Entropy Maximization and Meta Classification for Out-of-Distribution Detection in Semantic Segmentation

Robin Chan, Matthias Rottmann and Hanno Gottschalk
IZMD, Faculty of Mathematics and Natural Sciences, University of Wuppertal
{rchan,rottmann,hgottsch}@uni-wuppertal.de

Abstract

Deep neural networks (DNNs) for the semantic segmentation of images are usually trained to operate on a predefined closed set of object classes. This is in contrast to the "open world" setting where DNNs are envisioned to be deployed to. From a functional safety point of view, the ability to detect so-called "out-of-distribution" (OoD) samples, i.e.*, objects outside of a DNN's semantic space, is crucial for many applications such as automated driving. A natural baseline approach to OoD detection is to threshold on the pixel-wise softmax entropy. We present a two-step procedure that significantly improves that approach. Firstly, we utilize samples from the COCO dataset as OoD proxy and introduce a second training objective to maximize the softmax entropy on these samples. Starting from pretrained semantic segmentation networks we re-train a number of DNNs on different in-distribution datasets and consistently observe improved OoD detection performance when evaluating on completely disjoint OoD datasets. Secondly, we perform a transparent post-processing step to discard false positive OoD samples by so-called "meta classification." To this end, we apply linear models to a set of hand-crafted metrics derived from the DNN's softmax probabilities. In our experiments we consistently observe a clear additional gain in OoD detection performance, cutting down the number of detection errors by 52% when comparing the best baseline with our results. We achieve this improvement sacrificing only marginally in original segmentation performance. Therefore, our method contributes to safer DNNs with more reliable overall system performance.*

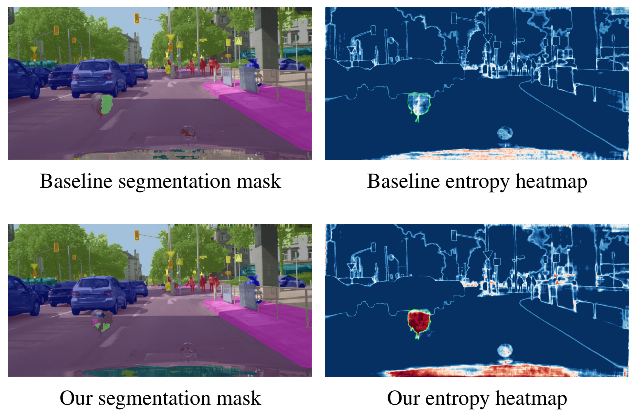


Figure 1: Comparison of segmentation mask and softmax entropy before our OoD training (*top row*) and after (*bottom row*). While there are minor differences in the segmentation masks, the annotated unknown object (marked with green contours) becomes clearly recognizable in the entropy heatmap due to our OoD training. In the heatmap high values are red, low ones are blue.

1. Introduction

In recent years spectacular advances in the computer vision task semantic segmentation have been achieved by deep learning [47, 51]. Deep convolutional neural networks (CNNs) are envisioned to be deployed to real world applications, where they are likely to be exposed to data that is substantially different from the model's training data. We consider data samples that are not included in the set of a model's semantic space as *out-of-distribution* (OoD) samples. State-of-the-art neural networks for semantic segmentation, however, are trained to recognize a predefined closed set of object classes [13, 32], *e.g.* for the usage in environment perception systems of autonomous vehicles [24]. In open world settings there are countless possibly occurring objects. Defining additional classes requires a large amount of annotated data (cf. [12, 52]) and may even lead to performance drops [15]. One natural approach is to introduce a *none-of-the-known* output for objects not belonging to any of the predefined classes [49]. In other words, one uses a set of object classes that is sufficient for most scenarios and treats OoD objects by enforcing an alternative model output for such samples. From a functional safety point of view, it is a crucial but missing prerequisite that neural networks are capable of reliably indicating when they are operating out of their proper domain, *i.e.*, detecting OoD objects, in order to initiate a fallback policy.

As images from everyday scenes usually contain many different objects, of which only some could be out-of-distribution, knowing the location where the OoD object occurs is desired for practical application. Therefore, we address the problem of detecting anomalous regions in an image, which is the case if an OoD object is present (see figure 1) and which is a research area of high interest [6, 20, 33, 42]. This so-called *anomaly segmentation* [5, 20] can be pursued, for instance, by incorporating sophisticated uncertainty estimates [3, 18] or by adding an extra class to the model's learnable set of classes [49].

In this work, we detect OoD objects in semantic segmentation with a different approach which is composed of two steps: As first step, we re-train the segmentation CNN to predict class labels with low confidence scores on OoD inputs, by enforcing the model to output high prediction uncertainty. In order to quantify uncertainty, we compute the softmax entropy which is maximized when a model outputs uniform probability scores over all classes [29]. By deliberately including annotated OoD objects as *known unknowns* into the re-training process and employing a modified multi-objective loss function, we observe that the segmentation CNN generalizes learned uncertainty to unseen OoD samples (*unknown unknowns*) without significantly sacrificing in original performance on the primary task, see figure 1.

The initial model for semantic segmentation is trained on the Cityscapes data [13]. As proxy for OoD samples we randomly pick images from the COCO dataset [32] excluding the ones with instances that are also available in Cityscapes, cf. [19, 22, 37] for a related approach in image classification. We evaluate the pixel-wise OoD detection performance via entropy thresholding for OoD samples from the LostAndFound [42] and Fishyscapes [6] dataset, respectively. Both datasets share the same setup as Cityscapes but include OoD objects.

The second step incorporates a *meta classifier* flagging incorrect class predictions at segment level, similar as proposed in [34, 44, 45] for the detection of false positive instances in semantic segmentation. After increasing the sensitivity towards predicting OoD objects, we aim at removing false predictions which are produced due to the preceding entropy boost (cf. [9]). The removal of false positive OoD object predictions is based on aggregated dispersion measures and geometry features within segments (connected components of pixels), with all information derived solely from the CNN's softmax output. As meta classifier we employ a simple linear model which allows us to track and understand the impact of each metric.

To sum up our contributions, we are the first to successfully modify the training of segmentation CNNs to make them much more efficient at detecting OoD samples in LostAndFound and Fishyscapes. Re-training the CNNs with a specific choice of OoD images from COCO [32] clearly outperforms the natural baseline approach of plain softmax entropy thresholding [21] as well as many state-of-the-art approaches from image classification. In addition, we are the first to demonstrate that entropy based OoD object predictions in semantic segmentation can be meta classified reliably, *i.e.*, classified whether one considered OoD prediction is true positive or false positive without having access to the ground truth. For this meta task we employ simple logistic regression. Combining entropy maximization and meta classification therefore is an efficient and yet lightweight method, which is particularly suitable as an integrated monitoring system of safety-critical real world applications based on deep learning.

2. Related Work

Methods from prior works have already proven their efficiency in identifying OoD inputs for image data. The proposed methods are either modifications of the training procedure [19, 22, 29, 31, 37] or post-processing techniques adjusting the estimated confidence [16, 21, 29]. However, most of these works treat entire images as OoD.

When considering the semantic space to be fixed, one possible approach to anomaly segmentation, which we also pursue here, is to estimate uncertainty of CNNs. Early approaches to uncertainty estimation involve Bayesian neural networks (BNNs) yielding posterior distributions over the model's weight parameters [35, 40]. In practice, approximations such as Monte-Carlo dropout [18] or stochastic batch normalization [3] are mainly used due to cheaper computational costs. Frameworks using dropout for uncertainty estimation applied to semantic segmentation have been developed in [4, 26]. Other approaches to model uncertainty consist of using an ensemble of neural networks [28], which captures model uncertainty by averaging predictions over multiple models, and density estimation [6, 11, 39, 43] via estimating the likelihood of samples with respect to the training distribution. Methods for OoD detection in semantic segmentation based on classification uncertainty and processing only monocular images have been analyzed in [2, 7, 23, 25, 36, 41].

Using BNNs for estimating uncertainty in deep neural networks is associated with prohibitive computational costs. Uncertainty estimates that are generated by multiple models or by multiple forward passes are still computationally expensive compared to single inference based ones. In our approach, we unite semantic segmentation and OoD detection in one model without any modifications of the underlying CNN's architecture. Therefore, our re-training approach can be even combined with existing OoD detection techniques and potentially enhance their efficiency.

Works with similar training approaches as ours use a different OoD proxy and are presented in [6, 25]. They train neural networks on the unlabeled objects in Cityscapes as OoD approximation. However, in our experiments we observe that the unlabeled data in Cityscapes lacks in diversity and therefore tends to be too dataset specific. With respect to other OoD datasets, such as LostAndFound and Fishyscapes, on which we perform our experiments, we observe that these mentioned methods fail to generalize. Furthermore, in contrast to those works we incorporate a post-processing step that significantly improves the OoD detection performance.

Another line of work detects OoD samples in semantic segmentation by incorporating autoencoders [1, 5, 14, 33]. Training such a model only on specific samples from a closed set of classes, it is assumed that the autoencoder model performs less accurately when fed with samples from never-seen-before classes. The identification of an OoD input then relies on the reconstruction quality. In this way, no OoD data is required, except for further adjusting the sensitivity of the method.

Autoencoders are in fact deep neural networks themselves and usually do not include a segmentation model. For the goal of safe real-time semantic segmentation, *e.g.* necessary for automated driving [24], more lightweight approaches are favorable. We avoid incorporating deep auxiliary models at all and only employ a lightweight linear model instead. Usually the more complex a model, the greater the lack of interpretability. As monitoring systems are supposed to make deep learning models safer, one seeks for simpler and thereby more explainable approaches. We post-process our entropy boosted semantic segmentation CNN output via logistic regression whose computational overhead is negligible. This linear model is transparent as it allows us to analyze the impact of each single feature fed into the model and it demonstrates in our experiments to efficiently reduce the number of OoD detection errors.

3. Entropy based OoD Detection

In this section, we present our training method to improve the detection of OoD pixels in semantic segmentation via spatial entropy heatmapping.

3.1. Training for high Entropy on OoD Samples

Let $f(x) \in (0,1)^q$ denote the softmax probabilities after processing the input image $x \in \mathcal{X}$ with some deep learning model $f : \mathcal{X} \to (0,1)^q$ and let $q = |\mathcal{C}| \in \mathbb{N}$ denote the number of classes. For the sake of brevity, we omit the consideration of image pixels in this section. We compute the softmax entropy via

$$E(f(x)) = -\sum_{j\in\mathcal{C}} f_j(x) \log(f_j(x)) \, . \tag{1}$$

By $(x, y(x)) \sim \mathcal{D}_{in}$ we denote an "in-distribution" example with $y(x) \in \mathcal{C}$ being the ground truth class label of input x, and by $x' \sim \mathcal{D}_{out}$ we denote an "out-distribution" example for which no ground truth label is given. We aim at minimizing the overall objective

$$\begin{aligned} L := (1-\lambda)\, \mathbb{E}_{(x,y)\sim\mathcal{D}_{in}} [\ell_{in}(f(x), y(x))] \\ + \lambda\, \mathbb{E}_{x'\sim\mathcal{D}_{out}} [\ell_{out}(f(x'))] \ , \ \lambda \in [0,1] \end{aligned} \tag{2}$$

where

$$\ell_{in}(f(x), y(x)) := -\sum_{j\in\mathcal{C}} \mathbb{1}_{j=y(x)} \log(f_j(x)) \ \text{ and } \tag{3}$$

$$\ell_{out}(f(x')) := -\sum_{j\in\mathcal{C}} \frac{1}{q} \log(f_j(x')) \tag{4}$$

with the indicator function $\mathbb{1}_{j=y(x)} \in \{0,1\}$ being equal to one if $j = y(x)$ and zero else. In other words, for in-distribution samples we apply the commonly used empirical cross entropy loss, *i.e.*, the negative log-likelihood of the target class. For out-distribution samples, we consider the negative log-likelihood averaged over all classes.

By that choice of out-distribution loss function, minimizing $\ell_{out}(f(x'))$ is equivalent to maximizing the softmax entropy $E(f(x))$, see equation (1). Since the softmax definition implies $f_j(x) \in (0,1)$ and $\sum_{j\in\mathcal{C}} f_j(x) = 1$, Jensen's inequality yields $\ell_{out}(f(x)) \geq \log(q)$ and $E(f(x)) \leq \log(q)$, with equality (for both inequalities) if $f_j(x) = 1/q \; \forall \, j \in \mathcal{C}$, *i.e.*, if the softmax probabilities are uniformly distributed over all classes.

In order to control the impact of each single objective on the overall objective L, the convex combination between expected in-distribution loss and expected out-distribution loss is included, which can be adjusted by varying the parameter λ, see equation (2).

3.2. OoD Object Prediction in Semantic Segmentation via Entropy Thresholding

The softmax probabilities output of CNNs for semantic segmentation $f(x) \in (0,1)^{|\mathcal{Z}|\times q}, x \in \mathcal{X} \subseteq [0,1]^{|\mathcal{Z}|\times 3}$ can be viewed as pixel-wise probability distributions that express how likely each potential class affiliation $j = 1,\ldots,q$ at a given pixel $z \in \mathcal{Z}$ is, according to the model f. Let $f^z(x) \in (0,1)^q$ denote the softmax output at pixel location z which we implicitly considered throughout the previous section. In semantic segmentation one minimizes the averaged pixel-wise classification loss over the image, cf. equation (2). For the sake of simplicity, we consider the normalized entropy $\bar{E}(f^z(x))$ at pixel location z in the following, that is $E(f^z(x))$ divided by $\log(q)$. One pixel is then assumed to be out-of-distribution (OoD) if the normalized entropy $\bar{E}(f^z(x))$ at that pixel location z is greater than a threshold $t \in [0,1]$, *i.e.*, *z is predicted to be OoD* if

$$z \in \hat{\mathcal{Z}}_{out}(x) := \{z' \in \mathcal{Z} : \bar{E}(f^{z'}(x)) \geq t\} \, . \tag{5}$$

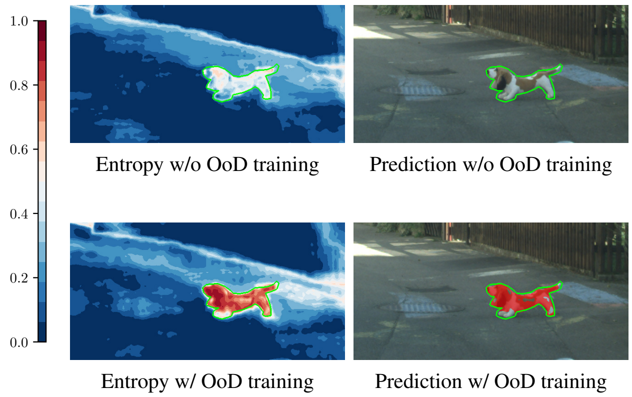


Figure 2: Comparison of softmax entropy heatmap and OoD prediction mask with our OoD training (*bottom row*) and without (*top row*). The green contours in the entropy heatmaps mark the annotation of the OoD object. The OoD object prediction is obtained by simply thresholding on the entropy heatmap (in this example at $t = 0.7$ yielding the red pixels in the OoD prediction masks).

A connected component $k \in \hat{\mathcal{K}}(x) \subseteq \mathcal{P}(\hat{\mathcal{Z}}_{out}(x))$ (the latter being the power set of $\hat{\mathcal{Z}}_{out}(x)$) consisting of neighboring pixels fulfilling the condition in equation (5) gives us an *OoD segment / object prediction*. An illustration can be viewed in figure 2. Obviously, the better an in-distribution pixel can be separated from an out-distribution pixel by means of the entropy, the more accurate the OoD object prediction will be.

4. Meta Classifier in Semantic Segmentation

By training the segmentation CNN to output uniform confidence scores as presented in section 3, we increase the sensitivity towards predicting OoD objects, aiming for an "entropy boost" on OoD samples. However, it is not guaranteed that only OoD samples have a high entropy. Therefore, detecting OoD samples via entropy boosting potentially comes along with a considerable number of false OoD predictions, resulting in an unfavorable trade-off.

In this context, we consider one entire OoD object prediction (see section 3.2) as true positive if its intersection over union (IoU, [17]) with a ground truth OoD object is greater than zero. More formally, let $\mathcal{Z}_{out}(x)$ be the set of pixel locations in x which are labeled OoD according to ground truth. Then $k \in \hat{\mathcal{K}}(x)$ *is true positive* (TP) if

$$\begin{aligned} & \mathrm{IoU}(k, \mathcal{Z}_{out}(x)) > 0 \\ \Leftrightarrow \; & \exists\, z \in k : \bar{E}(f^z(x)) \geq t \wedge z \in \mathcal{Z}_{out}(x)\,. \end{aligned} \tag{6}$$

One could also set a higher threshold on the IoU score, however in this work we treat every single pixel as a potential road hazard as this results in the least possible amount of overlooked OoD objects.

In [9] it has been demonstrated that false-positives due to increased prediction sensitivity can be removed based on a meta classifier's decision, achieving improved trade-offs between error rates. This meta classifier is essentially a binary classification model added on top of a segmentation CNN [34, 44, 45]. We construct hand-crafted metrics per connected component of pixels by aggregating different pixel-wise uncertainty measures derived from the softmax probabilities, one of which is the entropy. The entropy metric has proven to be highly correlated to the segment-wise IoU and therefore contributes greatly to the meta classifier's performance, cf. [44]. Therefore, we expect the learned entropy maximization on OoD objects to improve the meta classification performance. In contrast to existing approaches, that consider neighboring pixels sharing the same class label as segment, we generate metrics for segments above the given entropy threshold t to adapt meta classification to OoD detection. Moreover, we additionally consider the variances within segments when aggregating pixel-wise measures instead of the means only.

Given the softmax output, further pixel-wise measures we integrate into the meta classifier are the variation ratio $V(f(x)) = 1 - f_{\hat{c}}(x), \hat{c} = \arg\max_{j\in\mathcal{C}} f_j(x)$ and probability margin $M(f(x)) = V(f(x)) + \max_{j\in\mathcal{C}\setminus\{\hat{c}\}} f_j(x)$. Moreover, we also consider geometry features, such as the segment's size or its ratio between interior and boundary [44]. These metrics serve as inputs for the *meta* model that classifies into *true positive* and *false positive* (FP) OoD object prediction, *i.e.*, classifying $k \in \hat{\mathcal{K}}(x)$ into the sets

$$\begin{aligned} C_{\mathrm{TP}} &:= \{k' \in \hat{\mathcal{K}}(x) : \mathrm{IoU}(k', \mathcal{Z}_{out}(x)) > 0\} \;\text{ and} \\ C_{\mathrm{FP}} &:= \{k' \in \hat{\mathcal{K}}(x) : \mathrm{IoU}(k', \mathcal{Z}_{out}(x)) = 0\}\,. \end{aligned} \tag{7}$$

The outlined hand-crafted metrics form a structured dataset of features where the rows correspond to predicted segments and the columns to metrics.

5. Setup of Experiments

We consider the semantic segmentation of the Cityscapes data [13] as original task, *i.e.*, we consider Cityscapes as in-distribution $\mathcal{D}_{in}$. The training split consists of 2,975 pixel-annotated urban street scene images. As original model, we use the state-of-the-art semantic segmentation DeepLabv3+ model with a WideResNet38 backbone trained by Nvidia [51]. This model is initialized with publicly available weights and serves as our *baseline* model. For testing, we evaluate the OoD detection performance on two datasets comprising street scene images and unexpected objects. We consider images from the LostAndFound test split [42], containing 1,203 images with annotations of road and small obstacles in front of the (ego-)car, and Fishyscapes Validation [6], containing 30 images with annotated anomalous objects extracted from Pascal VOC [17] which are then

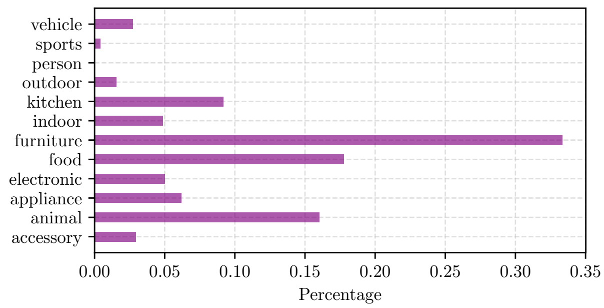
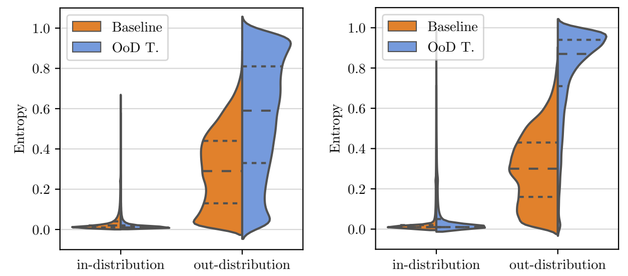


Figure 3: Relative number of pixels per supercategory in the COCO OoD proxy. In every epoch during OoD training 297 out of 46,751 images in total are randomly included.



(a) LostAndFound

(b) Fishyscapes

Figure 4: Relative pixel frequencies of (a) LostAndFound and (b) Fishyscapes OoD pixels, respectively. The density at different entropy values is displayed for the baseline model, *i.e.*, before OoD training, and after OoD training. The inner lines of the violins represent the quartiles.

overlayed in Cityscapes images. Both datasets share the same setup as Cityscapes but include some unknown road objects.

In order to perform the *OoD training* as proposed in section 3.1, we approximate the out-distribution via images from the COCO [32] dataset. This dataset contains images of objects captured in everyday scenes. Besides, we only consider COCO images with instances that are not included in Cityscapes (no persons, no cars, no traffic lights, *etc.*) and images that have a minimum height and width of at least 480 pixels. After filtering, there remain 46,751 images serving as our proxy for $\mathcal{D}_{out}$. The pixel frequencies per class is visualized in figure 3. We emphasize that none of the OoD objects in the test data have been seen during our OoD training since we use disjoint datasets for training and testing, that are originally also designed for completely different applications. The used OoD proxy is a mixture of true unknown unknowns (pylon, bloated plastic bag, styrofoam, *etc.*) as well as known unknowns in terms of visual similarities (*e.g.* dogs are available in the test data and share some visual features of cats which are available in the OoD proxy). Employing this COCO subset as approximation of $\mathcal{D}_{out}$ is motivated by works on OoD detection [22, 37] where 80 million tiny images [46] serve as proxy for all possible images.

We finetune the DeepLabv3+ model with loss functions according to equation (3) and equation (4). As training data we randomly sample 297 images from our COCO subset per epoch and mix them into all 2,975 Cityscapes training images (1:10 ratio of out-distribution to in-distribution images). We train the model's weight parameters on random squared crops of height / width of 480 pixels for 4 epochs in total and set the (out-distribution) loss weight $\lambda = 0.9$ (see equation (2)). As optimizer we use Adam [27] with a learning rate of 10^{-5}.

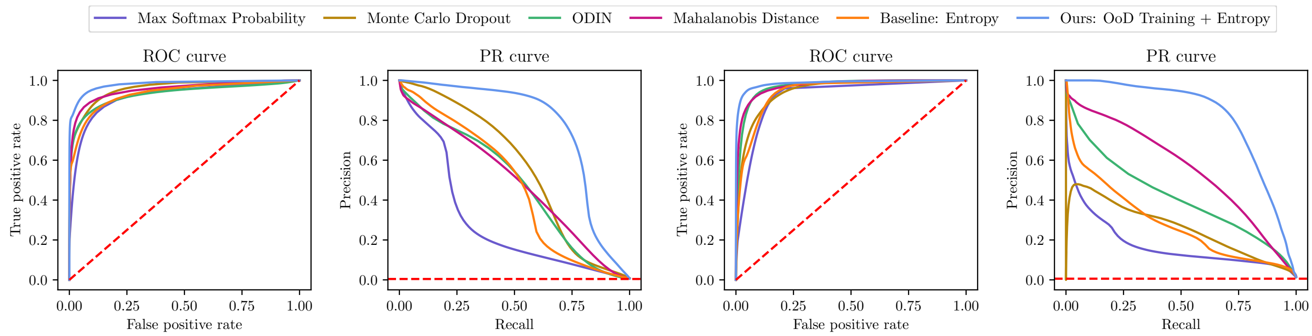
6. Pixel-wise Evaluation

Based on the softmax probabilities, we compute the normalized entropy $\bar{E}$ for all pixels in the respective test dataset. This gives us per-pixel anomaly / OoD scores which we compare with the ground truth anomaly segmentation. For the sake of clarity, in this section we refer to in-distribution pixels as *samples of the negative class* and to out-distribution pixels as *samples of the positive class*.

6.1. Separability by means of Area Under Curve

On basis of the violin plots in figure 4, one already notices the beneficial effect of our OoD training over the baseline in separating in-distribution and out-distribution pixels as large masses of the distributions corresponding to the respective classes can be well separated for a larger range of entropy thresholds. This effect can be further quantified with the aid of receiver operating characteristic (ROC) curves and precision recall (PR) curves. The area under the curve (AUC) then represents the degree of separability. The higher the AUC, the better the separability. In addition to the baseline, we include further scores of standard OoD detection methods. Namely these are: MSP [21], MC dropout [18], ODIN [31] and Mahalanobis distance [30].

By comparing the ROC curves for LostAndFound (figure 5 (a) left), we observe that there is a performance gain over the baseline model when OoD training is applied. The baseline curve indicates that the corresponding model has a lower true positive rate across various fixed false positive rates, *i.e.*, our model after OoD training assigns higher uncertainty / entropy values to OoD samples which is beneficial for OoD detection. Furthermore, also with respect to all other tested methods, entropy thresholding after OoD training shows the best degree of separability measured by the AUC of ROC curves (AUROC) with a score of 0.98. We observe the same effects for Fishyscapes (figure 5 (b)



(a) LostAndFound (*left:* AUROC, *right:* AUPRC)

(b) Fishyscapes (*left:* AUROC, *right:* AUPRC)

Figure 5: Detection ability of LostAndFound (a) and Fishyscapes (b) OoD pixels, respectively, evaluated by means of receiver operating characteristic curve (*a & b left*) and precision recall curve (*a & b right*). The red lines indicate the performance according to random guessing, *i.e.*, in the PR curves the red line indicate the fraction of OoD pixels.

left). From the Fishyscapes violins, the discrimination performance after OoD training seems already close to perfect. This is confirmed by the AUROC of 0.99, again outperforming all other tested methods.

As the AUROC essentially measures the overlap of distributions corresponding to negative and positive samples, this score does not place more emphasis on one class over the other in case of class imbalance. As there is a considerably strong class imbalance in LostAndFound and Fishyscapes (0.7% and 1.3% OoD pixels), respectively, we also consider the PR curves, see figure 5 (a) & (b) right. Thus, true negatives are ignored and the emphasis shifts to the detection of the positive class (OoD samples). Now the AUC of PR curves (AUPRC) serves as measure of separability. For LostAndFound as well as for Fishyscapes OoD pixels, the model after OoD training is superior not only over the baseline model but also any other tested method in terms of precision when we fix recall to any score. The AUPRC quantifies this performance gain and further clarifies the improved capability at detecting OoD pixels. Regarding LostAndFound, the OoD training increases the AUPRC over the baseline by 0.30 up to a score of 0.76. Regarding Fishyscapes, the performance gain is even more significant. We raise the AUC from 0.28 up to 0.81. We conclude that, measured by AUROC and AUPRC, our OoD training is highly beneficial for detecting OoD samples.

Moreover, we conducted the same experiments as for the DeepLabv3+ model [51] also for the weaker DualGCNNet [48] which is re-trained with $\lambda = 0.25$ for 11 epochs in total. We report all benchmark scores of all tested methods in table 1. Besides AUPRC, we also provide the false positive rates at 95% true positive rate (FPR_{95}) and the mean intersection over union (mIoU) for the semantic segmentation of the Cityscapes validation set. For further comparison, we additionally included scores of methods based on an autoencoder [33] and on density estimation [6].

	FPR_{95} ↓	AUPRC ↑	mIoU ↑
Network architecture and OoD score	LostAndFound Test		Cityscapes Val.
DualGCN [48] + Entropy	0.30	0.36	0.80
Ours: DualGCN + OoD T. + Entropy	0.12	0.51	0.76
PSPNet [50] + Image Resynthesis [33]	N/A	0.41	0.80
DeepV3W + Max Softmax [21]	0.32	0.27	**0.90**
DeepV3W + ODIN [31]	0.45	0.46	**0.90**
DeepV3W + MC Dropout [18]	0.21	0.55	0.88
DeepV3W + Mahalanobis [30]	0.27	0.48	**0.90**
Baseline: DeepV3W [51] + Entropy	0.35	0.46	**0.90**
Ours: DeepV3W + OoD T. + Entropy	**0.09**	**0.76**	0.89
	Fishyscapes Val.		Cityscapes Val.
DualGCN [48] + Entropy	0.46	0.07	0.80
Ours: DualGCN + OoD T. + Entropy	0.21	0.38	0.76
DeepV3W + Max Softmax [21]	0.21	0.17	**0.90**
DeepV3W + ODIN [31]	0.12	0.39	**0.90**
DeepV3W + MC Dropout [18]	0.23	0.26	0.88
DeepV3W + Mahalanobis [30]	0.14	0.55	**0.90**
Baseline: DeepV3W [51] + Entropy	0.18	0.28	**0.90**
Ours: DeepV3W + OoD T. + Entropy	**0.05**	**0.81**	0.89
	Fishyscapes Static[1]		Cityscapes Val.
DeepV3P [10] + Image Resynthesis [33]	0.27	0.30	0.80
DeepV3S [51] + Learned Density [6]	0.17	0.62	0.81
Ours: DeepV3W + OoD T. + Entropy	**0.09**	**0.87**	0.89

Table 1: Results for LostAndFound and Fishyscapes.

6.2. Original Task Performance

In order to monitor that the baseline model does not unlearn its original task due to OoD training, we evaluate the model's performance on in-distribution data with OoD predictions at different entropy thresholds. The original task is the semantic segmentation of the Cityscapes images and we evaluate by means of the most commonly used performance metric *mean Intersection over Union* (mIoU, [17]). Additionally to the Cityscapes class predictions, that is obtained via the standard maximum a posteriori (MAP) decision principle [8, 38], we consider an extra OoD class pre-

[1] 2nd best in public benchmark results: https://fishyscapes.com/results

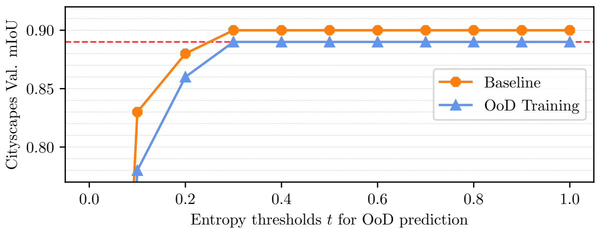


Figure 6: Mean intersection over union (mIoU) for the Cityscapes validation split with OoD predictions at entropy thresholds t. The dashed red line indicates the performance loss considered to be “acceptable” (1 percent point).

diction if the softmax entropy is above the given threshold t. We compute the mIoU for the Cityscapes validation dataset, but average only over the 19 Cityscapes class IoUs.

The state-of-the-art DeepLabv3+ model [51], which serves as our baseline throughout our experiments, achieves an mIoU score of 0.90 on the Cityscapes validation dataset without OoD predictions (implying $t = 1.0$). By re-training the CNN with entropy maximization on OoD inputs, we observe improved OoD-AUPRC scores. This gain at detecting OoD samples comes with a marginal drop in Cityscapes validation mIoU down to 0.89. These two mIoU scores remain nearly constant (deviations less than 1 percent point) for the thresholds $t = 0.3, \ldots, 1.0$. In general, the lower the entropy threshold, the more pixels are predicted to be OoD. For $t = 0.2$ this results in a noticeable performance decrease, 0.05 for the baseline model and 0.03 for the retrained model, respectively. As displayed in figure 6 further lowering the threshold leads to an even more significant sacrifice of original performance. Consequently, we consider in the following entropy thresholds of at least $t = 0.3$ since the performance loss seems acceptable, especially in view of a substantially improved OoD detection capability.

7. Segment-wise Evaluation

In this section we evaluate the meta classification performance on LostAndFound. The main metrics for the segment-wise evaluation are the numbers of FPs and FNs with respect to an OoD object prediction, cf. equation (6). The F_1-score $F_1 = 2\mathrm{TP}/(2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}) \in [0,1]$ summarizes the error rates into an overall score. As the removal of FP OoD predictions should not come at cost of a significant loss in original performance, see figure 7, we additionally consider the *miss rate of road pixels*:

$$\varepsilon := 1 - \left| \bigcup_{x \in \mathcal{X}} \left(\hat{\mathcal{Z}}_{in}(x) \cap \mathcal{Z}_{in}(x) \right) \right| \left| \bigcup_{x \in \mathcal{X}} \mathcal{Z}_{in}(x) \right|^{-1} \quad (8)$$

with pixel locations predicted to be in-distribution in $\hat{\mathcal{Z}}_{in}$ and annotated as in-distribution in $\mathcal{Z}_{in}$. The road miss rate

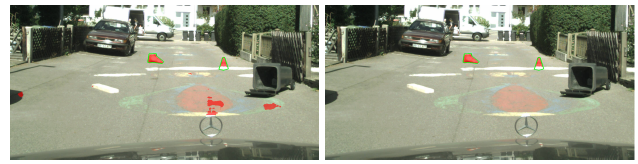


Figure 7: OoD detection with $t = 0.5$ after OoD training and meta classification. The green contours mark the annotations of OoD objects. OoD predictions in the background according to the ground truth are ignored (this includes *e.g.* the garbage bin even though it has been detected).

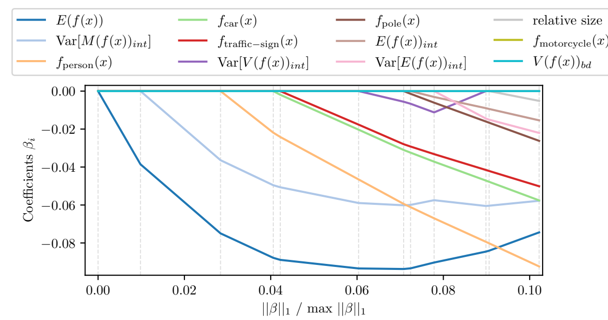


Figure 8: Least angle regression for the meta classifier with OoD training and an entropy threshold of $t = 0.3$. The 12 features becoming active first are displayed. The suffixes int and bd refer to the restriction of a metric on the segment's interior and boundary, respectively. See section 4 for a description of the metrics.

ϵ measures the fraction of actual road pixels in the whole dataset which are incorrectly identified.

We compute per-segment metrics as outlined in section 4 for OoD object predictions in the LostAndFound test set and feed them through meta classification models, which are simple logistic regressions throughout our experiments. The segments are then leave-one-out cross validated whether they are TP or FP, see equation (7). Via least angle regression we analyze the metrics having the most impact on the meta classification. The analysis shows that after OoD training the entropy metric $E(f(x))$ has the most impact, see *e.g.* figure 8 for $t = 0.3$.

In general, the higher the entropy threshold, the less OoD objects are predicted and consequently less data is fed through the linear models. This explains the observation that meta classifiers identify FPs more reliably the lower t. Due to our OoD training, the meta classifiers demonstrate to be more effective, being most superior when $t = 0.7$. In our experiments, OoD training in combination with meta classification at $t = 0.3$ turns out to be the best OoD detection approach achieving the best result with only 598 errors in total and $F_1 = 0.82$ while having a road miss rate of

Entropy Threshold	Baseline				Baseline + Meta Classifier				OoD Training				OoD Training + Meta Classifier			
$\bar{E} \geq t$	FP ↓	FN ↓	F_1 ↑	ε in % ↓	FP ↓	FN ↓	F_1 ↑	ε in % ↓	FP ↓	FN ↓	F_1 ↑	ε in % ↓	FP ↓	FN ↓	F_1 ↑	ε in % ↓
$t = 0.10$	33,584	77	0.09	7.60	386	314	0.80	3.24	21,967	99	0.12	5.22	245	302	0.83	2.70
$t = 0.20$	19,456	136	0.13	2.48	454	307	0.78	0.93	17,000	127	0.15	2.14	271	303	0.83	0.18
$t = 0.30$	7,349	218	0.28	0.38	412	302	0.79	0.09	8,068	**191**	0.26	0.30	**290**	308	**0.82**	**0.06**
$t = 0.40$	3,214	377	0.42	0.08	280	435	0.77	0.03	4,035	**289**	0.39	0.11	**251**	359	**0.81**	**0.03**
$t = 0.50$	809	662	0.58	0.01	**94**	686	0.71	**< 0.01**	1,215	**415**	0.60	0.04	145	447	**0.80**	0.02
$t = 0.60$	158	1,084	0.69	< 0.01	**26**	1,093	0.50	**< 0.01**	327	**613**	0.69	0.02	49	619	**0.76**	0.02
$t = 0.70$	10	1,511	0.16	< 0.01	**3**	1,512	0.16	**< 0.01**	135	**879**	0.61	0.01	21	881	**0.63**	0.01

Table 2: Detection errors for LostAndFound OoD objects at different entropy thresholds t. We consider the road miss rate ε, see equation (8), as further measure of loss in original performance (for Cityscapes mIoU, see figure 6). Below the horizontal line, *i.e.*, $t \geq 0.3$, we consider the loss in original performance to be acceptable, see section 6.2 for further details.

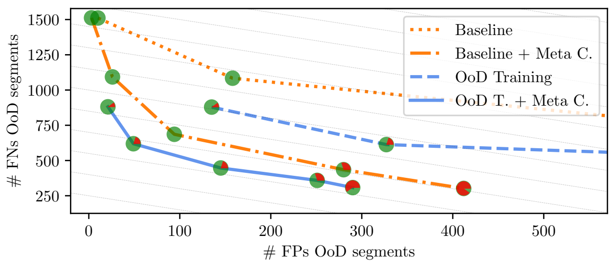


Figure 9: Detection errors of LostAndFound OoD objects. In this plot, the number of errors when $t = 0.7, \ldots, 0.3$ are displayed (when in the axes' range). The pie-chart markers indicate the road miss rate ε, being entirely red if $\varepsilon \geq 0.001$. See also table 2 for exact numbers.

Entropy Threshold t	Baseline + MSP [21] AUROC	Baseline + MSP [21] AUPRC	Baseline + Meta C. AUROC	Baseline + Meta C. AUPRC	OoD T. + Meta C. AUROC	OoD T. + Meta C. AUPRC
$t = 0.10$	0.8509	0.9817	0.9894	0.9993	0.9915	0.9993
$t = 0.20$	0.6470	0.9119	0.9859	0.9980	0.9898	0.9980
$t = 0.30$	0.5333	0.7376	0.9742	0.9884	**0.9847**	**0.9953**
$t = 0.40$	0.3847	0.4671	0.9715	0.9740	**0.9808**	**0.9807**
$t = 0.50$	0.4172	0.2286	0.9628	0.9214	**0.9665**	**0.9536**
$t = 0.60$	0.4906	0.1228	0.9291	0.7252	**0.9511**	**0.8405**
$t = 0.70$	0.5932	0.1334	0.9140	0.5283	**0.9444**	**0.7185**

Table 3: Meta classification performance on LostAndFound at different entropy thresholds t. As comparison to the meta classifier, we include the detection of OoD prediction errors via the maximum softmax probability (MSP, [21]).

marginally 0.06%, see also figure 9. Compared to the best baseline at $t = 0.6$ with $F_1 = 0.69$, we decrease the number of total errors by 52% from 1,242 down to 598. More safety-relevantly, at the same time we significantly reduce the number of overlooked OoD objects by 70% from 1,084 down to 308.

The numbers of detection errors, F_1 scores and road miss rates ε at different entropy thresholds t are summarized in table 2. The FP OoD removal efficiency is given in table 3.

8. Conclusion & Outlook

In this work, we presented a novel re-training approach for deep neural networks that unites improved OoD detection capability and state-of-the-art semantic segmentation in one model. Up to now, only a small number of prior works exist for anomaly segmentation on LostAndFound and Fishyscapes, respectively. We demonstrate that our OoD training significantly improves the detection efficiency via softmax entropy thresholding, leading to superior performance over existing OoD detection approaches.

Moreover, we introduced meta classifiers for entropy based OoD object predictions. By applying lightweight logistic regressions, we have demonstrated that entire LostAndFound OoD segments are meta classified reliably. This observation already holds for the tested CNN in its plain version. Due to the increased sensitivity of OoD predictions via entropy maximization, the meta classifiers' efficiency is even more pronounced. In view of emerging safety-critical deep learning applications, the combination of OoD training and meta classification has the potential to considerably improve the overall system's performance.

For future work, we plan to apply OoD training for the retrieval of OoD objects in order to assess the importance of their occurrence and whether a new concept is required to be learned. Our code is publicly available at https://github.com/robin-chan/meta-ood.

Acknowledgement. The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project "KI Absicherung – Safe AI for Automated Driving", grant no. 19A19005R. The authors would like to thank the consortium for the successful cooperation. The authors gratefully also acknowledge the Gauss Centre for Supercomputing e.V. (https://www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre (JSC).

References

[1] Samet Akçay, Amir Atapour-Abarghouei, and Toby P Breckon. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 3

[2] Matt Angus, Krzysztof Czarnecki, and Rick Salay. Efficacy of pixel-level OOD detection for semantic segmentation. *CoRR*, abs/1911.02897, 2019. 2

[3] Andrei Atanov, Arsenii Ashukha, Dmitry Molchanov, et al. Uncertainty estimation via stochastic batch normalization. In *Advances in Neural Networks – ISNN 2019*, pages 261–269, Cham, 2019. Springer International Publishing. 2

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 57.1–57.12. BMVA Press, September 2017. 2

[5] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *International MICCAI Brainlesion Workshop*, pages 161–169. Springer, 2018. 2, 3

[6] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, et al. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2019. 2, 3, 4, 6

[7] Dominik Brüggemann, Robin Chan, Matthias Rottmann, Hanno Gottschalk, and Stefan Bracke. Detecting Out of Distribution Objects in Semantic Segmentation of Street Scenes. In *The 30th European Safety and Reliability Conference (ESREL)*, 2020. 2

[8] Robin Chan, Matthias Rottmann, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Application of Maximum Likelihood Decision Rules for Handling Class Imbalance in Semantic Segmentation. In *The 30th European Safety and Reliability Conference (ESREL)*, 2020. 6

[9] Robin Chan, Matthias Rottmann, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Controlled false negative reduction of minority classes in semantic segmentation. In *2020 IEEE International Joint Conference on Neural Networks (IJCNN)*, 2020. 2, 4, 11

[10] Liang Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *The European Conference on Computer Vision (ECCV)*, 9 2018. 6

[11] Hyunsun Choi and Eric Jang. Generative ensembles for robust anomaly detection. *ArXiv*, abs/1810.01392, 2018. 2

[12] Pascal Colling, Lutz Roese-Koerner, Hanno Gottschalk, and Matthias Rottmann. MetaBox+: A new Region Based Active Learning Method for Semantic Segmentation using Priority Maps. In *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM*, pages 51—-62. SciTePress, 2021. 1

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, et al. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 4

[14] Clement Creusot and Asim Munawar. Real-time small obstacle detection on highways using compressive rbm road reconstruction. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 162–167, 2015. 3

[15] Jia Deng, Alexander C. Berg, Kai Li, and Li Fei-Fei. What does classifying more than 10,000 image categories tell us? In *Computer Vision – ECCV 2010*, pages 71–84, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. 1

[16] Terrance DeVries and Graham W. Taylor. Learning Confidence for Out-of-Distribution Detection in Neural Networks, Feb 2018. 2

[17] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, et al. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan 2015. 4, 6

[18] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. 2, 5, 6

[19] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[20] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. A benchmark for anomaly segmentation. *arXiv preprint arXiv:1911.11132*, 2019. 2

[21] 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. 2, 5, 6, 8

[22] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019. 2, 5

[23] S. Isobe and S. Arai. Deep convolutional encoder-decoder network with model uncertainty for semantic segmentation. In *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 365–370, 2017. 2

[24] Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308, 2020. 1, 3

[25] Nicolas Jourdan, Eike Rehder, and Uwe Franke. Identification of uncertainty in artificial neural networks. In *Proceedings of the 13th Uni-DAS e.V. Workshop Fahrerassistenz und automatisiertes Fahren*, July 2020. 2, 3

[26] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?

In *Advances in Neural Information Processing Systems 30*, pages 5574–5584. Curran Associates, Inc., 2017. 2

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[28] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc., 2017. 2

[29] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018. 2

[30] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems*, volume 31, pages 7167–7177. Curran Associates, Inc., 2018. 5, 6

[31] 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. 2, 5, 6

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, 2014. 1, 2, 5

[33] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2, 3, 6

[34] Kira Maag, Matthias Rottmann, and Hanno Gottschalk. Time-dynamic estimates of the reliability of deep semantic segmentation networks. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, November 2020. 2, 4, 11

[35] David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992. 2

[36] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, pages 1–1, 2020. 2

[37] Alexander Meinke and Matthias Hein. Towards neural networks that provably know when they don't know. In *International Conference on Learning Representations*, 2020. 2, 5

[38] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012. 6

[39] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019. 2

[40] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012. 2

[41] Philipp Oberdiek, Matthias Rottmann, and Gernot A. Fink. Detection and retrieval of out-of-distribution objects in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[42] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016. 2, 4

[43] Jie Ren, Peter J. Liu, Emily Fertig, et al. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems 32*, pages 14707–14718. Curran Associates, Inc., 2019. 2

[44] Matthias Rottmann, Pascal Colling, Thomas Paul Hack, et al. Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. In *2020 IEEE International Joint Conference on Neural Networks (IJCNN)*, 2020. 2, 4, 11

[45] Matthias Rottmann and Marius Schubert. Uncertainty measures and prediction quality rating for the semantic segmentation of nested multi resolution street scene images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2, 4, 11

[46] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. 30(11):1958–1970, Nov. 2008. 5

[47] Jingdong Wang, Ke Sun, Tianheng Cheng, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, PP, April 2020. 1

[48] Li Zhang, Xiangtai Li, Anurag Arnab, Kuiyuan Yang, Yunhai Tong, and Philip H. S. Torr. Dual graph convolutional network for semantic segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019. 6

[49] Xiang Zhang and Yann LeCun. Universum prescription: Regularization using unlabeled data. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 1, 2, 14

[50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 6

[51] Yi Zhu, Karan Sapra, Fitsum A. Reda, et al. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 4, 6, 7

[52] Aleksandar Zlateski, Ronnachai Jaroensri, Prafull Sharma, and Frédo Durand. On the importance of label quality for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1