matej Grcić, Josip Šarić, and Siniša Šegvić. On advantages of mask-level recognition for outlier-aware segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2937–2947, June 2023.

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[16] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

[17] Nicolas Jourdan, Eike Rehder, and Uwe Franke. Identification of uncertainty in artificial neural networks. 2019.

[18] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[19] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

[21] Quan Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *European Conference on Computer Vision*, 2022.

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

[23] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *J. Open Source Softw.*, 3:861, 2018.

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

[25] Nazir Nayal, Mısra Yavuz, João F. Henriques, and Fatma Güney. Rba: Segmenting unknown regions rejected by all. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[26] Nazir Nayal, Youssef Shoeb, and Fatma Güney. A likelihood ratio-based approach to segmenting unknown objects, 2024.

[27] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015.

[28] Jeremy Nixon, Jeremiah Z. Liu, and David Berthelot. Semi-supervised class discovery. *ArXiv*, abs/2002.03480, 2020.

[29] Jitendra Parmar, Satyendra Singh Chouhan, Vaskar Raychoudhury, and Santosh Singh Rathore. Open-world machine learning: Applications, challenges, and opportunities. *ACM Computing Surveys*, 55:1 – 37, 2021.

[30] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[31] 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.

[32] 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.

[33] Walter J. Scheirer, Anderson Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1757–1772, 2013.

[34] Erich Schubert. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. *SIGKDD Explor. Newsl.*, 25(1):36–42, jul 2023. ISSN 1931-0145.

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

[36] Colin Troisemaine, Joachim Flocon-Cholet, Stéphane Gosselin, Sandrine Vaton, Alexandre Reiffers-Masson, and V. Lemaire. A method for discovering novel classes in tabular data. *2022 IEEE International Conference on Knowledge Graph (ICKG)*, pages 265–274, 2022.

[37] Svenja Uhlemeyer, Matthias Rottmann, and Hanno Gottschalk. Towards unsupervised open world semantic segmentation. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

[38] Zifeng Wang, Batool Salehi, Andrey Gritsenko, Kaushik R. Chowdhury, Stratis Ioannidis, and Jennifer G. Dy. Open-world class discovery with kernel networks. *2020 IEEE International Conference on Data Mining (ICDM)*, pages 631–640, 2020.

[39] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero- and few-label semantic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8248–8257, 2019.

[40] Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8344–8353, 2022.

[41] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.